Explainable t-SNE for single-cell RNA-seq data analysis

Henry Han*[1], Tianyu Zhang[2], Mary Lauren Benton[1], Chun Li[3], Juan Wang[4], Junyi Li[5]

[1] Department of Computer Science, Rogers School of Engineering and Computer Science, Baylor University, Waco, TX 76798 USA

[2] Center for Cognitive and Behavioral Brain Imaging, Ohio State University, Columbus, OH, 43210 USA

[3] Key Laboratory of Data Science and Intelligence Education, Ministry of Education, Hainan Normal University, 571158, China

[4] School of Computer Science, Qufu Normal University, Rizhao 276826, China

[5] College of Computer Science and Engineering, Harbin Institute of Technology, Shenzhen, 518005 China

Corresponding email: Henry_Han@baylor.edu

Abstract. Single-cell RNA (scRNA-seq) sequencing technologies trigger the study of individual cell gene expression and reveal the diversity within cell populations. To measure cell-to-cell similarity based on their transcription and gene expression, many dimension reduction methods are employed to retrieve the corresponding low-dimensional embeddings of input scRNA-seq data to conduct clustering. However, the methods lack explainability and may not perform well with scRNA-seq data because they are often migrated from other fields and not customized for high-dimensional sparse scRNA-seq data. In this study, we propose an explainable t-SNE: cell-driven t-SNE (c-TSNE) that fuses the cell differences reflected from biologically meaningful distance metrics for input scRNA-seq data. Our study shows that the proposed method not only enhances the interpretation of the original t-SNE visualization for scRNA-seq data but also demonstrates favorable single cell segregation performance on benchmark datasets compared to the state-of-the-art peers. The robustness analysis shows that the proposed cell-driven t-SNE demonstrates robustness to dropout and noise in dimension reduction and clustering. It provides a novel and practical way to investigate the interpretability of t-SNE in scRNA-seq data analysis. Unlike the general assumption that the explainanbility of a machine learning method needs to compromise with the learning efficiency, the proposed explainable t-SNE improves both clustering efficiency and explainanbility in scRNA-seq analysis. More importantly, our work suggests that widely used t-SNE can be easily misused in the existing scRNA-seq analysis, because its default Euclidean distance can bring biases or meaningless results in cell difference evaluation for high-dimensional sparse scRNA-seq data. To the best of our knowledge, it is the first explainable t-SNE proposed in scRNA-seq analysis and will inspire other explainable machine learning method development in the field.

Keywords: Explainable AI, scRNA-seq, high-dimensional data, t-SNE, clustering, dimension reduction

1 Introduction

The recent emergence of the single-cell RNA sequencing (scRNA-seq) enables the study of gene expression at the level of individual cells, which can give more insight into the unexplained heterogeneity among cell populations [1]. scRNA-seq unveils transcriptomic landscapes by detecting the high-resolution differences between cells. Compared to bulk RNA-seq that assesses bulk cell populations, it measures gene expression on the level of individual cells, and deciphers essential cell-to-cell variability, and uncovers the dynamics of cell fate decision making by precisely comparing the transcriptome of individual cells. It brings more insight into the unexplained heterogeneity among cell populations by unveiling latent subtle biological behaviours compared to bulk RNA-seq and other traditional profiling techniques. The revelation of heterogeneity across each cell contributes to the identification of new cell subpopulation, which is particularly important in unveiling the mechanism of complex diseases, immune systems, and neural systems [2].

However, to reveal the heterogeneity among cell populations, it is essential to quantify variations between each cell's gene expression profile via clustering scRNA-seq phenotypes. In other words, the critical question that must be answered is which cells are similar or different based on their transcription and gene expression via clustering. scRNA-seq clustering (segregation) remains a challenge in machine learning and bioinformatics because scRNA-seq data is high-dimensional sparse data with built-in random noise from sequencing and experimental design artifacts. Its number of variables is much larger than that of observations. Each observation (sample) contains of a large number of zeros or near-zero values because of dropouts, where expressed transcripts may not be detected and assigned zero expression values in sequencing [2]. The noise can be rooted in batch effects, low sequencing depth, dropouts, or biological factors such as the stochastic mechanism in gene expression. Although different methods (e.g., ZIFA) have been developed to with deal the zero-inflation problem, it remains unclear how effectively they can enhance scRNA-seq clustering compared to other methods [3]**.** Therefore, the high-dimensionality, high-nonlinearity, and zero-inflation along with noise present a hurdle for effective scRNA-seq clustering, especially because existing normalization methods are still premature in most studies.

Many scRNA-seq studies have employed state-of-the-art dimension reduction algorithms such as t-SNE (t-distributed stochastic neighbor embedding) in scRNA-seq clustering [4-6]. They map scRNA-seq samples in an alternative low-dimensional space to detect subtle differences across samples and seek subpopulation similarities [7]. Although exceptions exist, the dimension reduction techniques provide a transformed feature extraction procedure to de-noise data, reduce redundancy between variables, and visualize data from different perspectives to unveil latent data behaviors. For example, t-SNE seeks a low-dimensional embedding for input data to keep the original intrinsic structure by maintaining the relative entropy between the probability distributions induced by the pairwise distances in the input and low-dimensional space [8]. Compared to the holistic dimension reduction methods such as PCA, t-SNE along with peer UMAP (uniform manifold approximation and projection) are good at capturing subtle local data behaviors. Sun et al provided detailed evaluations and comparisons about 18 different dimension reduction methods on 30 public scRNA-seq datasets in scRNA-seq clustering [4].

However, most methods do not perform as well as expected in scRNA-seq clustering. Many researchers believe the special characteristics of scRNA-seq data play an important role in the issue, because scRNA-seq data do not match the methods well [8]. Another important but rarely mentioned factor is that the dimension reduction methods are not explainable AI methods. They act as a black-box for users rather than providing understandable data-driven interpretations. For example, t-SNE is a widely employed dimension reduction method in scRNA-seq clustering with decent performance. However, it can be hard to explain why there are almost no methodological differences when applying t-SNE to a low-dimensional high-frequency trading dataset versus a high-dimensional sparse scRNA-seq dataset. The former has 10,000+ observations and <12 variables with almost zero sparseness; the latter has less than 500 observations and more than 20,000 variables, with at least 20% sparseness, in which at least 20 precents of entries are zeros [8].

Although most t-SNE applications employ the default Euclidean distance in t-SNE to compute the required pairwise distance matrix, the Euclidean distance complicates interpretation and can even introduce bias for high-dimensional sparse scRNA-seq data. This is because Euclidean distance gives the same importance to each direction of an input sample, which is usually expected to be dense data that has close to zero sparseness. It is doubtful that this metric can be applied well to high-dimensional sparse scRNA-seq data without generating biases. Previous work also pointed out the Euclidean distance can be meaningless when applied to high-dimensional data because of 'curse of high-dimensionality' as well as a high computing burden [9]. Therefore, the sample similarity calculation should be more customized for scRNA-seq data for the sake of interpretation and accuracy. In summary, though the state-of-the-art t-SNE method is widely used in scRNA-seq clustering and even achieves acceptable performance, it is still not an explainable machine learning method for scRNA-seq data because it is simply migrated from machine learning without considering the nuances of the input data.

Although explainable machine learning is still a new topic in biomedical data science , these methods are essential for scRNA-seq data analysis. Explainable machine learning will contribute to transparency and accuracy in clustering and other downstream analysis by avoiding biased or even wrong results, besides enhancing trustworthiness and transparency of the machine learning (e.g., clustering) results. Furthermore, scRNA-seq data analysis is closely bundled with disease diagnosis and disease subtype discovery; in this high-stakes application domain, we have higher requirements for the interpretation of a machine learning method (e.g., t-SNE).

How can we develop an explainable t-SNE method for the sake of scRNA-seq clustering? The t-SNE embedding is generally used for single cell segregation rather than the original data for its advantage in clustering accuracy and complexity. Enhancing the explainability of t-SNE for the sake of scRNA-seq clustering will make t-SNE more applicable to single cell data analysis and provide more accurate and robust cell segregations. But the explainability of t-SNE is rarely investigated in almost all existing scRNA-seq analysis. It is probably because the explainability of dimension reduction is seldom mentioned in the existing machine learning literature because most research efforts are investigated in the explainability of supervised machine learning [10-11].

We believe an explainable t-SNE will be a t-SNE customized by considering the special characteristics of scRNA-seq data for the sake of meaningful embedding calculation. For example, it should take more explainable distance metrics to evaluate scRNA-seq sample similarity rather than use the default Euclidean distance. The explainable distance metrics would bring better cell discrimination that contributes to better segregation. Therefore, in contrast with the assumption that there is a trade-off between explainability and learning efficiency in machine learning, an explainable t-SNE should be able to enhance low-dimensional embedding quality for the sake of clustering compared to the original t-SNE because it is customized according to input data [11-12].

On the other hand, the performance of many dimension-reduction and following clustering algorithms depend critically on the distance metric to calculate sample similarity in the original high dimensional space. It remains unknown which distance metrics will be appropriate for scRNA-seq data. Wang *et al*. proposed a SIMIL (single-cell interpretation via multi-kernel learning) approach to learn an appropriate similarity metric between samples by employing multi-kernel learning rather than selecting a particular distance metric in sample similarity evaluation [8]. The learned distance matrix in SIMLR can represent sample similarities automatically through an optimization framework. The results show that their model would outperform most of current methods in scRNA-seq clustering [8]. This seminal work suggests that cell diversity should be evaluated by more advanced distances rather than default Euclidean distance.

However, SIMLR lacks good explainability because the results may not be explained well on behalf of scRNA-seq data sample diversity. It is still unknown whether the distance metric learned from SIMLR is biological meaningful. There is also no guarantee that SIMLR will gain an optimal distance for sample similarity discrimination via multi-kernel learning. It is unknown whether the inferred distance matches high-dimensional sparse data well because their methods theoretically can be applied to any generic data rather than only designed for high-dimensional sparse scRNA-seq data. Besides, they may not explain well how the multi-kernel functions reveal the internal difference between cells' gene expressions [8]. On the other hand, quite a few imputation methods were proposed for the sake of scRNA-seq clustering and other down-stream analysis from different perspectives (e.g., low-rank approximation) [13-14]. They provide an alternative way to overcome the dropout issue in scRNA-seq analysis. In this study, we view dropout as a built-in characteristic in sampling transcriptome for each scRNA-seq data rather than conduct imputation in clustering.

In this study, we propose an explainable t-SNE method: cell-driven t-SNE (c-TSNE) for scRNA-seq clustering. It is a customized t-SNE augmented with cell-driven distance metrics and provides biologically meaningful interpretations to cell discriminations and more explainable low-dimensional embeddings. The cell-driven distance metrics make more relevant samples mapped as the closest neighbors to each other in the low-dimensional embedding space. The customized cell-driven metrics are designed to reflect cell differences between scRNA-seq samples from a biological viewpoint.

Unlike the general t-SNE, the proposed cell-driven t-SNE (c-TSNE) should only work well for scRNA-seq data. It is not intended to apply to other similar data, even bulk RNA-seq data.

Cell-driven t-SNE (c-TSNE) first assumes three important biological factors hierarchically contributing to cell diversity. The first is 'which genes are expressed in single cell sequencing?'. Such a factor can be essential for discriminating two scRNA-seq samples because a large number of genes are not expressed due to the dropout issue. The second is 'what are the expression levels of the top-expressed genes in sequencing?' Since the top-expressed genes can contribute to distinguishing cells more than the other genes, their impacts should be accounted for. The third is 'what are the expression levels of whole genes?'. The third factor plays a role in evaluating the cell diversity, provided two cells have the same or similar levels of contributions from the previous two factors.

The cell-driven t-SNE (c-TSNE) then addresses the three biological factors by using appropriate and explainable distance metrics including yule (Yule's Y), low rank approximation with Chebyshev (L-Chebyshev), and fractional distance metrics. The distance metrics aim to capture the differences between cells by analyzing the impacts from the gene expression status, the expression of the top-expressed genes, and the expression of whole genes jointly. The Yule metric captures the cell differences caused by whether certain genes are expressed or not. Such a metric is especially employed to model high-dimensional sparse scRNA-seq data, in which a zero entry in a scRNA-seq sample is interpreted as 'gene not expressed' and a non-zero entry is interpreted as 'gene expressed'. The low-rank approximation with Chebyshev (L-Chebyshev) metric is designed to capture the difference of the top expressed genes. The fractional distance metric is designed to calculate the cell difference by considering the expression of the whole genes in a more accurate and meaningful manner to avoid the bias from the Euclidean distance [9]. The cell-driven t-SNE (c-TSNE) fuses the three distance metrics to build a more explainable and representative pairwise matrix according to the sparse degree of input data. The fused pairwise distance matrix models the cell similarity between the input scRNA-seq samples. It is further employed to calculate the Gaussian distribution P in the input space and the student t-distribution Q in the embedding space, which measure the similarity between input scRNA-seq samples, before computing the final cell-driven t-SNE embedding [10].

The proposed cell-driven t-SNE (c-TSNE) is more explainable than the original t-SNE for its targeted customization. It interprets cell diversity as the fusion of the three biologically meaningful, i.e., 'cell-driven' distances. The distances are designed by considering the special characteristics of high-dimensional sparse scRNA-seq data and overcome the evaluation bias of the Euclidean distance [9]. The experimental results show that scRNA-seq clustering with c-TSNE can not only greatly outperform the original t-SNE, but also the specifically designed sophisticated peer algorithms such as SIMIL [8]. More importantly, it is much easily understood and explainable than its peers from scRNA-seq application perspective besides machine learning. To the best of our knowledge, it is the first explainable t-SNE in scRNA-seq segregation that integrates good interpretation and decent performance. Compared to other scRNA-seq clustering methods, the proposed method is efficient and easy to implement but with good performance. The robustness analysis shows that the proposed cell-driven t-SNE (c-TSNE) demonstrate robustness to dropout and noise in clustering. Thus, it will inspire more explainable dimension reduction algorithm development for the sake of scRNA-seq analysis or even other biomedical data science fields. On the other hand, this work also points out that biologically meaningful metrics outperform general metrics in identifying subpopulations of scRNA-seq cells.

2. Cell-driven t-distributed stochastic neighbor embedding(c-TSNE)

Before describing the proposed cell-driven t-SNE (c-TSNE), we introduce cell-driven diversity distance metrics as follows. The metrics include the Yule metric (Yule'Y), low-rank approximation Chebyshev (L-Chebyshev) metric, and fractional distance metric [15-17]. The cell-driven metrics are used to evaluate the impacts from the three biological factors on cell difference: 'which genes are expressed in single cell sequencing', ' the expression levels of the top-expressed genes', and 'the expression levels of whole genes' respectively.

2.1 The cell difference under the Yule metric

The Yule metric, namely *Yule's Y*, developed by George Udny Yule in 1912, is a measure of association between two binary variables and known as the coefficient of colligation [15]. To measure the cell difference under the Yule's metric, we first map each scRNA-seq sample $x = (x_1, x_2, x_3, \dots, x_N)^t$, to its binary vector $x' = f(x) = (x_1', x_2', \dots, x_N')^t$ by using the mapping function $f(t) = \begin{cases} 1, & t \neq 0 \\ 0, & t = 0 \end{cases}$: $x_j' = f(x_j), j = 1,2, \dots N$. The binarization preprocess maps all zeros in a scRNA-seq sample to 0 ('not expressed') and all non-zero entries to 1 ('expressed'), i.e., it creates a corresponding 'binary map' for each cell to model its gene expression status in sequencing.

Given the binary vectors of two scRNA-seq cells across N genes $U = (u_1, u_2, u_3, \dots, u_N)^t$, $Z = (v_1, v_2, v_3, \dots, v_N)^t$, where u_i, $v_i \in \{0,1\}, i \in \{0,1,2, \dots, N\}$, then the cell difference between two cells caused by the statuses of gene expressed and not-expressed can be presented by the following Yule's Y as follows,

$$Yule's\ Y(U,V) = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}} \tag{1}$$

The parameter a is the frequency of both $v_i = 0$ and $u_i = 0$: $a = |\{i: u_i = 0 \wedge v_i = 1\}|$. Similarly, $b = |\{i: u_i = 1 \wedge v_i = 0\}|$, $c = |\{i: u_i = 0 \wedge v_i = 1\}|$, $d = |\{i: u_i = 1 \wedge v_i = 1\}|$. The range of Yule metric is between -1 and 1, where −1 and 1 reflect total negative (positive) association, whereas 0 reflects no association.

If the distance between two cells A and B under the Yule metric is 0.92, it means they are similar with respect to each other. Those genes expressed in the cell A will also express themselves in the cell B. Similarly, if the between two cells A and B under the Yule metric is -0.88, it means they are dissimilar with each other 88%. Those genes expressed in the cell A mostly will not express themselves in the cell B.

2.2 The cell difference under low-rank approximation with Chebyshev metrics (L-Chebyshev)

The top-expressed genes in scRNA-seq data can contribute to a large portion of the whole data variances and play an important role in contributing to cell difference. The top-expressed genes can be those differentially expressed with a large-fold expression change relevant to the other genes in one or more cells. Biologically, they may be important candidates in the pathway interacting with other genes mostly or even affecting other genes' expressions. We design a low-rank approximation with Chebyshev metric (L-Chebyshev) to model the expression levels of the top-expressed genes as follows.

We first employ single value decomposition (SVD) to approximate the original data in the subspace spanned by the top singular vectors [16]. Given $X = \{x_{1,} x_{2,\dots} x_n\}$, $x_i \in \Re^p$, $p \gg n$, we have the SVD decomposition at the selected rank $l < p$

$$X' = \sum_{i=1}^{l} u_i \ \ s_i v_i^t \tag{2}$$

The selected approximation rank can be decided by $l = \underset{k}{\operatorname{argmin}} \frac{\sum_{i=1}^{k} s_i}{\sum_{i=1}^{\min(p,n)} s_i} > \rho(80\%)$. The approximated gene expression matrix $X' \in \Re^{n \times l}$ would include the expressions of the top-expressed genes and the contributions of those genes with low expressions are ignored in the SVD approximation. The maximum approximation rank l will be determined by checking the variance ratio $\frac{\sum_{i=1}^{k} s_i}{\sum_{i=1}^{\min(p,n)} s_i}$ so that singular values of the selected approximation rank can hold at least 80% total variance ratio. As a result, each column vector in the matrix X' is the approximated cell containing the expressions of the top-expressed genes corresponding to each original cell.

We then use the Chebyshev metric to capture the difference between two approximated cells to evaluate the impacts of top-expressed genes on the cell difference and such metric is named as the low-rank approximation with Chebyshev metric (L-Chebyshev). Given two cells x, y $x_i \in \Re^p$ in X, the distance given by L-Chebyshev metric is calculated as follow:

$$D_{L-chebyshev}(x,y) = \lim_{k \to \infty} (\sum_{i=0}^{p} |x'_i - y'_i|^k)^{1/k} \tag{3}$$

where the $x', y' \in \Re^p$ are the corresponding approximated cells of x, y in X'. The reason we use the Chebyshev distance rather than the default Euclidean distance is to avoid the bias from the Euclidean distance because the approximated cells are still high-dimensional data. In fact, the Chebyshev distance between two points is always less than the Euclidean distance for two approximated cells. Thus, it avoids the amplified Euclidean distances for the sake of more accurate evaluations of the impact of the expression of top-expressed genes on cell difference.

2.3 The cell difference under fractional distance metrics.

As we mentioned before, Aggarwal *et al.* mathematically proved that in the high dimensional space, the distance or similarity calculated by standard metrics, such as Euclidean metric, is likely to be meaningless or biased and named this phenomenon as 'the curse of dimensionality.' [9]. When the dimensionality becomes higher, the difference between the maximum and minimum distances, for certain data points, does not increase as fast as the minimum distance to other data points. When the dimensionality becomes extreme high, the ratio of the difference between the maximum and minimum distances to the minimum distances will be close to 0, which makes it hard to tell which data points are similar or dissimilar.

Inspired by Aggarwal et al's work, we extend the normal Minkowski distance to a specific fractional distance metric to deal with this problem to evaluate the cell difference caused by all genes. Here, the fractional distance between the two single-cell samples $U, V' \in \Re^p$ is defined as:

$$\mathrm{D}_d^f(U,V) = \left(\sum_{i=0}^{p} |u_i - v_i|^f\right)^{1/f} \tag{4}$$

where $f \in (0,1)$. Thus, we can augment the difference between the nearest point and the farthest point and give a better measurement of similarity in the high dimensional space. Technically, we can set any value for f in the interval, but a too small f value may bring some unnecessary oscillations in distance calculations for a relatively dense sample that with less zeros and another sample with more zeros. We empirically choose $f = 1/4$ in this study because it demonstrates a better sensitivity in detecting cell difference than the other values.

Although the fractional distance theoretically may not follow the triangle inequality for a distance metric for some very special points such as $x = (0,0, \dots 0)^t$, $y = (1,0, \dots 1)^t$, $z = (0,0, \dots 1)^t$, such exceptional cases are unlikely to happen for real scRNA-seq samples because there is no a real scRNA-seq sample with all zero expressions. Actually, the good performance from the fraction distance in clustering demonstrates that it would be an effective distance metric to discriminate cell difference better than the traditional measures practically. Thus, we can still treat it as a distance metric in evaluating the similarities between cells.

2.4 Cell-driven distance fusion

The three different distance metrics applied to input scRNA-seq data generate three pairwise distance matrices to reflect the cell difference from the three different biological aspects. It needs to fuse the three matrixes to build a new pairwise distance matrix to model the cell difference more accurately, because the three biological factors contribute to the cell diversity jointly. A normalization process needs to apply to each matrix to standardize their scales so that they can be fused on the same page without losing their own characteristics. We find that standardizing each distance matrix to zero mean and one standard deviation will not be a good way because the distances may not be subject to a normal distribution, which is especially true for the distance matrix generated under the Yule and fractional distance metrics. Therefore, we adopt the approach to normalize each distance matrix by employing its largest eigenvalue as a scaling factor in normalization [16]. Given a distance matrix D, we conduct eigenvalue decomposition as follows,

$$D = Q\Lambda Q^{-1} \tag{5}$$

Then the normalized distance matrix $D_{normalized} = \lambda_1^{-1} \times D$, where $\lambda_1 = eig_{max}(D)$ is the largest eigenvalue of the pairwise distance matrix D under a metric. It is noted that such a normalization factor scales each matrix well but can be expensive because the complexity of the eigenvalue decomposition can reach $O(n^3)$. But such a complexity should not be a concern in implementation because n, which represents the number of cells, will not be a number larger than 10^3, for almost all scRNA-seq datasets.

After all pairwise distance matrices complete normalization, we fuse the three matrices to a new pairwise distance matrix for t-SNE. The ideal way is to assign exact weights to each matrix so that their impacts to the cell difference can be counted accurately. However, we lack enough prior knowledge to implement it because we do not know which factors will have more impact on the final cell diversity because of biological complexities, stochastic nature of sampling, dropout, and other artifacts. It is possible to do individual study for each dataset and their medical/biological background to determine a more appropriate fusing model for each dataset, but such a sophisticated fusing may let cell-driven t-SNE lose generalization for generic scRNA-seq data and increase algorithm complexity. Thus, we just take the following simple but effective fusion mechanisms: max-fusion and sum-fusion described in the following two equations.

Biologically, the max-fusion assumes the most important discriminated distance between two cells can come from any source no matter the top-expressed genes or others. It picks the ρ_{max} value of corresponding entries from the three normalized distance matrices: $D_{yule}, D_{fractional}$, and $D_{L\_chebyshev}$ as the corresponding entry of the fused matrix D^f. The ρ_{max} by default is the maximum function, but it also can be a percentile function to pick a high percentile (e.g., 80th percentile) for a scRNA-seq dataset with a high sparsity degree (e.g. ≥ 50%).

$$D^f\,(i,j) = \rho_{max}\left(D_{yule}(i,j), D_{fractional}(i,j), D_{L\_chebyshev}(i,j)\right) \tag{6}$$

Sparsity is defined to measure the sparseness degree of an input dataset $\mathrm{X} = \{x_{1,}x_{2,\dots}x_n\}$, $x_i \in \Re^p$, the sparsity of X is defined as the ratio between the number of zero entries over the total number of entries in the dataset. Generally, the sparsity of a typical scRNA-seq dataset falls between 25% and 85% without imputation processing.

$$\beta(\mathrm{X}) = \frac{|\{x_{ij}: x_{ij} = 0\}|}{np} \tag{7}$$

The summation fusion, i.e., sum-fusion, assumes the three sources contribute to cell difference jointly in a weighted way:

$$D^f\,(i,j) = w_{yule}D_{yule}(i,j) + w_{fractional}D_{fractional}(i,j) + w_{L_{chebyshev}}D_{L_{chebyshev}}(i,j) \tag{8}$$

The weights $w_{yule}, w_{fractional}, w_{L_{chebyshev}}$cannot be determined in an accurate way. But when input data has a high sparsity value (e.g., >0.5), it is recommended that the weight for the Yule metric w_{yule} to have more weights (e.g., 0.7). The default weight selection for them should be 1 for each when input data does not have a high sparsity or its other data characteristics are not quite clear. Without loss of the generality, we define $d_{ij} = D^f\,(i,j) = \|x_i - x_j\|$ that represents the fused distance between single cell samples x_i and x_j.

Generally speaking, the sum-fusion is recommended for input scRNA-seq datasets with relatively small sparsity values, but max-fusion is recommended to handle the datasets with relatively large sparsity degrees. In some special cases, it is recommended to only take a single cell-driven metric such as the Yule metric for some datasets with a high degree of sparsity because gene expressed or not can be a more important factor in discriminating the cell difference compared to the other factors. More details can be found in the Algorithm 1.

2.5 Cell-driven t-SNE (c-TSNE), an explainable t-SNE

The key difference between proposed c-TSNE and original t-SNE lies in that c-TSNE calculates a biologically meaningful and cell-driven pairwise matrix D^f fused by explainable distance matrices to evaluate cell similarity. We describe c-TSNE with more details by following the original t-SNE framework.

Given an input scRNA-seq dataset $X = \{x_{1,}x_{2,\cdots}x_n\}$, $x_i \in \Re^p, p \gg n$, c-TSNE calcualtes the corresponding low-dimensional embedding $Y = \{y_{1,}y_{2,\cdots}y_n\}$, $y_i \in \Re^l, n \gg l$, generally $l = 2$, by minimizing the Kullback-Leibler (K-L) divergence between a Gaussian distribution P and a normalized Student's t-distribution Q,

$$\varphi(P,Q) = KL(P \parallel Q) = \sum_{i=1}^{n} \sum_{j=1,j\neq i}^{n} p_{ij} \, log \frac{p_{ij}}{q_{ij}} \tag{9}$$

where p_{ij} models the pairwise similarity between points x_i and x_j in the original high-dimensional manifold and q_{ij} models the pairwise similarity of their corresponding low-dimensional embeddings: y_i and y_j [10]. $p_{ij} = \frac{p_{i|j}+p_{j|i}}{2n}$ is defined as the average of two conditional probabilities, in which $p_{j|i} = \frac{\exp(-d_{ij}{}^2/2\sigma_i^2)}{\sum_{k\neq i} \exp(-\frac{d_{ij}{}^2}{2\sigma_i^2})}$, where x_j, x_k are the neighbors of x_i, σ_i^2 is the variance of all neighbor data points of x_i, and $d_{ij} = D^f\,(i,j)$ is the fused distance between samples x_i and x_j. Similarly $q_{ij} = \frac{(1+\left\|y_i-y_j\right\|^2)^{-1}}{\sum_{k\neq l}(1+\left\|y_k-y_l\right\|^2)^{-1}}$. Due to the different input pairwise matrices, the distributions P and Q in the proposed c-TSNE will be totally different from those in the original t-SNE.

The c-TSNE embedding y_j is calculated by using the following gradient learning scheme,

$$y_{j+1} = y_j - r_j \frac{\partial \varphi}{\partial y_j} \tag{10}$$

where $\frac{\partial \varphi}{\partial y_j} = 4\sum_{j\neq i}^{n}(p_{ij} - q_{ij})q_{ij} \sum_{k\neq l}(1 + \|y_k - y_l\|^2)^{-1}\,(y_i - y_j)$, and r_j is the learning rate , which is recommened to set as about 1/10 of the learning rate of the classic t-SNE. Although small learning rate might 'condense' the observations in small regions, we find the small learning rate will contrbute to finding better local optmimum under the cell-driven distances rather than got stuck in one local optimum easily though it takes more time [19,26].

It is noted that the embeddding from c-TSNE should be more representative and robust compared to the embedding from the original t-SNE because the fused distance matrix D^f is more representative and robust compared to the one from the original t-SNE. The proposed c-TSNE still shows randomness as the classic t-SNE because of the nature of the non-unique solutions from the non-convex optimization [10]. However, c-TSNE tends to be more stable than the classic t-SNE under the change of the perplexity values.

The proposed c-TSNE has a higher complexity than the classic t-SNE: $O(n^3 + nlogn)$. It is due to the high complexities from distance matrix normalization and L-Chebyshev distance calculations, both of which need $O(n^3)$ complexity. However, since c-TSNE mainly handles high-dimensional scRNA-seq data, the number of n is mostly less than 1000. Therefore, c-TSNE is a technically workable algorithm though it appears as a high-complexity algorithm before going through complexity optimization strategies [18-19].

2.6 Cell-driven t-SNE segregation (clustering)

The proposed cell-driven t-SNE (c-TSNE) can be employed to conduct single cell segreation. It is also an important way to validate its effectiveness and superiroity. The first is to employ c-TSNE to calculate the low-dimensional embedding of input scRNA-seq data. Then a clustering algorithm, say K-means, is employed to cluster the embedding to discover the subpopulations. Although different clustering approaches can be employed, K-means is selected for the conveience of comparisons because it is widely used in the existing literature for its simplicity [8]. Similarly, the other peer dimenson reduction methods' single cell clusterings follow the same scheme by replacing c-TSNE with other dimension reduction methods such as t-SNE, UMAP or kernel principal component analysis (KPCA) etc.[27-30].

The effectivness of clustering is evaluated by using classic clustering evalaution metrics such as normalized mutual information (NMI) [20]. The NMI, a ratio between 0 and 1, is defined as follows . Given two clustering results U and V for input data. then NMI is the ratio between the mutual information of U and V and the average entropy of U and V.

$$NMI = \frac{MI(U,V)}{\text{mean}\{H(U),H(V)\}} \tag{11}$$

where $MI(U,V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} P(i,j) \log\left(\frac{P(i,j)}{P(i)P'(j)}\right)$, $P(i,j) = \frac{|U_i \cap V_j|}{|U|}$, $P(i) = \frac{|U_i|}{|U|}$, $P'^{(j)} = \frac{|V_j|}{|V|}$, $H(U) = -\sum_{i=1}^{|U|} P(i) \log P(i)$, $H(V) = -\sum_{i=1}^{|V|} P'^{(i)} \log P'(i)$. The larger the NMI indicates the better quality scRNA-seq data clustering [21]. Although other clustering measures are also available, the NMI demonstrates a good stability across different scRNA-seq datasets. Thus, we only choose NMI as the clustering quality index in this work.

The following algorithm describes the cell-driven t-SNE clustering, where c-TSNE is also described in a detailed way. The proposed c-TSNE seeks different fusion methods and even different weights according to the sparsity values of input data. When the input dataset has a sparsity $\beta >$ the sparsity cutoff: β_*, which is set as 75% in this study, the max-fusion is employed to fuse the three pairwise distance matrices; when the sparsity falls in the interval [50%, 75%], more weights should be given to the weight of the pairwise distance matrix generated from the Yule metric.

The cell-driven t-SNE (c-TSNE) clustering has the complexity $O(n^3 + nlogn + kn)$, where k is the number of clusters and n is the number of observations. Compared to the complexity of the t-SNE-based clustering: $O(\ nlogn + kn)$, it is slightly an expensive clustering algorithm, although it still can be done in a real-time because the scale of $n < 10^3$.

Algorithm 1: Cell-driven t-SNE clustering

```
Input:
        scRNA-seq dataset $X \in \Re^{n \times p}$
        scRNA-seq label information: label
        sparsity cutoff: $\beta_*$ (default 75%)
Output:
        NMI of $X$ clustering

// Calculate data sparsity
  $\beta \leftarrow \frac{|\{x_{ij}: x_{ij}=0\}|}{np}$
// Calculate Yule, fractional, and L_Chebyshev metrics
  $D_{yule}, D_{fractional}, D_{L_{chebyshev}} \leftarrow CalculateDistanceMatrices(X, 'yule', 'fracional', 'L_{chebyshev}')$
// Pairwise distance matrix normalization
  $D_{yule} \leftarrow D_{yule}/eig_{max}(D_{yule})$
  $D_{fractional} \leftarrow D_{fractional}/eig_{max}(D_{fractional})$
  $D_{L_{chebyshev}} \leftarrow D_{L_{chebyshev}}/eig_{max}(D_{L_{chebyshev}}$
if $\beta > \beta_*$ //max fusion
    $D^f(i,j) = \rho_{max}\left(D_{yule}(i,j), D_{fractional}(i,j), D_{L\_chebyshev}(i,j)\right)$
else
    if $\beta > 50\%$
        $adjustWeights(w_{yule}, w_{fractional}, w_{L_{chebyshev}})$
    $D^f(i,j) = w_{yule}D_{yule}(i,j) + w_{fractional}D_{fractional}(i,j) + w_{L_{chebyshev}}D_{L_{chebyshev}}(i,j)$
//Compute the embedding of the cell-driven t-SNE
    $embeddding \leftarrow$ CalculateCellDrivenTSNEEmbedding($D_f$)
//K-means clustering
    $NMI \leftarrow Kmeans(embedding, label)$
Return $NMI$
```

3. Benchmark single-cell RNA-seq datasets

In this study, we use four benchmark datasets to analyze our method to evaluate the proposed cell-driven t-SNE method in scRNA-seq clustering. Table 1 summarizes the basic information of the four datasets along with their sparsity values.

Table 1 scRNA-seq data basic information

Dataset	No. of Cells	No. of Genes	No. of classes	Sparsity (%)
Beutter	182	8989	3	37.9
Kolod	704	10685	3	27.9
Pollen	249	14805	11	51.0
Usoskin	622	17772	4	78.1

It is noted that data sparsity was rarely addressed in the previous scRNA-seq studies, but it is interesting to find that the Kolod datatset with the smallest sparsity has the best clustering result in our study [22]. As mentioned before, the sparsity of a scRNA-seq data affects the final fusion matrix in the proposed cell-driven t-SNE. We provide more details about each dataset as follows.

1. The Buettner dataset consisting of 182 cells across 3 classes was obtained from a controlled experiment which studied the effect of the cell cycle on the gene expression levels in individual mouse embryonic stem cells (mESCs) [23].
2. The Kolod dataset, which has 704 cells across 11 cell subpopulations, was obtained from a controlled experiment that studied the effect of the cell cycle on the gene expression level in individual mouse embryonic stem cells (mESCs) [22].
3. The Pollen dataset that consists of 249 cells across 11 cell populations including neural cells and blood cells (Pollen data set [24]). It was designed to test the utility of low-coverage single-cell RNA-seq in identifying distinct cell populations, and thus contained a mixture of diverse cell types: skin cells, pluripotent stem cells, blood cells, and neural cells.
4. The Usoskin dataset contains 622 neuronal cells from the mouse dorsal root ganglion across different 4 neuronal cell types, with an average of 1.14 million reads per cell [25].

4. Results

This section includes c-TSNE clusteritng performance on the benchmark datasets in comparison with its peer methods that include t-SNE, UMAP, KPCA, and SIMIL. The first three are dimension-reduction-based scRNA-seq clustering that use the Euclidean distance to evaluate cell difference. The last one is a multi-kernel learning-based scRNA-seq clustering approach that use the learned optimal distance metric to evaluate cell similarity [8].

Peer methods. Besides t-SNE, we employ UMAP and KPCA as the peer methods in comparison with the proposed c-TSNE in scRNA-seq clustering. This is mainly because UMAP and KPCA are state-of-the-art or effective dimension reduction methods but rarely appeared in the literature of scRNA-seq clustering compared to the other peers [8]. Both UMAP and KPCA demonstrate better performance than widely-used PCA (data not shown) according to our study. Besides, we also include SIMIL as one peer methods because its importance in scRNA-seq clustering. More details about it can be found in [8]. UMAP can be viewed as a method closely related to t-SNE because it can be somewhat viewed as a manifold learning method to fix some weakness of t-SNE (e.g., low convergence) [27-28]. KPCA is one of few methods to conduct dimension reduction in a high-dimensional Hilbert space by using kernel tricks [29]. We choose the cosine kernel $k(x, y) = (x \cdot y)/||x|| \times ||y||$ in KPCA for its automatic L2 normalizataion and good performance compared to the other popular kernels (e.g., Gaussian kernels) [29].

4.1 scRNA-seq clustering under t-SNE with individual cell-driven metrics.

To demonstrate the advantage of the proposed cell-driven distances over the traditional metrics (e.g., Euclidean distance) in single cell RNA-seq visualization and following clustering, we integrate each cell-driven metric with t-SNE to the four benchmark datasets to compare its performance with those of the classic metrics in t-SNE clustering. Like other low-dimensional-embedding-based clustering, t-SNE clustering employs the classic K-means to cluster the t-SNE embedding of input data.

Figure 1 shows the t-SNE visualizations under the Yule, L-Chebyshev, and Euclidean distance (default metric) across the four datasets, in which each point is colored by its true label in the corresponding dataset. It is interesting to see the t-SNE visualizations are more meaningful under the two cell-driven metrics than the Euclidean distance because their more representative embeddings. For example, t-SNE with the Yule and L-chebyshev metrics improve interpretation by discovering better clustering results. For example, it clearly shows the well-separated three cluster separations for the Beutter and Kolod data. Similarly, the Usoskin and Pollen datasets continue to demonstrate the advantage of two cell-driven metrics in clustering, in which the Pollen data clusters are more orthogonal and have larger distances with each other under the cell-driven metrics. On the other hand, t-SNE with the Euclidean distance cannot indicate three clusters of the Buettner and Kolod data, in which cells are 'jammed together'. It makes the t-SNE visualization more inexplainable because they cannot display their correct clusters due to the poor cell diversity discriminability from the Euclidean distance.

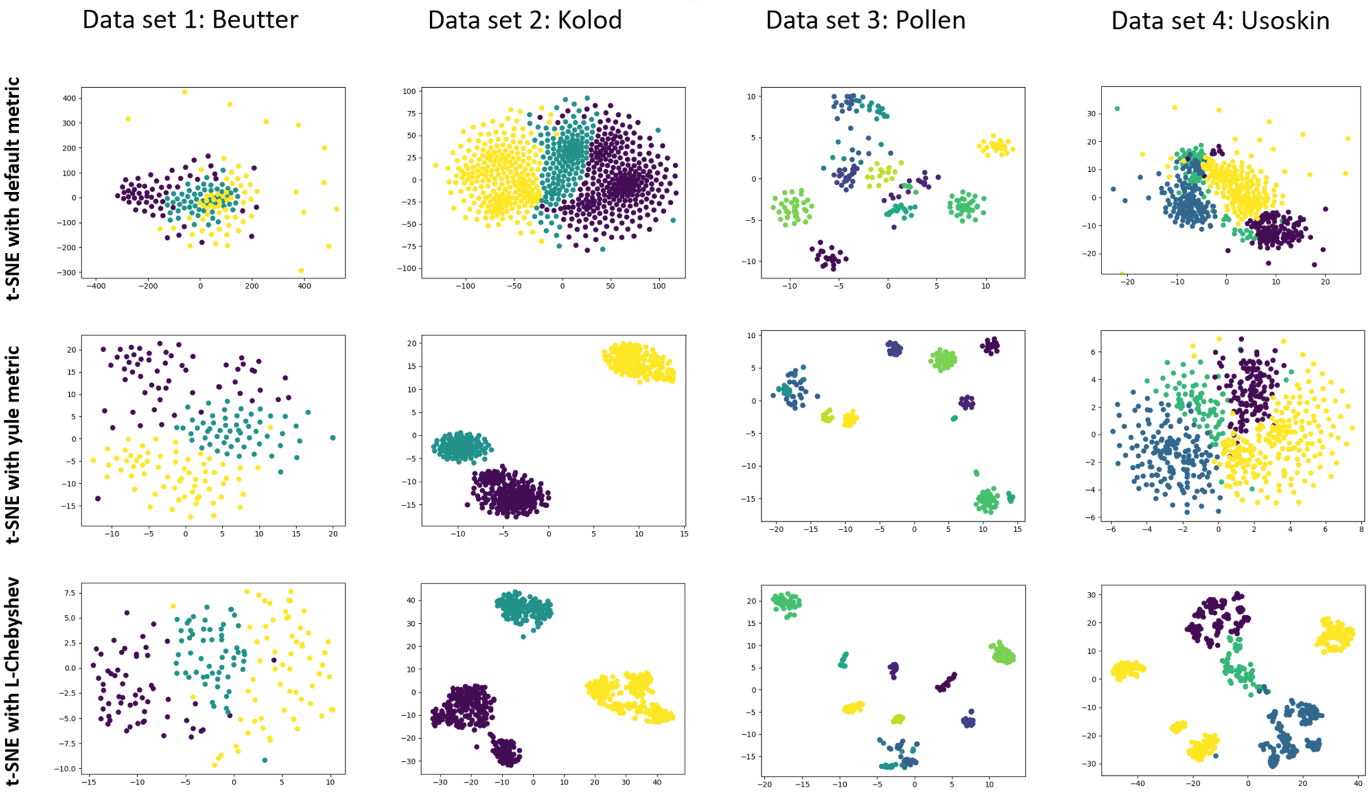

Fig 1. The t-SNE visualization comparisons under Yule, L-Chebshev, and Euclidean distances (default metrics) on the four benchmark datasets that have 3,3,11, and 4 clusters. The Yule and L-Chebshev distance metrics demonstrate more meaningful separations for different subpopulations than the Euclidean distance for their more representative embeddings.

Table 3. The NMIs of t-SNE clustering under different metrics

Datasets	*Yule*	*L-chebyshev*	*Euclidean*	*Fractional*
Beutter	0.73	0.68	0.18	0.79
Kolod	1.0	1.0	0.47	0.68
Pollen	0.9	0.93	0.86	0.91
Usoskin	0.45	0.77	0.2	0.54

Table 3 shows the corresponding NMI values of clustering different t-SNE embeddings under the three cell-driven metrics and Euclidean distance. It echoes the visualization results in Figure 1 and suggests that the two cell-driven metrics not only enhance the clustering performance, but make the results are more explainable: the more targeted distance metric choice leads to better cell segregation than the default Euclidean distance that may bring biases in cell similarity evaluation. It indicates that the widely used Euclidean distance has the worst performance for the four benchmark datasets compared to the proposed cell-driven distances. For example, the NMI values reach 1.0 under the Yule and L-chebyshev metrics, but the corresponding NMI under the Euclidean distance is only 0.18. The other three datasets show similar results. It also echoes the previous result that the Euclidean distance can be meaningless or biased for high-dimensional data. It indicates that the proposed metrics can unveil the cell differences in an explainable and effective way: the more biologically meaningful metrics, the more representative embeddings, and the better single cell segregation performance. Similarly, the fractional distance also achieves the leading advantage to the Euclidean distance in clustering as the Yule and L-chebyshev metrics. The corresponding t-SNE visualizations can be found in the supplemental materials.

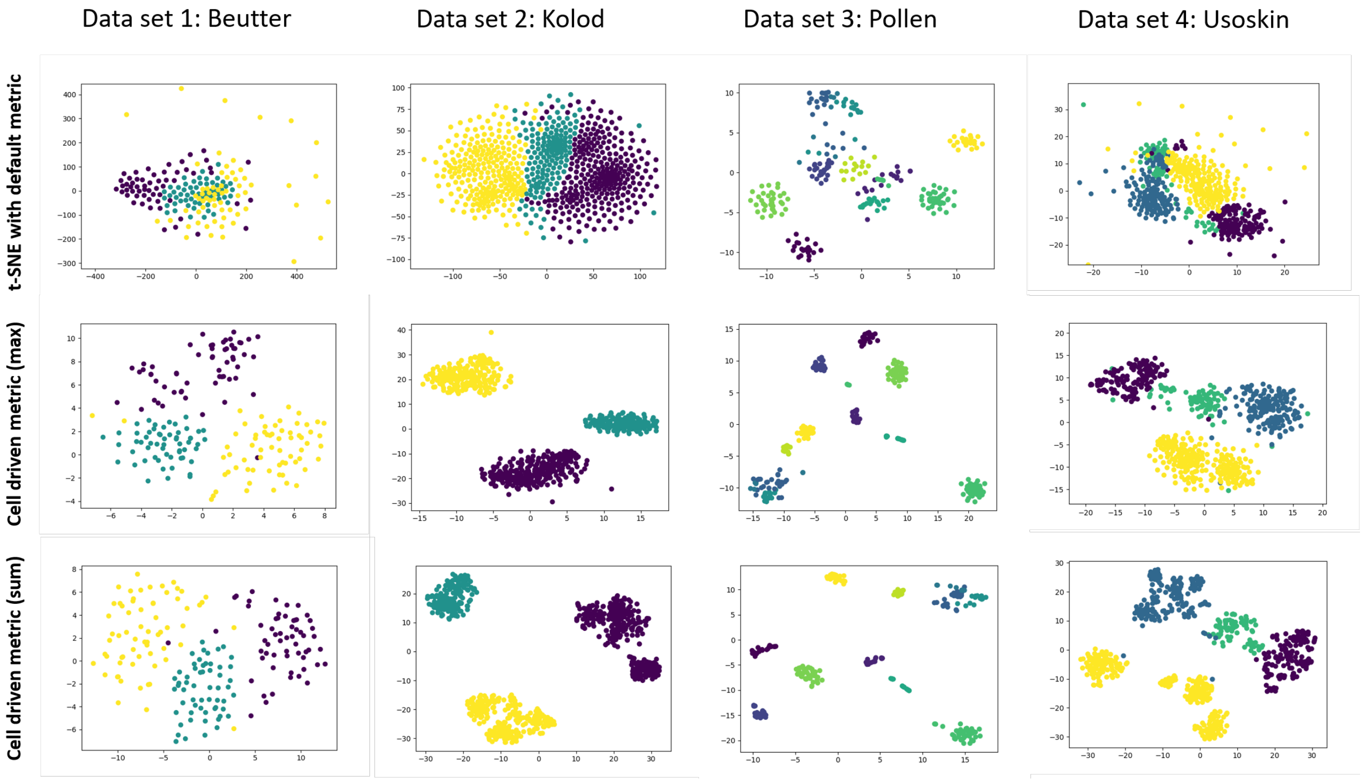

Fig 2. The comparison of the visualizations of cell-driven t-SNE and classic t-SNE with the default Euclidean metric on the four benchmark datasets that have 3, 3, 11, and 4 clusters respectively. The cell-drive t-SNE creates more meaningful clusters than the classic t-SNE in visualization for the four datasets no matter with the sum or max fusion.

4.2 Cell-driven t-SNE segregation

Since the three biological factors will contribute to the cell diversity jointly, it would be hard to rely on only one cell-driven metric to achieve the desirable results across all the datasets though they may work well for individual ones. Thus, we employ the proposed cell-driven t-SNE clustering algorithm that combines different cell-driven metrics to unveil scRNA-seq segregations.

Figure 2 shows the embedding visualizations of c-TSNE in comparison with those of the classic t-SNE on the four datasets. The proposed c-TSNE produces more meaningful embeddings that generally show correct clustering than the classic t-SNE no matter with the sum or max fusion. Furthermore, it demonstrates advantages over the t-SNE with single cell-drive metrics. For example, the Beutter, Kolod, Pollen, and Usoskin datasets all show better class bounds in visualization under c-TSNE than those under t-SNE with individual cell-driven metrics. It suggests that fusion procedure in c-TSNE contributes to improving the representative quality of the embeddings. Furthermore, it seems that max-fusion works well for the datasets with high sparsity. For example, the max-fusion produces the well-grouped four clusters for the Usoskin dataset with a 78.1% sparsity compared to the sum-fusion. On the other hand, the sum fusion seems to work well on the other datasets without high sparsity values.

Table 4 compares the NMI values from cell-driven t-SNE clustering with its peers: t-SNE, SIMILR, KPCA, and UMAP, where the final NMI values from the cell-driven t-SNE clustering are marked in bold. It shows that c-TSNE clustering shows leading and stable performance for all the four datasets. The NMIs from the t-SNE, KPCA, and UMAP clustering all demonstrate unstable performance across the four datasets, i.e., good performance can be only achieved on an individual dataset but cannot be extended to the others. It alternatively suggests biased and poor performance from using the Euclidean distance metric evaluate the cell similarity, because all the three methods employ the Euclidean metric in calculating their distances. Only the performance of the complicated and expensive SIMLR method can compare with those of the cell-driven t-SNE clustering. However, the proposed c-TSNE clustering obviously demonstrates a leading performance in comparison with SIMILR on the Kolod, Pollen, and Usoskin datasets in terms of NMI values. Additionally, the proposed c-TSNE clustering is more transparent, explainable, and implementation-friendly compared to SIMIL [8].

Table 4. The cell-driven t-SNE clustering performance and its peers' performance

Datasets	*c-TSNE*	*c-TSNE*	*t-SNE*			
	(max)	*(sum)*		*SIMLR*	*KPCA*	*UMAP*
Beutter	0.79	**0.83**	0.18	0.89	0.47	0.27
Kolod	1.0	**1.0**	0.47	0.99	0.49	0.28
Pollen	0.92	**0.97**	0.86	0.95	0.91	0.84
Usoskin	**0.87**	0.77	0.21	0.74	0.30	0.41

4.2.1 The entry-usage percentage in c-TSNE

We examine the entry-usage percentage in the final fusion matrix of c-TSNE. The entry-usage percentage is ratio between the number of entries appears in the final fusion distance matrix. For example, the entry-usage percentage for the Yule metric is defined as

$$\rho_{yule} = |\operatorname{argmax}_{yule,i,j}\left(D_{yule}(i,j), D_{fractional}(i,j), D_{L_{chebyshev}}(i,j)\right)|/n^2 \tag{12}$$

provided the max-fusion scheme is employed for a dataset. Examining the entry-usage percentage can provide more information about which metric will have more contribution to the final pairwise distance matrix of c-TSNE. Table 5 summarizes the percentages of entry-usage of the three metrics. Interestingly, Kolod, the dataset with the lowest sparsity among the four datasets, has the most important contribution from the L-chebyshev metric. It suggests the impact of the top-expressed genes on the final pairwise matrix is more important than whether some genes are expressed or not for a dataset with a low sparsity value. In contrast, Usoskin, the dataset with the largest sparsity, has the most important contribution from the Yule metric. It suggests the impact of whether the certain genes are expressed or not on the final pairwise matrix can be more important than the other two factors.

Table 5 The percentage of entry usage in the fused distance matrix of c-TSNE

Dataset	ρ_{yule}	$\rho_{fractional}$	$\rho_{L_{chebyshev}}$
Beutter	34.3	34.8	30.9
Kolod	25.6	32.9	41.5
Pollen	26.7	38.2	35.1
Usoskin	41.7	23.6	34.7

4.2 Cell-driven t-SNE clustering robustness tests

It is necessary to examine the robustness of the proposed c-TSNE clustering to validate its robustness to dropouts and noise. Because the c-TSNE clustering mainly relies on the embeddings produced from c-TSNE, it is actually the robustness test of c-TSNE. Since dropouts are a built-in nature of scRNA-seq data, it is highly likely that the sparsity of a given dataset would vary from the original one provided the sequencing were conducted the second time. Thus, it needs to answer how robust the proposed c-TSNE will under different dropout events.

On the other hand, noise can be involved in the gene expression quantification of scRNA-seq data from different sources that range from experimental design, human error, or even unknown system or biological complexity issues [14,25]. Therefore, it also needs to answer how robust the proposed c-TSNE will behave under the involvement of possible noise. These queries are rarely investigated in the previous low-dimensional methods-based scRNA-seq clustering where each dataset is assumed as the final deterministic one. However, answering these questions will help us know whether the proposed c-TSNE cand its following clustering can be applied to generic scRNA-seq datasets rather than only provide a case study.

To test the robustness of the proposed c-TSNE on behalf of single cell clustering, we conduct the robustness tests for each dataset to evaluate the impacts of dropouts and noise involvement on the c-TSNE clustering in comparison with the classic t-SNE clustering. In the robustness test for dropout, we view each dataset as the population and randomly drop varying fractions that range from 5% to 50% of the original scRNA-seq data to analyze the performance of the proposed c-TSNE clustering by observing the final change NMI values.

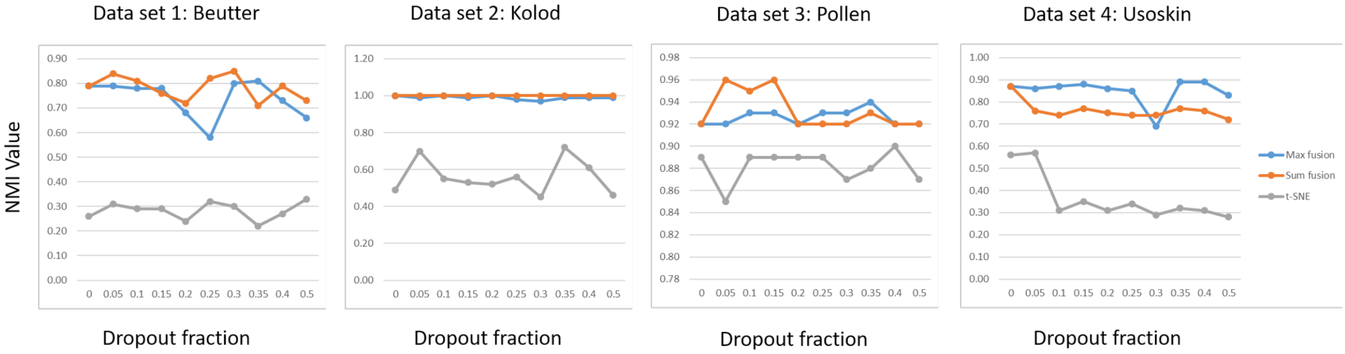

Figure 3 The robustness test of dropouts on the proposed c-TSNE clustering and the classic t-SNE clustering. The proposed c-TSNE clustering stably keeps the leading performance over the t-SNE clustering in a large margin for all the four datasets under different dropout fractions.

Figure 3 illustrates that the robustness test results on dropouts for c-TSNE and t-SNE clustering on the four datasets. It shows that proposed c-TSNE clustering always keeps the leading advantages over the t-SNE clustering in a large margin under various dropout fractions. The results suggest the robustness of the c-TSNE clustering to the dropouts, which is also the robustness of the proposed c-TSNE to the dropouts, compared to the traditional t-SNE clustering. It further implies that the proposed c-TSNE clustering can achieve robust clustering performance when extending to the other scRNA-seq datasets.

Moreover, it indicates that c-TSNE under sum-fusion generally have better robustness than c-TSNE under the max-fusion. Both show strong clustering advantages over the classic t-SNE clustering, which demonstrates high oscillations for all datasets. However, the c-TSNE clusterings under the sum-fusion and max-fusion both show good stability for the Kolod dataset that has the smallest sparsity among all datasets. The proposed method has the most stable performance for the Kolod data no matter how much the dropout fraction increases. On the other hand, the classic t-SNE clustering has a large ups and downs under the same situation.

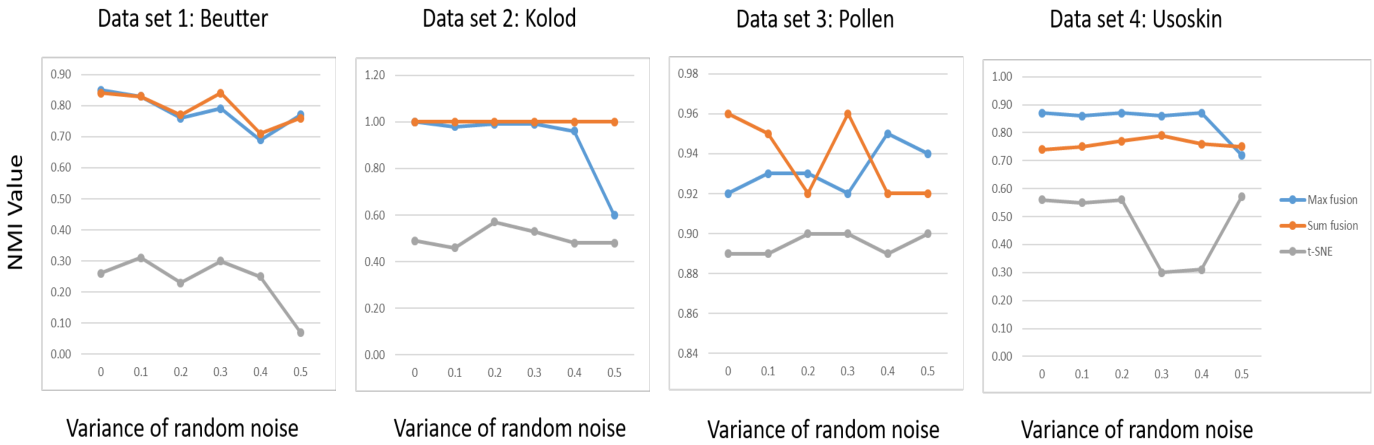

Figure 4 The robustness test of noise involvement on the proposed c-TSNE clustering and the classic t-SNE clustering. The c-TSNE clustering demonstrates consistent leading advantages over the t-SNE clustering under the different variances of random noise. The c-TSNE clustering under the sum-fusion seems to be more robust to the noise involvement than its peer.

Figure 4 illustrates the robustness test of the noise involvement on the proposed c-TSNE clustering in comparison with the t-SNE clustering in terms of NMI values. It shows that the c-TSNE clustering holds consistent leading advantages over the t-SNE clustering under the different variances of random noise, although both show oscillations. To conduct this robustness test with respect to the noise involvement, we add independent zero-mean Gaussian noise with different variance (sigma from 0.1 to 0.7) $x_{ij} = x_{ij} + N(0, \sigma_{ij})$, $0.1 \le \sigma_{ij} < 0.7$, to the original expression matrixes of scRNA-seq data.

We find no matter how the variance of random noise increases, the proposed c-TSNE clustering always leads the classic t-SNE clustering by a large margin. This is likely because the metric fusion weakens the impacts of the noise on the clustering. Furthermore, the c-TSNE clustering with the sum-fusion seems to show an advantage over the c-TSNE with max-fusion, especially for the Kolod dataset, but for the Usoskin data with high sparsity the c-TSNE with the max-fusion leads its peers. It echoes that the c-TSNE with the max-fusion should be the better choice in cell segregation for those with high sparsity values. In summary, the we highlight the robustness of the proposed c-TSNE clustering with respect to dropouts and noise, suggesting it is would perform well when applied to the other scRNA-seq datasets.

4.3 Cell-driven t-SNE clustering on additional scRNA-seq datasets

We include two additional scRNA-seq datasets: Goolam and Patel to further verify the superiority of the c-TSNE clustering [31-32]. Table 6 illustrates their basic information. The Goolam dataset contains 124 blastomeres from 5 different stages of mouse embryos [26]. In the Goolam dataset, transcriptomes were determined for all blastomeres of 28 embryos at the 2-cell, 4-cell and 8-cell (n=4) stages, and individual cells taken from 16- and 32- cell stage embryos [32]. The Patel dataset consists of 430 cells from five heterogeneous primary glioblastomas according to their transcriptional expressions [31].

Table 6. Additional two scRNA-seq datasets information

Dataset	No. of Cells	No. of Genes	No. of classes	Sparsity (%)
Patel	543	5947	5	53.3
Goolam	124	41480	5	68.6

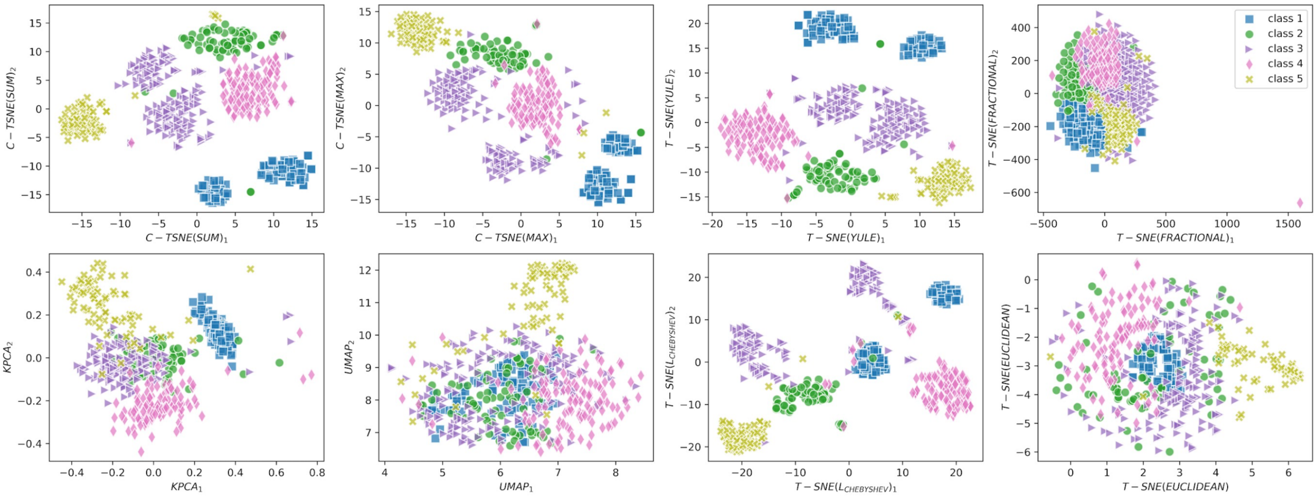

Fig 5. The comparisons of the c-TSNE visualizations with its peers: t-SNE with the three cell-driven metrics, KPCA, UMAP, and t-SNE visualizations for the Patel dataset. The c-TSNE with sum-fusion achieves the best performance for their good separations of the five clusters of samples. But t-SNE with the default Euclidean distance and UMAP have the worst cluster separations in their visualizations among all the methods. But the KPCA, t-SNE with the L-Chebyshev and fractional metrics all have decent separations though not the best ones.

Figure 5 compares the c-TSNE (with the sum-fusion and max-fusion) visualizations for the Patel data with it peers. The peer methods include t-SNE under Yule, Fractional, and L-Chebyshev metrics, KPCA, UMAP, and the classic t-SNE. It seems that the visualization from c-TSNE with the sum-fusion achieve the best performance among all the methods, the five classes are well separated compared to the other methods. According to the proposed c-TSNE clustering algorithm, we adjust the weights $w_{yule} = 0.8, w_{fractional} = 0.1, w_{L_{chebyshev}} = 0.1$ in the fusion because of the relatively high sparsity of input data. The c-TSNE with the max-fusion and the t-SNE under the Yule metric itself can achieve good separations for the five classes of samples as well. Table 7 shows their corresponding NMIs are 0.89, 0.87, and 0.86, respectively. This suggests that, for a high-sparsity dataset, whether certain genes are expressed or not can provide the most important information to group data compared to the other factors. It also explains why the t-SNE under the Yule metric achieves the second best visualization and clustering performance.

On the other hand, similar to UMAP, t-SNE under the Euclidean distance has quite poor performance, in which different groups of cells do not have clear boundaries (Fig 5). Compared to the other peers, the UMAP and t-SNE embeddings have the smallest value ranges that limit them to extract more meaningful latent data intrinsic structures as the others on the Patel dataset. Furthermore, both KPCA and t-SNE with the L-Chebyschev metric achieve relatively decent separations for the dataset with NMIs 0.75 and 0.61 respectively.

Figure 6 compares the cell-driven t-SNE (c-TSNE) visualizations for the Goolam data with it peers. Like the previous results, c-TSNE achieves the best performance. For example, the c-TSNE with the sum-fusion demonstrates the obvious advantage in visualization, in which almost all subgroups are separated well. The corresponding NMI achieves 0.79, which is slightly higher than the 0.76 NMI achieved under the t-SNE with the Yule metric. It echoes that the Yule metric can be more important than the others for the Goolam dataset with a sparsity 68.6% as the Patel dataset. Like all the other cases, t-SNE with the default Euclidean metric has the worst performance. Unlike before, t-SNE with the L-Chebyschev metric also has a poor performance, but UMAP, KPCA, and t-SNE with the fractional metric achieve relatively fair cluster separations in their visualizations, although they may not separate all groups well for the Goolam dataset.

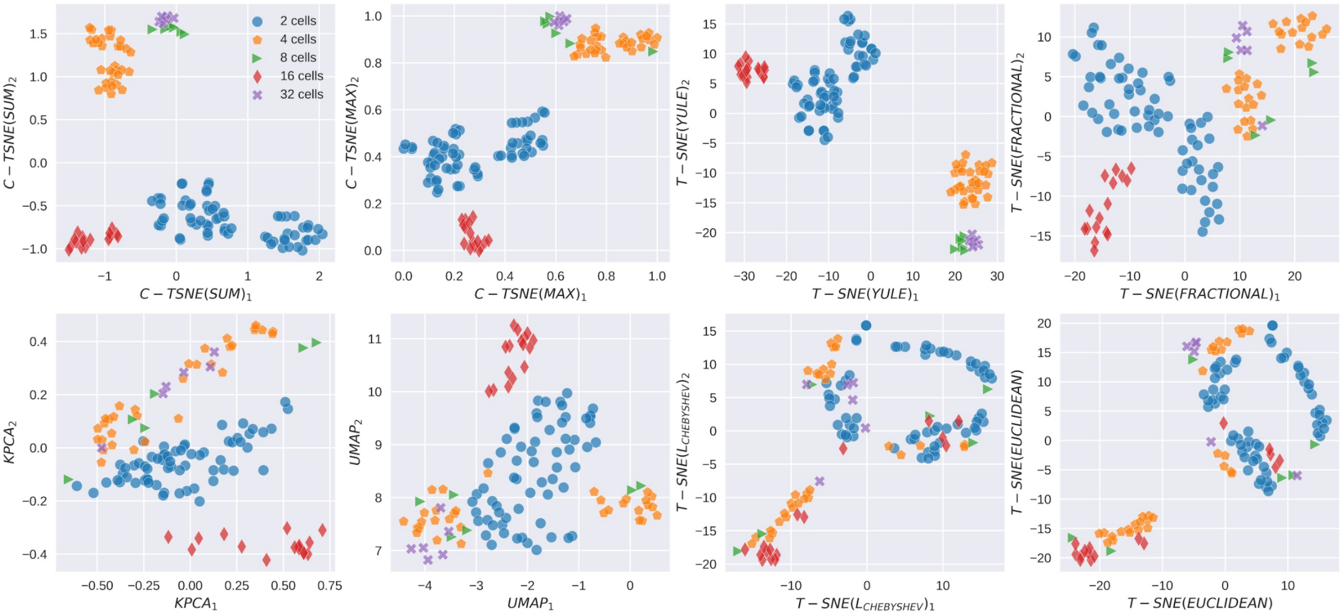

Fig 6. The comparisons of the c-TSNE visualizations with its peers: t-SNE with the three cell-driven metrics, KPCA, UMAP, and t-SNE visualizations for the Goolam dataset. The c-TSNE with the sum-fusion/max-fusion and t-SNE with the Yule metric achieve almost the same level of good cluster separations in their visualizations. The t-SNE with the default Euclidean distance and L-Chebshev metric both have the worst level of cluster separations. But KPCA, UMAP, and t-SNE with the fractional metric all have relatively fair cluster separations, especially the t-SNE with the fractional metric.

Table 7 compares the proposed c-TSNE clustering and its peers for the Patel and Goolam datasets, in which the default cell-driven t-SNE performance is in bold. Like before, the c-TSNE clustering shows the best performance and stability among its peers. We notice that the Yule metric seems to outperform other metrics in clustering with an obvious advantage because of the relatively high sparsity of input datasets. In fact, almost all the cell-driven metrics demonstrate the obviously leading advantages over the default Euclidean distance under t-SNE clustering. For example, t-SNE with the Yule, fractional, and L-Chebshev metrics achieve 0.86, 0.53, and 0.61 NMI values in the Goolam data clustering, and 0.76, 0.65, and 0.26 in the Patel data clustering respectively. The best results from the Yule metric are even better than those from the scRNA-seq clustering method SIMLR [8]. On the other hand, KPCA and UMAP fail to achieve consistently good performance. For example, UMAP achieves a fair level NMI on the Goolam data, but it provides a very poor clustering with NMI 0.27.

Table 7. The cell-driven t-SNE clustering performance and its peers' performance

Data sets	c-TSNE	c-TSNE	t-SNE	t-SNE	t-SNE	t-SNE	SIMLR	KPCA	UMAP
	(max)	(sum)	(Yule)	(fract.)	($L_{chebyshev}$)				
Patel	0.87	**0.89**	0.86	0.53	0.61	0.36	0.82	0.75	0.27
Goolam	0.65	**0.79**	0.76	0.65	0.26	0.24	0.69	0.41	0.59

To the best of our knowledge, the proposed c-TSNE clustering achieves the best performance or at least equivalent performance for all the six datasets in this study in an explainable and efficient way [33-34]. The classic t-SNE with the default Euclidean distance has the worst performance for almost all datasets. It suggests the popularity using t-SNE with the default Euclidean distance as a visualization and clustering tool in scRNA-seq analysis may not be an optimal choice. The other peer methods all demonstrate instability in achieving good cell clustering. They are unable achieve decent clustering performance for different datasets. For example, KPCA achieves 0.91 NMI for the Pollen data but only 0.41 NMI for the Goolam data. It alternatively suggests highly nonlinear and complexity of scRNA-seq data itself as well as the effectiveness and repeatability of the proposed cell-driven t-SNE.

5. Discussion

The cell-driven t-SNE (c-TSNE) provides an explainable visualization and clustering tool for scRNA-seq analysis. Unlike the traditional t-SNE, it employs the cell-driven metrics to distinguish the cell differences in the low-dimensional embedding space. The Yule, fractional, and L-chebyshve metrics model the most relevant biological factors in scRNA-seq cell differences. The fused distance matrix from the metrics models the cell similarity more representative, biologically meaningful, and easily understood and explainable.

Moreover, the proposed cell-driven t-SNE (c-TSNE) breaks the general myth about the trading-off between machine learning efficiency and explainability [11]. It strongly demonstrates that enhancing the interpretability of a machine learning model (e.g., t-SNE) can contribute to its efficiency as well. In out context, it means to improve clustering performance and robustness to dropouts and possible noise involvement. More importantly, our method stays explainable by avoiding multi-kernel learning or possible feature selection tuning that makes single cell segregation ambiguous or even a black-box procedure [8,33]. By combining different sample diversity revealed by different metrics, the proposed cell-driven t-SNE also shed light on the still unknown mechanism of cellular differentiation from an explainable machine learning perspective [3,11].

Unlike other existing scRNA-seq clustering methods, the proposed cell-driven t-SNE conducts clustering from the raw scRNA-seq data rather than normalized one by viewing the raw data as the final data [35]. It views the dropout issue in the scRNA-seq sampling as a normal procedure in sampling the whole transcriptome. It considers the sparsity as an important parameter in deciding the final pairwise matrix fusion. It is also one reason we choose the raw data rather than the normalized data or data after imputation because data may lose its original sparsity after normalization or imputation [14]. To the best of our knowledge, it is the first single cell segregation method considering the sparseness degree of input data. That the results of c-TSNE clustering are superior to its peers suggests the effectiveness of using the raw data for the sake of clustering. Alternatively, it may suggest imputation processing may not be an essential one for determining cell-to-cell difference. The recent work reported that most imputation methods had no impact on scRNA-seq analysis compared to non-imputed data [36]. It would be interesting for us to evaluate the role of normalization in scRNA-seq clustering by employing normalized data using the widely-used normalization methods (e.g., SCron) in the proposed cell-driven t-SNE (c-TSNE) from a clustering model selection perspective [12,33,36].

Furthermore, the proposed method does not rely on feature selection to determinate single cell data dimension reduction and following segregation because we believe feature selection may not be able to provide a good interpretation. In other words, we invite all transcripts involved in the single cell segregation because we believe the whole transcriptome will determine the cell segregation rather than few important genes selected from some feature selection methods. However, previous results reported that appropriate feature selection may contribute to scRNA-seq segregation [33]. It would be worthwhile to compare the proposed c-TSNE clustering with the whole transcriptome and feature selection [33,36].

c-TSNE Speedup. The weakness of the proposed cell-driven t-SNE (cTSNE) can be its high complexity though it is still implementable because most scRNA-seq datasets have their sample size <1000. However, it is desirable to decrease the complexity of c-TSNE so that it can handle datasets with a large number of cells, especially because of the increasing number of cells assayed per experiment in scRNA-seq [37,38].

We can optimize the existing complexity of the proposed c-TSNE to $O(n^2 + nlogn)$ with the following speedup techniques. The first is to use less expensive method to conduct normalization for the three different pairwise distance matrices before fusion. Without the normalization procedure for three different distance matrices, c-TSNE has almost the same complexity as the classic t-SNE: $O(nlogn)$. But the eigenvalue decomposition involved in the distance matrix normalization generally take $O(n^3)$ in practice. It is somewhat expensive to conduct such normalization by doing an eigenvalue decomposition though it does achieve good performance [16]. Using alterative less expensive normalization scaling factor $trace(D)/n$ would greatly decrease the complexity cost because the trace is the sum of the eigenvalues, but it can be calculated by the sum of all. Thus, the complexity of cell-driven t-SNE should be decreased by using this technique.

The second is to use truncated SVD to calculate the L-chebyshev distance metric. The L-chebyshev distance that needs an SVD-based reconstruction may take more costs because SVD complexity can be $O(n^3)$. However, such a weakness can be fixed by using truncated SVD that has $O(n^2)$ [39]. Thus, the optimized cell-driven t-SNE is an implementable-favor method for its $O(n^2 + nlogn)$ complexity and will work well for possible datasets with a large number of cells.

7 Conclusion

We provide an explainable t-SNE: cell-driven t-SNE (c-TSNE) for scRNA-seq data visualization and clustering. Its explainability lies in the fact it is a customized dimension reduction method designed only for high-dimensional sparse scRNA-seq data. The more explainable and biologically meaningful distance metrics are employed in c-TSNE to discriminate the cell difference rather than use the default Euclidean distance that can bring biased or meaningless results for high-dimensional scRNA-seq data. It has no intention to generalize to other datasets for bulk RNA-seq data or even for normalized or imputed scRNA-seq data. The proposed explainable t-SNE outperforms classic t-SNE and its other peer methods in meaningful visualization and following segregation in an easily understood and interpretable way. The poor performance from the classic t-SNE in visualization and clustering may suggest that t-SNE can be easily misused in the existing scRNA-seq analysis. On the other hand, the poor performance of the classic t-SNE highlights the importance of developing explainable machine learning methods in scRNA-seq analysis. The customized machine learning methods should be designed from an explainable perspective to enhance efficiency in each application domain, especially because most machine learning methods employed in biomedical data science fields are methods migrated from the other fields [40].

Furthermore, we observe the power of using the Yule metric in discriminating single cells besides the fractional and $\mathrm{L}_{\mathrm{chebyshve}}$ metrics. It is possible to employ the Yule metric alone for some datasets with high sparsity for the sake of efficiency according to the existing experimental results. However, not all high-sparsity datasets can achieve the best performance in visualization and following clustering. For example, the Usoskin dataset has the highest sparsity (78.1%) but it only achieves 0.45 NMI under the Yule metric alone. It is a possible to figure out a Yule-similar but an explainable pseudo-metric learned from machine learning models to reflect the cell similarity for the high-sparsity datasets from a nonlinear optimization [41].

Besides comparing the proposed c-TSNE clustering with other state-of-the-art peers such as SOAP [42], we plan to apply the c-TSNE to single cell classification to further validate its superiority and possible enhancement as well as compare it with the existing cell classification methods [43-44]. For example, it is possible to make the existing distance fusion more accurate and adaptive by considering the possible impacts from other factors (e.g., data entropy) in the specific classification procedure, besides the existing data sparsity because it is likely that other latent data characteristics (e.g., data entropy) can also affect fusion besides sparsity [45-46].

The existing clustering procedure for the embedding from cell-driven t-SNE and its peers is K-means. It has the advantage of simplicity and efficiency because of its explainability, but it requires the input embeddings to be convex, which may not be guaranteed each time for c-TSNE, t-SNE, or UMAP dimension reduction. Also, K-means needs to know the number of clusters in advance to have a good estimation. It may be desirable to try different clustering methods such as DBSCAN that can overcome the weakness of the K-means in the clustering quality evaluation [47]. At the same time, we are also interested in investigating to extend the proposed cell-driven t-SNE (c-TSNE) to its quantum version to exploit its quantum advantage to handle larger and more complicated scRNA-seq datasets in a fast and accurate way [48].

References

1. Huang et al. Saver: gene expression recovery for single-cell RNA sequencing. Nature Methods, 15(7):539, 2018.
2. Leng, N., Chu, L.-F., Barry, C., Li, Y., Choi, J., Li, X. Kendziorski, C. (2015). Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nature Methods*, (10), 947.
3. E. Pierson, & C. Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biology 16, 241, 2015.
4. Sun, S., Zhu, J., Ma, Y. *et al.* Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol* **20,** 269 (2019). https://doi.org/10.1186/s13059-019-1898-6
5. Townes FW, Hicks SC, Aryee MJ, Irizarry RA. Feature selection and dimension reduction for single cell RNA-Seq based on a multinomial model. BioRxiv. 2019;574574:574574
6. Linderman GC, Rachh M, Hoskins JG, Steinerberger S, Kluger Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. Nat Methods. 2019;16:243–5
7. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, Ginhoux F, Newell EW. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37:38–44.
8. Wang, et al Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. Nature Methods, 14(4), 414–416. 2017
9. Aggarwal, C., Hinneburg, A., & Keim, D. On the surprising behavior of distance metrics in high dimensional space (Vol. 1973). Springer Verlag, 2001
10. Van Der Maaten, L. , & Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research, 9, 2579–2625. 2008
11. Vilone, G, Longo, L (2020) Explainable Artificial Intelligence: a Systematic Review, arXiv:2006.00093
12. Pouyan, M. B., & Kostka, D. (2018). Random forest based similarity learning for single cell RNA sequencing data. *Bioinformatics 34*(13), i79–i88.
13. V. Li and J. Li. An accurate and robust imputation method scImpute for single-cell RNA-seq data, Nature communication, 9:997, 2018
14. G. Linderman, et al Zero-preserving imputation of scRNA-seq data using low-rank approximation, bioRxiv, 2018
15. Yule, G. On the Methods of Measuring Association Between Two Attributes. Journal of the Royal Statistical Society. 75 (6): 579–652, 1912
16. Strang G., *Introduction to Linear Algebra* (3rd ed.). Wellesley-Cambridge Press, 1998
17. Abdi, H., O'Toole, A. J., Valentin, D., & Edelman, B. (2005). DISTATIS: The Analysis of Multiple Distance Matrices. 2005 IEEE Computer Society Conference on Computer Vision & Pattern Recognition (CVPR'05), 42.
18. Li, et al Application of t-SNE to human genetic data, *Journal of Bioinformatics and Computational Biology,* 15:04, 1750017, 2017
19. Dmitry,K, Berens, P: The art of using t-SNE for single-cell transcriptomics, Nature Communications volume 10: 5416, 2019
20. Andrew Rosenberg and Julia Hirschberg: V-Measure: A conditional entropy-based external cluster evaluation measure 2007
21. Vinh, Epps, and Bailey, (2010). Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. JMLR, 2837–2854 2010
22. Kolodziejczyk, et al. Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell Stem Cell*, *17*(4), 471–485. 2015
23. Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., … Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nature Biotechnology, (2), 155.
24. Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., … West, J. A. A. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nature Biotechnology, (10), 1053.
25. Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lonnerberg, P., Lou, D., … Ernfors, P. (2015). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nature Neuroscience*, (1), 145.

26. Belkina, A. et al (2019). Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. Nature Communications, 10(1), 1-12. 2019
27. Becht et al (2019) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, Nature Biotech, **37**: 38–44 (2019)
28. Mclnnes, L, Healy, J, Melville, J: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv 2018, arXiv:1802.03426
29. Bernhard Schoelkopf, Alexander J. Smola, and Klaus-Robert Mueller. 1999. Kernel principal component analysis. In Advances in kernel methods, MIT Press, Cambridge, MA, USA 327-352
30. Han, X (2010) Nonnegative Principal component Analysis for Cancer Molecular Pattern Discovery, IEEE/ACM Transaction of Computational Biology and Bioinformatics 7:(3), p537-549
31. Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., … Bernstein, B. E. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (New York, N.Y.)*, *344*(6190), 1396–1401.
32. Goolam et al. Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. *Cell*, *165*(1), 61–74.
33. Sun Z, et al. (2017) DIMM-SC: A Dirichlet mixture model for clustering droplet-based single cell transcriptomic data. Bioinformatics 34:139–146.
34. Zhu et al (2019) Semisoft clustering of single-cell data, PNASS 2019, 116 (2) 466-471
35. Cole, M. B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., … Yosef, N. (2019). Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. Cell Systems, 8(4), 315–328.
36. Hou et al. A systematic evaluation of single-cell RNA-sequencing imputation methods. Genome Biol 21, 218 (2020)
37. Tran et al. Fast and precise single-cell data analysis using a hierarchical autoencoder. Nat Commun 12, 1029 (2021).
38. Wan, S., Kim, J. & Won, K. J. SHARP: hyperfast and accurate processing of single-cell RNA-seq data via ensemble random projection. Genome Res. 30, 205–213 (2020).
39. Kalantis et al Projection techniques to update the truncated SVD of evolving matrices, arXiv:2010.06392
40. Han, H, The challenges of explainable AI in biomedical data science, BMC Bioinformatics, 2021,DOI: 10.1186/s12859-021-04368-1
41. Xing, E. P., Ng, A. Y., Jordan, M. I., & Russell, S. (n.d.). Distance metric learning, with application to clustering with side-information. Advances in Neural Information Processing Systems.
42. Zhu et al (2019) Semisoft clustering of single-cell data, PNAS 2019, 116 (2) 466-471
43. Zhao et al (2019) Evaluation of single-cell classifiers for single-cell RNA sequencing data set, Briefings in Bioinformatics, Volume 21, Issue 5, September 2020, Pages 1581–1595
44. Alquicira-Hernandez, J., Sathe, A., Ji, H.P. et al. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome Biol 20, 264 (2019).
45. Han et al (2021): Predict High-Frequency Trading Marker via Manifold Learning, Knowledge-based system, 213:106662, 2021
46. Han and Men (2018) How does normalization impact RNA-seq disease diagnosis? JBI, 85: 78-92
47. Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS), 42(3)
48. Huang, HY., Broughton, M., Mohseni, M. et al. Power of data in quantum machine learning. Nat Commun 12, 2631 (2021)