Shape-aware Stochastic Neighbour Embedding for Robust Data Visualisations

Tobias Wängberg
Stockholm University
Chun-Biu Li (cbli@math.su.se)
Stockholm University
Joanna Tyrcha
Stockholm University

Research Article

Keywords: t-distributed Stochastic Neighbour Embedding (t-SNE), High Dimensional (HD) data, clusters, spurious patterns

Posted Date: July 21st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-668207/v1

Shape-aware Stochastic Neighbour Embedding for Robust Data Visualisations

Tobias Wängberg[1], **Joanna Tyrcha**[1,*], **and Chun-Biu Li**[1,*]

[1]Department of Mathematics, Stockholm University, 10691, Stockholm, Sweden
[*]cbli@math.su.se, joanna@math.su.se

ABSTRACT

The t-distributed Stochastic Neighbour Embedding (t-SNE) method has emerged as one of the leading methods for visualising High Dimensional (HD) data in a wide variety of fields, especially for revealing cluster structure in HD single cell transcriptomics data. However, several shortcomings of the algorithm have been identified. Specifically, t-SNE is often unable to correctly represent hierarchical relationships between clusters and spurious patterns may arise in the embedding due to incorrect parameter settings, which could lead to misinterpretations of the data. Here we incorporate t-SNE with shape-aware graph distances, a method termed shape-aware stochastic neighbour embedding (SASNE), to mitigate these limitations of the t-SNE. The merits of the SASNE are first demonstrated using synthetic data sets, where we see a significant improvement in embedding imbalanced and nonlinear clusters, as well as preservation of hierarchical structure, based on quantitative validation in clustering and dimensionality reductions. Moreover, we propose a data-driven parameter setting which we find consistently optimal in all test cases. Lastly, we demonstrate the superior performance of SASNE in embedding the MNIST image data and the single cell transcriptomics gene expression data.

1 Introduction

Analysing high dimensional (HD) data is an important challenge in a wide variety of fields. In particular, Dimensionality Reduction (DR) techniques have been increasingly used for visualising HD data by projecting them onto a low dimensional (LD), usually 2D, space. The aim is to reveal the key hidden structures in the HD data, such as clusters or other geometrical arrangements. One of the most frequently used methods for this purpose is the t-distributed Stochastic Neighbour Embedding (t-SNE) [1]. The t-SNE is able to create compelling data visualisations with hundreds of dimensions in fields ranging from image processing [1], speech recognition [2], immuno-profiling of COVID-19 patients [3], etc. One important area of application is cell biology where data are collected on gene expressions in individual cells [4, 5, 6, 7]. Cells are often characterised by expressions of thousands of different genes, where the t-SNE has enabled visual analysis of the data. One of the main successes of t-SNE is its ability to capture discrete patterns even for data with very high dimensions compared with traditional DR methods [1], such as principal component analysis (PCA) [8], locally linear embedding [9], ISOMAP [10] and Laplacian eigenmaps [11]. The approach taken by t-SNE is to focus on preserving local structures, usually characterised by Euclidean distances (ED), not taking into account the global structures. Despite the merits of t-SNE, there have been drawbacks identified in the literature. Specifically, the t-SNE requires the user to define what is meant by local, this is often difficult to assess in practise and an incorrect notion of local can result in spurious patterns appearing in the LD embedding.

To alleviate these limitations, we proposed to incorporate graph-based distances into the framework of t-SNE. The method first constructs a graph in a data-driven way to represent the HD data by only connecting points in small local neighborhoods. Information about the global structures of the constructed graph can then be captured by shape-aware graph distances (the commute time distance in this study), which evaluate distances between any pair of points by summing over local connections between the two points in the graph. In contrast to conventional distance measures, such as the ED, shape-aware graph distances are able to learn the global shapes of the underlying manifold or structure on which the HD data reside. We term the t-SNE applied to the shape-aware distances SASNE, short for Shape-Aware Stochastic Neighbour Embedding. The original t-SNE applied to conventional distance measures, e.g., ED, is simply referred as t-SNE hereafter.

In order to confirm the advantages of SASNE compared to t-SNE, we apply the methods to embed both synthetic and real data sets that demonstrate imbalanced, nonlinear and hierarchical structures. The real data set are, respectively, the MNIST data of handwritten digits and the gene expressions from cells of the mouse brain. Instead of judging the embedding performance simply by visual inspecting the LD embedding as in some of the previous works [4, 5, 6, 7, 3], the embedding qualities are scrutinized in terms of quantitative validation methods for clustering (the silhouette indices and plots) and for dimensionality reduction (rank-based methods). It was found that SASNE not only shows significantly improvement in preserving both clustering and hierarchical structures at all scales, but also allows us to fix the hyper-parameter of the method, which is commonly chosen by default [4], in a data-driven way.

2 Shape-aware Stochastic Neighbor Embedding

Overview of t-SNE The t-SNE [1, 2] is a DR method that takes a HD data set X as input and returns the LD (usually 2D) coordinates Y enabling the visualization of data patterns and organizations. The basic idea of the method is to transform the distances between data points in both of the HD and LD spaces into probability distributions. How well the distances are preserved are then quantified in terms of a dissimilarity measure (or cost function), with the Kullback-Leibler divergence commonly used, between the two distributions. Variants of t-SNE [12, 13, 14] differ from each other in the probability distributions and the dissimilarity measure used in the methods.

The t-SNE directly takes as inputs the distances between points without the need to know the coordinates of the HD feature space. It proceeds by first converting the HD distances into a probability distribution p_{ij}, usually defined by a Gaussian kernel, over all pairs of points x_i and x_j, such that close points have high probability. A key parameter in t-SNE is the 'perplexity' which corresponds to the effective number of neighbours covered by the Gaussian kernel (see Methods). The perplexity therefore controls the variable widths of the Gaussian kernel (or the neighborhood ranges) around data points in the HD space such that points separated beyond this range are considered as faraway.

Another key idea of t-SNE is the use of long-tailed t-distribution for the probability distribution q_{ij} associated with y_i and y_j in the LD space. As a result of the mismatching of the two distributions p_{ij} and q_{ij} at large distances, faraway points tend to map to much larger distances in the LD space. This is a special claim of t-SNE to mitigate the crowding problem in DR [1]. Moreover, points within the neighbourhood ranges set by the perplexity in the HD space tend to map to points also close in the LD space. These together amplify and better reveal discrete cluster structures provided that an appropriate value of perplexity is chosen. In practice, a default perplexity value of around 50 is often used with the hope that it defines reasonable neighborhood ranges that match with the spatial extents of clusters in the data.

On the other hand, the Kullback-Leibler divergence, given by $\sum_{ij} p_{ij} \log \frac{p_{ij}}{q_{ij}}$, as the cost function is asymmetric in p_{ij} and q_{ij}. This means that short distances in the HD space with large p_{ij} contribute significantly to the cost function, whereas long

distances with small p_{ij} contribute less. Consequently, this asymmetric property tends to prevent close points in the HD space from getting separated in the LD space (i.e., extrusions are discouraged). However, it does not prevent distant points in the HD space from being mapped close in the LD space (i.e., intrusions can occur) despite the mismatching of p_{ij} and q_{ij} at large distances mentioned above. The optimisation of t-SNE to find the configuration of points Y that minimizes the cost function are generally performed using gradient-based methods (see Methods for details).

Graph distance motivation The t-SNE [1, 4] is commonly employed to embed HD data based on, e.g., the ED in the HD space. However, many conventional distance measures, such as the ED, Hamming distance for string comparison [15], negative binomial distance for comparison of gene count vectors [13], etc., are often good distance measures only in local neighbourhoods that are small compared to the extents of nonlinear structures in the data. For instance, the ED can only be used locally for data points lying on a hemi-sphere since it fails to capture the nonlinear shape of the underlying manifold. This poses a problem when the common perplexity value of 50, which can connect moderately remote points, is used to produce LD embedding of distinct clusters, e.g., for the MNIST data set [1]. However, a choice of small perplexity that focuses only on preserving small local structures could result in a LD embedding composed of many small spurious clusters that do not exist in the HD data [16]. Furthermore, global structure and hierarchical organization of clusters are likely lost when a small perplexity is used [4, 16].

It is therefore generally difficult to choose an appropriate perplexity that is small enough for the convention distance measures to be useful, but large enough to be able to capture global structures in the HD data. Here we propose to employ the graph distances of the HD data as inputs to t-SNE to resolve the above shortcomings. Graph distances, sometimes called shape-aware distances [17], measure distances by summing over many short local distances, such as ED, and can therefore better capture the global nonlinear structures where the HD data reside. As will be shown later in Results, this naturally leads to a choice of large perplexity values that cannot only mitigate the problem with spurious clusters, but also largely preserve global and hierarchical structures.

To evaluate the graph distances, the first step is to construct a graph to represent the HD data where each node in the graph corresponds to a data point and edges represent the local relationships between points. We follow the common approach [18] which defines local neighborhoods by only connecting each data point x_i to its k nearest neighbors based on, e.g., the ED (see Methods for details). A graph similarity matrix w_{ij} with $i, j = 1, \cdots, n$ between data points x_i and x_j is defined as the inverse of the squared ED, $1/||x_i - x_j||^2$, and similarities of disconnected data points are simply set to zero.

Commute time distance Various graph distances, such as the geodesic distance [19], commute time distance (CTD) [20], diffusion distance [21], exist in defining relationships between nodes that capture the intrinsic geometry of the data. Here, we employ the CTD that is also known as the resistance distance [22]. The CTD between any pair of nodes i and j is defined as the expected time for a random walker to travel from node i to node j and back. Hence, the CTD connects remote nodes (data) by summing over many local connections along the way between nodes i and j in the graph, facilitating its ability to learn about the global nonlinear shape of the underlying manifold.

Several advantages of the CTD are as follows: (i) The CTD between points from the same (different) clusters are usually very small (large) due to the strong within-cluster (weak between-cluster) connectivity in the graph, meaning that CTD makes discrete structure exaggerated and easier to detect. (ii) Compared with the geodesic distance, the CTD is robust to random noise due to the averaging over all paths between nodes. (iii) The CTD has simple interpretation as a random walk on the graph. This allows us to express and compute it in terms of the eigenvalues and eignevectors of the graph Laplacian (see Methods) that is one of the fundamental concepts in graph theory [18]. (iv) Unlike, e.g., the diffusion distance, the CTD involves no additional parameter and therefore reduces the subjective input from users.

Validation of low dimensional embedding In order to monitor the preservation of cluster and hierarchical structures by t-SNE and SASNE, we advocate the use of quantitative validation indices to compare and evaluate the quality of the LD embedding. In previous studies [1, 2], quality of the embedding are often carried out by simple visual inspections, but this may lead to misleading conclusions about the data by interpreting spurious patterns created by t-SNE [16]. To quantitative account for the merits of SASNE compared with the t-SNE at the point-wise, cluster-wise (or intermediate), and inter-cluster (or global) scales, we introduce two complementary validation indices, one for clustering and another for DR, as follows.

Cluster validation In this study, we evaluate how faithfully the embedding preserves the underlying clusters using the silhouette index [23]. For a given point x_i assigned to the cluster C_k ($k = 1, \cdots, K$ with K the number of clusters) containing N_k points, the cohesion a_i is defined as $a_i = \frac{1}{N_k} \sum_{j:j \in C_k} \delta_{ij}$ where δ_{ij} denotes the distances between points x_i and x_j and the sum runs over all points in the same cluster C_k. Here δ_{ij} is the conventional distance measure, e.g. ED, when the t-SNE is used and the CTD when the SASNE is used.

To quantify separation, we first define a point-to-cluster distance $\delta(x_i, C_l) = \frac{1}{N_l} \sum_{j:j \in C_l} \delta_{ij}$ where the sum runs over all points in the cluster C_l. For a given point x_i in the cluster C_k, the separation b_i is defined as the distances from x_i to the closest

cluster that x_i does not belong to, i.e., $b_i = \min_{l \neq k} \delta(x_i, C_l)$. Combining the cohesion and separation, the point-wise *silhouette value* s_i for point x_i can then be defined as

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}. \tag{1}$$

One can see that $-1 \leq s_i \leq 1$ and s_i is close to 1 (-1) for a good (bad) clustering with large (small) separation b_i and small (large) cohesion a_i. Furthermore, the cluster-wise silhouette score $\bar{s}_k$ can be naturally evaluated as the average silhouette value over all points in the cluster C_k,

$$\bar{s}_k = \frac{1}{N_k} \sum_{i: x_i \in C_k} s_i \tag{2}$$

Finally, an overall silhouette coefficient $\bar{S}$ is evaluated by averaging over all clusters,

$$\bar{S} = \frac{1}{K} \sum_{k=1}^{K} \bar{s}_k. \tag{3}$$

We first note that the silhouette index is primarily designed to validate clustering (i.e., unsupervised learning) methods in which the data do not come with labels. Nevertheless, we will apply the silhouette index in Results below to our test and real data sets whose clusters C_k are known, to evaluate how well clustering structures are preserved from the HD space to the LD embedding.

To correctly evaluate clustering results with non-spherical clusters, conventional distance measures, e.g., the ED, which does not contain any shape information, should not be used as the distances δ_{ij} in the silhouette index. Instead, we will show in Results that the use of the CTD is more appropriate. On the other hand, the separation b_i in the silhouette index only considers the closest cluster to the data point under consideration. This means that the silhouette index cannot validate how well hierarchical organizations of clusters at the inter-cluster scales are preserved by the LD embedding. This leads us to introduce a complement validation method that takes the relative placement of the data points into account.

Dimensionality reduction validation In DR, preservation of exact distances is too restrictive that can seriously hamper the flexibility of the nonlinear mapping from the HD to the LD space [24]. Instead, it is more desirable for the embedding to only impose a monotonic relationship between the HD and LD distances that corresponds to the preservation of distance rank ordering [25, 26, 27]. In addition, unlike classical methods such as PCA and multidimensional scaling, t-SNE is not aimed at preserving exact distances.

In this study, a rank-based validation scheme for DR is formulated as follows. For each point x_i in the HD space, the rank vector $r_i^x = (r_{ij}^x)_{j \neq i}$ is defined, where $r_{ij}^x = r$ if x_j is the rth closest point to x_i. Similarly, the rank vector r_i^y is defined in the same way for the LD space. We then define a point-wise quality measure, $\bar{r}_i$, for the point x_i as the mean absolute rank error,

$$\bar{r}_i = \frac{1}{n-1} \sum_{j: j \neq i} |r_{ij}^x - r_{ij}^y|, \tag{4}$$

to quantify how well the embedding from the HD to LD space preserves the distance ordering relative to the point x_i. Likewise, an overall quality measure of preservation of rank ordering that we term 'average rank error', $\bar{R}$, can simply be evaluated by averaging the point-wise quality over all data points,

$$\bar{R} = \frac{1}{n} \sum_{i=1}^{n} \bar{r}_i. \tag{5}$$

In addition to the quality measures, it is informative to create the rank residual plot (RRP) that allows us to visually inspect the distribution of the rank residuals $r_{ij}^x - r_{ij}^y$. The RRP is a 2D density plot whose ordinate and abscissa are the value of the rank residuals $r_{ij}^x - r_{ij}^y$ and the index j ($j = 1, \cdots, n$), respectively. As we will see in Results, the RRP also tells us at what scale and to what degree the distance ordering are distorted in the embedding.

3 Results

3.1 Synthetic data sets

To provide insights for our graph based approach and demonstrate the advantages of SASNE, we apply both t-SNE and SASNE to three synthetic test sets (see Fig. 1) with known clustering structures. These test cases are: i) 'Imbalanced' data (Fig. 1**a**)

with two equally distributed clusters of points but different number of sampled points. ii) 'Nonlinear' data (Fig. 1**b**) where the clusters are sampled on a curved 1D manifold. iii) 'Hierarchical' data (Fig. 1**c**) containing 6 equally distributed clusters, where clusters 1-3 and clusters 4-6 are two distinct 'super-clusters', respectively. These test sets aim to represent different data features that are often found in real data. A good LD embedding should be able to reveal distinct clusters and display the underlying nonlinear and hierarchical data structures. Before applying embedding, it is informative to see the advantage of using the CTD over ED in highlighting clustering structures in terms of the silhouette plots (Fig S1).

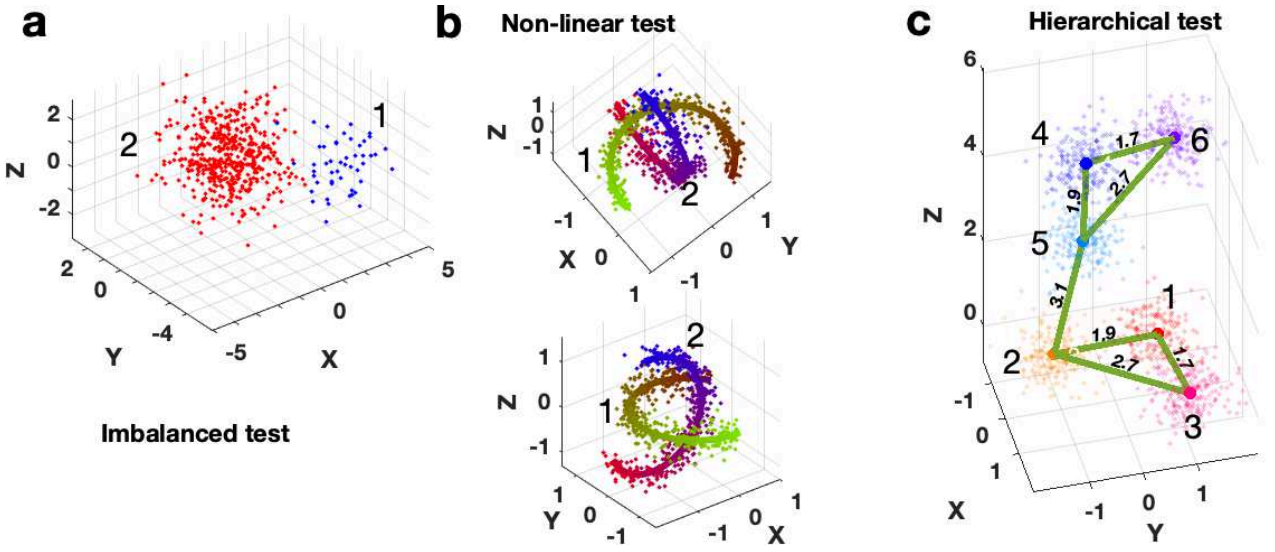

Figure 1. Three synthetic data sets. **a** Data sampled from two Gaussians with equal covariance matrix but different means. The red and blue clusters contains 1000 and 50 points, respectively. **b** Data sampled uniformly along two non-overlapping 1D nonlinear curves with Gaussian noise added. Each cluster contains 400 points. **c** Data contains 6 clusters with 100 points each. Data are sampled from Gaussians with equal covariance. The cluster means are arranged in two major groups, each containing 3 sub-clusters. The green lines are included for clarity. The numbers next to the lines indicate the ED between the cluster means.

Dimensionality reduction validation and perplexity choice Using the rank-based validation defined above, we first examine the quality of the t-SNE and SASNE for the test cases with different perplexity values. Fig. 2 shows the RRPs and the average rank errors, $\overline{R}$, for the three test cases embedded by the t-SNE and SASNE. Three perplexity values are examined for each test case, the default value of 50, an in-between value and a high perplexity close to the number of data points. The RRPs show the degree of distance rank preservation at all scales. In particular, the local and global scales locate, respectively, on the left and right sides in the RRP, and distortion of small (large) ranks corresponds to error on the local (global) scale.

From the RRPs in Fig. 2, both t-SNE and SASNE demonstrate a consistent pattern of better preservation of the distance rankings for the large perplexity. The default and in-between perplexity values tend to distort the intermediate and global scales, with a gradual decrease in rank error for increasing perplexity values for both t-SNE and SASNE. Furthermore, there are pronounced improvements in the preservation of CTD ranks by SASNE compared to the preservation of ED ranks by t-SNE. This can be understood from the fact that the CTDs are small within clusters and in directions normal to the underlying nonlinear manifold (e.g. directions normal to the nonlinear 1D lines in Fig. 1**b**). This potentially generates a data structure with lower dimensionality before subjecting to the subsequent t-SNE, leading to a more efficient embedding compared to the original t-SNE that tries to preserve, such as the ED. For all three cases, we also provide plots for the average rank error for the full perplexity range $\overline{R}$ in Fig. S2, showing a consistently decreasing trend of the average rank error as perplexity increases. From these trends, we suggest the choice of large perplexity (∼80-90 % of the number of points) for the SASNE and proceed with this setting in the following analysis, unlike the default value of 50.

Embeddings of test data sets The t-SNE and SASNE of the test cases are shown in Fig. 3. For comparison, the t-SNEs of the test sets with the default perplexity of 50 are given in Fig. S3. Both the embedding and the corresponding silhouette plots shown in Fig. 4 demonstrate that the SASNE has almost perfect clustering qualities for all test cases. In particular, the SASNE in Fig. 3**e** for the nonlinear data not only untangles the clusters, but also reveal their hidden 1D structure. Furthermore, the RRPs in Fig. 2**a** show that the t-SNE introduces distortion at all scales that cannot be easily detected by eye in Fig. 3**a** and **d**. On the other hand, the SASNE achieves much lower distance rank distortion at all scales as shown in Fig. 2**b**, implying that the hierarchical structure of the clusters is well preserved in the embeddings in Fig. 3**d**-**f**.

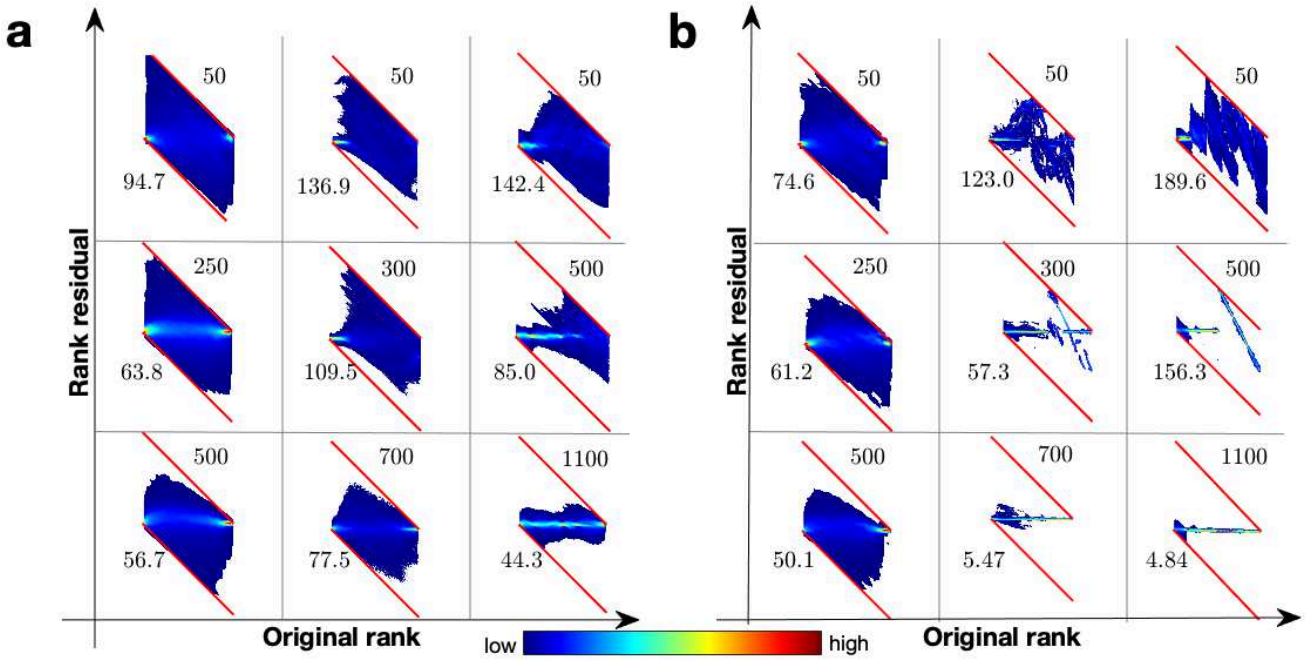

Figure 2. Rank residual plots (RRP) for the three simulated test cases. The perfect situation in which all distance rank orderings are preserved in the embedding implies that all residuals equal to zero. In that case, RRP shows a shape peak along the horizontal line in the middle of the plot. The residuals are visualised via a 2D histogram, where each bin is colored according to the relative density of points, according to the colormap located at the bottom of the plot. Empty bins are colored white. The red lines indicate the maximum rank distortion. The values on the top right and bottom left of each RRP correspond to the perplexity and the average rank error $\overline{R}$, respectively. The test cases are arranged per column, with the same order as in Fig. 1. The results for t-SNE and SASNE are shown in **a** and **b**, respectively. The superior performance of SASNE in embedding nonlinear and hierarchical data at all scales is evident from **b**.

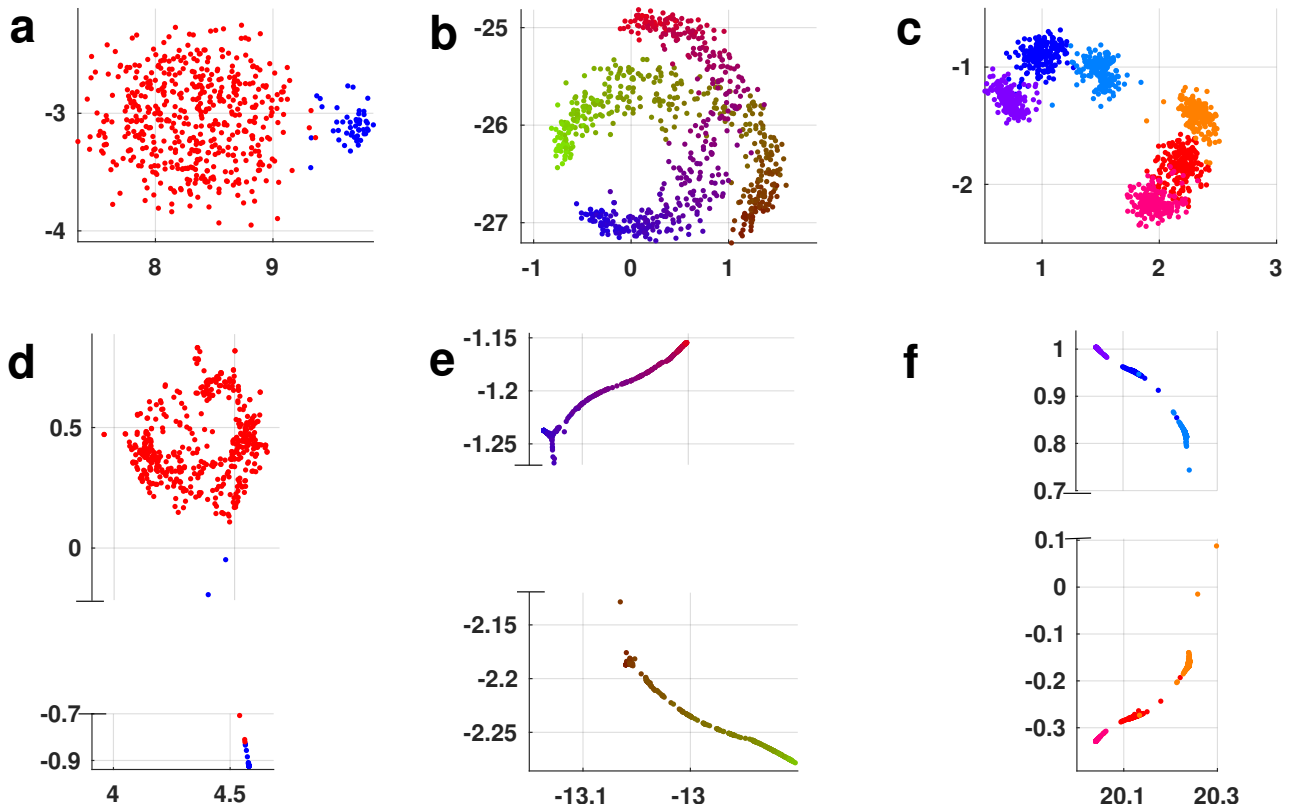

Figure 3. 2D embedding of the test cases in Fig. 1. **a**-**c** t-SNE of the imbalanced, nonlinear and hierarchical data sets. **d**-**f** SASNE of the imbalanced, nonlinear and hierarchical data sets. Notice the axis breaks in these figures, indicated by the cut-off in the graphs. The color scheme of the clusters are the same as in Fig. 1. From **a** and **d** we see the improved cluster separation of SASNE, clearly revealing the discrete structure. From **b** and **e** we see that SASNE not only correctly separates the clusters but also reveals their 1D structure, whereas t-SNE is not able to untangle the clusters correctly. From **c** and **f** we see that SASNE clearly reveal 6 clusters, with the hierarchies of sub-clusters within the blueish and reddish groups visible. On the other hand, t-SNE is not able to reveal the clusters within each group. At first glance, the t-SNE may be preferred as the spherical shape of the clusters from the original 3D data (Fig. 1**c**) are retained. However, the RRPs in the right column of Fig. 2**a** show that the t-SNE introduces distortion at all scales that cannot be easily detected by eye.

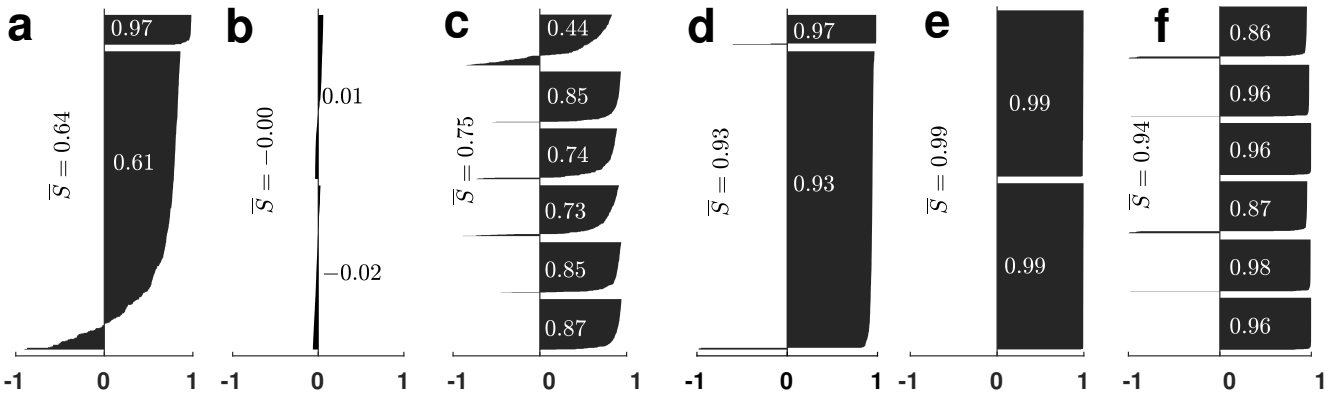

Figure 4. Silhouette plots for the test cases showing point-wise (horizontal dark bars), cluster-wise (white numerical values) and the overall ($\overline{S}$) silhouette scores. The point-wise silhouette scores are sorted in descending order per cluster. The clusters ordered from top to bottom in the silhouette plots are numbered according to Fig. 1. The x-axes are the point-wise silhouette scores. **a**-**c** Silhouette plots for t-SNE for imbalanced, nonlinear and hierarchical data, respectively. **d**-**f** Silhouette plots for SASNE for imbalanced, nonlinear and hierarchical data, respectively.

In summary, the above test cases demonstrate that the SASNE can resolve the shortcomings of t-SNE to reliably embed and reveal clusters with imbalanced, arbitrarily shaped and hierarchical structures based on the qualities of both clustering and preservation of distance ranks. Since the shape-aware CTD provides us with a valid global distance measure, the choice of a larger perplexity value allows us to consistently fix the only hyper-parameter of the embedding method in a data-driven way. To demonstrate the superior performance of SASNE for real HD data, we consider the following two data sets.

3.2 MNIST handwritten digit dataset

We first apply the SASNE to the MNIST data consisting of 70 000 images of handwritten digits [28]. Each image is represented by a 784 (28×28) dimensional vector whose entries correspond to the pixels of the image. The images are labelled based on which digit, from 0 to 9, it corresponds to. This enables us to evaluate how well the images are grouped according to their labels in the LD embedding without the need for extra clustering procedures. For computational convenience, we randomly downsample the MNIST to 5000 images. The MNIST data has known hierarchical structures, for example, digits 4 and 9 look more alike to each other compared to digits 4 and 1 (see Fig. S4). Some digit clusters are also non-spherically shaped (see Fig. S5) indicating the advantage of using shape-aware distance measures.

The resulting 2D SASNE is shown in Fig. 5**a**. The embedding shows that digits 0, 1 and 6 form relatively distinct clusters, whereas, for example, digit 7 and 9 show strong overlap that are consistent with the overlapping plots in Fig. S4. In terms of the clustering quality, the SASNE gives comparable results to t-SNE with default perplexity 50, see Fig. 5**e** and Fig. S8**a** respectively. We note that the overall silhouette coefficients of both t-SNE and SASNE are not high for the MNIST data due to the overlapping of some digits, leading to the small separations, b_i, in the point-wise silhouette value in Eq. 1. Nevertheless, for digits that are quite distinct from the others, e.g., digits 0, 1 and 6 (see Fig. S4), the cluster-wise silhouette score are higher in the SASNE than in the t-SNE by comparing Fig. 5**e** and Fig. S8**c**. This manifests the ability of the CTD to amplify discrete structures in the data.

Although the t-SNE with a small perplexity of 50 shows similar clustering quality as the SASNE, it is expected that the hierarchical organization of the digit clusters are largely distorted in the t-SNE for such small perplexity. Indeed, the RRPs for the SASNE and t-SNE (with perplexities equal to 50 and 4000) shown in Fig. 5**c** and Fig. S8**a-b**, respectively, confirm the significant improvement in preservation of the relative placement of the clusters in the LD embedding by SASNE. Furthermore, the optimal perplexity for the MNIST data is again achieved at high perplexity values (see Fig. S7), which is consistent with our conclusion learned from the test cases. To sum up from the analyses of the MNIST data, the SASNE with a large perplexity performs well simultaneously in clustering quality and preservation of distance ranks. On the other hand, the t-SNE cannot preserve both the clustering and their hierarchical structures well at the same time regardless of the choice of perplexity (see Fig. S8).

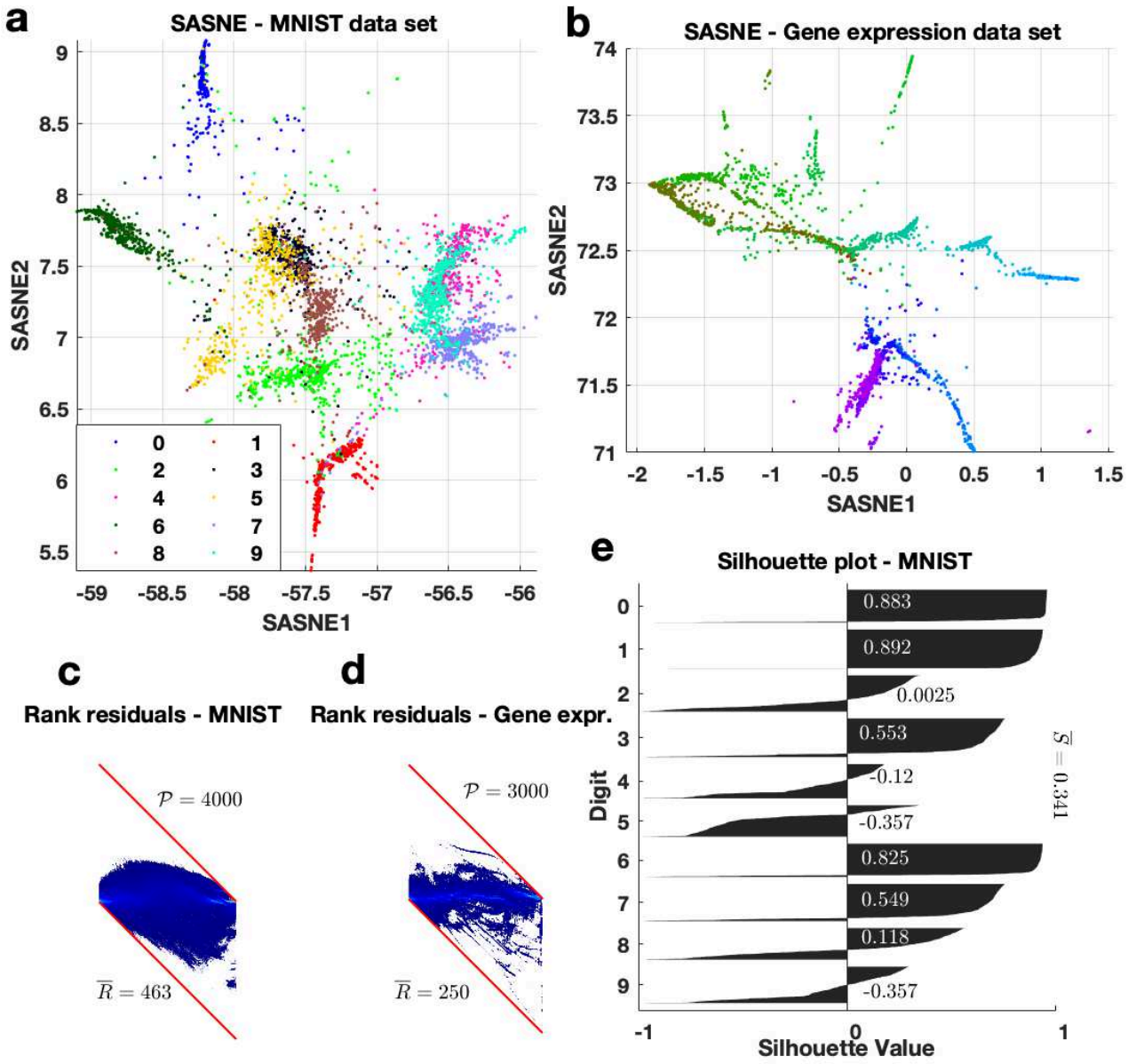

Figure 5. The SASNE of the MNIST and gene expression data sets along with the RRPs and silhouette plot. **a** SASNE of the MNIST data set. Inset shows the color corresponds to each digit. **b** The SASNE of the gene expression data set. The coloring is according to the clustering result of [13] obtaining 49 classes. Similar color indicate clusters that belong to the same hierarchy. **c**-**d** show the RRPs for the MNIST and gene expression data sets, respectively. **e** Silhouette plot for the MNIST data set.

3.3 Gene expression data

We next consider a data set of gene expressions from 3663 cells taken from the hippocampal area of a mouse brain [13]. Each cell is characterised by a gene count vector, indicating the expression frequency of the sequenced genes. With the gene count vector as coordinates of the HD space, the data set allows us to identify groupings of cells that correspond to distinct cell types based on their gene expression profiles. In contrast to the MNIST data, the gene expression data is unlabelled, i.e., the corresponding clusters, or cell types, to which the cells belong to are unknown beforehand. Therefore, an additional clustering procedure (not performed here) is needed to group the data points in the LD embedding. Since no cluster label is available, we focus only on how well distance ranks are preserved in the LD embedding.

Before applying SASNE, we follow the same procedures performed by Kobak *et. al.* [4] to reduce the number of features and produce comparable results to those reported by the original work [13]. Specifically, we select 1000 representative genes out of 27 998 in total that show high expression levels in a smaller subset of cells, indicating their capability of being good molecular indicators to distinguish cell types (see Methods). The resulting SASNE of the gene expression data is shown in

Fig. 5**b**. For comparison, the data is colored according to a previous clustering result performed by Harris *et. al.* [13] that gave rise to a total of 49 clusters by fitting a mixture of negative binomial distributions. It has been reported that these cell clusters form hierarchies, where clusters close to each others are indicated by similar colors in Fig. 5**b**. Therefore, a good preservation of distance ranks in the LD embedding is important to correctly embed these hierarchies in order to provide meaningful biological interpretations.

From Fig. 5**b**, the SASNE corroborates the previous clustering result that cell groups colored similarly also fall into nearby regions in the SASNE space. For the preservation of hierarchical structures, the RRP shown in Fig. 5**d** confirms a relatively low degree of rank distortion across all scales, with pronounced improvement compared to the t-SNE according to the RRPs and average rank errors shown in Fig. 5**d** and Fig. S9. Although clustering validation was not performed for this data set, one can still see from Fig. S9 that the t-SNE with low perplexity of 50 displays better discrete data structures but a large distortion of distance ranks across all scales. With a large perplexity of 3000, no apparent discrete structures can be visualized in the t-SNE. The SASNE, on the other hand, is able to retain both the discrete and hierarchical structures of the data set.

4 Discussion

By incorporating the concept of shape-aware distances, we proposed in this paper the SASNE and showed how it can mitigate some of the shortcomings of the t-SNE in a data-driven way that can consistently fix the hyper-parameter, perplexity, of the method. In terms of quantitative validation methods in both clustering and DR, the advantages of SASNE in embedding imbalanced, nonlinear and hierarchically structured data were first demonstrated with simulations with known ground-truth. The SASNE were then applied to two real data sets, the MNIST handwritten digits and the single cell gene expression data set, showing its superior performance compared with the t-SNE in capturing discrete and hierarchical structures hidden in the HD feature spaces.

There exist some related studies that also make use of graph-based methods to improve the performance of t-SNE. In particular, Parviainen *et al.* proposed the Graph-SNE (GSNE) method [29] that considers the probability for a random walker to reach data point i from point j and vice versa in a fixed time τ. This probability was then used as the HD distribution p_{ij} in the t-SNE procedures. GSNE has the advantage that speeds up the evaluation of p_{ij} without the need to perform matrix diagonalisation. However, there is no good strategy in choosing the hyper-parameter τ that is crucial in determining the 'scale' of the regions in the graph explored by the random walker. Therefore, it was suggested [29] to examine a wide range of diffusion times τ when using GSNE to capture hierarchical structures in the data, which in turn requires several runs of the t-SNE optimizations with possibly higher computational cost.

A recent popular alternative to t-SNE in performing DR of HD data is the Uniform Manifold Approximation and Projection (UMAP) proposed by McInnes *et al.* [30]. It has been claimed in certain cases that the UMAP can outperform t-SNE in computational speed and preservation of global structures [31]. Nevertheless, it was found [4] that the performance of the two methods depends highly on the hyper-parameter settings, and their results could be similar for certain choices of hyper-parameters. On the other hand, the UMAP works similarly [4, 30] to t-SNE by transforming the HD and LD distances to probability distributions based on a defined neighborhood size k similar to the perplexity. To preserve distances with longer ranges in the embedding, the UMAP minimises the cross-entropy, instead of the KL divergence, between the probability distributions. Therefore, it is expected that the shape-aware CTD, as the HD distance, can be readily applied to the UMAP and the idea presented in this study can be employed directly by replacing the t-SNE scheme with that of the UMAP. As a future study, it will be interesting to compare the performance of SASNE and UMAP in terms of the quantitative validations in clustering and DR.

5 Methods

Formalism of t-SNE Here we provide some mathematical details of the t-SNE method. Suppose there are n data points, the first step is to transform the distances in the HD space into a probability distribution. Specifically, a 'directed' measure of similarity from point x_i to point x_j in the HD space (with $i, j = 1, \ldots, n$) is defined as a conditional probability in terms of the Gaussian kernel and the softmax function,

$$p_{i|j} = \exp\left(-\frac{\delta_{ij}^2}{2\sigma_j^2}\right) \Big/ \sum_{k\neq j} \exp\left(-\frac{\delta_{kj}^2}{2\sigma_j^2}\right), \qquad i \neq j. \tag{6}$$

The self similarity $p_{i|i}$ is set to 0. Here δ_{ij} denotes the distance between points x_i and x_j, which is the conventional distance measure, e.g. ED, in the t-SNE and the CTD in the SASNE. The variable standard deviations σ_j (with $j = 1, \cdots, n$) can be fixed by choosing a constant value for the perplexity, $\mathscr{P}$, defined by

$$\mathscr{P} = 2^{H(p_{\cdot|j})}. \tag{7}$$

In Eq. (7), $H(p_{\cdot|j})$ denotes the Shannon entropy [32] of the probability distribution $p_{\cdot|j}$, defined as $H(p_{\cdot|j}) = -\sum_{i\neq j} p_{i|j} \log p_{i|j}$.

The perplexity can vary between 1 and n and it corresponds to the effective number of neighbors around a point x_j covered by the Gaussian kernel with standard deviation σ_j. Points beyond the perplexity range will simply be counted as 'faraway'. When perplexity equals 1, it corresponds to the case $\sigma_j \to 0$ that all probability mass is placed on the nearest neighbor. On the other hand, when perplexity equals $n-1$, it corresponds to the case $\sigma_j \to \infty$ in which all neighbors are weighted equally. The perplexity is the main hyper-parameter of t-SNE methods that needs to be determined. Moreover, the probability distribution is symmetrised as $p_{ij} = \frac{p_{i|j}+p_{j|i}}{2n}$ for computational convenience.

Similarly, the distances in the LD embedding space are also transformed into a probability distribution in terms of the long-tailed t-distribution with one degree-of-freedom as follows

$$q_{ij} = \frac{(1+||y_i - y_j||^2)^{-1}}{\sum_{k\neq l}(1+||y_k - y_l||^2)^{-1}}, \qquad i \neq j. \tag{8}$$

Here in the LD embedding space, the ED is used for both the t-SNE and SASNE. The self-similarity q_{ii} is again set to zero.

The LD embedding coordinates y_i are then obtained by minimizing the Kullback-Leibler (KL) divergence, $\mathrm{KL}(p||q) = \sum_i \sum_{j\neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$, as a cost function between the probability distributions, p_{ij} and q_{ij}, using gradient-based methods. The KL divergence has the property that $\mathrm{KL}(p||q) = 0$ if and only if $p_{ij} = q_{ij}$ for all i and j.

t-SNE optimisation The optimisation procedures of t-SNE are as follows: In both the t-SNE and SASNE methods, one minimises numerically the KL divergence between the probability distributions, p_{ij} and q_{ij}, as described above by gradient descent. Since the cost function is not convex, the optimisation may converge to a local minimum and therefore the solution may depend on the initialisation, i.e., the initial configuration of the coordinates y_i with $i = 1, \cdots, n$ in the LD space.

In case of optimizing t-SNE, we follow the protocol of Kobak and Berens [4] that the optimisation is initialised with the two leading principal components of the HD data set, normalised by the standard deviation of the corresponding principal component. The initial configuration is further multiplied by a factor of 10^{-4} which was shown empirically to speed up the convergence. For the SASNE optimisation, we use a similar initialisation procedure as in the case of the t-SNE but apply it to the CTD.

We also adopted the optimisation trick to multiply all HD probabilities p_{ij} by a constant $\alpha = 12$, called early exaggeration, for the first 250 iterations. which was shown to lead to better cluster separation [2]. Moreover, as originally suggested by Belkina *et al.* [33], the learning rate in the gradient descent is set to $\eta = n/\alpha$ where n is the number of points which has shown to lead to improved convergence behaviour in terms of stability and speed. Given the above settings, the optimisation was performed by the `tsne` function provided by the `MATLAB` Statistics and Machine Learning Toolbox.

Remarks on graph construction Although the Gaussian kernel for w_{ij} are used in some studies as the weights for the graph similarity matrix, we suggest the inverse of squared ED in this study to avoid introducing the Gaussian width as an additional parameter. With the similarity matrix w_{ij}, the constructed graph can also be viewed as a Markov network with transition probability $w_{ij}/\sum_k w_{ik}$ for a transition from node i to node j.

Different from the perplexity, the parameter k in the graph construction specifies the extent of the local neighourhoods where conventional distance measures, e.g., ED, can be used. We therefore choose a value for k that is as small as possible, just to keep the graph connected, that is, for each point x_i one can reach any other data point x_j using only the local connections. Commonly k is found to be around 5 with this method. If the data consists of highly disconnected regions, k may end up being very large to maintain connectivity in the graph. Nevertheless, this case can be handled by first locating the disconnected regions with a small k and then connecting the regions by placing edges between the k nearest points.

Computing the commute time distance Given a graph G defined by q $n \times n$ similarity matrix W with elements $w_{ij} = 1/\|x_i - x_j\|^2$, one can compute the graph Laplacian $L = D - W$ [18]. Here D is the diagonal degree matrix with elements $d_i = \sum_k w_{ik}$ that is the degree of the node i. The CTD between the points x_i and x_j can be expressed [18] in terms of the eigendecomposition of L as $C_{ij} = \mathrm{Vol}(G)\sum_{k=2}^{n}(v_{ik} - v_{jk})^2/\lambda_k$, where λ_k is the kth eigenvalue, v_{ik} is the ith element of the kth eigenvectors of L, and and $\mathrm{Vol}(G) = \sum_i d_i$ is the volume of the graph G. This expression also shows that the CTD has the form of an ED, i.e., sum of squares $\sum_{k=2}^{n}(z_{ik}^2 - z_{jk}^2)$, with the $(n-1)$D Euclidean coordinates for the ith data point given by $z_{ik} = v_{ik}\sqrt{\mathrm{Vol}(G)/\lambda_k}$ $(k = 2, \cdots, n)$. A convenient property of these coordinates are that the corresponding covariance matrix is diagonal. Therefore the PCA initilisation based on the CTD is simply the leading coordinates with largest corresponding eigenvalues, in this Euclidean space. Hence, after computation of the eigendecomposition of L, these coordinates can be directly input to any standard t-SNE implementation where we keep optimisation scheme consistent with the original t-SNE algorithm.

Pre-processing of single cell data We follow the same pre-processing procedures in [4] as follows: Let n_c and n_g be, respectively, the number of cells and the number of genes under consideration. We denote x_{ig} as the expression level of gene g

$(g = 1, \cdots, n_g)$ in cell i $(i = 1, \cdots, n_c)$. The fraction of cells that do not express the gene g is given by $d_g = \frac{1}{n_c} \sum_{i=1}^{n_c} I(x_{ig} = 0)$, where the indicator function $I(x_{ig} = 0) = 1$ when $x_{ig} = 0$, and zero otherwise. Furthermore, the mean log-expression level of the gene g can be expressed as $m_g = \frac{1}{n_{c\neq 0}} \sum_{i:x_{ig}\neq 0} \log x_{ig}$ where $n_{c\neq 0} = \sum_{i=1}^{n_c} I(x_{ij} > 0)$ is the number of cells with non-zero expression of gene g. The next step adopts a heuristic approach from [4] to select 1000 genes by finding a value of b such that there are exactly 1000 genes that exhibit high fraction of zero-expression levels across cells in relation to its mean expression value, which has shown to be able to select biologically relevant genes [34]. Mathematically, this is done by finding a value b such that exactly 1000 genes satisfying the relation $d_g > \exp\left[-\frac{3}{2}(m_g - b)\right] + 0.02$ can be selected. The coefficient $\frac{3}{2}$ and 0.02 are chosen for a good distributional fit [4]. This selected subset of 1000 genes is then kept for the analysis, whereas the others are discarded. Finally, the $\log(1 + x_{ig})$ transformation is applied to the counts of the 1000 selected genes to even out the variance of the larger expression levels. That is, the relative expression difference is considered as opposed to the absolute difference so that, for example, an expression difference from 1 to 5 is considered equal to the difference between 100 and 500.

Code availability The codes and test data sets are available at https://github.com/tobiaswangberg/SASNE.git.

References

[1] Maaten, L. van der and Hinton, G. "Visualizing data using t-SNE." In: *Journal of machine learning research* **9**.11 (2008).

[2] Maaten, L. van der. "Accelerating t-SNE using tree-based algorithms". In: *The Journal of Machine Learning Research* **15**.1 (2014), pp. 3221–3245.

[3] Mathew, D. et al. "Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications". In: *Science* **369**.6508 (2020).

[4] Kobak, D. and Berens, P. "The art of using t-SNE for single-cell transcriptomics". In: *Nature communications* **10**.1 (2019), pp. 1–14.

[5] Scala, F. et al. "Phenotypic variation of transcriptomic cell types in mouse motor cortex". In: *Nature* (2020), pp. 1–7.

[6] Wagner, D. E. et al. "Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo". In: *Science* **360**.6392 (2018), pp. 981–987.

[7] Scala, F. et al. "Layer 4 of mouse neocortex differs in cell types and circuit organization between sensory areas". In: *Nature communications* **10**.1 (2019), pp. 1–12.

[8] Pearson, K. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**.11 (1901), pp. 559–572.

[9] Roweis, S. T. and Saul, L. K. "Nonlinear dimensionality reduction by locally linear embedding". In: *science* **290**.5500 (2000), pp. 2323–2326.

[10] Tenenbaum, J. B., De Silva, V., and Langford, J. C. "A global geometric framework for nonlinear dimensionality reduction". In: *science* **290**.5500 (2000), pp. 2319–2323.

[11] Belkin, M. and Niyogi, P. "Laplacian eigenmaps for dimensionality reduction and data representation". In: *Neural computation* **15**.6 (2003), pp. 1373–1396.

[12] Lee, J. A. et al. "Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation". In: *Neurocomputing* **112** (2013), pp. 92–108.

[13] Harris, K. D. et al. "Classes and continua of hippocampal CA1 inhibitory neurons revealed by single-cell transcriptomics". In: *PLoS biology* **16**.6 (2018), e2006387.

[14] Yang, Z. et al. "Heavy-tailed symmetric stochastic neighbor embedding". In: *Advances in neural information processing systems* **22** (2009), pp. 2169–2177.

[15] Waggener, B., Waggener, W. N., and Waggener, W. M. *Pulse code modulation techniques*. Springer Science & Business Media, 1995.

[16] Wattenberg, M., Viégas, F., and Johnson, I. "How to Use t-SNE Effectively". In: *Distill* (2016). DOI: 10.23915/distill.00002. URL: http://distill.pub/2016/misread-tsne.

[17] Yaron, L., Raif M., R., and Thomas A., F. "Biharmonic distance". In: *ACM Transactions on Graphics* **29** (2010), pp. 1–11.

[18] Luxburg, U. von. "A tutorial on spectral clustering". In: *Statistics and computing* **17**.4 (2007), pp. 395–416.

[19] Bouttier, J., Di Francesco, P., and Guitter, E. "Geodesic distance in planar graphs". In: *Nuclear physics B* **663**.3 (2003), pp. 535–567.

[20] Lipman, Y., Rustamov, R., and Funkhouser, T. "Biharmonic Distance". In: *ACM Transactions on Graphics* **29**.3 (June 2010).

[21] Coifman, R. R. and Lafon, S. "Diffusion maps". In: *Applied and computational harmonic analysis* **21**.1 (2006), pp. 5–30.

[22] Klein, D. J. and Randić, M. "Resistance distance". In: *Journal of mathematical chemistry* **12**.1 (1993), pp. 81–95.

[23] Rousseeuw, P. J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* **20** (1987), pp. 53–65.

[24] Lee, J. A. and Verleysen, M. *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

[25] Lee, J. A. and Verleysen, M. "Quality assessment of dimensionality reduction: Rank-based criteria". In: *Neurocomputing* **72**.7-9 (2009), pp. 1431–1443.

[26] Mokbel, B. et al. "Visualizing the quality of dimensionality reduction". In: *Neurocomputing* **112** (2013), pp. 109–123.

[27] Gracia, A. et al. "A methodology to compare dimensionality reduction algorithms in terms of loss of quality". In: *Information Sciences* **270** (2014), pp. 1–27.

[28] LeCun, Y. et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* **86**.11 (1998), pp. 2278–2324.

[29] Parviainen, E. and Saramäki. “Drawing clustered graphs by preserving neighborhoods”. In: *Pattern Recognition Letters* **100** (2017), pp. 174–180.

[30] McInnes, L., Healy, J., and Melville, J. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018).

[31] Becht, E. et al. “Dimensionality reduction for visualizing single-cell data using UMAP”. In: *Nature biotechnology* **37**.1 (2019), pp. 38–44.

[32] Shannon, C. E. “A mathematical theory of communication”. In: *ACM SIGMOBILE mobile computing and communications review* **5**.1 (2001), pp. 3–55.

[33] Belkina, A. C. et al. “Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets”. In: *Nature communications* **10**.1 (2019), pp. 1–12.

[34] Andrews, T. S. and Hemberg, M. “M3Drop: dropout-based feature selection for scRNASeq”. In: *Bioinformatics* **35**.16 (2019), pp. 2865–2867.

Figure 1 Three synthetic data sets. **a** Data sampled from two Gaussians with equal covariance matrix but different means. The red and blue clusters contains 1000 and 50 points, respectively. **b** Data sampled uniformly along two non-overlapping 1D nonlinear curves with Gaussian noise added. Each cluster contains 400 points. **c** Data contains 6 clusters with 100 points each. Data are sampled from Gaussians with equal covariance. The cluster means are arranged in two major groups, each containing 3 sub-clusters. The green lines are included for clarity. The numbers next to the lines indicate the ED between the cluster means.

Figure 2 Rank residual plots (RRP) for the three simulated test cases. The perfect situation in which all distance rank orderings are preserved in the embedding implies that all residuals equal to zero. In that case, RRP shows a shape peak along the horizontal line in the middle of the plot. The residuals are visualised via a 2D histogram, where each bin is colored according to the relative density of points, according to the colormap located at the bottom of the plot. Empty bins are colored white. The red lines indicate the maximum rank distortion. The values on the top right and bottom left of each RRP correspond to the perplexity and the average rank error $\overline{R}$, respectively. The test cases are arranged per column, with the same order as in Fig. 1. The results for t-SNE and SASNE are shown in **a** and **b**, respectively. The superior performance of SASNE in embedding nonlinear and hierarchical data at all scales is evident from **b**.

Figure 3 2D embedding of the test cases in Fig. 1. **a**-**c** t-SNE of the imbalanced, nonlinear and hierarchical data sets. **d**-**f** SASNE of the imbalanced, nonlinear and hierarchical data sets. Notice the axis breaks in these figures, indicated by the cut-off in the graphs. The color scheme of the clusters are the same as in Fig. 1.

Figure 4 Silhouette plots for the test cases showing point-wise (horizontal dark bars), cluster-wise (white numerical values) and the overall ($\overline{S}$) silhouette scores. The point-wise silhouette scores are sorted in descending order per cluster. The clusters ordered from top to bottom in the silhouette plots are numbered according to Fig. 1. The x-axes are the point-wise silhouette scores. **a**-**c** Silhouette plots for t-SNE for imbalanced, nonlinear and hierarchical data, respectively. **d**-**f** Silhouette plots for SASNE for imbalanced, nonlinear and hierarchical data, respectively.

Figure 5 The SASNE of the MNIST and gene expression data sets along with the RRPs and silhouette plot. **a** SASNE of the MNIST data set. Inset shows the color corresponds to each digit. **b** The SASNE of the gene expression data set. The coloring is according to the clustering result of [13] obtaining 49 classes. Similar color indicate clusters that belong to the same hierarchy. **c**-**d** show the RRPs for the MNIST and gene expression data sets, respectively. **e** Silhouette plot for the MNIST data set.

Acknowledgements

We thank M. Nilsson and C. M. Langseth for many helpful discussions and comments that helped improve this work. This work was supported by the pair-doctoral program at Stockholm University, titled ‘Statistical methods for spatial tissue profiling’.

Author contributions statement

C.-B.L. and J.T. designed and supervised the study. T.W. performed the analyses. All authors involved in developing the method, discussing the results and preparing the manuscript.

Competing interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformationSASNE.pdf