Yousuff *et al.*
Journal of Analytical Science and Technology (2024) 15:1
https://doi.org/10.1186/s40543-023-00414-0

Journal of Analytical Science and Technology

RESEARCH ARTICLE

Open Access

Nonlinear dimensionality reduction based visualization of single-cell RNA sequencing data

Mohamed Yousuff[1], Rajasekhara Babu[1*] and Anand Rathinam[2]

Abstract

Single-cell multi-omics technology has catalyzed a transformative shift in contemporary cell biology, illuminating the nuanced relationship between genotype and phenotype. This paradigm shift hinges on the understanding that while genomic structures remain uniform across cells within an organism, the expression patterns dictate physiological traits. Leveraging high throughput sequencing, single-cell RNA sequencing (scRNA-seq) has emerged as a powerful tool, enabling comprehensive transcriptomic analysis at unprecedented resolution. This paper navigates through a landscape of dimensionality reduction techniques essential for distilling meaningful insights from the scRNA-seq datasets. Notably, while foundational, Principal Component Analysis may fall short of capturing the intricacies of diverse cell types. In response, nonlinear techniques have garnered traction, offering a more nuanced portrayal of cellular relationships. Among these, Pairwise Controlled Manifold Approximation Projection (PaCMAP) stands out for its capacity to preserve local and global structures. We present an augmented iteration, Compactness Preservation Pairwise Controlled Manifold Approximation Projection (CP-PaCMAP), a novel advancement for scRNA-seq data visualization. Employing benchmark datasets from critical human organs, we demonstrate the superior efficacy of CP-PaCMAP in preserving compactness, offering a pivotal breakthrough for enhanced classification and clustering in scRNA-seq analysis. A comprehensive suite of metrics, including Trustworthiness, Continuity, Mathew Correlation Coefficient, and Mantel test, collectively validate the fidelity and utility of proposed and existing techniques. These metrics provide a multi-dimensional evaluation, elucidating the performance of CP-PaCMAP compared to other dimensionality reduction techniques.

Keywords Dimensionality reduction, Compactness preservation, Machine learning, Single-cell RNA sequencing, scRNA-seq data visualization, Single cell data analysis

Introduction

It is a widely accepted and proven scientific fact that cells are the fundamental building blocks of all living organisms. They play a vital role in the structure and function of these organisms. In recent years, there has been a significant shift in cell biology research due to the development of single-cell multi-omics technology. Despite the fact that the genome structure of every cell in a given individual is essentially the same, the expression pattern of this genome determines the cell's physiological characteristics. The diverse range of physical traits observed in different organisms is a result of both the genotype and the expression pattern of the genome, and deviations from the norm in these patterns can lead to various diseases. To fully understand the relationship between genotype and phenotype, it is necessary to analyze transcriptomic information at a high resolution, and advances in high throughput sequencing technologies have made it possible to do so at the level of single cell (Nayak and Hasija 2021; Battenberg et al. 2022).

*Correspondence:
Rajasekhara Babu
mrajasekharababu@vit.ac.in
[1] School of Computer Science and Engineering, Vellore Institute of Technology, Vellore Campus, Vellore, Tamilnadu 632014, India
[2] Abercrombie & Fitch, Florida 32256, USA

Recent single-cell RNA sequencing (scRNA-seq) technologies can create data for multitudes of cells in a single experiment, a portion of which are open to the public over the internet. This surge in throughput has allowed researchers to use scRNA-seq for a wide variety of tissues and even whole organisms (Ghazanfar et al. 2016). As the technology advances, it is anticipated that scRNA-seq will become more precise, dependable, and cost-effective per cell, making it possible for a vast array of studies. scRNA-seq has unleashed a plethora of opportunities in biomedical research; however, we have only touched a small portion of the possibilities of such a huge and varied dataset (Wang et al. 2021). scRNA-seq transcriptome profiles have opened up the possibility for recognition of unusual and peculiar cell types in organs or tissues, resolving the fate of a cell (Grün et al. 2015), cell lineage connections in early stages of development (Petropoulos et al. 2016), differentiating normal and abnormal cells (Shalek et al. 2013), antigen sensitivity and specificity of immune cells (Tu et al. 2019), deducing cellular trajectory (Miragaia et al. 2019), finding regulatory signatures in malignant tumors (Granja et al. 2019), decoding immune repertoire for contagious diseases (Yao et al. 2019), knowing and interpreting tumor heterogeneity (Wagner et al. 2019), enlightening the pathway for drug resistance and various stages of cancer treatment including relapse of tumors (Shaffer et al. 2017). More applications are being uncovered as a result of improved analysis techniques.

The data collected from hundreds of thousands of cells, each with numerous genes, results in a dataset with a large number of data points and high dimensionality. While this vast amount of data has the potential to reveal valuable insights, extracting useful information from it can be difficult (Babjac et al. 2022). To address this challenge, Dimensionality Reduction (DR) techniques have been developed to simplify the data and create lower-dimensional representations that are easier to understand and interpret. DR methods involve reconstructing the underlying distributions of the data in the "gene space" and providing a more intuitive way to analyze single-cell data. Researchers are seeking ways to represent high-dimensional scRNA-seq datasets in a Low Dimensional Space (LDS) while preserving patient similarities and differences (Xiang et al. 2021).

The goal is to create an LDS representation that captures the relationships between patients, such that those with the same disease have similar patterns of expression. DR techniques are used to map High Dimensional Space (HDS) data to a 2-dimensional (2D) or 3-dimensional (3D) space, which makes it easier to visualize connections between data points that would be difficult or impossible to identify in the HDS (Carter et al. 2008; Yousuff and Babu 2022). The key principle of the DR approach is that it maintains the proximity of similar data points and keeps distant data points separated. Retention of local structure refers to maintaining the proximity of elements in the HDS in the LDS. In broader terms, the local structure is maintained when the neighboring elements in the HDS correspond to those in the LDS. On the other hand, preserving global structure implies maintaining relationships between clusters and larger-scale structures (Heiser and Lau 2020).

Principal Component Analysis (PCA), a linear DR technique, is commonly used in unsupervised data reduction by identifying linear feature combinations that have the highest variance. However, linear DR approaches are not always reliable for scRNA-seq analysis as they may not fully capture the complexity of diverse cell types and can result in an inadequate representation of the data (Tsuyuzaki et al. 2020). In contrast, nonlinear DR techniques have become popular for scRNA-seq data visualization because of their ability to identify both local and global patterns while avoiding coordinate overlap. These techniques are particularly useful for scRNA-seq data, which is often highly diverse and has complex associations between cell types and states. Additionally, nonlinear DR techniques are more effective in reducing the dimensionality of scRNA-seq data with many features per cell (Pierson and Yau 2015).

Several nonlinear dimensionality reduction algorithms have been proposed for visualizing and generating LDS for scRNA-seq data. Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018), t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton 2008), TriMap (Amid and Warmuth 2019), Potential of Heat-diffusion for Affinity-based Trajectory Embedding (PHATE) (Moon et al. 2019), and IVIS (Szubert et al. 2019) are commonly used among these algorithms. Each of these methods has limitations; for example, t-SNE is sensitive to the perplexity hyperparameter and may create clusters that are not real, t-SNE and UMAP are good at retaining local structures but have difficulty maintaining global structures. TriMap is a triplet model to reach the performance of UMAP and t-SNE, but it also has limitations; at times, it struggles with preserving local structures. Additionally, it is not possible to regulate these techniques, such as t-SNE, UMAP, or TriMap, effortlessly from local to global structure retention through any apparent modification of parameters (Coenen et al. 2019; Wang et al. 2022). PHATE is also a recently proposed alternative approach, but it is sensitive to initialization values, and it is liable to serious deformations when attempting to maintain pairwise associations or distances from HDS data in 2D or 3D (Moon et al. 2019). IVIS, on the other hand, uses Siamese neural

networks, which can lead to high computational cost, limited interpretability, confined ability to handle variations, limited scalability and the need for a large amount of labeled data for effective training (Chicco and Cartwright 2021).

Selecting which points to attract and which to repel is crucial in maintaining both local and global structures. Pairwise Controlled Manifold Approximation Projection (PaCMAP) is a recent nonlinear DR algorithm that claims to achieve this by using a unique loss function and graph components. PaCMAP is demonstrated on synthetic, benchmark and real-time datasets and it has been proven to preserve local and global structures. It is quite reliable in hyperparameter choices and exhibit considerably faster runtime compare to other DR algorithms (Wang et al. 2022). This paper aims to present an augmented version of PaCMAP termed as Compactness Preservation Pairwise Controlled Manifold Approximation Projection (CP-PaCMAP) which can additionally preserve compactness property of HDS datapoints into LDS. CP-PaCMAP is remarkable in order to visualize scRNA-seq data. Further, the LDS obtained through CP-PaCMAP can be effectively utilized for better classification or clustering of scRNA-seq data.

Research gap

In spite of the vast potential inherent in scRNA-seq data, the colossal size and soaring dimensionality of these datasets introduce formidable hurdles. The quest for gleaning meaningful insights from such data has spurred the evolution of dimensionality reduction (DR) techniques. DR methodologies strive to reshape the high-dimensional landscape of gene expression into a more manageable, lower-dimensional form, facilitating streamlined analysis and visualization (Babjac et al. 2022). While Principal Component Analysis (PCA), a linear DR approach, has enjoyed widespread use, its applicability in scRNA-seq investigations is somewhat constrained, as it may fall short of encapsulating the full spectrum of cell diversity (Tsuyuzaki et al. 2020). Consequently, nonlinear DR techniques have risen in prominence within the realm of scRNA-seq data, primarily due to their adeptness in unveiling both local and global patterns within data characterized by intricate relationships among cell types and states (Pierson and Yau 2015).

Nonetheless, the prevailing nonlinear DR techniques, including UMAP, t-SNE, TriMap, PHATE, and IVIS, exhibit variable degrees of sensitivity to hyperparameters and encounter obstacles in preserving both local and global data structures (McInnes et al. 2018; Maaten and Hinton 2008; Amid and Warmuth 2019; Moon et al. 2019; Szubert et al. 2019). The specific research gap targeted by this study comes to the forefront: the demand for an enhanced nonlinear DR methodology tailored for scRNA-seq data analysis. This method should exhibit unwavering proficiency in effectively capturing both local and global data structures while concurrently preserving compactness, offering a holistic solution to the challenges presented by high-dimensional single-cell transcriptomics data.

Materials and methods

This section will discuss a comprehensive overview of the scRNA-seq datasets utilized in this study. We will also describe the preprocessing procedures carried out on the datasets to ensure the quality and reliability of the data. Finally, details about the proposed CP-PaCMAP approach are presented.

scRNA-seq data collection and preprocessing

Benchmark scRNA-seq datasets belonging to three vital human organs, the pancreas, skeletal muscles, and heart, are gathered and used in this study and the dataset are available from https://hemberg-lab.github.io/scRNA.seq.datasets/. The human pancreas dataset consists of 16,382 cells and 19,093 genes from 14 different classes of cells. The human skeletal muscle dataset contains 52,825 cells and 33,538 genes belonging to 8 unique categories of cells. A set of 38,929 cells and 27,420 genes categorized under 13 labels of cells of the human heart are present in the third dataset. Initially, all the datasets are subjected to a doublets removal process. Then, other preprocessing techniques, such as filtering, quality control, and normalization, are utilized to prepare the data for nonlinear DR and LDS visualization. All the preprocessing tasks are carried out in Python language using the Scanpy library (Wolf et al. 2018).

Doublets in scRNA-seq data indicate two separate cells combined by unexpected events during the sequencing procedure. In a droplet-based sequencing approach, this can occur if, for instance, two cells reside in the same droplet. Doublets can significantly influence the processing of scRNA-seq data, leading to skewed results and inaccurate inferences. This is due to the fact that the combined gene expression readings of the doublets do not adequately represent the genuine gene expression of either individual cell (Weber et al. 2021). Therefore, identifying and eliminating doublets from single-cell data is essential before undertaking subsequent analysis. This ensures that the results of the study are based on reliable and representative measurements of individual cells rather than on measurements of cells that have been artificially blended.

Single-Cell Variational Inference (SCVI) is a method that can be used to model and analyze scRNA-seq data. SCVI is based on a variational autoencoder architecture consisting of two main components: an encoder network and a decoder network. The encoder network maps each cell's HDS gene expression data to an LDS, while the decoder network maps the LDS representation back to the actual HDS. The training process of SCVI involves minimizing a reconstruction loss, which measures the difference between the input data and the reconstructed data generated by the decoder network. In addition, SCVI uses a regularization term in the loss function to encourage the learned latent representation to be smooth and continuous and to prevent overfitting to the training data (Gayoso et al. 2022). We have incorporated SCVI to identify doublets by calculating the reconstruction error of each cell in the data and setting a threshold based on this error. Cells with a high reconstruction error are considered doublets and can be removed from the data before further analysis. Table 1 displays the number of doublets identified from each dataset.

Filtering is a crucial preprocessing step in analyzing scRNA-seq data because it helps eliminate low-quality or undesirable cells and low-quality genes or irrelevant features. This can enhance subsequent analysis and improve the precision of the results. Moreover, by deleting redundant data points, filtering might lower the computing load of downstream analysis (McCarthy et al. 2017). A general filtering criterion for cells is given in Eq. 1, whereas filtering criteria for each dataset with specific values are given in Table 1.

$$C\prime = \{\text{cinC} : \text{n(c)} \geq \text{X}\} \tag{1}$$

where C is the set of all cells in the dataset, $C\prime$ is the filtered set of cells with at least X expressed genes, $n(c)$ is the number of expressed genes in cell c, and the colon (:) represents a filter operation that retains only the cells that meet the specified criteria. Let Z be the gene expression matrix, where each row corresponds to a cell and each column corresponds to a gene. The element $Z[i, j]$ represents the expression level of gene j in cell i. To remove genes that are found in fewer than Y cells, we have applied a filter based on the number of non-zero entries in each column of Z. Let $M[j]$ be the number of non-zero entries in column j of Z, i.e., the number of cells where gene j is expressed. Then, it can be defined as a filtered gene expression matrix $Z\prime$ as given in Eq. 2, i.e., $Z\prime$ consists of the columns of Z where the corresponding gene is expressed in at least Y cells. The colon (:) symbol in Eq. 2 is a notation for all the rows and columns of Z which satisfies the condition.

$$Z\prime = Z[:, (M >= Y)] \tag{2}$$

Our scRNA-seq data were meticulously obtained using cutting-edge sequencing platforms to guarantee exceptional data quality and reliability. The iPSC and TMWC libraries were sequenced on an Illumina NextSeq 500 platform, employing a 150-cycle NextSeq High Output Reagent Kit v2.5. The sequencing protocol consisted of specific parameters: 26 bp for Read 1, 8 bp for the Index, and 98 bp for Read 2. The sequencing process on the NextSeq 500 platform was managed by the skilled team at the Institute of Molecular Bioscience Sequencing Core Facility.

Furthermore, the two PBMC libraries underwent sequencing on the Illumina NovaSeq 6000 instrument, featuring a 2×150 cycle S4 flow cell, operating in standalone mode. The libraries were loaded at a concentration of 8 nM, with each sample having a volume of 350 µL. The proficient execution of the NovaSeq 6000 sequencing procedure was carried out by the Kinghorn Centre for Clinical Genomics Sequencing Core Facility.

Libraries generated using the 10×Genomics Chromium system underwent a critical conversion process employing the MGIEasy Universal Library Conversion kit (App-A) before being sequenced on the MGISEQ-2000 instrument. For each library, 10 ng of material underwent 10 cycles of polymerase chain reaction (PCR) to introduce a 5' phosphorylation exclusively on the forward strand. Following this, the purified PCR

Table 1 Description of datasets including number of singlets, doublets and cells filtering criteria

scRNA-seq dataset	No. of cells	No. of genes	No. of cell type	No. of singlet	No. of doublet	Filtering criterion
Human pancreas	16,382	19,093	14	16,373	9	$C\prime = \{\text{cinC} : \text{n(c)} \geq 200\}$ $Z\prime = Z[:, (M \geq 10)]$
Human skeletal muscle	52,825	33,538	8	52,743	82	$C\prime = \{\text{cinC} : \text{n(c)} \geq 100\}$ $Z\prime = Z[:, (M \geq 10)]$
Human heart	38,929	27,420	13	37,450	1479	$C\prime = \{\text{cinC} : \text{n(c)} \geq 200\}$ $Z\prime = Z[:, (M \geq 10)]$

product was subjected to denaturation, after which it was combined with a 'splint' oligonucleotide. This oligonucleotide is homologous to the P5 and P7 adapter regions of the library, facilitating the formation of a circular single-stranded DNA molecule. A ligase reaction was subsequently carried out to produce a complete single-stranded DNA circle of the forward strand. An exonuclease digestion step was executed to remove single-stranded non-circularized DNA molecules. The circular single-stranded DNA molecules then underwent Rolling Circle Amplification (RCA), generating 300–500 precise copies of the libraries, forming DNA Nanoballs (DNB). Each DNB library was loaded onto a 1500 M feature patterned array flow cell in preparation for sequencing, utilizing the MGISEQ-2000RS High-Throughput Sequencing Set (App-A). The sequencing process entailed 26 bp for Read 1 and 100 bp for Read 2 cycles, without an index barcode read, as only one sample was run per flow cell. FASTQ files were locally generated on the instrument, and sequencing was expertly conducted at the BGI Shenzhen, MGI R&D facility.

Filtering out mitochondrial and ribosomal genes can enhance the reliability of scRNA-seq data, as high expression levels of these genes can signal poor data quality caused by technical issues like mitochondrial stress or cell lysis. In addition to reducing technical differences between cells, the removal of these genes can also improve downstream analysis and interpretation (McCarthy et al. 2017). Owing to the stochastic nature of RNA sequencing, various cells in a collection may have differing degrees of RNA sequenced, resulting in varying total read counts per cell. Normalization aids in compensating for these variations in sequencing depth by scaling the gene expression values for each cell by a

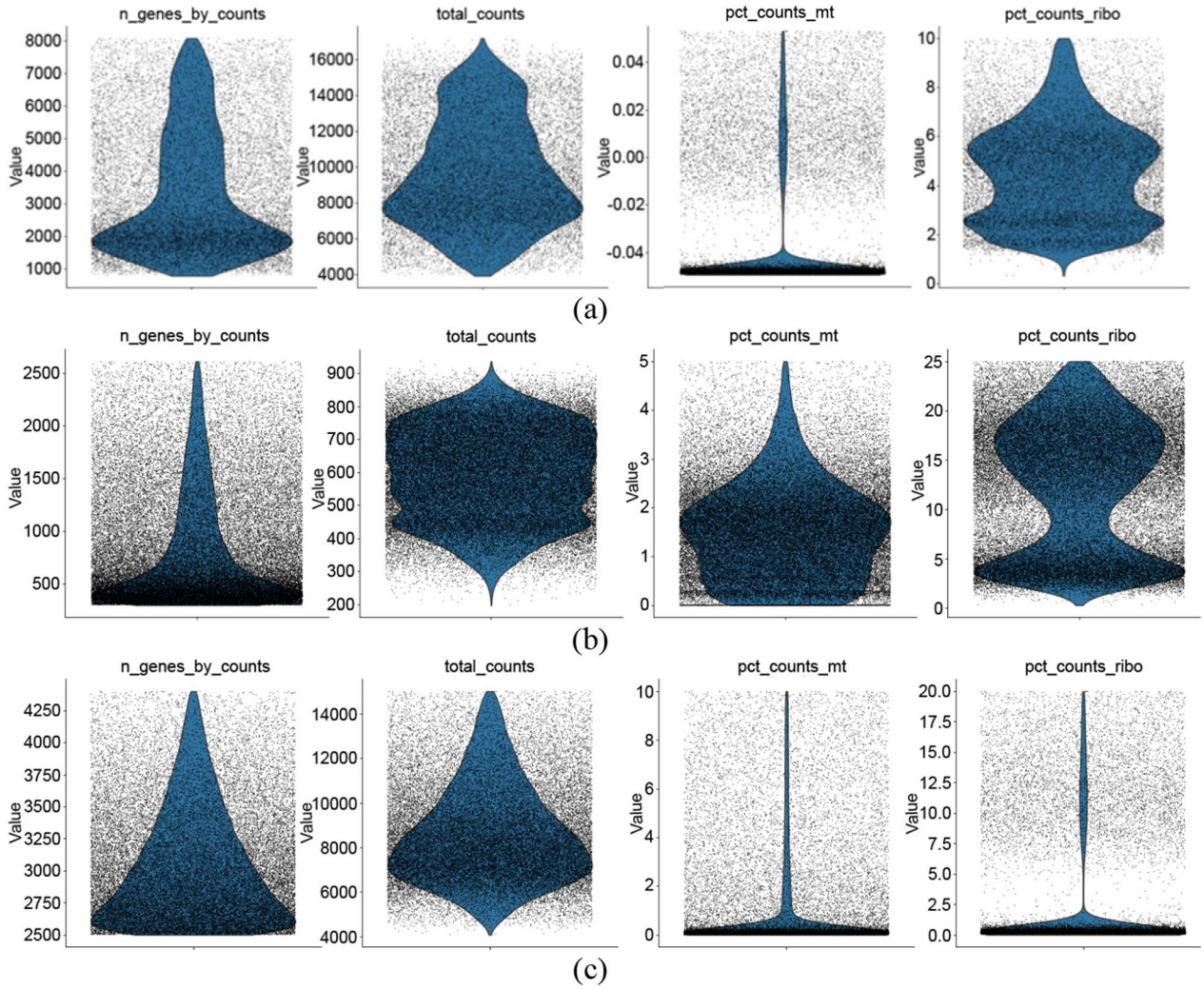


Fig. 1 Violin plot of preprocessed data to visualize the distribution of four metrics across the cells in three scRNA-seq human tissue datasets: **a** pancreas, **b** skeletal muscle, **c** heart

factor that corresponds to the total amount of reads for that cell (Vallejos et al. 2017).

Data derived from scRNA-seq may be vulnerable to technical biases such as batch effects, variances in cell capture efficiency, or gene-specific effects such as amplification bias or content bias. Normalization can assist in accounting for these technological biases, allowing for a more accurate comparison of gene expression levels across cells. During sample preparation, sequencing, and data processing, technical noise can be created during scRNA-seq. By leveling the gene expression data of each cell, scaling can lessen the influence of technical noise. Scaling can enhance the display and clustering of scRNA-seq data by lowering the influence of genes with high expression values, which can control the analysis and obscure the signal from other genes with lower expression levels (Lytal and Ran 2020). Figure 1 depicts the preprocessed data in terms of a violin plot, explaining the distribution of four metrics across the cells in all three scRNA-seq datasets. The four metrics are: (i) the number of genes detected in each cell based on read counts, (ii) the total number of reads sequenced for each cell, (iii) the percentage of reads mapped to mitochondrial genes for each cell, (iv) the percentage of reads mapped to ribosomal genes for each cell. The y-axis of the plot shows the distribution of the metric values, with the width of the violin indicating the density of cells at that value.

Methodology

The loss function regulates the attractive and repulsive forces between each pair of data points; thus, fine-tuning the loss function helps to maintain local structure. The PaCMAP aims to bring together the neighbors from the HDS in the LDS and push away further points in the original space in the LDS. Specifically, it highlights the significance of having forces on non-neighbors. PaCMAP algorithm prioritizes global structure: neighbors and mid-near pairings are attracted, whereas distant points are repelled. After the global structure is in place, the attractive force on mid-near edges reduces, stabilizes, and eventually vanishes over time, leaving the algorithm to fine-tune the local structure. PaCMAP has a primary objective with three kinds of pairwise loss elements, each related to a certain kind of graph section: nearest neighbor edges (NE), mid-near edges (ME), and repulsion edges with additional points (RE) (Wang et al. 2022). The loss function of PaCMAP is given Eq. 3.

$$Loss^{PaCMAP} = \omega_{NE}.Loss_{NE} + \omega_{ME}.Loss_{ME} + \omega_{RE}.Loss_{RE} \quad (3)$$

where

$$Loss_{NE} = \sum_{i,N} \frac{đ_{iN}}{10 + đ_{iN}}, Loss_{ME} = \sum_{i,M} \frac{đ_{iM}}{10000 + đ_{iM}}, Loss_{RE} = \sum_{i,R} \frac{1}{1 + đ_{iR}}$$

and $đ_{pq} = d_{pq}^2 + 1 = \| x_p - x_q \|^2 + 1$. The edges are additionally weighted by the coefficients ω_{NE}, ω_{ME}, and ω_{RE}, which collectively represent the total loss. As part of the optimization process, the weights are dynamically adjusted. The Student's t-distribution utilized in the similarity functions of t-SNE and TriMap is the reason for the decision to employ the scaled distance đ (Wang et al. 2022).

UMAP employs a binary search for the scale of each point, comparable to t- SNE, which utilizes entropy as perplexity for a similar search. UMAP and t-SNE imply that data points are distributed uniformly on an inherent LDS manifold since the search makes the neighborhoods of several data points behave identically. PaCMAP discards the data compactness surrounding each point by nullifying the influence of compactness with the search for scales of data points. CP-PaCMAP regularizes the cost function of PaCMAP to account for and return the compactness information surrounding each data point. Empirical evidence demonstrates that this incorporation of compactness information yields a remarkable embedding despite requiring additional calculation for the regularization element. If a data point's neighbors are relatively close, the surrounding area is compact for that point. Consequently, the local radius, determined as the mean distance between neighbors, can serve as a measure of local compactness.

A method for producing LDS that preserves compactness information at individual data points is proposed. This is achieved by defining a local radius, which formalizes the concept of spatial compactness. The proximity of nearest neighbors is often used to determine whether a data point belongs to a compact or sparse region. Specifically, a data point is considered to be in a compact area if its nearest neighbors are in close proximity to it. In contrast, a data point is deemed to be in a sparse area if its nearest neighbors are located at a considerable distance from it. The level of compactness for a given data point is determined by utilizing the average distance to nearest neighbors. In order to formalize this

concept, it is necessary to have two elements for a given data point a_i. The proposed methodology involves the use of a pairwise distance function, indicated as $d(a_i, a_j)$, and a probability distribution, denoted as $p_{j|i}$, which assigns weights to each data point a_j depending on its distance from a_i. The weights assigned to distant points are comparatively lower than those assigned to nearby points. The local radius at a given data point a_i, represented as $C_p(a_i)$, is defined as the expected value of the distance function on all other data points a_j, with respect to the conditional probability $p_{j|i}$. This measure effectively captures the average distance between a_i and its neighboring points as given in Eq. 4. The CP-PaCMAP approach leverage the probability distributions of PaCMAP, which is capable of capturing local associations. To determine the local radius in the input HDS, we perform a renormalization of the edge probabilities H_{ij}. To obtain a conditional distribution $p_{j|i}$, $H_{ij}/\Sigma_{j=1}^{N} H_{ij}$ can be calculated, and then determine the local radius as given in Eq. 5.

$$C_p(a_i) := \mathbb{E}_{j\sim p_{ij}}\left[\| a_i - a_j\|^2\right] \tag{4}$$

$$C_H(a_i) = \frac{1}{\sum_{j=1}^{N} H_{ij}} \sum_{j=1}^{N} H_{ij} \| a_i - a_j\|^2 \tag{5}$$

Subsequently, the local radius is determined within the LDS. Let b_i denote the embedding coordinates of data point a_i. A distribution that is corresponding to H is required to compute the probable distance between b_i and its neighboring data points in the LDS. It is appropriate for the distribution in examining to possess adaptive length-scales analogous to those of H. This is necessary to ensure that a consistent number of nearest neighbors are incorporated in the calculation of the local radius at various data points within the dataset. The variable L is indicative of a total average across various length-scales. By defining $p_{j|i}$ as $L_{ij}/\Sigma_{j=1}^{N} L_{ij}$ and $d(b_i, b_j)$ as $\| b_i - b_j\|^2$, the local radius in the LDS can be determined using Eq. 6.

$$C_L(b_i) = \frac{1}{\sum_{j=1}^{N} L_{ij}} \sum_{j=1}^{N} L_{ij} \| b_i - b_j\|^2 \tag{6}$$

Let us consider a data point from the input high (H) dimensional data $x \in \mathbb{R}^H$ with K neighborhood data points uniformly distributed in a sphere of radius ζ_H and volume $v \propto \zeta_H^H$. Both structure and compactness should be preserved in LDS ($L < H$), this implies that x and its neighbors should be mapped to an L-dimensional sphere of uniform density with radius ζ_L, and, to retain the compactness of x's K-neighborhood, the volume of the L-dimensional sphere should also remain as v such that $v \propto \zeta_L^L$, this indicates that ζ_L and ζ_H have a power law association i.e. $\zeta_L \propto \zeta_H^{H-L}$. Applying logarithms will result in $\log\zeta_L = (\mathrm{H} - \mathrm{L})\log\zeta_H + \beta$ `for some values of` β. Driven by the exponential scaling of compactness with regards to dimensionality of the feature vectors, we seek for a power law association between the local radius in the input HDS dataset and in the output LDS for some hyperparameters α and β in order to retain the compactness. This is reformulated as an affinal connection between the logarithms of the local radii as given in Eq. 7.

$$C_L(b_i) = \alpha(C_H(a_i))^{\beta} \Rightarrow c_L^i = \beta c_H^i + \gamma \tag{7}$$

where $c_L^i = \ln(C_L(b_i)), c_H^i = \ln(C_H(b_i))$, and $\gamma = \ln(\alpha)$. Our compactness retention objective is to select the LDS in such a way that the correlation between the logarithmic local radii of the input HDS data points and the output LDS is maximized. This method basically resembles canonical correlation analysis (Andrew et al. 2013). Thus, it can be stated that there exists an affine relationship between the logarithms of local compactness. Correlation serves as a means of measuring linear or affine interdependence; therefore, the correlation of the logarithms of local compactness is implemented as given in Eq. 8, whereas the covariance and variance of compactness can be computed using Eq. 9 and Eq. 10, respectively.

$$Corr(c_L, c_H) = \frac{Cov(c_L, c_H)}{\sqrt{Var(c_L)Var(c_H)}} \tag{8}$$

$$Cov(c_L, c_H) = \frac{1}{n-1}\sum_{i=0}^{n}\left[(c_L^i - \mu_L)((c_H^i - \mu_H)\right] \tag{9}$$

$$\begin{aligned} Var(c_L) &= \frac{1}{n-1}\sum_{i=0}^{n}(c_L^i - \mu_L)^2, Var(c_H) \\ &= \frac{1}{n-1}\sum_{i=0}^{n}(c_H^i - \mu_H)^2 \end{aligned} \tag{10}$$

where $\mu_L = (1/n)\Sigma_{j=1}^{N} c_L^j$, $\mu_H = (1/n)\Sigma_{j=1}^{N} c_H^j$. The PaCMAP's cost function is regularized by maximizing the correlation of local compactness to create the CP-PaCMAP's cost function, which needs to be minimized. The CP-PaCMAP algorithm is given in Algorithm 1 and its loss function is stated in Eq. 11. η is the regularization parameter that weights the correlation in respect to the initial cost of the PaCMAP. Similar to PaCMAP, CP-PaCMAP optimizes via stochastic gradient descent

Algorithm 1 Compactness Preservation Pairwise Controlled Manifold Approximation Projection.

Definition:

H - HDS input observations matrix.
n_{NBR} - the number of neighbor data points (default $n_{NBR} = 15$).
Ratios - Ratio between the number of mid-near pairs, further pairs and average distant pairs (default: $ME\_ratio = 0.5, RE\_ratio = 2, AE\_ratio = 0.5$).
n_{iters} - the number of gradient steps (default $n_{iters} = 450$).
$init$ - initializing method for the LDS (default $init = PCA$).
L - LDS data matrix.

Execution:

for $i = 1\ to\ N$ do
Construct neighbor edges by computing the n_{NBR} of H_i.
Construct $n_{ME} = \lfloor n_{NBR} * ME\_ratio \rfloor$ mid-near pairs.
Construct $n_{FE} = \lfloor n_{NBR} * FE\_ratio \rfloor$ further pairs by sampling non-neighbor points.
Construct $n_{AE} = \lfloor n_{NBR} * AE\_ratio \rfloor$ average distant pairs.
end for
Implement the initializing method ($init$) to determine the preliminary values of L.
Run AdamOptimizer for n_{iters} to optimize the cost function $Loss^{CP-PaCMAP}$,while parallelly updating the weights.
return L.

$$Loss^{CP-PaCMAP} = Loss^{PaCMAP} - \eta Corr(c_L, c_H) \quad (11)$$

To ensure transparency and reproducibility, we are committed to providing details of the specific non-default parameters employed in our methodology. Below, we outline the key non-default parameters along with their values:

Number of Neighbor Data Points (n_NBR):
Default value: 15
Used value: 20

Ratios for mid-near pairs, further pairs, and average distant pairs (ME_ratio, RE_ratio, AE_ratio):
Default values: ME_ratio = 0.5, RE_ratio = 2, AE_ratio = 0.5
Used values: ME_ratio = 0.6, RE_ratio = 1.5, AE_ratio = 0.6

Number of Gradient Steps (n_iters):
Default value: 450
Used value: 600

Initializing Method for the LDS (init):
Default value: PCA
Used value: Random

Regularization Parameter (η) for CP-PaCMAP:
Default value: 0.01
Used value: 0.005

Experiments

DR techniques such as UMAP, TRIMAP, and t-SNE are very commonly used for scRNA-seq data visualization. PaCMAP is a recently proposed approach to visualize high-dimensional feature vectors. PHATE and IVIS are occasionally used for data visualization. We have considered all these techniques in our study to compare with the proposed CP-PaCMAP approach. Initially, all the DR techniques are implemented on 2D-generated data to understand the necessity and idea behind compactness preservation. The 2D data is generated in such a way that it contains linear data points belonging to 4 class labels. Each category is meant to hold a different degree of compactness within the cluster, as shown in Fig. 2a. The base cluster of data points is very compactly placed compared to the second cluster. The third cluster contains a lesser level of compactness compared to the second. The fourth cluster is the one with more sparse data points.

The DR techniques are applied to the 2D generated data, and their corresponding 2D embeddings are depicted in Fig. 2, 3, 4 and 5. PaCMAP and UMAP visualization shown in Fig. 2b, c clearly prove that both the local structure and global structure of the original data are well preserved in LDS. TRIMAP and IVIS are able to retain the global structure but slightly struggle

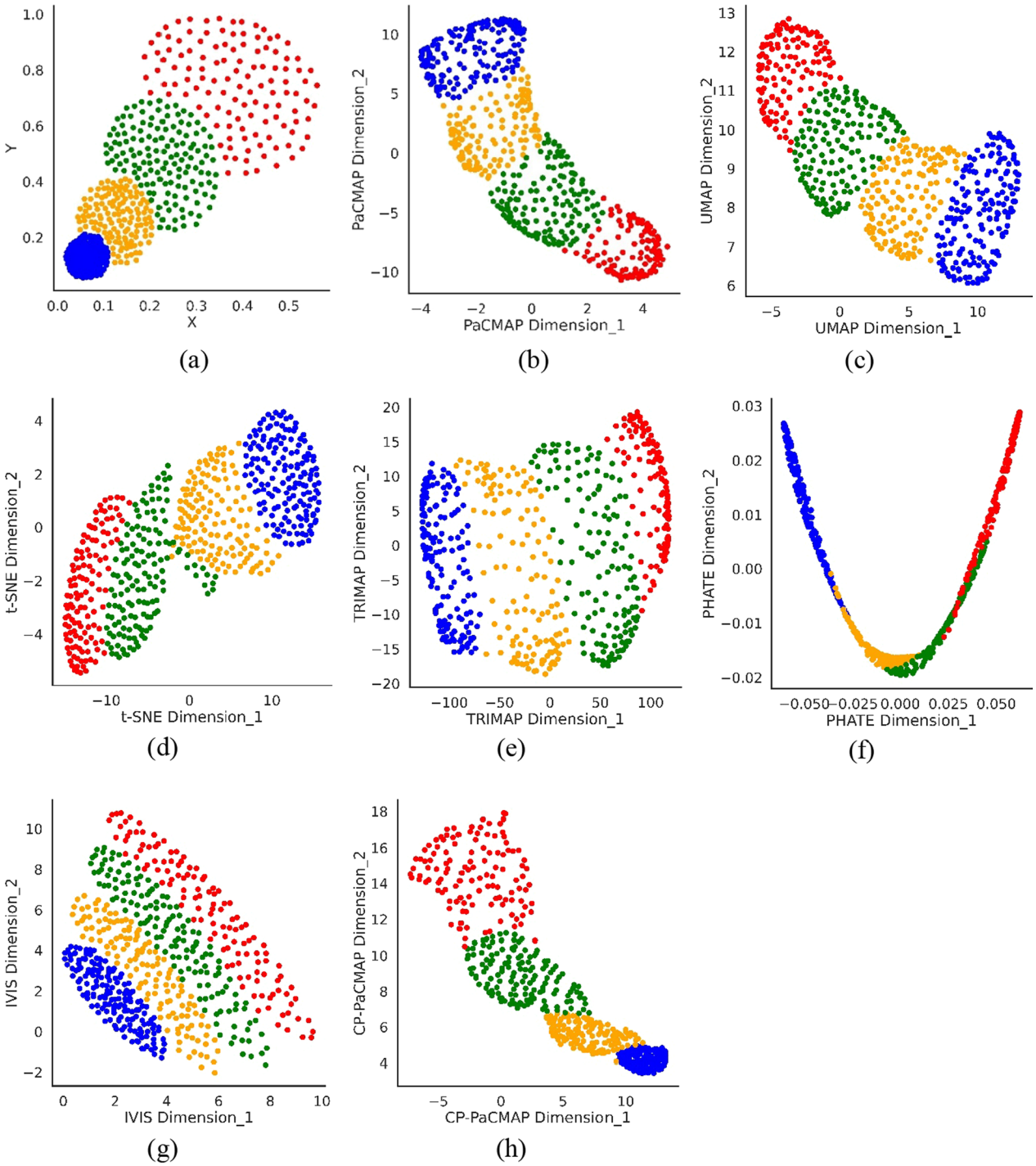


Fig. 2 2D visualization of LDS generated by various DR techniques. **b** PaCMAP, **c** UMAP, **d** t-SNE, **e** TRIMAP, **f** PHATE, **g** IVIS, **h** CP-PaCMAP on **a** generated linear data, with different level of compactness at each cluster

to maintain the local structure, as shown in Fig. 2e, g. PHATE (Fig. 2f) has an issue in preserving both structures, while t-SNE (Fig. 2d) is able to retain local structure in the LDS but has minor deviations in maintaining the global structure. Among all the seven DR techniques examined, the proposed CP-PaCMAP approach performs as well as PaCMAP in preserving both local and global structures in the LDS. It is also able to hold the compactness aspect present in all the clusters of the original data, as shown in Fig. 2h.

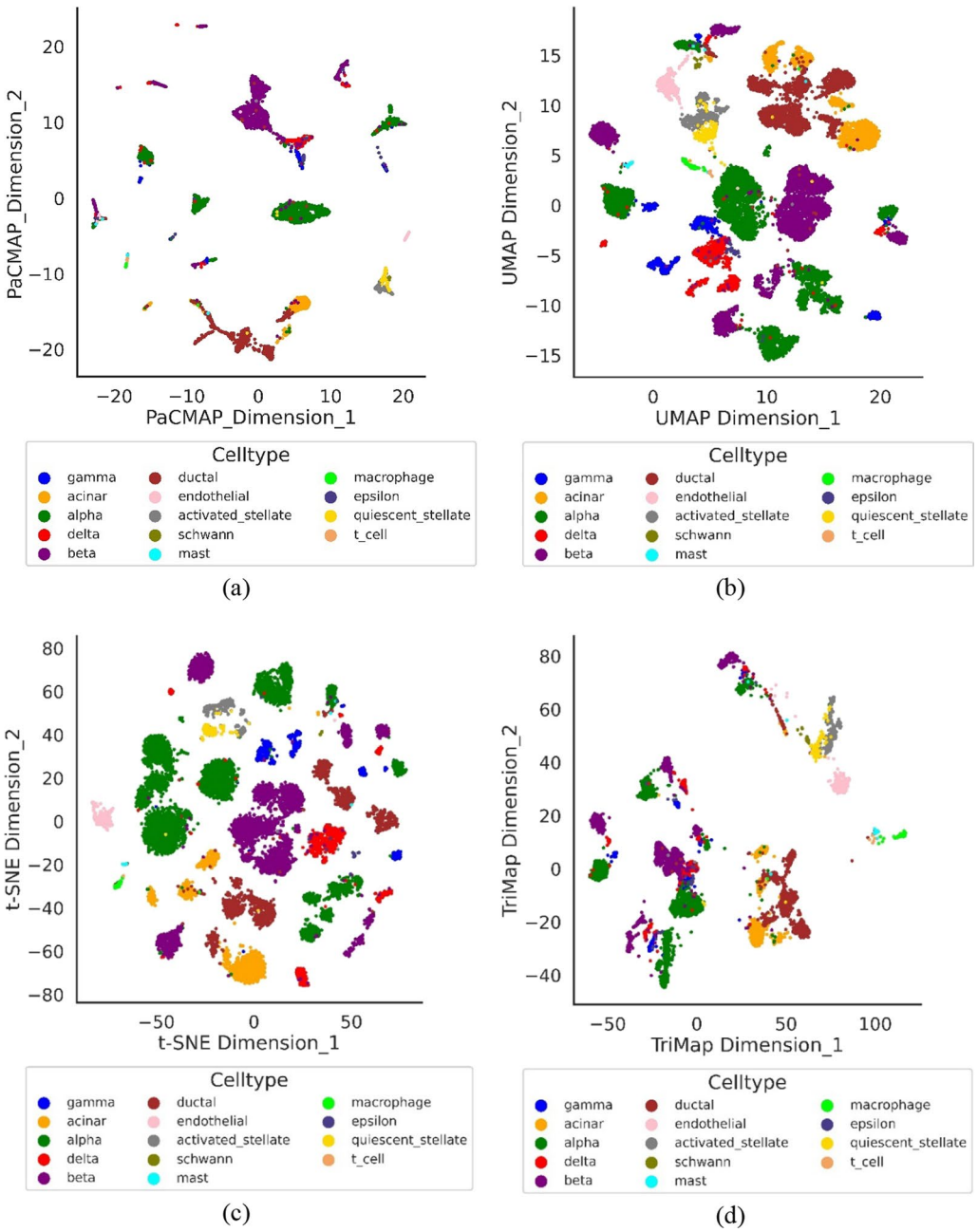


Fig. 3 2D visualization of LDS generated by various DR techniques. **a** PaCMAP, **b** UMAP, **c** t-SNE, **d** TRIMAP, **e** PHATE, **f** IVIS, **g** CP-PaCMAP on human pancreas scRNA-seq data containing 14 categories of cell type

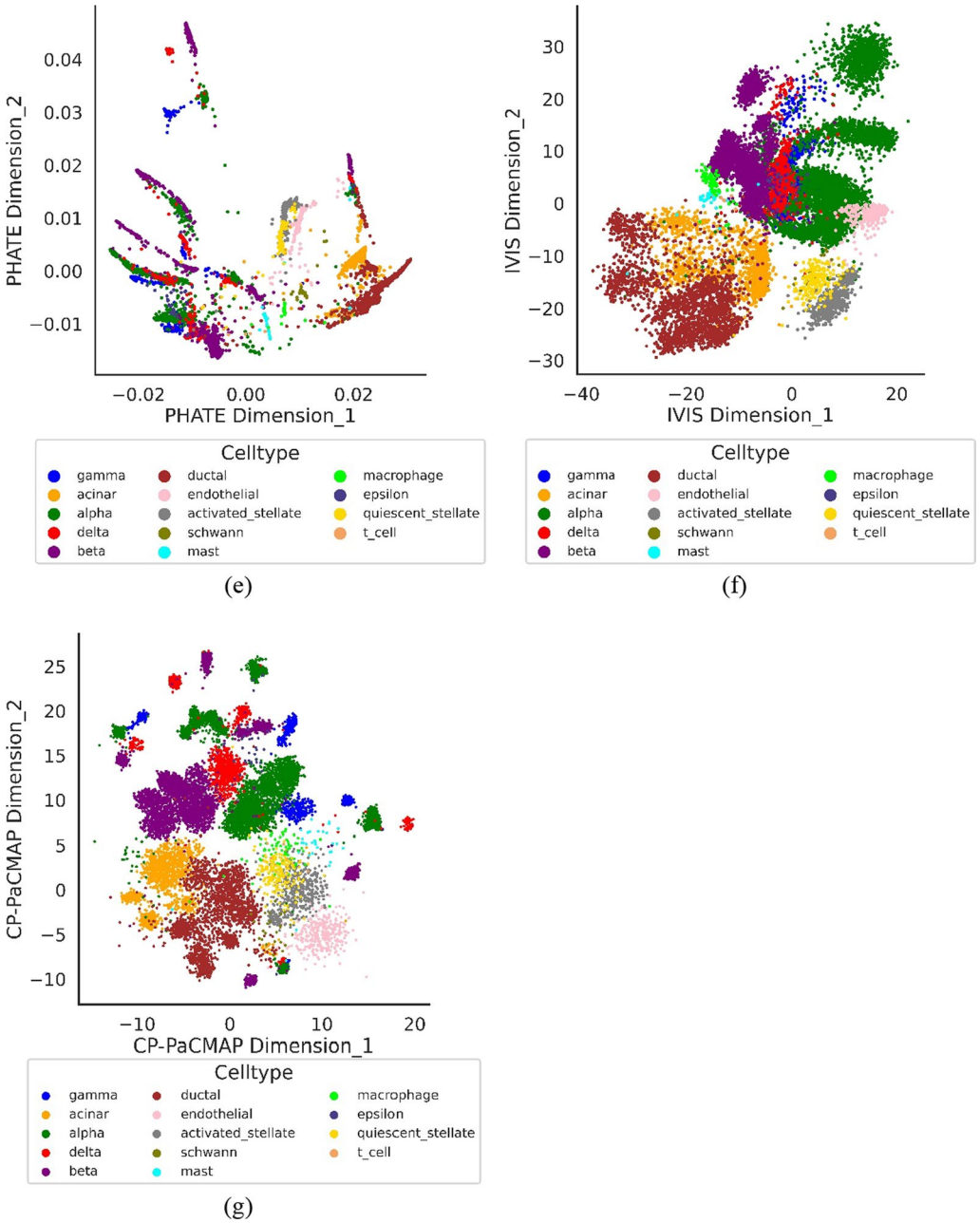


Fig. 3 continued

Results and discussions

We utilized a diverse array of assessment criteria to appraise the efficacy of the proposed approach and various DR methodologies across three distinct scRNAseq datasets—Human pancreas, skeletal muscle, and heart. Trustworthiness and continuity metrics are leveraged to scrutinize the fidelity of local and global structures within the reduced-dimensional representations (Andrew et al. 2013; Lee and Verleysen 2009; Jurman et al. 2012; Yousuff and Babu 2023; Allen et al. 2021; Gatin et al. 2019). The Mathew Correlation Coefficient metric provides the assessment of classification task (with imbalanced cell

types classes) performed on the LDS generated by all the DR techniques, while the Mantel test helps to evaluate the preservation of pairwise relationships between cells in the original HDS and their corresponding LDS. Furthermore, a runtime analysis is done to visualize the computational efficiency of each technique. This comprehensive suite of metrics collectively furnishes a multi-dimensional evaluation, elucidating both the merits and potential limitations of each approach within the diverse landscape of scRNAseq data analysis.

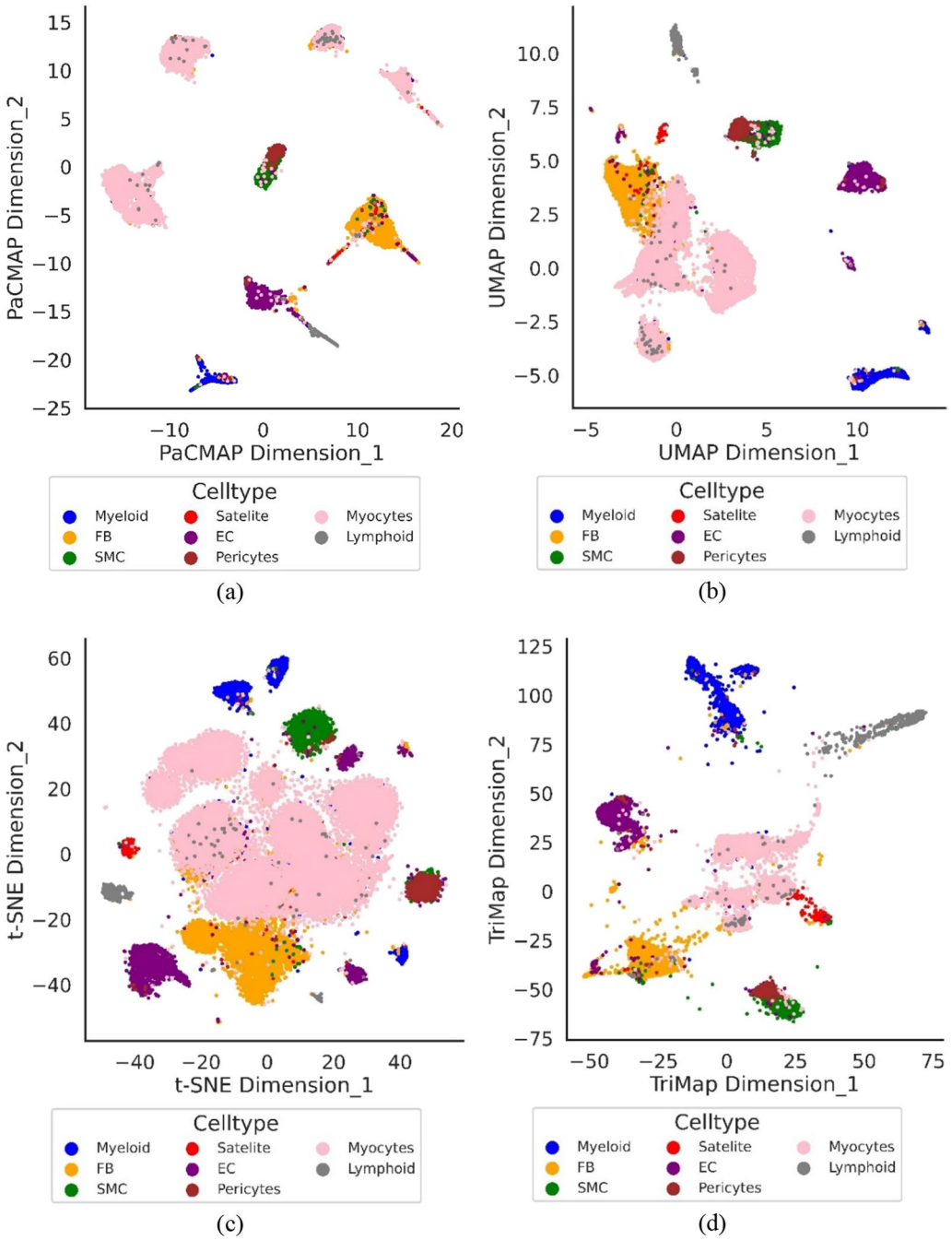


Fig. 4 2D visualization of LDS generated by various DR techniques. **a** PaCMAP, **b** UMAP, **c** t-SNE, **d** TRIMAP, **e** PHATE, **f** IVIS, **g** CP-PaCMAP on human skeletal muscle scRNA-seq data containing 8 categories of cell type

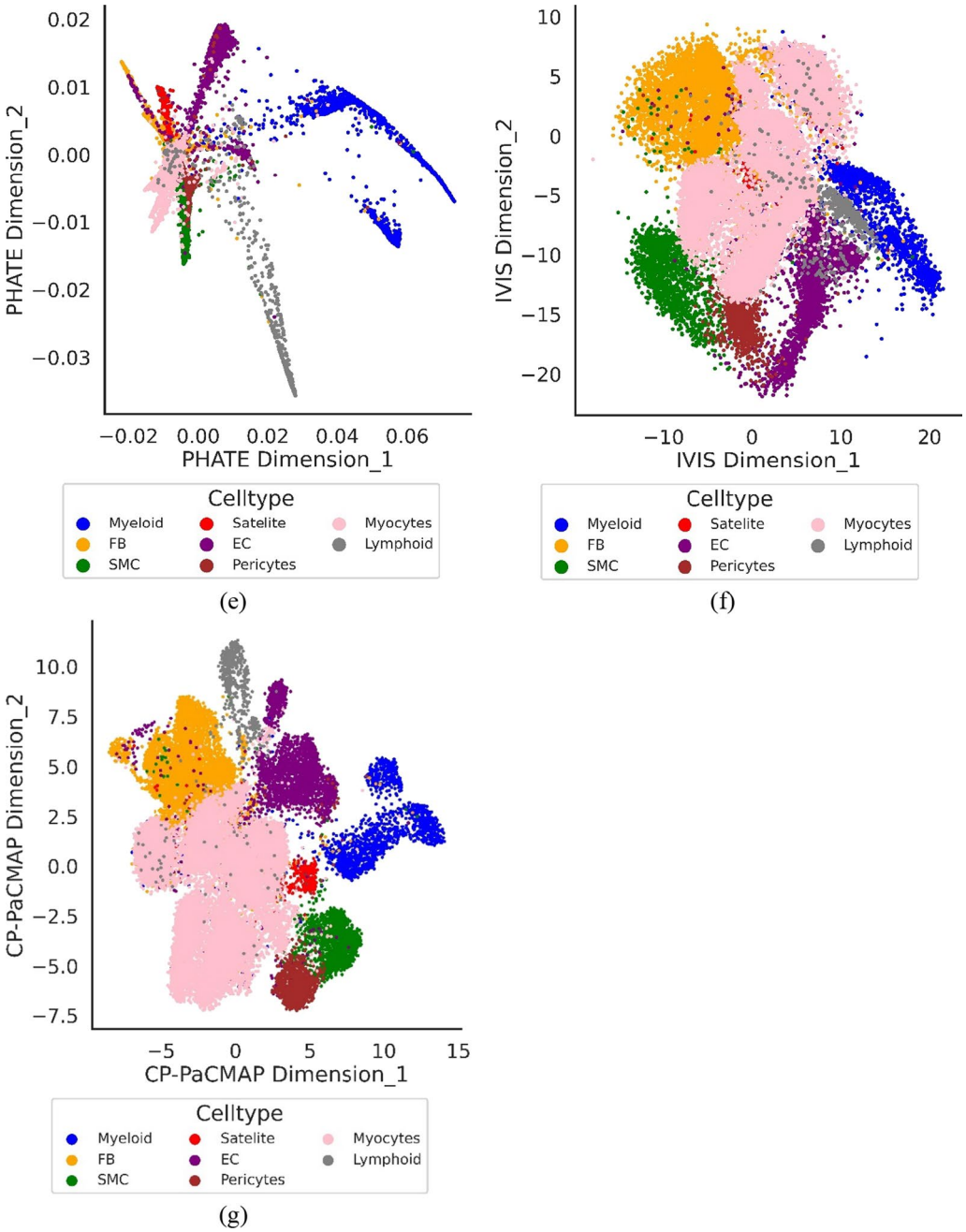


Fig. 4 continued

Trustworthiness and continuity

Trustworthiness (*TW*) helps us understand how well local relationships are preserved. It focuses on the nearest neighbors of each data point and checks if they remain close in the LDS. This is particularly important for methods that aim to capture local structures and clusters. Continuity (*CN*) helps us understand the preservation of global data patterns and the overall structure. It ensures that data points that were far apart or close in the HDS retain their relative distances in the LDS. This is essential for methods that aim to maintain the broader structure of the data (Wulfman et al. 2010; Ribaut et al. 2007; Sharini et al.

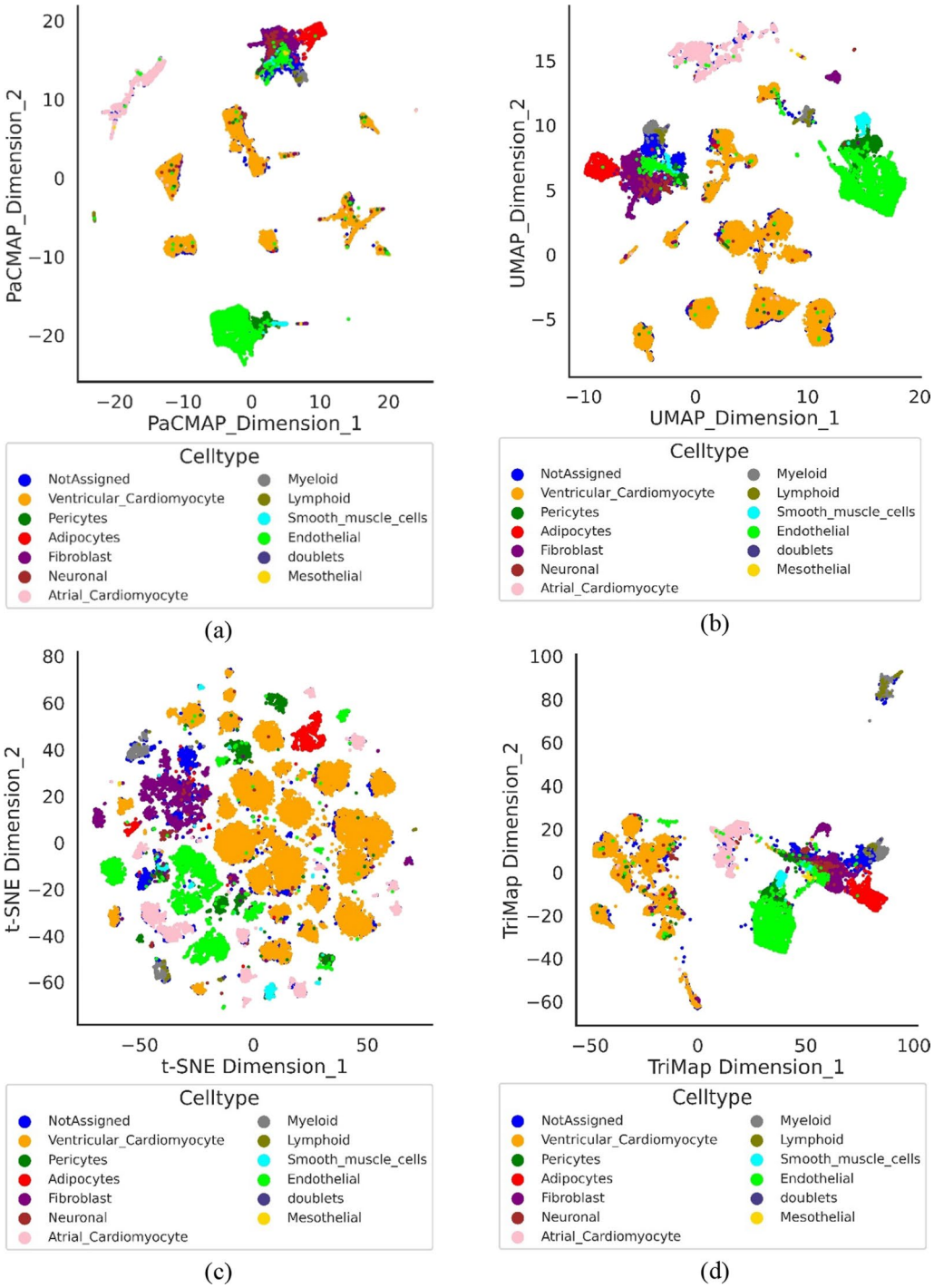
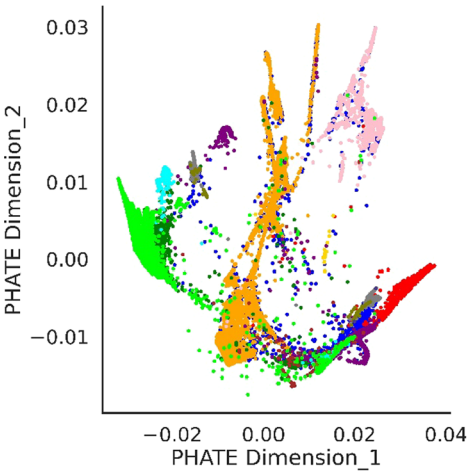
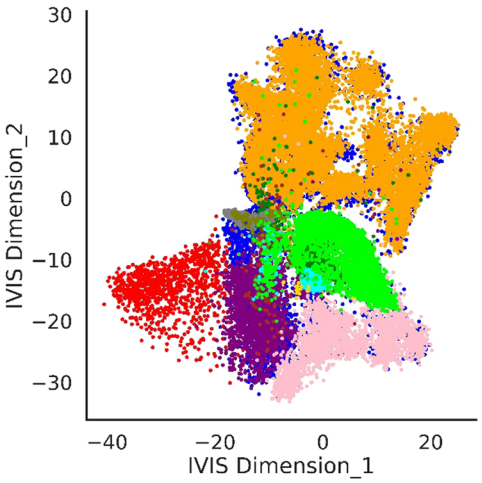


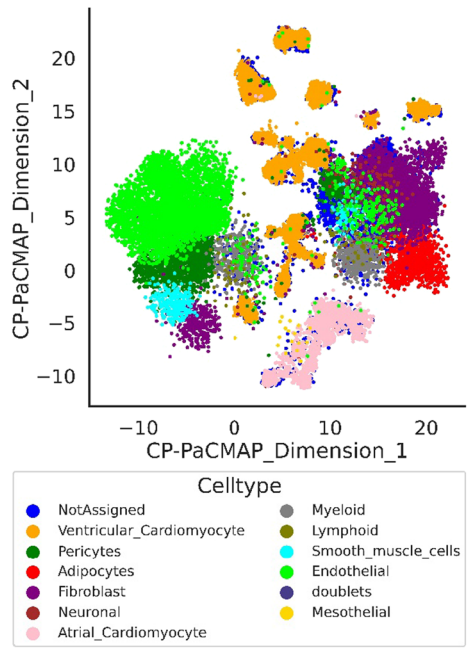
Fig. 5 2D visualization of LDS generated by various DR techniques. **a** PaCMAP, **b** UMAP, **c** t-SNE, **d** TRIMAP, **e** PHATE, **f** IVIS, **g** CP-PaCMAP on human heart scRNA-seq data containing 13 categories of cell type



(e)



(f)



(g)

Fig. 5 continued

2018; Pouard and Collange 2007; Bonnet et al. 2012; Sun et al. 2023; Dong et al. 2022). *TW* and *CN* play a crucial role in validating and selecting appropriate DR techniques for scRNA-seq data analysis. They provide quantitative measures of how well the DR technique preserves the biological structure, ultimately leading to more reliable and interpretable results. The *TW* and *CN* score of a DR technique can be calculated using the formulae given in Eqs. 12 and 13 (Lee and Verleysen 2009).

$$TW = 1 - \frac{2}{C(C-1)} \sum_{i=1}^{C} \sum_{j=1}^{C} \left(r_m(i,j) - m_{neighbors}(i,j) \right) \tag{12}$$

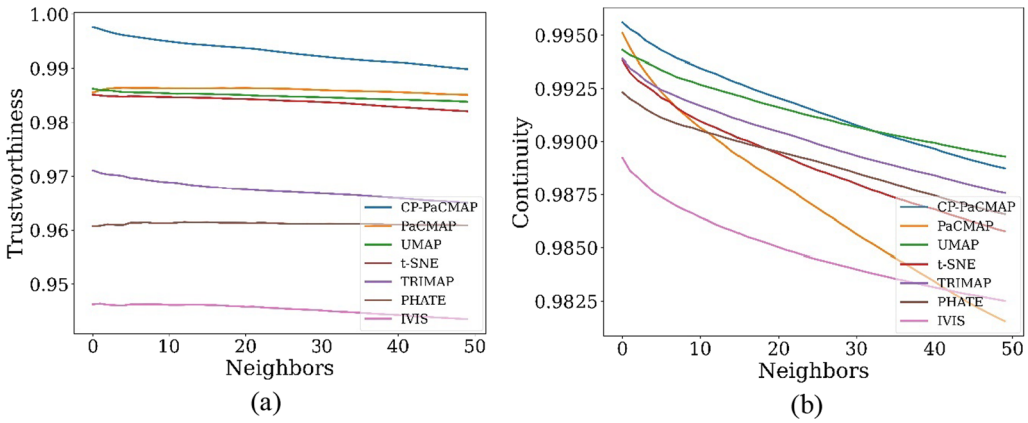


Fig. 6 **a** Trustworthiness and **b** Continuity scores for Human pancreas scRNA-seq dataset

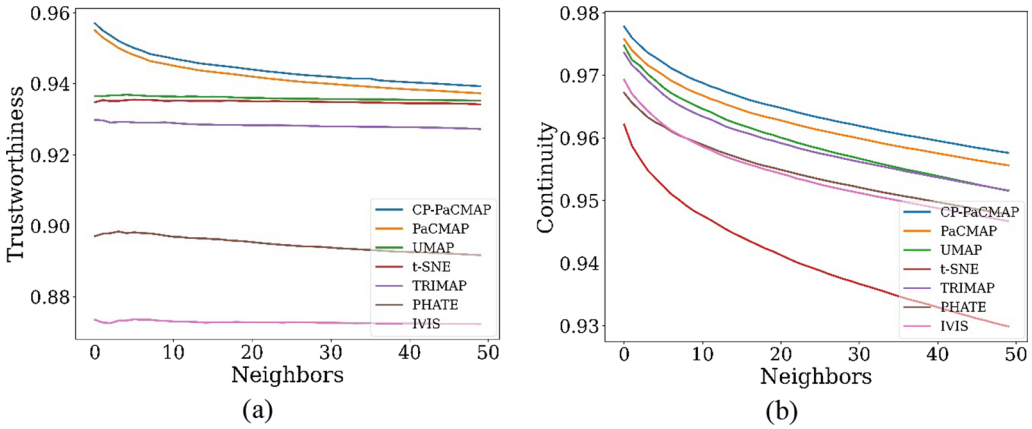


Fig. 7 **a** Trustworthiness and **b** Continuity scores for Human skeletal muscles scRNA-seq dataset

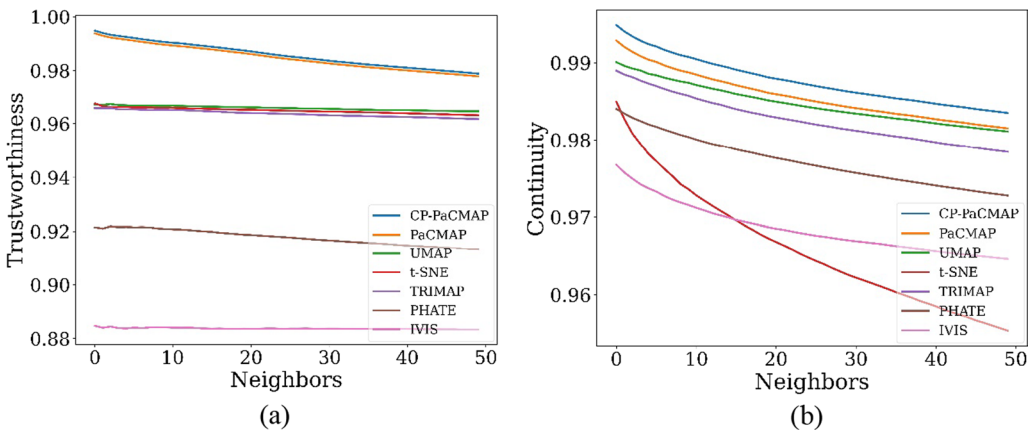


Fig. 8 **a** Trustworthiness and **b** Continuity scores for Human heart scRNA-seq dataset

$$CN = 1 - \frac{2}{C(C-1)} \sum_{i=1}^{C} \sum_{j=1}^{C} \left(m_{neighbors}(i,j) - r_m(i,j) \right) \quad (13)$$

- TW is the trustworthiness score and CN is the continuity score for a given m, which represents the number of nearest neighbors to consider.
- N is the total number of cells (data points).
- $r_m(i,j)$ represents the rank of cell j among the m nearest neighbors of cell i in the HDS. This indicates how close cell j is to cell i in the original space considering m neighbors.
- $m_{neighbors}(i,j)$ is a binary indicator function. It takes a value of 1 if cell j is among the m nearest neighbors of cell i in the LDS. It checks whether the proximity relationship is maintained in the reduced space.

The TW score ranges from 0 to 1 ($TW \in [0,1]$), where 0 indicates that the local structures are not preserved well in the LDS, and 1 indicates perfect preservation of local structures. The continuity score ranges from -1 to 1 ($CN \in [-1,1]$). A score of -1 means that the global structure is perfectly preserved in reverse order (what is close in the original space is far in the reduced space), 0 means no preservation, and 1 means perfect preservation of the global structure. TW and CN scores are computed on various m values for all the three different scRNAseq datasets. Figures 6, 7, and 8 demonstrate that CP-PaCMAP is performing comparatively fine with respect to all other DR techniques. Hence, compactness can be well preserved along with local and global structures of HDS into LDS without any compensation in performance.

Classification model and Matthew's correlation coefficient

In this study, we applied the K-nearest neighbor (KNN) classification algorithm to analyze scRNA-seq data. To assess the algorithm's performance and ensure robustness, we employed tenfold cross-validation. The scRNA-seq dataset consisted of gene expression profiles for individual cells, with the target variable being the cell type. By utilizing the KNN algorithm with a k value of 25, we aimed to predict the cell types based on the similarity of gene expression profiles among neighboring cells. The tenfold cross-validation approach allowed us to evaluate the algorithm's performance by splitting the data into 10 subsets, training the model on nine of them, and testing it on the remaining subset. This process was repeated 10 times, ensuring that each subset served as training and testing data. Confusion matrices are obtained for all the DR techniques applied to each scRNA-seq dataset. Finally, all the confusion matrices are utilized to compute Matthew's correlation coefficient (MCC), the classification performance metric.

The MCC is a widely utilized performance metric for assessing prediction precision in multi-class classification tasks (Zegarra Flores and Radoux 2023; Dine et al. 2022; Lee and Park 2022; Thakur et al. 2023; Zhang and Leatham 2019; Zhou et al. 2018). The overall assessment of classification accuracy is determined by considering the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). TP refers to the count of positive instances that have been exactly classified. TN refers to the count of negative instances that are accurately classified. FP refers to the quantity of instances that are erroneously classified as positive. FN refers to the quantity of instances that are erroneously classified as negative. MCC is a metric that takes into account the distribution of true positives, true negatives, false positives, and false negatives in order to yield a singular value that serves as a comprehensive indicator of the classifier's predictive performance. A higher MCC score signifies superior performance, where a value of 1 represents the ideal result, and -1 represents the lowest outcome (Jurman et al. 2012; Yousuff and Babu 2023).

The MCC is a valuable metric in the context of multi-class classification tasks due to its ability to consider the disparities in class distributions. This metric offers a more dependable assessment of the classifier's effectiveness, particularly when confronted with unequal class proportions or imbalanced datasets. In the context of scRNA-seq classification, the MCC is a valuable metric. It is beneficial because it takes into account the differences in class distributions, which are often encountered in scRNA-seq data. The MCC provides a reliable measure of the classifier's performance, especially when dealing with imbalanced datasets or variations in the proportions of different cell types. It helps assess the accuracy and robustness of the classification algorithm in handling the complexities of scRNA-seq data (Jurman et al. 2012). In the multiple cell types (categories or classes) scenario, the MCC can be mathematically expressed by utilizing a confusion matrix M that represents the classification outcomes for each category C as given in Eq. 12. The MCC value of the DR techniques computed on all three scRNA-seq data is depicted in Fig. 9. The proposed CP-PaCMAP technique demonstrates slight improvement in MCC metric, compared to existing DR techniques (Jurman et al. 2012).

$$MCC = \frac{o \times d - \sum_c^C p_c \times t_c}{\sqrt{\left(d^2 - \sum_c^C p_c^2\right) \times \left(d^2 - \sum_c^C t_c^2\right)}} \quad (12)$$

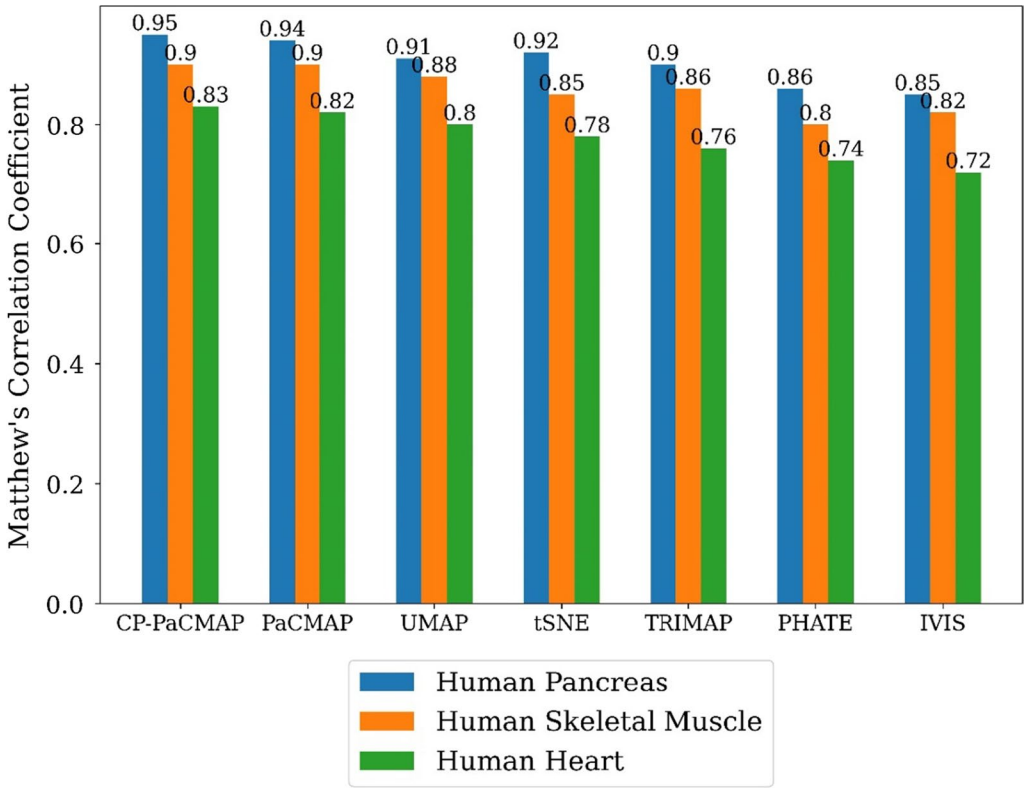


Fig. 9 MCC performance metric computed using confusion matrices of KNN model for all the DR techniques implemented on each scRNA-seq dataset

where, $t_c = \sum_i^C M_{ic}$ the number of times category C actually occurred, $p_c = \sum_i^C M_{ci}$ the number of the times category C got predicted, $o = \sum_c^C M_{cc}$ the total count of observations rightly predicted, $d = \sum_i^C \sum_j^C M_{ij}$ the total count of data points.

Mantel test

The Mantel test can be utilized along with the Pearson Correlation Coefficient (PCC) to evaluate the preservation of pairwise relationships between cells in the original HDS and their LDS representations. The PCC is a measure of relationship between two sets of data, is commonly used as the correlation coefficient in the Mantel test (Zhou et al. 2018; Mushtaq et al. 2020; Singh et al. 2023; Fakhfakh et al. 2020; Gupta 2022; Zhao 2021). By comparing the PCC obtained from the Mantel test, it is possible to determine how well the DR technique preserves the pairwise relationships between cells. A higher PCC value (+1) indicates a stronger correlation and suggests better preservation of the relationships in the LDS. To create a distribution of correlation values, Mantel test procedure was performed multiple times on randomly chosen subsamples of the scRNA-seq data points (n = 500 cells per subsample picked without replacement). Mantel test on cluster centroid distance matrices exposes potential similarities or variations in the underlying grouping patterns (Szubert et al. 2019). PCC values obtained for various DR techniques on three different scRNA-seq Datasets demonstrated a strong correlation between the actual HDS and LDS cluster centroid distances. Mean and Median PCC values for all the DR techniques on scRNA-seq Human pancreas, skeletal muscle, and heart are listed in Table 2.

The RainCloud plot is a smart combination of a Strip plot, a split-half violin plot, a boxplot with whiskers, and a point plot. In the case of a strip plot, the data points are represented as individual dots distributed evenly along the categorical axis, providing a more granular view. Violin plots reveal data distribution shape, density, and spread. Width signifies density; wider areas have more data, and narrower areas have less. Longer violins suggest a broader range, while shorter ones imply a narrower range. Outliers are shown when data points extend beyond the violin's range. The box in the box plot represents the middle 50% of the data (interquartile range—IQR), with the median shown as a central line. The box length reflects the data spread, longer indicating a larger spread and shorter suggesting a narrower spread.

Table 2 Mean and median PCC values obtained from the Mantel test for various DR techniques on all three scRNA-seq Datasets

scRNA-seq dataset	DR techniques	Mean PCC value	Median PCC value
Human pancreas	CP-PaCMAP	0.85	0.85
	PaCMAP	0.83	0.84
	UMAP	0.78	0.80
	t-SNE	0.80	0.82
	TRIMAP	0.76	0.76
	PHATE	0.75	0.74
	IVIS	0.72	0.73
Human skeletal muscle	CP-PaCMAP	0.80	0.79
	PaCMAP	0.77	0.76
	UMAP	0.77	0.77
	t-SNE	0.76	0.77
	TRIMAP	0.74	0.75
	PHATE	0.73	0.74
	IVIS	0.76	0.77
Human heart	CP-PaCMAP	0.77	0.78
	PaCMAP	0.75	0.76
	UMAP	0.74	0.76
	t-SNE	0.73	0.74
	TRIMAP	0.73	0.72
	PHATE	0.72	0.73
	IVIS	0.70	0.72

Whiskers extend to 1.5 times the IQR, covering the data's range from minimum to maximum values. The flag of the point plot is meant to represent the mean of data in the context of the RainCloud plot (Allen et al. 2021).

PCC values collected after the permutations of the Mantel Test on different HDS scRNA-seq datasets and their corresponding LDS are plotted using RainCloud plots, as shown in Figs. 10, 11, and 12, respectively. We are able to observe a higher density of PCC values towards + 1 in the case of CP-PaCMAP, as depicted using a split-half violin plot. The Median, minimum, and maximum values of PCC are also comparatively better in CP-PaCMAP, which is observed in the box plot. THE mean PCC values of CP-PaCMAP are also high compared to other DR techniques demonstrated using point plot flags (Fasil and Rajesh 2023; Gupta et al. 2023a, 2023b; Sénéchal et al. 2005; Mukherjee et al. 2021; Gupta 2023; Kaur and Khehra 2021).

Runtime analysis

We have comprehensively explored several DR techniques, including PaCMAP, UMAP, t-SNE, TRIMAP, PHATE, and IVIS, alongside the proposed technique termed CP-PaCMAP. These techniques are pivotal in scRNA-seq analysis, revealing large-scale datasets' intrinsic structures and relationships. The primary objective was to assess these DR techniques' runtime

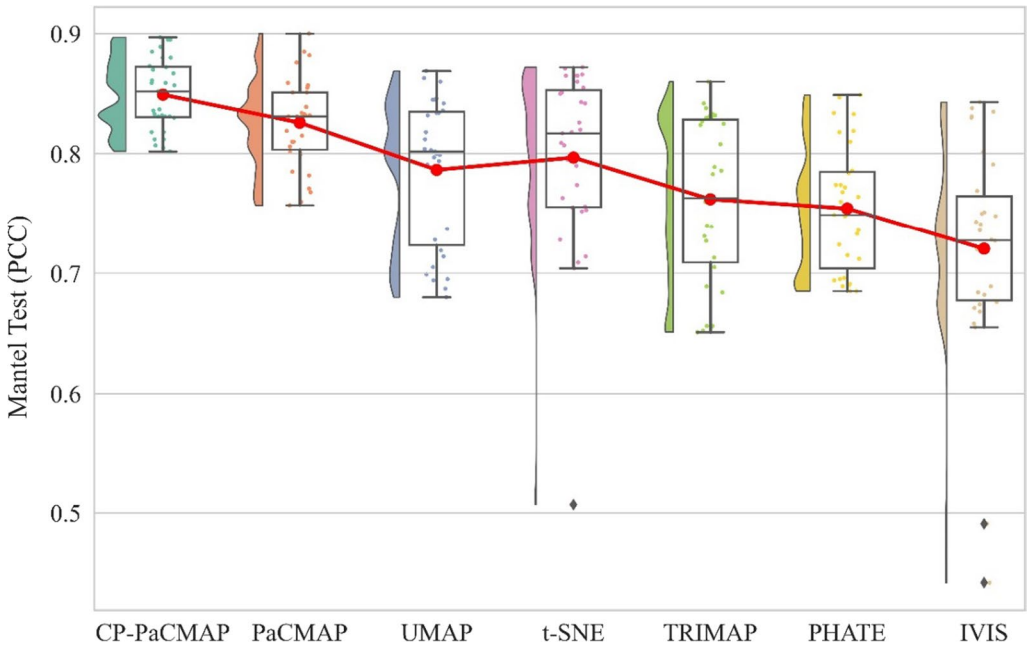


Fig. 10 Statistical analysis of PCC values of Human pancreas scRNA-seq dataset: Raincloud plot incorporating Strip plot (PCC values distribution), Split-Half Violin plot (Density of distribution), Box plot (Outliers, Min, Max, Median), and Point plot (Mean)

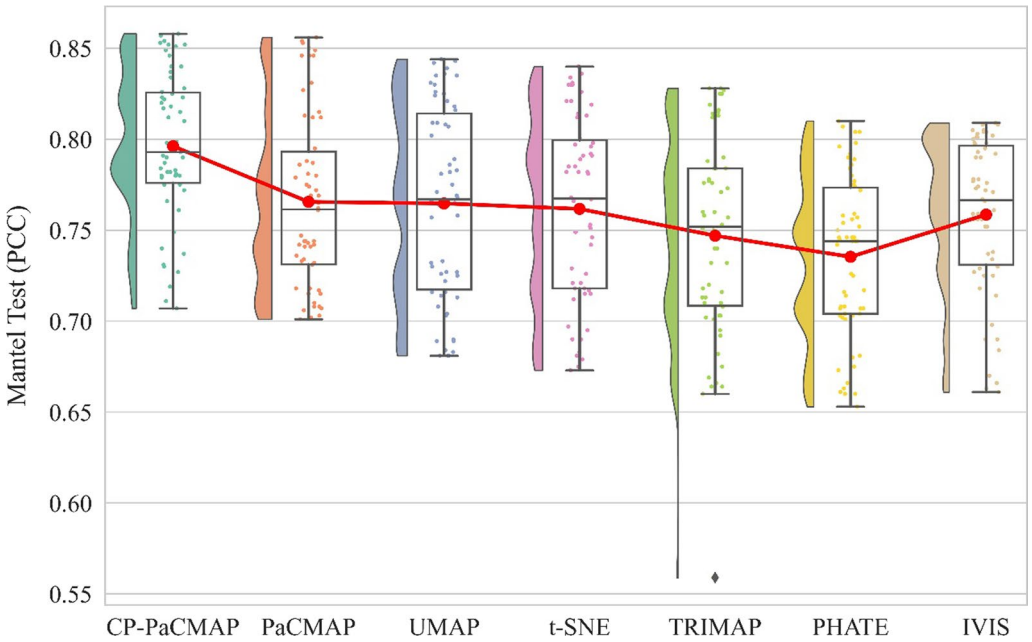


Fig. 11 Statistical analysis of PCC values of Human skeletal muscle scRNA-seq dataset: Raincloud plot incorporating Strip plot (PCC values distribution), Split-Half Violin plot (Density of distribution), Box plot (Outliers, Min, Max, Median), and Point plot (Mean)

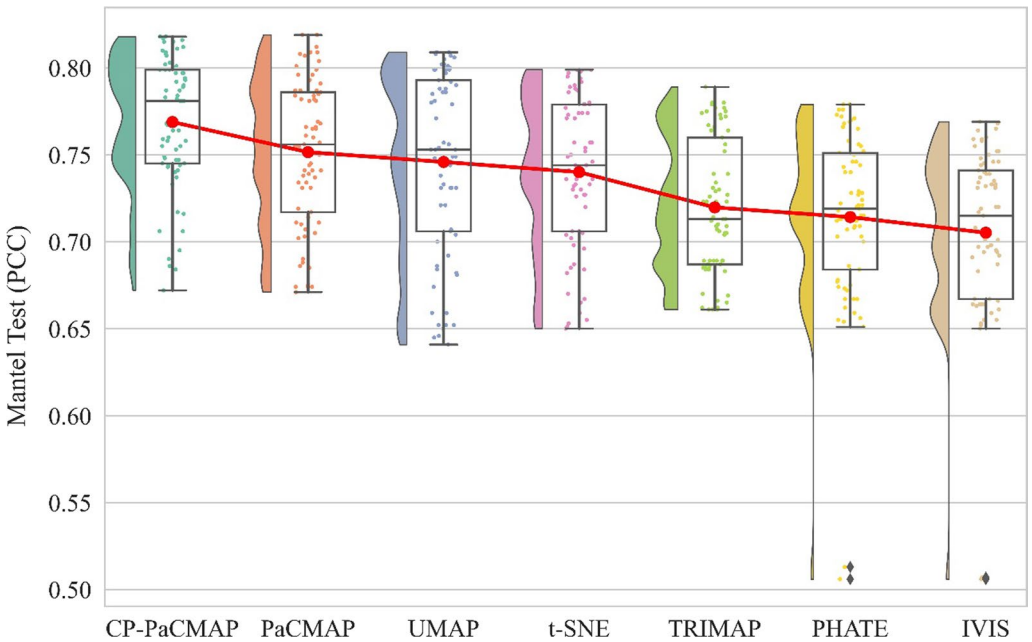


Fig. 12 Statistical analysis of PCC values of Human heart scRNA-seq dataset: Raincloud plot incorporating Strip plot (PCC values distribution), Split-Half Violin plot (Density of distribution), Box plot (Outliers, Min, Max, Median), and Point plot (Mean)

(computational efficiency) across various data point magnitudes ranging from 5000 to 30,000. To accomplish this, the execution times of each technique in seconds are recorded and subsequently visualized through a line graph, as shown in Fig. 13. Upon scrutinizing the outcomes, it is apparent that PaCMAP exhibited remarkable performance across all scenarios. It consistently outperformed its counterparts, showcasing its prowess in runtime. Intriguingly, CP-PaCMAP emerged as a notable approach, securing the second position in terms of

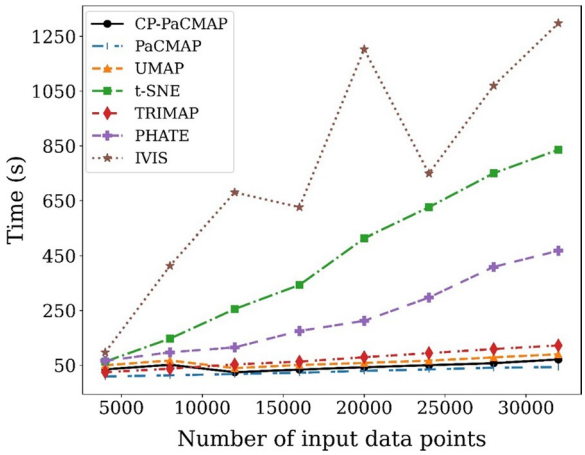


Fig. 13 Runtime analysis of DR techniques on scRNA-seq data with varying data point sizes

runtime. The slight overhead incurred by CP-PaCMAP can be attributed to its endeavor to maintain compactness within the transformed LDS. CP-PaCMAP is built upon PaCMAP, so it additionally involves computing average distance pairs (while inducing a minimal delay), preserving essential structural integrity.

Conclusion

Our study highlights the pivotal role of DR techniques in unraveling the intricate relationships within scRNA-seq data. While PCA remains a stalwart in linear DR, the limitations of this approach are evident in the face of diverse cell types. Nonlinear techniques like UMAP, t-SNE, TriMap, PHATE, and IVIS have emerged as powerful alternatives, each with unique strengths and constraints. Our introduction of the CP-PaCMAP algorithm addresses many challenges, providing a robust solution for visualizing and analyzing scRNA-seq data. Its ability to preserve both local and global structures, coupled with its enhanced computational efficiency, positions CP-PaCMAP as a promising tool for researchers seeking to gain deeper insights into cellular heterogeneity.

Future work

Looking ahead, several avenues for further exploration and refinement can be implemented. Firstly, extending CP-PaCMAP to accommodate even larger and more diverse datasets could enhance its applicability across a broader spectrum of biological systems. Additionally, incorporating CP-PaCMAP into integrated workflows for scRNA-seq analysis, potentially in conjunction with advanced machine learning techniques, holds promise for uncovering novel biological insights. Exploring the potential of CP-PaCMAP in the context of multi-modal single-cell omics data could further expand its utility in deciphering complex cellular landscapes. Furthermore, investigating the algorithm's performance in scenarios of perturbed cellular states or rare cell type identification could yield valuable insights for various biomedical applications. Finally, efforts towards enhancing the interpretability of the resulting low-dimensional representations and developing user-friendly interfaces will be crucial for enabling the broader adoption of CP-PaCMAP in the scientific community. By pursuing these directions, we aim to advance the capabilities of DR techniques in scRNA-seq analysis and contribute to a more comprehensive understanding of cellular biology.

Acknowledgements
Not Applicable.

Author contributions
All authors contributed equally in formulation and execution of this work. RB was involved in planning and supervising the work and wrote the first draft of the manuscript. RB and AR performed data collection, processed the experimental data, data analysis, and designed the figures. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
Available on request.

Declarations

Competing interests
Authors declare no conflict of interest.

Received: 2 October 2023 Accepted: 15 December 2023
Published online: 11 January 2024

References

Allen M, Poggiali D, Whitaker K, Marshall TR, van Langen J, Kievit RA. Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Res. 2021;4:63.

Amid E, Warmuth MK. TriMap: Large-scale Dimensionality Reduction Using Triplets. arXiv Prepr. 2019.

Andrew G, Arora R, Bilmes J, Livescu K. Deep Canonical Correlation Analysis. In: Dasgupta S, McAllester D, editors. Proceedings of the 30th International Conference on Machine Learning [Internet]. Atlanta, Georgia, USA: PMLR; 2013. p. 1247–55. (Proceedings of Machine Learning Research; vol. 28). https://proceedings.mlr.press/v28/andrew13.html

Babjac A, Royalty T, Steen AD, Emrich SJ. A Comparison of Dimensionality Reduction Methods for Large Biological Data. In: Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. Association for Computing Machinery; 2022. (BCB '22).

Battenberg K, Kelly ST, Ras RA, Hetherington NA, Hayashi M, Minoda A. A flexible cross-platform single-cell data processing pipeline. Nat Commun. 2022;13(1):6847.

Bonnet, S., Bêche, J.-F., Gharbi, S., Abdoun, O., Bocquelet, F., Joucla, S., Guillemaud, R. NeuroPXI: A real-time multi-electrode array system for recording, processing, and stimulation of neural networks and the control of high-resolution neural implants for rehabilitation [NeuroPXI : un système multi-électrode temps-réel pour l'enregistrement, le traitement et la stimulation de réseaux neuronaux et le contrôle d'implants à haute résolution spatiale pour la réhabilitation]. IRBM, 2012;33(2), 55–60.

Carter KM, Raich R, Finn WG, Hero AO. Dimensionality reduction of flow cytometric data through information preservation. In: 2008 IEEE Workshop on Machine Learning for Signal Processing. 2008;462–7.
Chen W, Wahiduzzaman M, Li Q, Li Y, Zheng G, Huang T. Comparative analysis of NovaSeq 6000 and MGISEQ 2000 single-cell RNA sequencing data. Quant Biol. 2022;10(4):333–40. https://doi.org/10.15302/J-QB-022-0295.
Chicco D. Siamese neural networks: an overview. In: Cartwright H, editor. Artificial neural networks. Springer: US; 2021. p. 73–94.
Coenen A, Reif E, Yuan A, Kim B, Pearce A, Viégas F, et al. Visualizing and Measuring the Geometry of BERT. arXiv; 2019.
Dong B, Wang X, Qiang X, Du F, Gao L, Wu Q, Cao G, Dai C. A multi-branch convolutional neural network for screening and staging of diabetic retinopathy based on wide-field optical coherence tomography angiography. IRBM. 2022;43(6):614–20. https://doi.org/10.1016/j.irbm.2022.04.004.
El Dine KB, Nader N, Khalil M, Marque C. Uterine synchronization analysis during pregnancy and labor using graph theory, classification based on neural network and deep learning. IRBM. 2022;43(5):333–9. https://doi.org/10.1016/j.irbm.2021.09.002.
Fakhfakh M, Chaari L, Fakhfakh N. Bayesian curved lane estimation for autonomous driving. J Ambient Intell Hum Comput. 2020;11:4133–43. https://doi.org/10.1007/s12652-020-01688-7.
Fasil OK, Rajesh R. Epileptic seizure classification using shifting sample difference of EEG signals. J Ambient Intell Hum Comput. 2023;14:11809–22. https://doi.org/10.1007/s12652-022-03737-9.
Gatin E, Nagy P, Paun I, Dubok O, Bucur V, Windisch P. Raman spectroscopy: application in periodontal and oral regenerative surgery for bone evaluation. IRBM. 2019. https://doi.org/10.1016/j.irbm.2019.05.002.
Gayoso A, Lopez R, Xing G, Boyeau P, Valiollah Pour Amiri V, Hong J, et al. A Python library for probabilistic analysis of single-cell omics data. Nat Biotechnol. 2022.
Ghazanfar S, Bisogni AJ, Ormerod JT, Lin DM, Yang JYH. Integrated single cell data analysis reveals cell specific networks and novel coactivation markers. BMC Syst Biol. 2016;10(5):127. https://doi.org/10.1186/s12918-016-0370-4.
Granja JM, Klemm S, McGinnis LM, Kathiria AS, Mezger A, Corces MR, et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nat Biotechnol. 2019;37(12):1458–65.
Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015;525(7568):251–5.
Gupta V. Application of chaos theory for arrhythmia detection in pathological databases. Int J Med Eng Inf. 2022;15(2):191–202. https://doi.org/10.1504/IJMEI.2023.129353.
Gupta V. Wavelet transform and vector machines as emerging tools for computational medicine. J Ambient Intell Human Comput. 2023;14:4595–605. https://doi.org/10.1007/s12652-023-04582-0.
Gupta V, Mittal M, Mittal V, et al. ECG signal analysis based on the spectrogram and spider monkey optimisation technique. J Inst Eng India Ser B. 2023a;104:153–64. https://doi.org/10.1007/s40031-022-00831-6.
Gupta V, Mittal M, Mittal V, Gupta A. Adaptive autoregressive modeling based ECG signal analysis for health monitoring. In Optimization Methods for Engineering Problems. 2023b. https://doi.org/10.1201/9781003300731-1.
Heiser CN, Lau KS. A quantitative framework for evaluating single-cell data structure preservation by dimensionality reduction techniques. Cell Rep. 2020;31(5): 107576.
Jurman G, Riccadonna S, Furlanello C. A comparison of MCC and CEN error measures in multi-class prediction. PLoS ONE. 2012;7(8):1.
Kaur J, Khehra BS. Fuzzy logic and hybrid-based approaches for the risk of heart disease detection: state-of-the-art review. J Inst Eng (eng) Series B. 2021;103(2):1–17. https://doi.org/10.1007/s40031-021-00644-z.
Lee S, Park D. Abnormal beat detection from unreconstructed compressed signals based on linear approximation in ECG signals suitable for embedded IoT devices. J Ambient Intell Hum Comput. 2022;13:4705–17. https://doi.org/10.1007/s12652-021-03578-y.
Lee JA, Verleysen M. Quality assessment of dimensionality reduction: rank-based criteria. Neurocomputing. 2009;72(7):1431–43.
Lytal N, Ran D, An L. Normalization methods on single-cell RNA-seq data: an empirical survey. Front Genet. 2020;11:1.
McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017;33(8):1179–86.
McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. ArXiv e-prints. 2018 Feb;
Miragaia RJ, Gomes T, Chomka A, Jardine L, Riedel A, Hegazy AN, et al. Single-cell transcriptomics of regulatory T cells reveals trajectories of tissue adaptation. Immunity. 2019;50(2):493-504.e7.
Moon KR, van Dijk D, Wang Z, Gigante S, Burkhardt DB, Chen WS, et al. Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol. 2019;37(12):1482–92.
Mukherjee A, Kundu PK, Das A. Transmission line fault location using PCA-based best-fit curve analysis. J Inst Eng India Ser B. 2021;102:339–50. https://doi.org/10.1007/s40031-020-00515-z.
Mushtaq Z, Ali I, Shah R, et al. Detection, localization and analysis of oil spills in water through wireless thermal imaging and spectrometer based intelligent system. Wirel Pers Commun. 2020;111:679–98. https://doi.org/10.1007/s11277-019-06880-3.
Nayak R, Hasija Y. A hitchhiker's guide to single-cell transcriptomics and data analysis pipelines. Genomics. 2021;113(2):606–19.
Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, Codeluppi S, et al. Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. Cell. 2016;165(4):1012–26.
Pierson E, Yau C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16(1):241.
Pouard P, Collange V. Neuromonitoring by near infrared spectroscopy in pediatric cardiac surgery. IRDM. 2007. https://doi.org/10.1016/S1297-9562(07)78715-6.
Ribaut C, Reybier K, Torbiero B, Launay J, Valentin A, Reynes O, Fabre P-L, Nepveu F. Strategy of red blood cells immobilisation onto a gold electrode: characterization by electrochemical impedance spectroscopy and quartz crystal microbalance [Stratégie d'immobilisation de globules rouges sur électrode d'or : caractérisation par spectroscopie d'impédance électrochimique et microbalance à quartz]. Revue De Biologie Et De Médecine Expérimentales. 2007. https://doi.org/10.1016/j.rbmret.2007.12.009.
Sénéchal P, Perroud H, Kedziorek MAM, et al. Non destructive geophysical monitoring of water content and fluid conductivity anomalies in the near surface at the border of an agricultural. Subsurf Sens Technol Appl. 2005;6:167–92. https://doi.org/10.1007/s11220-005-0005-0.
Shaffer SM, Dunagin MC, Torborg SR, Torre EA, Emert B, Krepler C, et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature. 2017;546(7658):431–5.
Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature. 2013;498(7453):236–40.
Sharini H, Fooladi M, Masjoodi S, Jalalvandi M, Yousef Pour M. Identification of the pain process by cold stimulation: using dynamic causal modeling of effective connectivity in functional near-infrared spectroscopy (fNIRS). Innov Res Biomed Eng. 2018. https://doi.org/10.1016/j.irbm.2018.11.006.
Singh H, Kumar V, Saxena K, et al. Smart channel modelling for cloud and fog attenuation using ML for designing of 6G networks at D and G bands. Wirel Pers Commun. 2023;129:1669–92. https://doi.org/10.1007/s11277-023-10201-0.
Sun J, Liu Q, Wang Y, Wang L, Song X, Zhao X. Five-year prognosis model of esophageal cancer based on genetic algorithm improved deep neural network. IRBM. 2023;44(3): 100748. https://doi.org/10.1016/j.irbm.2022.100748.
Szubert B, Cole JE, Drozdov I. Structure-preserving visualisation of high dimensional single-cell datasets. Sci Rep. 2019;1:1–10. https://doi.org/10.1038/s41598-019-45301-0.
Thakur M, Dhanalakshmi S, Kuresan H, et al. Automated restricted Boltzmann machine classifier for early diagnosis of Parkinson's disease using digitized spiral drawings. J Ambient Intell Hum Comput. 2023;14:175–89. https://doi.org/10.1007/s12652-022-04361-3.
Tsuyuzaki K, Sato H, Sato K, Nikaido I. Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. Genome Biol. 2020;21(1):9.
Tu AA, Gierahn TM, Monian B, Morgan DM, Mehta NK, Ruiter B, et al. TCR sequencing paired with massively parallel 3′ RNA-seq reveals clonotypic T cell signatures. Nat Immunol. 2019;20(12):1692–9.
Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat Methods. 2017;14(6):565–71.

van der Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
Wagner J, Rapsomaniki MA, Chevrier S, Anzeneder T, Langwieder C, Dykgers A, et al. A single-cell atlas of the tumor and immune ecosystem of human breast cancer. Cell. 2019;177(5):1330-1345.e18.
Wang Z, Zhang P, Sun W, Li D. Application of data dimension reduction method in high-dimensional data based on single-cell 3D genomic contact data. ASP Trans Comput. 2021;1(2):1–6.
Wang Y, Huang H, Rudin C, Shaposhnik Y. Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization. J Mach Learn Res. 2022;22(1):1.
Weber LL, Sashittal P, El-Kebir M. doubletD: detecting doublets in single-cell DNA sequencing data. Bioinformatics. 2021;37(1):i214–21.
Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15.
Wulfman C, Sadoun M, Lamy de la Chapelle M. Interest of Raman spectroscopy for the study of dental material: The zirconia material example [Intérêt de la spectroscopie Raman dans l'étude d'un matériau dentaire : l'exemple de la zircone]. Innov Res Biomed Eng Biomech. 2010. https://doi.org/10.1016/j.irbm.2010.10.004.
Xiang R, Wang W, Yang L, Wang S, Xu C, Chen X. A comparison for dimensionality reduction methods of single-cell RNA-seq data. Front Genet. 2021;12:1.
Yao C, Sun H-W, Lacey NE, Ji Y, Moseman EA, Shih H-Y, et al. Single-cell RNA-seq reveals TOX as a key regulator of CD8+ T cell persistence in chronic infection. Nat Immunol. 2019;20(7):890–901.
Yousuff M, Babu R. Deep autoencoder based hybrid dimensionality reduction approach for classification of SERS for melanoma cancer diagnostics. J Intell Fuzzy Syst. 2022;43(6):7647–61.
Yousuff M, Babu R. Enhancing the classification metrics of spectroscopy spectrums using neural network based low dimensional space. Earth Sci Informatics. 2023;16(1):825–44.
Zegarra Flores J, Radoux JP. Catheter tracking using a convolutional neural network for decreasing interventional radiology X-ray exposure. IRBM. 2023;44(2): 100737. https://doi.org/10.1016/j.irbm.2022.09.004.
Zhang N, Leatham K. A neurodynamics-based nonnegative matrix factorization approach based on discrete-time projection neural network. J Ambient Intell Hum Comput. 2019. https://doi.org/10.1007/s12652-019-01550-5.
Zhao Q. Social emotion classification of Japanese text information based on SVM and KNN. J Ambient Intell Hum Comput. 2021. https://doi.org/10.1007/s12652-021-03034-x.
Zhou M, Du W, Qin K, et al. Distinguish crude and sweated chinese herbal medicine with support vector machine and random forest methods. Wireless Pers Commun. 2018;102:1827–38. https://doi.org/10.1007/s11277-017-5239-3.

Publisher's Note