DOI: 10.1111/cgf.70151

COMPUTER GRAPHICS *forum*
*Volume 0 (2025), number 0, e70151*

# Optimal Dimensionality Selection Using Hull Heatmaps for Single-Cell Analysis

Haejin Jeong,[1] Hyoung-oh Jeong,[2] Semin Lee[2] and Won-Ki Jeong[1]

[1]Korea University, Seoul, South Korea
{haejinjeong, wkjeong}@korea.ac.kr
[2]Ulsan National Institute of Science and Technology, Ulsan, South Korea
{hyoung-oh, seminlee}@unist.ac.kr

**Abstract**

*Single-cell RNA sequencing (scRNA-seq) has gained prominence as a valuable technique for examining cellular gene expression patterns at the individual cell level. In the analysis of scRNA-seq datasets, it is common practice to visualise a subset of principal components (PCs), obtained via principal component analysis (PCA), using dimensionality reduction techniques such as t-stochastic neighbour embedding (t-SNE). Determining the number of PCs (i.e. dimensionality) is a critical step that influences the outcome of single-cell analysis, and this process typically requires a labour-intensive manual assessment involving the inspection of numerous projection plots. To address this challenge, we present a visualisation system that assists analysts in efficiently determining the optimal dimensionality of scRNA-seq data. The proposed system employs two hull heatmaps, a cell type heatmap and a cluster heatmap, which offer comprehensive representations of target cells of multiple cell types across various dimensionalities through the utilisation of a convex hull-embedded colour map. The cell type heatmap shows overlaps between cell types, and the cluster heatmap compares cell clustering results. The proposed hull heatmaps effectively alleviate the labourious task of manually evaluating hundreds of projection plots for searching for the optimal dimensionality. Additionally, our system offers interactive visualisation of gene expression levels and an intuitive lasso selection tool, thereby enabling analysts to progressively refine the convex hulls on the hull heatmaps. We validated the usefulness of the proposed system through two quantitative evaluations and three case studies.*

**Keywords:** scientific visualisation, visual analytics, visualisation, visualisation

**CCS Concepts:** • Human-centred computing → Heat maps; Visual analytics;

## 1. Introduction

The advent of single-cell RNA sequencing (scRNA-seq) technology has made it possible to identify cell-to-cell differences, greatly enhancing our understanding of diseases such as cancer [LTX*21]. The scRNA-seq analysis workflow (Figure 1) consists of multiple stages, including data pre-processing, feature selection, dimensionality reduction, clustering and identification of differentially expressed genes (DEGs) [RSTK17, LT19, KAH19, ZLL*23, SKH*24]. Among these, dimensionality reduction has a substantial impact on the results of the analysis. Selecting a subset of principal components (PCs) derived from principal component analysis (PCA) for visualisation through dimensionality reduction methods such as t-stochastic neighbour embedding (t-SNE) [VdMH08] is a widely adopted practice in single-cell analysis [LT19, KB19, HL20, AKMH20, KL21, XLL24]. Through PCA, the initial high-dimensional scRNA-seq data, often comprising thousands of dimensions, is condensed into a more manageable form, typically fewer than 100 dimensions, while preserving global structural information in the subsequent 2D projection. It in turn, effectively characterises different cell types.

This work focuses on the challenging task of selecting the optimal (PCA intermediate) dimensionality, which is a time-consuming and labourious aspect of single-cell analysis. Finding a good embedding, that is grouping identical cell types and distinguishing diverse cell types within the embedding space, is difficult because the shape of the embedding changes depending on the selected dimensionality (Figure 2). Single-cell analysts identify several target cell types in a dimensionality reduction plot. However, in certain instances, these

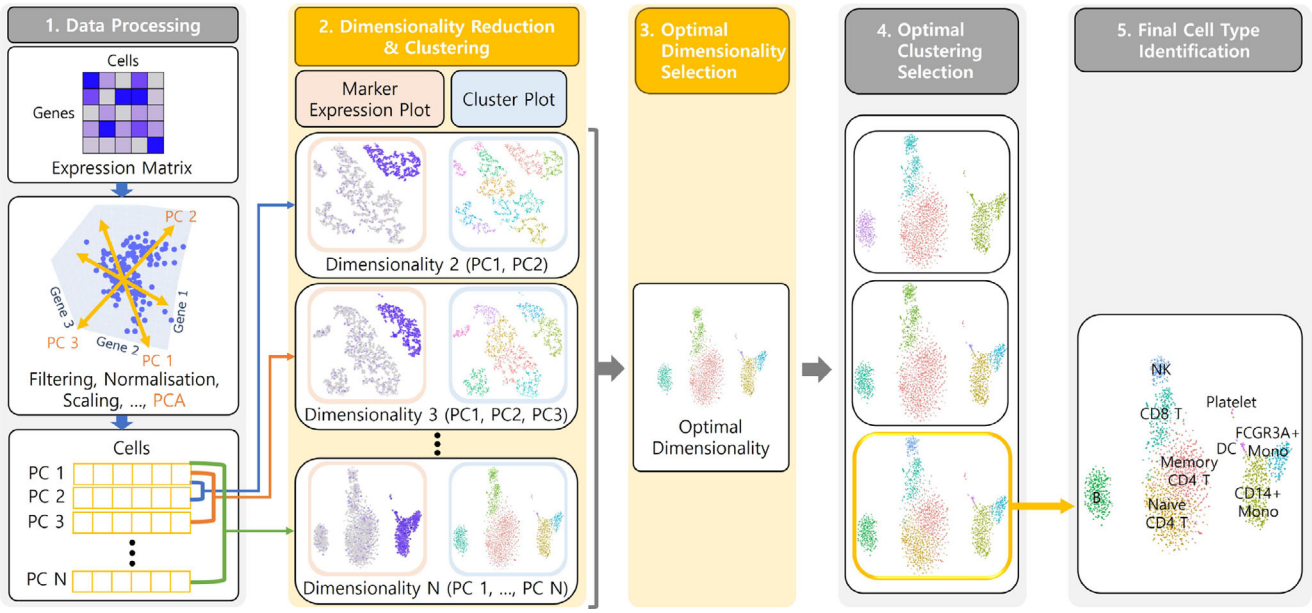


**Figure 1:** *Overview of the single-cell RNA sequencing (scRNA-seq) data analysis process. ScRNA-seq generates an expression matrix that includes thousands of cells and genes. The data undergoes processing steps, which include filtering, normalisation, scaling, and principal component analysis (PCA). Dimensionality reduction and clustering are applied to the top-N principal components (PCs), and the marker expression and cluster plots are created for each dimensionality (i.e. the number of used PCs). Analysts select the optimal dimensionality based on the plots. They subsequently conduct cluster analysis and finalise the identification of cell types. Notably, Steps 2 and 3, which is the process of the optimal dimensionality selection based on the marker expression and cluster plots, are the contributions of our work presented in this paper.*

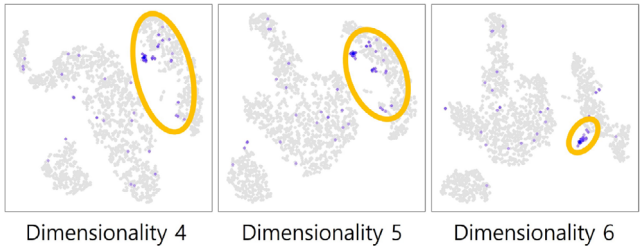


**Figure 2:** *Embedding changes of dendritic cells (DC) according to the dimensionality. In dimensionalities 4 to 5, DC are scattered, making it challenging to identify their cell type. However, in dimensionality 6, DC form a distinct cluster, which enables identifying their cell type. This example illustrates how the choice of dimensionality can influence the outcomes of cell type identification.*

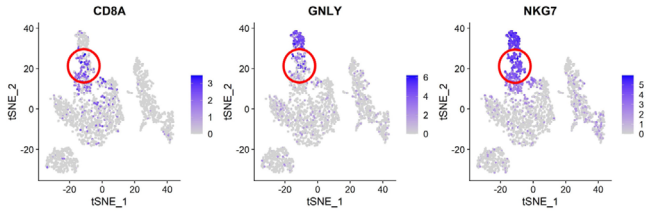


**Figure 3:** *CD8A is a cell marker of CD8+ T cells, and GNLY/NKG7 are cell markers of NK cells. Because there is an overlap (marked with red circles) between different cell types, this dimensionality is hardly considered optimal dimensionality.*

target cell type clusters overlap with each other (Figure 3) or do not align with the outcomes of the clustering algorithm (Figure 4), thereby introducing complexity into the analysis. Therefore, traditional dimensionality selection methods require exhaustive exploration of the dimensionality space to identify the dimensionality that clearly separates multiple target cell types while aligning with the clustering results. This process involves an extensive back-and-forth examination of various embeddings, relying on the analysts' memories for the final decision.

To address this challenge, we introduce a visualisation system that guides the selection of optimal dimensionality in single-cell analysis for improved cell type identification with reduced effort. This work makes **three main contributions**: **(1)** Through interviews with domain experts, we identified **the challenges currently present in scRNA-seq analysis** and defined **system requirements** to address them. **(2)** Based on the requirements, we propose **a novel visualisation technique called the hull heatmap, which provides a holistic overview of marker expression plots across multiple dimensionalities and cell markers**. This technique substantially reduces the labour-intensive process of manually reviewing diverse embeddings. Two variants of the hull heatmap are introduced: the cell type heatmap, which reveals overlaps in cell type areas across dimensionalities, simplifying the identification of the

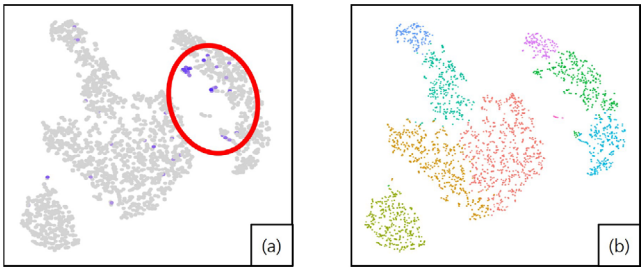


**Figure 4:** *(a) A small number of dendritic cells (DC) are scattered within the red ellipse. (b) However, in the cluster plot, DC and other adjacent cells are grouped into the same green cluster. Therefore, the accuracy of cell type identification in this dimensionality may be lower than in other dimensionalities.*

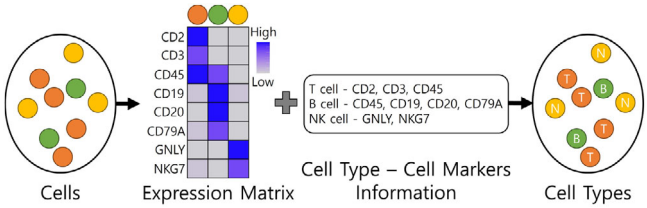


**Figure 5:** *An expression matrix is created using collected cells, and cell types are determined by analysing the matrix in combination with cell markers.*

dimensionality where target cell types separate distinctly, and the cluster heatmap, which facilitates the discovery of the dimensionality where user-defined cell type clusters and clusters generated by a clustering algorithm are the most similar. **(3)** We demonstrate the effectiveness of our proposed system through **two quantitative evaluations and three case studies**. In the quantitative evaluations, using the dimensionality determined by our method yielded results that were similar to or better than those obtained with other dimensionalities. In the case studies, we demonstrate how our method facilitates a more convenient determination of the optimal dimensionality. We also provide an online demo and source codes at https://github.com/hvcl/DESC.

## 2. Background

### 2.1. Overview of single-cell analysis workflow

Gene expression is the process of transcribing genes into RNA. Cell markers denote genes specifically expressed in a particular cell type (Figure 5). These unique combinations of cell markers are instrumental in the classification and identification of individual cells. To discern the diverse cell types and their characteristics, researchers must check the expression levels of cell markers within the cells. This task has been improved by scRNA-seq technique which quantifies gene expression at the individual cell level [TBW*09].

As depicted in Figure 1, the scRNA-seq analysis workflow begins with the scRNA-seq data, providing an expression matrix that details the gene expression levels in each cell. This matrix comprises thousands of cells and genes. Data processing steps such as filtering,

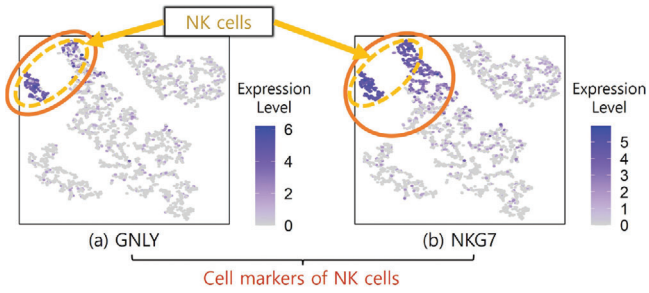


**Figure 6:** *Expression plots for cell markers (a) GNLY and (b) NKG7. The colour indicates the expression level, while the orange ellipses highlight cells where each marker is differentially expressed. Since GNLY and NKG7 are both markers for NK cells, the cells within the yellow dashed ellipses, where both genes are differentially expressed, are identified as NK cells.*

normalisation, and scaling are applied to the expression matrix, followed by dimensionality reduction and clustering. The goal of the single-cell analysis is to determine specific cell types based on the expression level of cell markers, as well as to identify cell markers and relationships between cell types [KAH19, YZC*20, ZLL*23].

Given that different cell types may express the same genes, analysts rely on *differentially* expressed cell markers (genes) to identify cell types. A gene is considered differentially expressed within a cell cluster if a significant difference in expression levels exists among multiple cell clusters. Typically, this examination of cell marker expression is accomplished through 2D projection plots of cells using dimensionality reduction techniques such as t-SNE and uniform manifold approximation and projection (UMAP) [MHM18]. Cells in the plots are colour-coded based on the expression levels of specific markers (Figure 6).

One of the key procedures is dimensionality reduction. It is necessary because of the inherently high dimensionality of scRNA-seq data, which renders clustering and 2D projection results unreliable (i.e. the curse of dimensionality) [KAH19]. In single-cell analysis, PCA is the most prevalent method employed for dimensionality reduction, primarily because it effectively reduces the technical noise associated with individual features in scRNA-seq data. The selected PCs are commonly used for 2D projection through methods such as using PCs as input for t-SNE or UMAP [LT19, KB19, HL20, AKMH20, KL21, XLL24].

Selecting the appropriate number of PCs (i.e. dimensionality) is essential because this choice directly influences the 2D projection result, ultimately affecting cell type identification. The analysts gather valuable insights from the spatial proximity of cells in the 2D projection plot, assuming that similar cells tend to exhibit smaller feature distances. However, certain rare cell types, characterised by a relatively small number of cells, may only become evident in specific dimensionalities. Consequently, relying solely on a small subset of PCs may not effectively represent these cell types [KB19]. Furthermore, Raimundo *et al.* demonstrated that the selection of dimensionality significantly affects the performance scores, such as adjusted mutual information and silhouette, of the embedding of scRNA-seq data [RVV20]. Hence, determining the optimal num-

ber of PCs that efficiently capture critical information related to target cell types while eliminating technical noise and other undesired sources of variability is a challenging yet crucial task.

### 2.2. Optimal dimensionality selection

Conventional approaches to selecting the optimal dimensionality revolve around analysing the changes in data distribution across various dimensionalities. One method involves the use of an elbow plot (Figure S1a) to illustrate the fraction of variance explained by each PC. Analysts are tasked with visually identifying the point at which the curve displays a sharp bend, referred to as the "elbow," and retaining only those PCs preceding this point. Another commonly used method is the JackStraw plot [CS15] (Figure S1b), which presents the distribution of $p$-values for each PC. The best PC is typically observed near a sharp drop (change) in the $p$-values within this plot. However, determining a precise threshold in these plots is not always straightforward because subtle changes may occur, making accurate differentiation challenging. Thus, analysts perform clustering, cell type identification and other downstream analyses across various dimensionalities and compare the results. Throughout this process, they conduct visual inspections of dimensionality reduction plots [ZJ20, SCS*21].

The process of determining optimal dimensionality through the review of marker expression and cluster plots is as follows: Both marker expression and cluster plots share the same 2D embedding from the projection of the selected PCA intermediate dimensionality. First, target cell types and markers must be listed. Analysts typically search for known cell markers of target cell types through literature reviews. A *target cell* is a candidate for the target cell type in which a specific cell marker is highly expressed. Subsequently, the analysts determine the target cells for each target cell type across different dimensionalities based on marker expression plots. As shown in Figure 6, the plots display the expression levels of the genes (that is, the degree to which a particular gene is expressed) in individual cells, with each dot representing a cell, and colour signifying the gene's expression level. Analysts identify target cells where each marker is highly expressed (i.e. cells enclosed in the orange ellipses in Figure 6). Note that GNLY and NKG7 in this instance serve as cell markers for NK cells. Once analysts identify target cells for GNLY and NKG7, they aggregate cells shared between these target cells and assign the final cell type accordingly (as shown in Figure 6, within the yellow dashed ellipses). Some cell types may not be evident in certain dimensionalities, as illustrated in Figure 2, where a few target cells are dispersed across dimensionalities 4–5, complicating the definition of a target cell type cluster. In such cases, analysts need to explore better dimensionalities with a more cohesive cluster, such as dimensionality 6 in Figure 2.

Additionally, analysts evaluate dimensionalities through cluster plots (Figure 7b). These cluster plots are colour-coded based on clustering outcomes achieved at the chosen PCA intermediate dimensionality. If the target cells estimated as the same cell type exhibit disparate clustering or are clustered with cells from other cell types, it may indicate that some target cells have been misclassified (as seen in Figure 7). In such instances, analysts may reevaluate target cells based on the cluster plot or examine cluster plots from different dimensionalities. In conclusion, analysts search for the optimal dimensionality that yields the best target cell clusters, where all target cell types are identifiable, cells of the same cell type cluster closely together while different cell types remain distinctly separated, and the clusters on the cluster plots are most similar to the target cell type clusters defined by analysts. Therefore, the process of determining the optimal dimensionality requires the thorough examination of hundreds of plots, encompassing various dimensionalities and target cell markers.

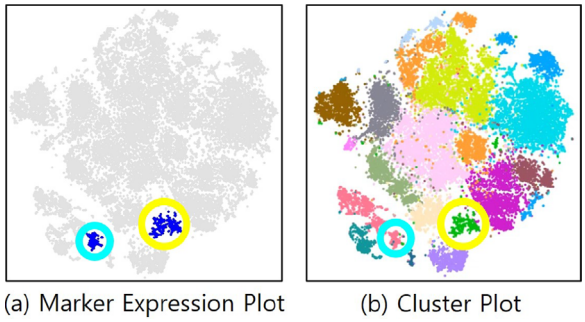


**Figure 7:** *Marker expression plot and cluster plot. (a) The marker expression plot reveals two distinct target cell clusters (cyan and yellow circles) for ILC1-like cells. (b) Nevertheless, on the cluster plot, it is evident that the cells within the cyan circle form a cluster with other non-target cells. Consequently, the cells within the yellow circle are more likely to represent ILC1-like cells.*

## 3. Related Work

### 3.1. Single-cell analysis methods

Clustering is a common approach used for handling the multidimensional nature of scRNA-seq data encompassing diverse cell types. Several publicly available software packages for single-cell analysis, such as Seurat [BHS*18], Scanpy [WAT18], and SINCERA [GWP*15], offer various clustering methods. These tools primarily focus on clustering rather than dimensionality considerations. Although some of these tools provide insights into different dimensionalities, analysts often resort to heuristic approaches to determine optimal dimensionality. These tools are valuable for identifying significant cell clusters and obtaining insights into cell types, especially following the selection of optimal dimensionality.

In single-cell analysis, using pcs as input for t-SNE has been the preferred choice for 2D projections [KAMK19], although more recent techniques, such as UMAP [MHM18], have emerged as alternatives. Deep learning-based dimensionality reduction methods, including parametric t-SNE [VDM09], parametric UMAP [SMG20], scvis [DCS18], DR-A [LMK20], and VASC [WG18], have also gained traction in single-cell analysis; however, they require substantial scRNA-seq datasets for good performance. Various dimensionality reduction methods specifically designed for single-cell analysis exist [VTMP20, RSTK17, VTDP19, WTZ18, TVGP18, AHB*15, SCMD19]. However, to employ these methods effectively, analysts must manually fine-tune the algorithm parameters and data dimensionality, which significantly influence the projection outcomes. The challenge of selecting the appropriate parameters remains unresolved [KAH19].

Most existing R or Python packages for single-cell analysis require analysts to write code. Furthermore, analysts must independently select suitable analytical algorithms and visualisation techniques. To overcome these challenges, various visual analytical tools have been developed to assist analysts in exploring single-cell data. Cerebro [HPL20] provides an intuitive graphical interface for investigating single-cell data. CyteGuide [HPvU*17] enables the exploration of single-cell data hierarchies within a single view. Cytosplore [HPvU*16] offers multiple linked views for the identification of known and unknown cell types. VDJView [SRGL20] visualises scRNA-seq data using metadata profiles, streamlining hypothesis testing, data interpretation, and the discovery of cellular heterogeneity. scQuery [ARP*18] automates the process of downloading and analysing publicly available scRNA-seq datasets. scSVA [TGR19] enables interactive three-dimensional visualisation and exploration of massive single-cell data. However, none of these visual analytical tools is specifically designed for dimensionality selection, making direct comparisons with our system inappropriate.

Widely used visualisation methods for selecting optimal dimensionality include not only the elbow plot and JackStraw plot, as discussed in Section 2.2, but also the DimHeatmap function in Seurat [BHS*18]. It visualises the expression of the top genes that contribute to each PC. Based on this, domain experts assess whether each PC captures sufficient variability to identify their target cell types, and they select an appropriate number of PCs accordingly. A method to automatically determine the optimal dimensionality is provided by findPC [ZWJ22], which is an R package including six methods that computationally detect the elbow point on the elbow plot using different heuristics. However, these methods do not guarantee the detection and separation of all target cell types. Therefore, we propose a method that allows for easier comparison of various dimensionality reduction plots using information about cell types and cell markers, in order to identify the optimal dimensionality that best represents different target cell types.

### 3.2. Visual analysis of dimensionality reduction and clustering

Various analytical methods use dimensionality reduction or clustering with visualisation. Poco *et al.* [PEP*11] and Bernard *et al.* [BHZ*17] used convex hulls to interact with projections. Eckelt *et al.* [EHA*22] introduced an interactive visual method for constructing and investigating the structural relationships in low-dimensional embeddings. DimBridge [MAR*24] employs predicate logic to enable users to identify relevant subspaces by interacting with projections. PRIM-9 [FFT74] enables analysts to explore multidimensional data by using continuously updated projections. Asimov [Asi85] proposed a sequence of orthogonal projections of multidimensional data and searched for suitable sequences to understand the data. Hierarchical Clustering Explorer [SS02] is a visualisation tool for hierarchical clustering that offers an overview of large datasets, dynamic query controls, coordinated displays, and cluster comparisons. Cluster Sculptor [NHM*07] is a cluster analysis framework that allows analysts to interactively fine-tune clustering parameters based on the visualisation of high-dimensional data characteristics. Clustrophile 2 [CD18] recommends diverse clustering parameters and iteratively refines clustering results based on user feedback. INCREMENT [Mit16] enhances clustering outcomes through user feedback by training a feature embedder to map the input features to a new feature space. TINDER [SZS16] introduced a Bayesian prior elicitation framework that incorporates user feedback, allowing analysts to reject clustering results and obtain new results.

Similarly, our method enables users to easily explore multiple projection plots and guides them in searching for optimal solutions based on user feedback. However, the key distinction lies in our approach, which leverages intuitive visualisation-guided techniques to compare over hundred projection plots in a single view (i.e. hull heatmaps and an interactive visual interface). Our method reduces the time and effort required to compare numerous marker expression plots across multiple dimensionalities and cell markers.

## 4. Requirement Analysis for System Design

In the process of designing our visualisation system, we interviewed five domain experts who regularly used Seurat [BHS*18] and SC3 [KKS*17] for single-cell analysis. The requirements were derived through structured interviews with the experts, during which we inquired about the existing analysis workflows, the methods for using the tools, the challenges encountered in current analyses, and the functionalities they would like to see added to the existing workflow. The responses received were then used to formulate the following requirements:

**R1: Visualisation of multiple cell marker expressions across various dimensionalities in a single view**: To determine the optimal dimensionality, analysts typically need to create and assess expression plots for tens of markers and dimensionalities. This task is time-consuming and demanding. In traditional workflows, analysts view multiple marker expression plots on a single screen. Owing to limited screen space, they frequently switch between views to examine hundreds of plots. Moreover, they must commit previous views to memory and mentally compare them, which is a challenging task given the large number of plots to analyse. Therefore, a compact visual representation that enables analysts to simultaneously evaluate the expression of multiple cell markers across various dimensionalities in a single view would significantly simplify their work.

**R2: Tracking target cells of target cell types**: Identifying a particular cell type requires analysts to check the expression of relevant cell markers. Within each dimensionality, expression plots for multiple cell markers must be compared to identify target cells where all markers associated with a specific cell type are highly expressed. This process involves tracking the cells across numerous dimensionalities to ascertain the optimal dimensionality, which requires substantial time and memory resources. Incorporating features that highlight these target cells alleviates the need for continuous comparisons across multiple plots.

**R3: Visual cues to help identify target cells**: Analysts currently rely on colour (expression levels) and spatial positioning in marker expression plots to identify target cells. Since there are no universally accepted criteria (e.g. intercellular distance and intensity thresholds) for defining target cells, analysts subjectively determine them by assessing the relative distances and differences in expression levels between cells. In such cases, visual cues that provide additional information about expression level distribution and local densities enhance the ability to make informed judgements.

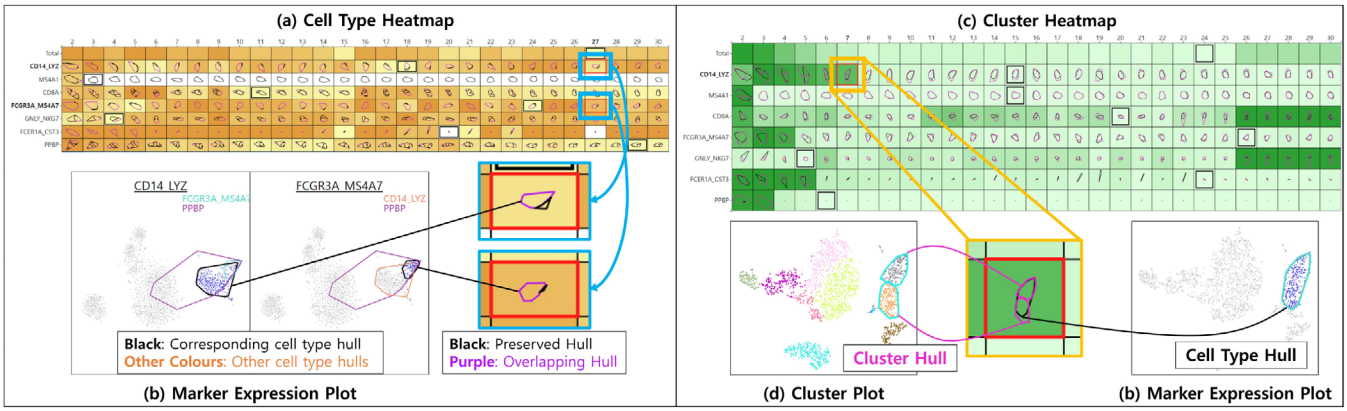


**Figure 8:** *An example of the proposed hull heatmaps. (a) A* cell type heatmap *illustrates the changes in the hull overlap. In each block of the heatmap, a cell type area without overlapping with other cell types is depicted as black convex hulls (referred to as "preserved hulls"), while the regions with overlapping are shown as purple convex hulls (referred to as "overlapping hulls"). (b) Users can examine the expression levels of individual cells and assess the overlap between cell types using a marker expression plot. (c) A* cluster heatmap *shows the changes in the consistency between cell type hulls and cluster hulls. The cell type hull represents the convex hull outlining the cell type region, whereas the cluster hull represents the convex hull of the cluster region on the cluster plot (d). By comparing these two types of hulls, the cluster heatmap aids in the identification of dimensionalities where the cell type estimations and clustering results are similar.*

**R4: Visualisation to help identify overlaps between target cell types**: Analysts aim to find a dimensionality in which all target cell types are distinctly separated. However, two challenges arise: First, cell positions vary with dimensionality, making it impossible for analysts to identify dimensionalities with overlaps until they have reviewed them all. Second, cells may express cell markers from multiple cell types. In such instances, analysts need to identify the cell type by examining the expression levels of various markers and the clustering results. To address these challenges, a visualisation tool that helps prevent unwanted overlap between cell types is essential.

**R5: Visualisation that shows whether the target cell type clusters judged by the analyst are similar to the cell clusters created by the clustering algorithm**: In the case studies from our previous work [JJLJ23], we found that the analysts identified the optimal dimensionality using a cell type heatmap and subsequently compared the defined cell type clusters with the cluster plot to make a final determination regarding the optimal dimensionality. If an algorithm-defined cluster contains different analyst-defined cell type clusters, the cell type identification may be inaccurate in that dimensionality. Therefore, the analyst usually selects the dimensionality at which the analyst-defined cell type cluster and the cluster obtained as a result of clustering are most similar. Thus, visualisation is required to make it easier to compare these clusters in each dimensionality.

## 5. Proposed Method

Our approach assists analysts in determining optimal dimensionality by introducing a novel visualisation technique to explore dimensionality space. We propose hull heatmaps that are convex hull-embedded colour maps to provide comprehensive views of multiple cell type regions across various dimensionalities. Analysts can update hull heatmaps interactively based on their expertise. All dimensionality reduction plots (t-SNE plots) utilised in our system were created using the RunTSNE function in Seurat [BHS*18] with default settings. In the following subsections, we describe the design rationale of the proposed system and how to determine optimal dimensionality using it.

### 5.1. Hull heatmap

We devised hull heatmaps to present the changes in marker expression across various dimensionalities and cell markers within a single view (**R1**). In the heatmaps, rows represent individual cell markers or cell types, whereas columns represent different dimensionalities (Figure 8). To show hundreds of target cell regions in a limited visualisation space, we chose convex hulls, which can visualise the regions in a relatively simpler manner than other methods, such as concave hulls and alpha-shapes. We call a cell in the hull heatmap a block to avoid confusion with the word 'cell', which is also used in the single cell we analyse. Hulls are drawn for each block. Thus, analysts can easily track the area of target cells over multiple dimensionalities without reviewing multiple plots. The hulls of the hull heatmaps consist of two types: cell type or marker hulls and cluster hulls. A cell type or marker hull is a convex hull of the target cells of a cell type or marker, and is defined based on the expression level and local density of the target cells. A cluster hull is a convex hull of cells from different clusters in the clustering results of all cells in a given dataset, which is adjusted by the local density of the cells. Based on these hull definitions, the hull heatmaps are divided into two categories: a cell type heatmap, which reveals overlaps between the cell type hulls, and a cluster heatmap, which compares the cell type and cluster hulls.

#### 5.1.1. *Cell type heatmap*

A cell type heatmap (Figure 8a) was designed to visualise the overlaps between multiple target cell types across several dimensionalities (**R4**). Each cell type block of the heatmap displays cell type hulls of the corresponding cell type and other cell type hulls that

overlap with it, allowing users to easily identify the overlap between cell types.

Users can create groups of target cell markers that track target cell types (e.g. IL7R_CCR7 is a marker group consisting of two markers, IL7R and CCR7); therefore, the blocks are divided into cell type blocks (marker group blocks) and marker blocks. When creating a cell type hull, we identify cells whose expression levels of all group member markers are higher than the expression level threshold, filter the cells based on local density, and create a convex hull of the cells. A marker hull is a convex hull of cells that are also filtered by expression level and local density threshold. These hulls can be modified interactively using a cell filtering function (Section 5.4).

To show the overlap information between the target cell types, a cell type hull is divided into preserved and overlapping hulls. The preserved hulls represent the areas of the cell type hulls not overlapped by the cell type hulls of other target cell types. Overlapping hulls represent the areas of cell type hulls overlapped by the cell type hulls of other target cell types. The preserved hulls are black, and the overlapping hulls are purple. Thus, analysts can easily identify the areas that overlap with other target cell types.

The colour of the block also provides information regarding the overlap. When a cell type block has only preserved hulls, it is white. When a cell type block has both preserved and overlapping hulls, its colour is determined by the preservation ratio. The ratio is defined as the area of the preserved hulls divided by the area of the cell type hull. For colouring, the ratio is normalised in each row of the heatmap. The cell type block with the lowest ratio in the row is dark orange, and the block with the highest ratio is light orange. The colours of the intermediate cells is obtained through linear interpolation. The colour of the marker block is determined by the area of the marker hull. The total row shows the average preservation ratio of all target cell types. Based on this colour visualisation, analysts can easily identify the extent to which different cell type hulls overlap in each dimensionality. The black inner edge, called the highlight edge, highlights the cell type block that has the largest preservation ratio in each row of the heatmap, so analysts can refer to it when they search for dimensionalities without overlap.

#### 5.1.2. *Cluster heatmap*

A cluster heatmap (Figure 8c) is a visualisation that compares the cell type and cluster hulls of each cell type (**R5**). On each block, while the cell type heatmap represents a cell type hull as preserved hulls and overlapping hulls, the cluster heatmap displays a cell type hull and corresponding cluster hulls. Therefore, users can easily find the dimensionality in which the areas indicated by the cell type hull and cluster hulls are the most similar.

To create cluster hulls, we use the clustering function of Seurat [BHS*18] with the default parameters. The function clusters cells using a shared nearest neighbour modularity optimisation-based clustering algorithm [WVE13]. Next, the set of clusters ($S$) corresponding to each cell type is found in each dimensionality, as follows:

$$S = \left\{ C_i \mid \left( \frac{|C_i \cap H|}{|C_i|} > \theta \right) \vee \left( \frac{|C_i \cap H|}{|H|} > \theta \right) \right\} \quad (1)$$

where $C_i$ is a cluster, and $H$ is the set of target cells of the corresponding cell type hull in the dimensionality. $\theta$ denotes a matching threshold and its default value is 0.6. Next, the local density of the cells in each cluster of $S$ is obtained using the Gaussian kernel density estimation [Sco15]. Cluster hulls are created by filtering the cells of each cluster with a density threshold and creating convex hulls of the cells. When the generated cluster hulls and the corresponding cell type hull do not match well, users can directly select the clusters on the cluster plot (Section 5.2) with mouse clicks. In this case, the set of clusters ($S$) is found in each dimensionality as follows:

$$S = \left\{ C_i \mid \exists U_j \in U, \left( \frac{|C_i \cap U_j|}{|C_i|} > \theta \right) \vee \left( \frac{|C_i \cap U_j|}{|U_j|} > \theta \right) \right\} \quad (2)$$

where $U$ is a set of user-selected clusters, $U_j$ is a user-selected cluster, and $C_i$ is a cluster with a dimensionality other than the dimensionality in which the user selects the clusters ($U$). New cluster hulls are then created from $S$ in the same manner as described above, and the cluster heatmap is updated. $\theta$ and the density threshold can be changed by the user.

Each block colour in the cluster heatmap represents the average IoU between the target cells of a cell type hull ($H$) and the cells of each corresponding cluster ($S_i$). The higher the IoU, the brighter the block colour, and the lower the IoU, the darker the block colour. The highlight edge indicates the block with the largest average IoU in the heatmap row. Analysts can observe how the IoU changes between the cell type hull and the corresponding cluster hulls by looking at the block colour and edges of the heatmap. They can identify whether the corresponding clusters are splitting or merging by visually checking the changes in the cluster hulls in the dimensionalities where the block colour changes significantly. Therefore, the user can easily find the optimal dimensionality where the cell type clusters defined by the user (cell type hulls) and the clusters obtained as a result of the clustering (cluster hulls) are the most similar.

### 5.2. Cluster plot

The cluster plot (Figure 8d) is a t-SNE plot that is coloured according to the clustering results of each dimensionality. The clustering results used to generate the cluster plot is the same as those used to create the cluster hulls. Analysts can change the dimensionality of the hull heatmap; subsequently, the cluster plot is updated into a dimensionality reduction plot for the selected dimensionality. After the optimal dimensionality is selected, analysts commonly explore the clustering results performed on the selected dimensionality with various clustering algorithms and parameters, and compare the cell type identification results. Because this step is beyond the scope of this paper, we used the default clustering algorithm and parameters provided in Seurat [BHS*18] to generate the cluster plots.

### 5.3. Marker expression plot

The marker expression plot (Figures 8b) is a dimensionality reduction plot in which the colouration is determined by the expression level of an individual cell marker or target cells belonging to a marker group (cell type). This plot shares the same embedding as the cluster plot. In the marker expression plot specific to a particular cell

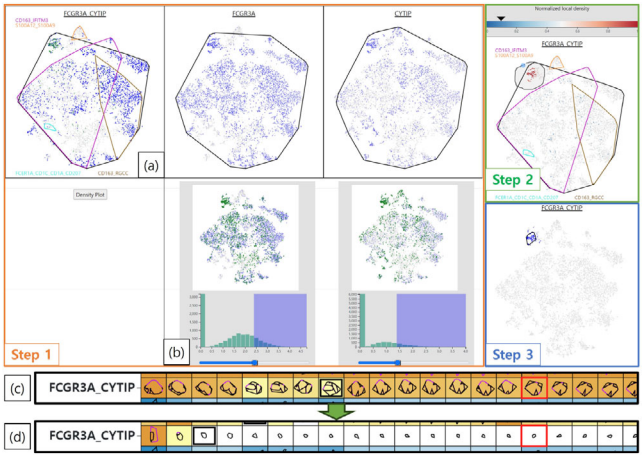


**Figure 9:** *Example of the cell filtering process for (a) the FCGR3A_CYTIP group. (b) In step 1, analysts can change each expression threshold with each cell filtering window of the member markers. (a) The target cells of the group are updated based on the new thresholds, and the colour of the cells changes to green. In step 2, the colour of the target cells changes based on their local densities. Analysts can select a new density threshold using the density colour bar. In step 3, the target cells and cell type hull are updated. (c) Before the cell filtering, there were several overlaps between the FCGR3A_CYTIP group and other cell types. (d) After the cell filtering, the overlaps disappeared and the background colour of the blocks changed to white in the dimensionalities larger than 3.*

type, the hull of that cell type is illustrated, along with any overlapping cell type hulls. These marker expression plots enable analysts to identify target cells and overlapping cell types. Furthermore, the plots show which cell type hulls need to be modified to what shape.

### 5.4. Cell filtering

Before applying cell filtering, the configuration of the cell type or marker hulls rely on the default expression level and density thresholds. Typically, analysts engage in subjective assessments to determine which cell markers are highly expressed within a given cell cluster on marker expression plots. Discrepancies between these subjective judgements and actual hulls can render hull heatmaps and marker expression plots less effective. Analysts must focus on particularly noteworthy cells and monitor their positional changes as the dimensionality evolves (**R2**).

To address this concern, we introduce a feature called *cell filtering* (Figure 9). A cell filtering window comprises three key components: a marker expression plot, an expression level histogram, and an expression threshold slider. A distinct cell filtering window is presented for each constituent marker when cell filtering is conducted for a marker group or cell type. The expression level histogram shows the distribution of marker expression levels across all cells within the dataset (**R4**). Analysts can readily modify the marker expression threshold by sliding the threshold slider, which subsequently results in the visualisation of target cells whose expression levels surpass the updated threshold. These target cells are colour-coded in green on the marker expression plot within the window. As the expression level threshold for any member marker is adjusted, the corresponding target cells are also updated and rendered in green on the marker expression plot (Figure 9a). When the user interacts with the system by clicking a specific button, the target cells are colour-coded based on their local densities (**R4**) (Figure 9 Step 2), and a density colour bar becomes visible. This tool equips analysts with the ability to readily discern regions in which target cells are closely clustered. Subsequently, analysts can select a new density threshold by selecting an appropriate colour from the density colour bar. This mechanism enables analysts to filter out a limited number of target cells that may be positioned far from the remaining cells, thereby preventing excessive hull expansion. An alternative method for selecting cells of interest is to employ a lasso tool for the direct selection of cells within the marker expression plot. This empowers analysts to efficiently filter out less relevant cells. Across all dimensionalities, our system updates the target cells and hulls of the group based on the new density threshold and selected cells (Figure 9 Step 3).

Analysts can leverage the cell filtering feature to explore the expression level distribution of a marker. For example, the expression of IFITM3 in the Clusters 1 and 8 in Figure S2a can be compared. In Figure S2b, the cells within Clusters 1 and 8 appear to exhibit similar expression levels, without any defined expression level threshold. In Figure S2c, cells with an expression level of zero are excluded, and the difference between the two clusters cannot be identified. In Figure S2d, after elevating the threshold, it becomes evident that IFITM3 is highly expressed in cells from Cluster 1 but not in those from Cluster 8. This observation suggests that the two clusters likely represent different cell types.

## 6. Evaluation

To demonstrate the usefulness of our system, we conducted two quantitative evaluations and three case studies with three domain experts (P1-3). The participants' information is listed in Table S1. The quantitative evaluations focus on a comparison of the dimensionality selected using our method and others, and the case studies focus on demonstrating how our method is useful for determining the optimal dimensionality.

### 6.1. Quantitative evaluations

The primary objective of the quantitative evaluations is to numerically assess the performance of the proposed method. To demonstrate that the optimal dimensionality derived from our method contributes to downstream analysis, we compared the performance of state-of-the-art single-cell clustering methods by comparing their scores on clustering evaluation metrics and their similarities of DEGs to the ground truth across clustering results at different dimensionalities.

We compared the differences in clustering performances resulting from variations in input dimensionality, while using the same clustering method. The state-of-the-art scRNA-seq data clustering techniques employed—scGAD [WZZ*24] and scDFC [HLZ*23]—both use PCA dimensionality as the input, similar to our method. Two publicly available datasets were used: epithelial cells and endothelial cells datasets [KKL*20]. For each clustering method and

**Table 1:** *Average clustering metric scores for each method using the epithelial and endothelial cells datasets.*

| | | scGAD | | | scDFC | | |
|---|---|---|---|---|---|---|---|
| | | Default | findPC | Ours | Default | findPC | Ours |
| Epithelial | ARI | **0.917** | 0.909 | 0.911 | 0.717 | 0.842 | **0.875** |
| | NMI | **0.856** | 0.848 | 0.846 | 0.653 | 0.786 | **0.818** |
| Endothelial | ARI | 0.340 | 0.471 | **0.546** | 0.214 | 0.439 | **0.473** |
| | NMI | 0.509 | 0.582 | **0.609** | 0.447 | **0.572** | 0.550 |

dataset, we compared the performance when using three different input dimensionalities: the default dimensionality of the clustering method (256 for scGAD and 512 for scDFC), the dimensionality suggested by findPC [ZWJ22] (4 for the epithelial dataset and 11 for the endothelial dataset), and the dimensionality determined by our approach (8 for both the epithelial and endothelial datasets).

The epithelial cells dataset comprises 3643 cells and 29 634 genes, whereas the endothelial cells dataset contains 2107 cells and 29 634 genes. In the epithelial cells dataset, the participants aimed to identify four target cell types: AT1, AT2, Club, and Ciliated, using the cell markers AGER, SFTPC/LAMP3, SCGB1A1, and FOXJ1/RFX2, respectively. In the endothelial cells dataset, they distinguished five target cell types: Tumor ECs, Tip-like ECs, Stalk-like ECs, Lymphatic ECs, and EPCs, using the cell markers HSPG2/INSR/VWA1, RAMP3/RGCC/ADM, SELP/ACKR1, CCL21/ LYVE1, and TYROBP/C1QB, respectively.

#### 6.1.1. *Clustering evaluation metrics*

We assessed the clustering performance using adjusted rand index (ARI) [HA85] and normalised mutual information (NMI) [DDGDA05] to gauge agreement with the ground truth. For these metrics, higher values indicate superior performance. We ran each clustering method ten times under the same conditions and calculated the average score.

The average clustering metric scores for each method using the epithelial and endothelial cells datasets are listed in Table 1. When using the endothelial dataset with scGAD and the epithelial dataset with scDFC, both ARI and NMI scores were highest when the dimensionality selected by our method was applied. When using the endothelial dataset with scDFC, the ARI score was highest when our method was applied. With the epithelial dataset and scGAD, both ARI and NMI scores were highest when using scGAD's default dimensionality; however, there was minimal difference when applying our method. When using the endothelial dataset with scDFC, the NMI was highest, with a slight difference, when the dimensionality determined by findPC was applied. In summary, our method yielded results that were either better than or comparable to those obtained with other methods.

#### 6.1.2. *DEG similarity*

Using the different dimensionalities, we examined how similar the DEGs detected from the clustering results were to those identified based on ground truth clusters, following the evaluation approach identical to one of those used to assess a scRNA-seq data clustering method [CWZD20]. We used the FindAllMarkers function in Seurat [BHS*18] with default parameters to identify the DEGs for each cluster. Next, we selected the top 100 DEGs from each cluster and calculated the similarity between the top 100 DEGs of clusters obtained through the clustering methods and those of the ground truth clusters. The similarity was calculated by dividing the number of overlapping DEGs by 100. The similarity heatmaps for each clustering result are shown in Figure 10. In each heatmap, the rows represent clusters from each clustering result, and the columns represent ground truth clusters. The heatmap colours range from blue to white to red as the similarity ranges from 0 to 1, with values closer to 0 appearing blue, those near 0.5 appearing white, and values closer to 1 appearing red.

In the scGAD/epithelial results (Figure 10a–c), across all three different dimensionalities, each cell type corresponds to a unique cluster with a single high similarity. In the scGAD/endothelial results (Figure 10g–i), when using scGAD's default dimensionality, there is no cluster matching EPCs (see the far-left column in Figure 10g). When comparing the results of findPC (Figure 10h) and ours (Figure 10i), the similarity of the cluster matching Tip-like ECs (the fourth column from the left) is higher in ours. In the scDFC/epithelial results (Figure 10d–f), when using scDFC's default dimensionality, there is no cluster with high similarity to Club (see the far-right column in Figure 10d). When comparing the results of findPC (Figure 10e) and ours (Figure 10f), the similarity of the cluster matching with AT1 is nearly identical, while the others are higher in ours. In the scDFC/endothelial results (Figure 10j–l), when using scDFC's default dimensionality (Figure 10j), cluster 0 (the first row from the top) shows very low similarity with any cell type, and no clusters display high similarity with cell types other than Lymphatic ECs (the second column from the left). When comparing the results from findPC (Figure 10k) and ours (Figure 10l), both show similar outcomes. In summary, for all results, using the dimensionality identified by our method yielded DEG lists with similarity to the ground truth that is either similar to or higher than when using the default dimensionalities of the two clustering methods or the dimensionality found by findPC.

### 6.2. Case studies

The goal of the case studies was to conduct an in-depth analysis of how our method can help analysts identify optimal dimensionality. Three domain experts analysed three publicly available scRNA-seq datasets [ZTB*17, LWB*18]: peripheral blood mononuclear cells (PBMC), T cells and myeloid-like cells in the lung tumour microenvironment. The analysis dimensionalities ranged from 2 to 30 for the PBMC dataset, and from 2 to 50 for the T cells and myeloid-like cells datasets.

Each case study consisted of the following four steps: (1) loading the dataset into our system, (2) searching the dimensionalities without overlap between target cell types using the cell type heatmap, (3) searching the dimensionalities where cell type and cluster hulls are most similar using the cluster heatmap, and (4) selecting the optimal dimensionality. In our previous work [JJLJ23], the same experts conducted case studies using the same datasets but without the cluster heatmap. The previous case studies were conducted, except

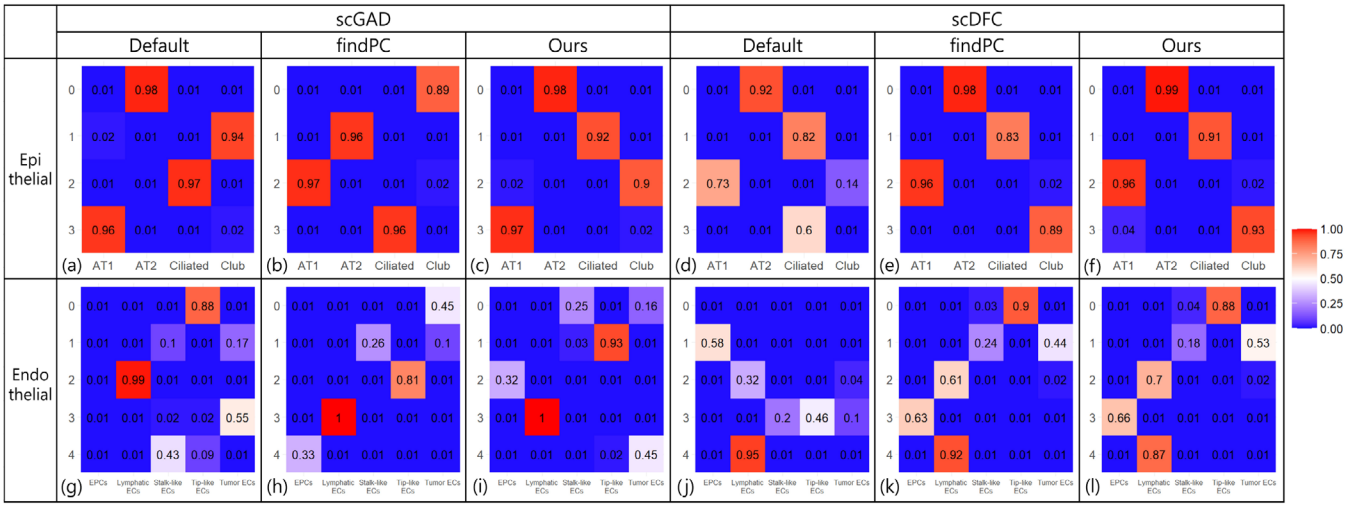


**Figure 10:** *Heatmaps showing the similarity between DEGs identified in each clustering result and those found in the ground truth clusters.*

for the third step of the four steps above. Nineteen months after the previous case studies were conducted, we re-ran the studies using a cluster heatmap and compared the results.

In each case study, domain experts created a cell type heatmap based on the given target cell types and target cell markers information. First, they identified the overlapping pattern of the target cell types based on the colour of each block and the shape of the cell type hulls. Because the dimensionality highlighted by the highlight edge of the total row is where the average overlap between all target cell types is the lowest, they usually started the analysis with that dimensionality. They generated marker expression plots for all target cell types and checked the cell type hulls and overlapping hulls drawn on the plots. For a cell type with many overlapping hulls drawn on a marker expression plot, there is a high probability that the cell type hull was created inappropriately. The experts modified the cell type hulls through cell filtering starting with cell types with many overlapping hulls. During the cell filtering process, they checked the expression level and local density distribution and modified the cell type hull mainly using lasso selection. After finishing all cell filtering and completing satisfactory cell type hulls, they identified the smallest dimensionality without overlap, as indicated by the highlight edge of the total row.

Next, they created a cluster heatmap and identified changes in the clusters by examining the block colour and shape of the hulls of the cluster heatmap, as they did when creating a cell type heatmap. For cell types with pink blocks for which no corresponding clusters were selected, they directly selected appropriate corresponding clusters from the cluster plot and modified the cluster hulls of those cell types. They found that the cluster hull was often too large and modified it to become tighter by increasing the density threshold. After modifying the cluster hulls, they checked the cluster heatmap for dimensionalities with sharp colour changes. They checked the cluster and marker expression plots to see if there was actually a cluster change in the dimensionalities, and then selected the last dimensionality in which a significant cluster change occurred as the optimal dimensionality. In the previous study, experts reviewed the cluster plot for most of the 49 dimensionalities; however, in this study, owing to the cluster heatmap, the experts checked the cluster plot for fewer than 10 dimensionalities. In addition, because there was no need to check all clusters on the cluster plot in each dimensionality, only the corresponding clusters visualised in the cluster heatmap were needed; thus, the time to look at one dimensionality was greatly reduced.

We identified several noteworthy points in the study using the myeloid-like cells dataset. The target cell types used in the study were as follows: Langerhans, Tumor-associated, Cross-presenting dendritic, Granulocytes, Lung-associated macrophage, and Monocyte-derived dendritic cells. The cell markers used to identify the cell types were as follows: FCER1A/CD1C/CD1A/CD207, CD163/IFITM3, CLEC9A/XCR1, S100A12/S100A9, CD163/RGCC, and FCGR3A/CYTIP.

With this dataset, the experts encountered more challenges when defining hulls compared to other datasets because of the larger overlapping areas among the target cell types. Consequently, they found that it was necessary to refer to cluster plots more frequently during the hull definition process. When they were adjusting the hull of FCGR3A_CYTIP, they observed FCGR3A and CYTIP were highly expressed in the middle of their respective marker expression plots. As a result, they initially assumed that the target cells of FCGR3A_CYTIP would be clustered in the middle when analysing the individual expression plots. However, our system's visualisation demonstrated that there were relatively few target cells in which both genes were highly expressed in the middle of the plot, as shown in Figure 9. This visualisation aided in locating the cell type area with greater precision.

In the previous study, after eliminating overlaps between different target cell types, the experts compared cell type hulls and cluster plots by increasing the dimensionality one step at a time. They observed that FCER1A_CD1C_CD1A_CD207 and CLE9A_XCR1

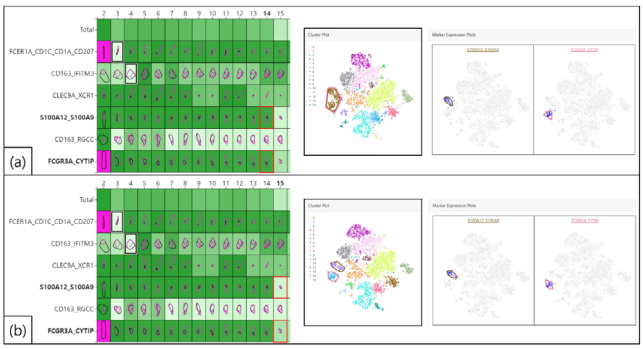


**Figure 11:** *(a) The cluster hull of S100A12_S100A9 and FCGR3A_CYTIP includes both cell type hulls of the two cell types in dimensionality 14. (b) The cluster hull is divided into two in dimensionality 15.*

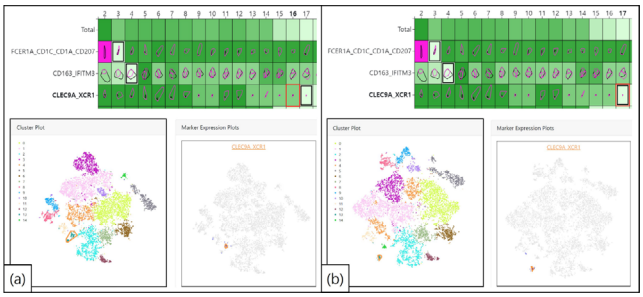


**Figure 12:** *(a) The cluster hull of CLEC9A_XCR1 includes the corresponding cell type hull and other areas in dimensionality 16. (b) The cluster hull is reduced to a similar shape as the corresponding cell type hull in dimensionality 17.*

appeared to be clustered together, and S100A12_S100A9 and FCGR3A_CYTIP were in the same cluster, as shown in Figure S3a. Subsequently, they aimed to identify the specific dimensionality in which all target cell types were consistently separated into distinct clusters in the cluster plot. Through their analysis, they determined that FCER1A_CD1C_CD1A_CD207 and CLE9A_XCR1 began to appear in different clusters with dimensionality 9 or higher, as shown in Figure S3b. Similarly, S100A12_S100A9 and FCGR3A_CYTIP were found in different clusters with dimensionality 15 or higher (Figure S3c). Consequently, the experts selected dimensionality 15 as optimal.

In this study, the cell type heatmap analysis results were the same, but the final optimal dimensionality was changed through cluster heatmap analysis. The cluster separation of S100A12_S100A9 and FCGR3A_CYTIP at dimensionality 15, which was discovered by increasing the dimensionality individually in the previous study, was observed as a sharp change in block colour in the cluster heatmap (Figure 11). Furthermore, a drastic change in the block colour was observed at dimensionality 17 in CLEC9A_XCR1 (Figure 12). It was observed that the cluster hulls became drastically smaller in the dimensionality, and when checking the cluster plots, it was confirmed that the corresponding clusters were split when the dimensionality changed. Therefore, the optimal dimensionality was 17, where the split in the corresponding cluster occurred last.

Descriptions and figures from the studies using the PBMC and T cells datasets, along with overall feedback comparing existing methods with ours—based on interviews with the experts involved in the case studies—can be found in Supplementary Section S2.

## 7. Discussion and Limitation

In this work, we focused on selecting the optimal number of PCs for t-SNE projection. Although it is feasible to directly employ raw input data for 2D projection using t-SNE, a straightforward application of t-SNE is not effective. This is mainly because of the high dimensionality (i.e. the substantial number of genes) of scRNA-seq data. Consequently, the distances between the cells tend to be quite similar, making it challenging to preserve global structures. Hence, the standard practice in single-cell analysis involves using PCs as input for t-SNE [LT19, KB19, HL20, AKMH20, KL21, XLL24].

T-SNE introduces various hyperparameters, including perplexity, learning rate, and number of iterations. Many single-cell analysis studies employ t-SNE with its default hyperparameters [KKS*17, CWZD20, XHZ*20, INE*22]. Xiang *et al.* applied t-SNE with perplexity values ranging from 1 to 50 on single-cell RNA-seq datasets, followed by clustering, and evaluated the results using ARI and NMI against the true cell types. ARI and NMI increased as the perplexity was raised from 2 to 5, after which both metrics tended to stabilise [XWY*21]. Kobak *et al.* have suggested that adjusting t-SNE hyperparameters might be necessary for extremely large datasets (e.g. with n >> 100,000) [KB19]. However, in many state-of-the-art methods for scRNA-seq data [CM22, CD22, HLZ*23, WZZ*23, WXW*23, WZZ*24, XRT*24], the sizes of the real-world datasets used for the evaluations do not exceed 100 000, and this is also the case for the dataset we utilised. There are lines of work targeting large scRNA-seq datasets [XHZ*20, ZHZ21, RLG*24], but this is beyond the scope of this paper. With reference to previous studies, we used Seurat's RunTSNE function with its default parameters (the default perplexity is 30). Further in-depth investigation into the sensitivity of t-SNE to its hyperparameters would be worthwhile.

We replicate the approach used by analysts in conventional methods, where they visually assess overlap between cell types, by recommending dimensionality based on the overlap of convex hulls for target cells within the cell type heatmap. However, a limitation exists: even when labelling through convex hulls does not accurately reflect relationships between cells in the original high-dimensional space, the dimensionality may still receive a high rating based on the degree of overlap. Label-T&C [JKM*23] measure the difference between cluster-label matching evaluated in both original and embedded spaces. Applying such an approach to our method could enhance its reliability. Our goal is to improve the ease of use of existing analyses that utilise t-SNE. We do not claim that t-SNE is entirely accurate, and addressing its limitations lies beyond the scope of our work. Our method shares the known limitations and criticisms associated with t-SNE.

In the case of rare cell types, we observed that choosing the optimal number of PCs could be challenging and complex using conventional methods because significant changes often occur in the most

important PCs. However, less-prominent PCs may convey vital information about infrequent cell types. Our approach provides analysts with a straightforward means of tracking positional changes in target cells across various dimensionalities, even aiding in the identification of less significant PCs where all target cell types are discernible. However, the limitation of our method is that it selects the top-N PCs and creates them within a single dimensionality, which restricts the comparison of various PC combinations.

## 8. Conclusion and Future Work

In this paper, we presented a novel visual analytics system designed for the interactive determination of optimal dimensionality in dimensionality reduction, particularly in the context of cell type identification. Through the introduction of novel visualisation tools, such as cell type and cluster heatmaps, along with several interactive cell filtering methods, our approach significantly streamlines the process of reviewing a substantial number of dimensionality reduction plots. We demonstrate the effectiveness and practicality of our proposed system by conducting quantitative evaluations and three case studies.

In the future, we aim to expand our system by incorporating a range of dimensionality reduction techniques, including UMAP. Currently, our system exclusively supports the default t-SNE hyperparameters. However, investigating the impact of various t-SNE hyperparameters in conjunction with PCA dimensionality is an intriguing avenue for future research. Additionally, we plan to apply various class separability evaluation measures, such as Sepme [AS16] and Label-T&C [JKM*23], and compare their results. Moreover, the combination of dimensionality reduction and clustering provides additional opportunities for visualisation and interaction [WCR*17]. Leveraging this advantage, we aim to explore methods for identifying the optimal combination of dimensionality reduction and clustering in a single step.

## Acknowledgements

This work was partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2021-NR060143 and RS-2024-00349697), the National Research Council of Science & Technology (NST) grant by MSIT (No. GTL24031-000), the ICT Creative Consilience program of the Institute for Information & communications Technology Planning & Evaluation (IITP) funded by MSIT (IITP-2025-RS-2020-II201819), and a Korea University Grant.

## Conflict of Interest

The authors declare that they have no conflict of interest.

## References

[AHB*15] Angerer P., Haghverdi L., Büttner M., Theis F. J., Marr C., Buettner F.: Destiny: Diffusion maps for large-scale single-cell data in R. *Bioinformatics 32*, 8 (December 2015), 1241–1243. URL: https://doi.org/10.1093/bioinformatics/btv715, https://academic.oup.com/bioinformatics/article-pdf/32/8/1241/16920896/btv715.pdf.

[AKMH20] Andrews T. S., Kiselev V. Y., McCarthy D., Hemberg M.: Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nature Protocols 16*, 1 (2020), 1–9.

[ARP*18] Alavi A., Ruffalo M., Parvangada A., Huang Z., Bar-Joseph Z.: A web server for comparative analysis of single-cell RNA-seq data. *Nature communications 9*, 1 (2018), 4768.

[AS16] Aupetit M., Sedlmair M.: Sepme: 2002 new visual separation measures. In *2016 IEEE Pacific Visualization Symposium (PacificVis)* (2016), IEEE, pp. 1–8.

[Asi85] Asimov D.: The grand tour: A tool for viewing multidimensional data. *SIAM Journal on Scientific and Statistical Computing 6*, 1 (1985), 128–143.

[BHS*18] Butler A., Hoffman P., Smibert P., Papalexi E., Satija R.: Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology 36*, 5 (2018), 411–420.

[BHZ*17] Bernard J., Hutter M., Zeppelzauer M., Fellner D., Sedlmair M.: Comparing visual-interactive labeling with active learning: An experimental study. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2017), 298–308.

[CD18] Cavallo M., Demiralp Ç.: Clustrophile 2: Guided visual clustering analysis. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2018), 267–276.

[CD22] Ciortan M., Defrance M.: GNN-based embedding for clustering scrna-seq data. *Bioinformatics 38*, 4 (2022), 1037–1044.

[CM22] Cheng Y., Ma X.: scGAC: A graph attentional architecture for clustering single-cell RNA-seq data. *Bioinformatics 38*, 8 (2022), 2187–2193.

[CS15] Chung N. C., Storey J. D.: Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics 31*, 4 (2015), 545–554.

[CWZD20] Chen L., Wang W., Zhai Y., Deng M.: Deep soft k-means clustering with self-training for single-cell RNA sequence data. *NAR Genomics and Bioinformatics 2*, 2 (2020), lqaa039.

[DCS18] Ding J., Condon A., Shah S. P.: Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nature Communications 9*, 1 (May 2018), 2002, https://doi.org/10.1038/s41467-018-04368-5. URL: https://europepmc.org/articles/PMC5962608.

[DDGDA05] Danon L., Diaz-Guilera A., Duch J., Arenas A.: Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment 2005*, 09 (2005), P09008.

[EHA*22] Eckelt K., Hinterreiter A., Adelberger P., Walchshofer C., Dhanoa V., Humer C., Heckmann M., Steinparz C., Streit M.: Visual exploration of relationships and structure in low-dimensional embeddings. *IEEE Transactions on Visualization and Computer Graphics 29*, 7 (2022), 3312–3326.

[FFT74] Fisherkeller M. A., Friedman J. H., Tukey J. W.: An interactive multidimensional data display and analysis system. Tech. rep., SLAC National Accelerator Lab., Menlo Park, CA, 1974.

[GWP*15] Guo M., Wang H., Potter S. S., Whitsett J. A., Xu Y.: Sincera: a pipeline for single-cell RNA-seq profiling analysis. *PLoS Computational Biology 11*, 11 (2015).

[HA85] Hubert L., Arabie P.: Comparing partitions. *Journal of Classification 2*, 1 (1985), 193–218.

[HL20] Heiser C. N., Lau K. S.: A quantitative framework for evaluating single-cell data structure preservation by dimensionality reduction techniques. *Cell Reports 31*, 5 (2020).

[HLZ*23] Hu D., Liang K., Zhou S., Tu W., Liu M., Liu X.: scDFC: A deep fusion clustering method for single-cell RNA-seq data. *Briefings in Bioinformatics 24*, 4 (2023), bbad216.

[HPL20] Hillje R., Pelicci P. G., Luzi L.: Cerebro: Interactive visualization of scRNA-seq data. *Bioinformatics 36*, 7 (2020), 2311–2313.

[HPvU*16] Höllt T., Pezzotti N., van Unen V., Koning F., Eisemann E., Lelieveldt B., Vilanova A.: Cytosplore: Interactive immune cell phenotyping for large single-cell datasets. In *Computer Graphics Forum* (2016), vol. *35*, Wiley Online Library, pp. 171–180.

[HPvU*17] Höllt T., Pezzotti N., van Unen V., Koning F., Lelieveldt B. P., Vilanova A.: Cyteguide: Visual guidance for hierarchical single-cell analysis. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2017), 739–748.

[INE*22] Imoto Y., Nakamura T., Escolar E. G., Yoshiwaki M., Kojima Y., Yabuta Y., Katou Y., Yamamoto T., Hiraoka Y., Saitou M.: Resolution of the curse of dimensionality in single-cell RNA sequencing data analysis. *Life Science Alliance 5*, 12 (2022).

[JJLJ23] Jeong H., Jeong H.-o., Lee S., Jeong W.-K.: Dimensionality explorer for single-cell analysis. In *2023 IEEE 16th Pacific Visualization Symposium (PacificVis)* (2023), IEEE, pp. 51–60.

[JKM*23] Jeon H., Kuo Y. H., Aupetit M., Ma K. L., Seo J.: Classes are not clusters: Improving label-based evaluation of dimensionality reduction. *IEEE Transactions on Visualization and Computer Graphics 30*, 1 (2023), 781–791.

[KAH19] Kiselev V. Y., Andrews T. S., Hemberg M.: Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics 20*, 5 (2019), 273–282.

[KAMK19] Kulkarni A., Anderson A. G., Merullo D. P., Konopka G.: Beyond bulk: A review of single cell transcriptomics methodologies and applications. *Current Opinion in Biotechnology 58* (2019), 129–136.

[KB19] Kobak D., Berens P.: The art of using t-SNE for single-cell transcriptomics. *Nature Communications 10*, 1 (2019), 1–14.

[KKL*20] Kim N., Kim H. K., Lee K., Hong Y., Cho J. H., Choi J. W., Lee J.-I., Suh Y.-L., Ku B. M., Eum H. H., et al.: Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nature communications 11*, 1 (2020), 2285.

[KKS*17] Kiselev V. Y., Kirschner K., Schaub M. T., Andrews T., Yiu A., Chandra T., Natarajan K. N., Reik W., Barahona M., Green A. R., Hemberg M.: SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods 14*, 5 (2017), 483–486.

[KL21] Kobak D., Linderman G. C.: Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology 39*, 2 (2021), 156–157.

[LMK20] Lin E., Mukherjee S., Kannan S.: A deep adversarial variational autoencoder model for dimensionality reduction in single-cell RNA sequencing analysis. *BMC Bioinformatics 21*, 1 (2020), 1–11.

[LT19] Luecken M. D., Theis F. J.: Current best practices in single-cell RNA-seq analysis: A tutorial. *Molecular Systems Biology 15*, 6 (2019).

[LTX*21] Lei Y., Tang R., Xu J., Wang W., Zhang B., Liu J., Yu X., Shi S.: Applications of single-cell sequencing in cancer research: progress and perspectives. *Journal of Hematology & Oncology 14*, 1 (2021), 91.

[LWB*18] Lambrechts D., Wauters E., Boeckx B., Aibar S., Nittner D., Burton O., Bassez A., Decaluwé H., Pircher A., Van den Eynde K., Weynand B., Verbeken E., De Leyn P., Liston A., Vansteenkiste J., Carmeliet P., Aerts S., & Thienpont B.: Phenotype molding of stromal cells in the lung tumor microenvironment. *Nature Medicine 24*, 8 (2018), 1277–1289.

[MAR*24] Montambault B., Appleby G., Rogers J., Brumar C. D., Li M., Chang R.: DimBridge: Interactive Explanation of Visual Patterns in Dimensionality Reductions with Predicate Logic. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[MHM18] McInnes L., Healy J., Melville J.: Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[Mit16] Mitchell L. A.: Increment-interactive cluster refinement. Brigham Young University (2016).

[NHM*07] Nam E. J., Han Y., Mueller K., Zelenyuk A., Imre D.: Clustersculptor: A visual analytics tool for high-dimensional data. In *2007 IEEE Symposium on Visual Analytics Science and Technology* (2007), IEEE, pp. 75–82.

[PEP*11] Poco J., Etemadpour R., Paulovich F. V., Long T., Rosenthal P., Oliveira M. C. F. d., Linsen L., Minghim R.: A framework for exploring multidimensional data with 3D projections. In *Computer Graphics Forum* (2011), vol. *30*, Wiley Online Library, pp. 1111–1120.

[RLG*24] Ren J., Lyu X., Guo J., Shi X., Zhou Y., Li Q.: CDSKNN$^{XMBD}$: A novel clustering framework for large-scale single-cell data based on a stable graph structure. *Journal of Translational Medicine 22*, 1 (2024), 233.

[RSTK17] Rostom R., Svensson V., Teichmann S. A., Kar G.: Computational approaches for interpreting scRNA-seq data. *FEBS Letters 591*, 15 (2017), 2213–2225.

[RVV20] Raimundo F., Vallot C., Vert J.-P.: Tuning parameters of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biology 21*, 1 (2020), 1–17.

[SCMD19] Szubert B., Cole J. E., Monaco C., Drozdov I.: Structure-preserving visualisation of high dimensional single-cell datasets. *Scientific Reports 9*, 1 (June 2019), 8914, https://doi.org/10.1038/s41598-019-45301-0. https://europepmc.org/articles/PMC6586841.

[Sco15] Scott D. W.: *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, Hoboken, 2015.

[SCS*21] Schneider I., Cepela J., Shetty M., Wang J., Nelson A. C., Winterhoff B., Starr T. K.: Use of "default" parameter settings when analyzing single cell RNA sequencing data using seurat: a biologist's perspective. *Journal of Translational Genetics and Genomics 5* (2021), 37–49.

[SKH*24] Sun Y., Kong L., Huang J., Deng H., Bian X., Li X., Cui F., Dou L., Cao C., Zou Q., Zhang Z.: A comprehensive survey of dimensionality reduction and clustering methods for single-cell and spatial transcriptomics data. *Briefings in Functional Genomics* (2024), elae023.

[SMG20] Sainburg T., McInnes L., Gentner T. Q.: Parametric UMAP embeddings for representation and semisupervised learning. *Neural Computation 33*, 11 (2021), 2881–2907.

[SRGL20] Samir J., Rizzetto S., Gupta M., Luciani F.: Exploring and analysing single cell multi-omics data with VDJView. *BMC Medical Genomics 13*, 1 (2020), 1–9.

[SS02] Seo J., Shneiderman B.: Interactively exploring hierarchical clustering results [gene identification]. *Computer 35*, 7 (2002), 80–86.

[SZS16] Srivastava A., Zou J., Sutton C.: Clustering with a Reject Option: Interactive Clustering as Bayesian Prior Elicitation. In *33rd International Conference on Machine Learning: ICML 2016* (2016, July), pp. 16–20.

[TBW*09] Tang F., Barbacioru C., Wang Y., Nordman E., Lee C., Xu N., Wang X., Bodeau J., Tuch B. B., Siddiqui A., Lao K., Surani M.A.: mRNA-seq whole-transcriptome analysis of a single cell. *Nature Methods 6*, 5 (2009), 377–382.

[TGR19] Tabaka M., Gould J., Regev A.: scSVA: An interactive tool for big data visualization and exploration in single-cell omics. *BioRxiv* (2019), 512582.

[TVGP18] Tasoulis S. K., Vrahatis A. G., Georgakopoulos S. V., Plagianakos V. P.: Visualizing high-dimensional single-cell RNA-sequencing data through multiple random projections. In *2018 IEEE International Conference on Big Data (Big Data)* (2018), pp. 5448–5450.

[VDM09] Van Der Maaten L.: Learning a parametric embedding by preserving local structure. In *Artificial Intelligence and Statistics* (2009), PMLR, pp. 384–391.

[VdMH08] Van der Maaten L., Hinton G.: Visualizing data using t-SNE. *Journal of Machine Learning Research 9*, 11 (2008).

[VTDP19] Vrahatis A. G., Tasoulis S. K., Dimitrakopoulos G. N., Plagianakos V. P.: Visualizing high-dimensional single-cell RNA-seq data via random projections and geodesic distances. In *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (2019), IEEE, pp. 1–6.

[VTMP20] Vrahatis A. G., Tasoulis S. K., Maglogiannis I., Plagianakos V. P.: Recent machine learning approaches for single-cell RNA-seq data analysis. In *Advanced Computational Intelligence in Healthcare-7*. Springer, 2020, pp. 65–79.

[WAT18] Wolf F. A., Angerer P., Theis F. J.: SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology 19*, 1 (2018), 15.

[WCR*17] Wenskovitch J., Crandell I., Ramakrishnan N., House L., North C.: Towards a systematic combination of dimension reduction and clustering in visual analytics. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2017), 131–141.

[WG18] Wang D., Gu J.: VASC: Dimension reduction and visualization of single-cell rna-seq data by deep variational autoencoder. *Genomics, Proteomics & Bioinformatics 16*, 5 (2018), 320–331.

[WTZ18] Wu Y., Tamayo P., Zhang K.: Visualizing and interpreting single-cell gene expression datasets with similarity weighted nonnegative embedding. *Cell Systems 7*, 6 (2018), 656–666.

[WVE13] Waltman L., Van Eck N. J.: A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B 86*, 11 (2013), 1–14.

[WXW*23] Wang J., Xia J., Wang H., Su Y., Zheng C.-H.: scDCCA: Deep contrastive clustering for single-cell RNA-seq data based on auto-encoder network. *Briefings in Bioinformatics 24*, 1 (2023), bbac625.

[WZZ*23] Wang S., Zhang Y., Zhang Y., Wu W., Ye L., Li Y., Su J., Pang S.: scASGC: An adaptive simplified graph convolution model for clustering single-cell RNA-seq data. *Computers in Biology and Medicine 163* (2023), 107152.

[WZZ*24] Wang S., Zhang Y., Zhang Y., Zhang Y., Pang S., Su J., Liu Y.: Graph attention autoencoder model with dual decoder for clustering single-cell RNA sequencing data. *Applied Intelligence 54* (2024), 5136–5146.

[XHZ*20] Xie K., Huang Y., Zeng F., Liu Z., Chen T.: scAIDE: Clustering of large-scale single-cell RNA-seq data reveals putative and rare cell types. *NAR Genomics and Bioinformatics 2*, 4 (2020), lqaa082.

[XLL24] Xia L., Lee C., Li J. J.: Statistical method scDEED for detecting dubious 2d single-cell embeddings and optimizing t-SNE and UMAP hyperparameters. *Nature Communications 15*, 1 (2024), 1753.

[XRT*24] Xie J., Ruan S., Tu M., Yuan Z., Hu J., Li H., Li S.: Clustering single-cell RNA sequencing data via iterative smoothing and self-supervised discriminative embedding. *Oncogene 43*, 9 (2024), 2279–2292.

[XWY*21] Xiang R., Wang W., Yang L., Wang S., Xu C., Chen X.: A comparison for dimensionality reduction methods of single-cell RNA-seq data. *Frontiers in Genetics 12* (2021), 646936.

[YZC*20] Yu X., Zhang L., Chaudhry A., Rapaport A. S., Ouyang W.: Unravelling the heterogeneity and dynamic relationships of tumor-infiltrating T cells by single-cell RNA sequencing analysis. *Journal of Leukocyte Biology 107*, 6 (2020), 917–932.

[ZHZ21] Zou Z., Hua K., Zhang X.: HGC: Fast hierarchical clustering for large-scale single-cell data. *Bioinformatics 37*, 21 (2021), 3964–3965.

[ZJ20] Zhou B., Jin W.: Visualization of single cell RNA-seq data using t-SNE in R. *Stem Cell Transcriptional Networks: Methods and Protocols 2117* (2020), 159–167.

[ZLL*23] Zhang S., Li X., Lin J., Lin Q., Wong K.-C.: Review of single-cell RNA-seq data clustering for cell-type identification and characterization. *RNA 29*, 5 (2023), 517–530.

[ZTB*17] Zheng G. X., Terry J. M., Belgrader P., Ryvkin P., Bent Z. W., Wilson R., Ziraldo S. B., Wheeler T. D., McDermott G. P., Zhu J., et al.: Massively parallel digital transcriptional profiling of single cells. *Nature Communications 8*, 1 (2017), 14049.

[ZWJ22] Zhuang H., Wang H., Ji Z.: findPC: An R package to automatically select the number of principal components in single-cell analysis. *Bioinformatics 38*, 10 (2022), 2949–2951.

## Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information

Video 1