BMC Bioinformatics

RESEARCH

Open Access

scSMD: a deep learning method for accurate clustering of single cells based on auto-encoder

Xiaoxu Cui[1,2,4†], Renkai Wu[3,4†], Yinghao Liu[1,2,4], Peizhan Chen[5], Qing Chang[4], Pengchen Liang[3*] and Changyu He[4*]

†Xiaoxu Cui and Renkai Wu contributed equally to this work.

*Correspondence: liangpengchen@shu.edu.cn; hechangyu.2008@163.com

[1] School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China
[2] Shanghai University of Medicine & Health Sciences, Shanghai, China
[3] School of Microelectronics, Shanghai University, Shanghai, China
[4] Department of Surgery, Shanghai Key Laboratory of Gastric Neoplasms, Shanghai Institute of Digestive Surgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[5] Clinical Research Center, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Abstract

Background: Single-cell RNA sequencing (scRNA-seq) has transformed biological research by offering new insights into cellular heterogeneity, developmental processes, and disease mechanisms. As scRNA-seq technology advances, its role in modern biology has become increasingly vital. This study explores the application of deep learning to single-cell data clustering, with a particular focus on managing sparse, high-dimensional data.

Results: We propose the SMD deep learning model, which integrates nonlinear dimensionality reduction techniques with a porous dilated attention gate component. Built upon a convolutional autoencoder and informed by the negative binomial distribution, the SMD model efficiently captures essential cell clustering features and dynamically adjusts feature weights. Comprehensive evaluation on both public datasets and proprietary osteosarcoma data highlights the SMD model's efficacy in achieving precise classifications for single-cell data clustering, showcasing its potential for advanced transcriptomic analysis.

Conclusion: This study underscores the potential of deep learning-specifically the SMD model-in advancing single-cell RNA sequencing data analysis. By integrating innovative computational techniques, the SMD model provides a powerful framework for unraveling cellular complexities, enhancing our understanding of biological processes, and elucidating disease mechanisms. The code is available from https://github.com/xiaoxuc/scSMD.

Keywords: ScRNA-seq, Deep clustering, Deep learning, Multi-dilated attention gate

Introduction

Single-cell RNA sequencing (scRNA-seq) technology has become an essential high-throughput genomics tool for uncovering cellular heterogeneity and complexity within tissues and systems [1, 2]. By enabling whole-genome or transcriptome analysis at the single-cell level, scRNA-seq offers unparalleled resolution in identifying the cellular diversity within organisms [3]. As single-cell sequencing technology advances, vast

amounts of single-cell genomics data have been generated. However, accurately identifying and characterizing the diverse cellular states within these data presents a significant technological challenge [4]. Specifically, the high dimensionality, inherent noise [5], and sparsity of single-cell data introduce substantial obstacles for individual cell sample identification and clustering, particularly in distinguishing cell types. Moreover, the quality of cell clustering directly impacts the accuracy of downstream analyses [6, 7].

Cell clustering is a pivotal step in scRNA-seq data analysis, as many subsequent analyses-such as constructing cell trajectories and identifying key differentially expressed genes-rely on the precise identification of cell subpopulations [8, 9]. Various clustering approaches have been applied to scRNA-seq data, including traditional methods like hierarchical clustering, spectral clustering, and k-means [10]. However, these methods face scalability limitations when applied to large-scale scRNA-seq datasets [11, 12].

In the realm of popular methods for single-cell analysis, Seurat [13] and SCANPY [14] are widely utilized tools. They employ community detection algorithms, such as the Louvain or Leiden algorithm, to classify large-scale datasets of individual cells. These tools also utilize dimensionality reduction techniques, including Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE [15]), and Uniform Manifold Approximation and Projection (UMAP [16]). However, these algorithms are prone to local optima, particularly when dealing with large and complex networks, which may hinder them from finding optimal clustering results.

scDeepCluster [17] improves clustering effectiveness by integrating autoencoders and clustering algorithms to optimize feature learning and clustering processes. Despite its effectiveness, scDeepCluster can be computationally expensive for large-scale datasets and may face challenges in distinguishing between closely situated cell populations.

scGMAI [18] enhances feature extraction and dimensionality reduction, although it may lose some critical information when processing highly heterogeneous data, potentially impacting its fine clustering accuracy. Several deep learning models are currently employed to improve clustering and extract latent information from single-cell data.

The Deep Convolutional Autoencoder (DCA) model is widely applied in single-cell data analysis for data denoising and feature extraction through an autoencoder architecture [19]. While DCA effectively identifies critical signals in the data, its primary focus on denoising may limit its ability to capture complex intercellular interactions, which are essential for clustering and classification.

scGMAI [18] is also based on autoencoder networks and integrates Fast Independent Component Analysis (FastICA) to reduce the dimensionality of data reconstructed by the autoencoder. Another model, scDCCA [20], combines an autoencoder with a double-contrastive learning module within a deep clustering framework to extract valuable features and enhance cell clustering. Additionally, scMRA [21] uses a knowledge graph to represent cell type features across different datasets, with graph convolutional networks serving as discriminators within this graph-based structure.

Each of these methods has unique strengths and limitations, and the choice of approach should be based on the specific characteristics of the dataset and the analytical objectives. After thoroughly analyzing the limitations of existing methods, we propose a novel

approach that integrates the strengths of various techniques. Specifically, we incorporate the Multi-Dilated Attention Gate [22], originally developed for image segmentation, into a convolutional autoencoder framework informed by the negative binomial distribution. This integration enhances annotation accuracy and robustness, effectively addressing the challenges faced by current clustering methods when applied to large-scale scRNA-seq datasets.

Our approach mitigates the scalability constraints of traditional methods by utilizing the flexibility provided by multi-dilated convolutional layers, each with a unique dilation rate. This flexibility allows the model to adaptively adjust the receptive field size, enabling it to capture interactions between genes across different scales. By employing multiple dilation rates, our method effectively captures gene expression patterns at both local and broader contexts, thereby reducing computational overhead while maintaining high performance. Additionally, by integrating centroid loss and soft clustering, our model is less prone to getting trapped in local optima, thereby providing more stable clustering results compared to conventional clustering techniques. These enhancements enable our approach to overcome issues related to scalability, local optima, and high computational costs often encountered with popular clustering techniques.

Furthermore, the robustness of our model to diverse types of noise during training enhances its generalization performance across different datasets, making the SMD model especially advantageous for analyzing large-scale and complex scRNA-seq data. To assess the performance of the SMD model, we benchmarked it against six other models. Our model consistently outperformed these alternatives across various metrics. Training and testing were conducted on multiple public datasets representing a wide range of tissue types, disease states, and biological processes, ensuring the model's ability to generalize within diverse and complex data environments. Experimental results demonstrate that the SMD model effectively addresses the generalization challenges in single-cell data annotation, particularly in the accurate identification and clustering of cell types.

The primary contributions of our work can be summarized as follows:

- As a high-throughput genomics method, scRNA-seq has become an indispensable tool for elucidating cellular heterogeneity and complexity within tissues and systems. However, the high dimensionality and inherent sparsity of single-cell data present significant challenges for precise cell clustering in current research.
- To address this challenge, we propose SMD, a deep learning model that seamlessly integrates nonlinear dimensionality reduction techniques with a porous dilated attention gate component, adapted from image segmentation, within a convolutional autoencoder framework informed by the negative binomial distribution.
- During training, the integrated SMD model automatically adjusts its weights to capture essential features for accurate cell clustering.
- In experimental validation, we evaluated the SMD model's performance across four publicly available datasets and supplemented the analysis with proprietary osteosarcoma data, further verifying the model's superior clustering accuracy.

Methods

Overview of ScSMD

This study introduces the scSMD model, a convolutional autoencoder-based framework designed to analyze single-cell RNA sequencing (scRNA-seq) data. The encoder-decoder architecture integrates a denoising convolutional autoencoder grounded in the negative binomial (NB) model, which is trained to extract a latent space representation and perform an initial clustering within this space. The encoder processes the input gene expression matrix into a latent space through convolutional layers and a fully connected layer, while the decoder reconstructs the data using a fully connected layer followed by deconvolutional layers.

The autoencoder architecture illustrated in Fig. 1A serves as the foundational framework, employing convolutional and fully connected layers to map the input gene expression matrix into a latent space, followed by a reconstruction step. This structure supports initial clustering through the extraction of key latent features.

In contrast, Fig. 1C expands on this foundational framework by incorporating our innovative Multi-dilated Attention Gate module into the encoder. This advanced architecture processes the input data through multiple dilated convolutional layers with varying dilation rates, enabling the model to capture relationships across diverse scales. The attention mechanism refines feature selection, emphasizing key patterns in the gene expression data. By integrating these enhancements, the autoencoder in Fig. 1C demonstrates superior clustering performance, particularly in datasets with complex and heterogeneous structures. This distinction between the two frameworks highlights the progressive development of the scSMD model, from the foundational autoencoder in Fig. 1A to the enhanced version in Fig. 1C, showcasing how the inclusion of the

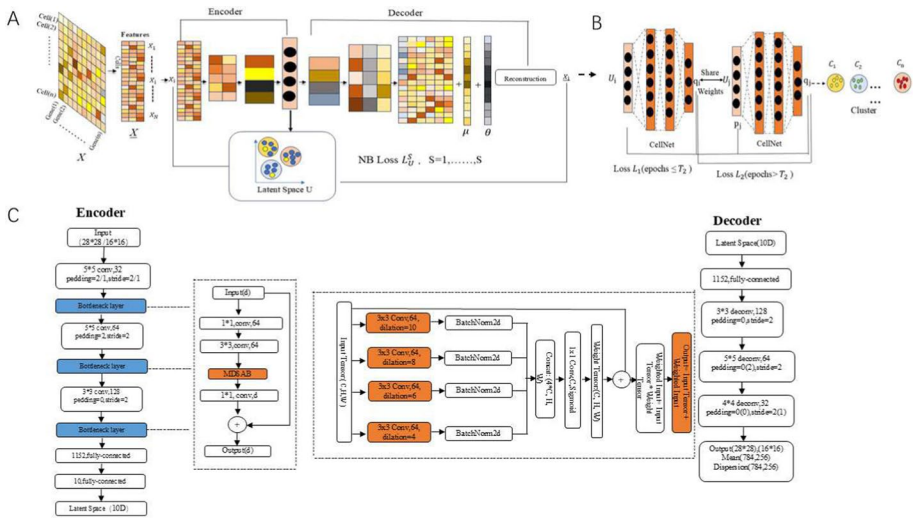

Fig. 1 Workflow of scSMD. **A** Encoder-Decoder Framework: A denoising convolutional autoencoder based on a Negative Binomial (NB) model is trained to obtain a latent space representation and perform preliminary clustering in the latent space. The encoder integrates a novel Multi-dilated Attention Gate to enhance feature selection and representation. **B** Cellnet Construction: Utilizes pairwise data similarity metrics to construct Cellnet, enabling the model to more effectively capture structural relationships within the data. **C** Algorithm Specifications and Implementation: Provides comprehensive details on the algorithm's specifications, emphasizing the Multi-dilated Attention Gate component. This component improves interpretability and contributes to the enhanced performance of scSMD

Multi-dilated Attention Gate module significantly improves the model's feature representation and clustering accuracy.

Additionally, the scSMD model integrates CellNet (Fig. 1B), a component that refines cell-cell similarity relationships within the latent space. CellNet is constructed using pairwise data similarity metrics and trained in two phases: first, by minimizing intra-cluster variance and maximizing inter-cluster distances, and second, by applying a self-supervised learning approach to fine-tune cell-type classification. This dual-phase training ensures that CellNet effectively captures structural relationships in the data, enabling the model to improve clustering robustness.

By combining the autoencoder (Fig. 1A), the multi-dilated attention gate (Fig. 1C), and CellNet (Fig. 1B), the scSMD model demonstrates high accuracy and robustness in cell clustering. This integration provides a powerful and reliable tool for single-cell RNA-seq analysis.

Datasets and evaluation metrics

Datasets

In this study, we applied our novel analytical approach across multiple single-cell RNA sequencing (scRNA-seq) datasets to assess its generalizability and applicability. Specifically, we analyzed the PBMC 4K dataset, which records gene expression patterns in 4,340 peripheral blood mononuclear cells. The Baron dataset, sourced from the NCBI GEO database, provided comprehensive transcriptomic data for pancreatic islet cells from four donors. Additionally, we utilized the Bhattacherjee dataset [23] and the Zeisel dataset [24], which offer cellular information from diverse regions of the mouse brain, enhancing our research with detailed analyses of neural cell types. Furthermore, osteosarcoma data from Ruijin Hospital [25] was employed to further validate the SMD model's effectiveness in cell clustering. Details of these datasets are presented in (Table 1).

Single-cell data is often sparse and contains a significant amount of low-quality information. To address these challenges and ensure the quality and consistency of single-cell RNA sequencing (scRNA-seq) data, we implemented a series of meticulous preprocessing steps. The raw datasets, available in formats such as CSV, TXT, and 10x MTX, were processed using the Scanpy library in Python. These datasets were initially converted into Scanpy's AnnData objects, enabling uniform handling and processing in all subsequent steps.

In the initial phase of data preprocessing, rigorous filtering was applied to both cells and genes to improve data quality. Specifically, cells were retained if they expressed between 200 and 5000 genes, thereby removing low-quality cells with insufficient RNA

Table 1 Summary of the real scRNA-seq datasets

Number	Datasets	Cell	Type
1	PBMC	4340	8
2	Bhattacherjee	24822	8
3	Zeisel	3005	9
4	Baron	1936	14
5	Osteosarcoma	100987	11

content and potential doublets exhibiting abnormally high gene counts. Additionally, cells with more than 10% mitochondrial gene content were excluded, as these are indicative of compromised cell integrity or apoptotic cells. This mitochondrial gene filtering was performed by first calculating the percentage of mitochondrial genes in each cell, and cells with a mitochondrial gene content greater than 10% were removed.

For gene filtering, genes expressed in fewer than three cells were removed to reduce noise and enhance the accuracy of downstream analysis. This ensures that only genes with sufficient expression across multiple cells are retained, which contributes to the robustness of clustering and cell-type identification.

Subsequently, gene expression levels for each cell were normalized using Scanpy's `sc.pp.normalize_total` function, ensuring standardized total expression across all cells. A logarithmic transformation (`sc.pp.log1p`) was then applied to stabilize variance, facilitating statistical analysis and clustering. After normalization, highly variable genes were identified using Scanpy's `sc.pp.highly_variable_genes` function. This step focused on retaining biologically informative genes, reducing data dimensionality while preserving signals essential for cell type identification and biological function analysis.

These preprocessing steps ensured that only high-quality cells and informative genes were included in subsequent analyses, effectively improving the reliability of clustering and other downstream results. The entire pipeline was implemented using the Scanpy library in Python, with the processed dataset saved in CSV format for model training. By enhancing data quality, minimizing noise, and emphasizing biologically relevant features, this pipeline provided a solid foundation for robust and reliable scRNA-seq analysis.

To evaluate the compatibility of the osteosarcoma dataset from Ruijin Hospital with the scSMD model, we performed an exploratory analysis using the R programming language. This included several dimensionality reduction and visualization techniques to investigate the underlying data structure. The Principal Component Analysis (PCA) Loadings Plot illustrated the contribution of individual genes to the first two principal components (PC1 and PC2), highlighting their influence on explained variance. The PCA Scores Plot provided an overview of sample relationships across principal components (PC1 to PC15), offering insights into sample clustering patterns. The Empirical Cumulative Distribution Function (ECDF) Plot evaluated the deviation of principal components from theoretical distributions, with p-values indicating the significance of observed differences. Finally, the Uniform Manifold Approximation and Projection (UMAP) visualization revealed the data distribution across two UMAP dimensions (umap_1) and (umap_2), offering insights into the intrinsic structure of cellular gene expression data.

These exploratory analyses were conducted solely to assess the quality and structure of the dataset, ensuring its suitability for subsequent analysis with the scSMD model. The Multi-dilated Attention Gate module, a key component of the scSMD model, was introduced to enhance the model's performance and plays a central role in improving feature extraction by capturing relationships across genes and cell types at multiple scales. Together, the preprocessing steps and exploratory analysis ensured that the data were appropriately structured and compatible with the scSMD model, while the Multi-dilated Attention Gate module enhanced the overall efficacy of the model in processing complex

gene expression data. Further details, including exploratory analyses such as PCA, ECDF, and UMAP visualizations, are provided in Supplementary Figure S1.

Evaluation metrics

By incorporating the multi-dilated attention gate component into a convolutional autoencoder grounded in the negative binomial distribution, our model demonstrated exceptional performance across a range of diverse datasets. In tests on public datasets, the model not only validated the feasibility of its approach but also highlighted its potential for practical applications in clinically relevant samples.

To rigorously assess the clustering algorithm, we introduced two evaluation metrics: Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI).

The Adjusted Rand Index (ARI), as defined in formula (1), is a widely used metric for evaluating the performance of clustering algorithms. It quantifies the similarity between clustering results and true categories, with values ranging from -1 to 1, where higher values indicate better clustering accuracy.

The Adjusted Rand Index (ARI), as defined in formula (1), is a widely used metric for evaluating the performance of clustering algorithms. It quantifies the similarity between clustering results and true categories, with values ranging from -1 to 1, where higher values indicate better clustering accuracy.

In the ARI formula, the following terms are used:

- $\sum_{ij} \binom{n_{ij}}{2}$ represents the sum of pairwise combinations within each cluster, indicating the degree of matching among samples in the same cluster according to the clustering results.
- $\sum_{i} \binom{n_{i}}{2}$ represents the sum of pairwise combinations within each true category, indicating the degree of matching among samples in the same category based on the ground truth labels.
- $\sum_{j} \binom{n_{j}}{2}$ reflects the sum of pairwise combinations for the predicted clusters, representing the clustering match for sample pairs.
- $\binom{n}{2}$ is the total number of pairwise combinations across all samples, providing the total possible matches for evaluation.

These components contribute to the overall ARI calculation, ensuring that both matching pairs and expected matching pairs are considered, thereby providing a robust evaluation metric for clustering quality.

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\left(\sum_i \binom{n_i}{2}\right)\left(\sum_j \binom{n_j}{2}\right)}{\binom{n}{2}}}{\frac{1}{2}\left(\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}\right) - \frac{\left(\sum_i \binom{n_i}{2}\right)\left(\sum_j \binom{n_j}{2}\right)}{\binom{n}{2}}}. \tag{1}$$

Normalized Mutual Information(NMI), as shown in formula (2), is a metric for evaluating the similarity between clustering results and true labels, utilizing the principles of mutual information and entropy. In this formula, n_{ij} denotes the number of samples shared by true label i and predicted label j, n_i indicates the sample count for true label i, n_j represents the sample count for predicted label j, and n is the total sample count.

$$\mathrm{NMI} = \frac{\sum_i \sum_j n_{ij} \log_2 \frac{n \cdot n_{ij}}{n_i \cdot n_j}}{\max\left(-\sum_i n_i \log_2 \frac{n_i}{n}, -\sum_j n_j \log_2 \frac{n_j}{n}\right)} \tag{2}$$

Auto-encoder embedding module

We developed an autoencoder network based on the negative binomial (NB) model [22] to estimate its parameters. The number of clusters, denoted as K, was predetermined for the model, and it is the same as the final number of clusters used for analysis. The rationale for selecting K was based on both domain knowledge and an initial evaluation of the data. Specifically, we conducted a preliminary analysis of the dataset's structure using dimensionality reduction techniques such as Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP). These exploratory analyses provided insights into the intrinsic clustering tendencies of the data, guiding an initial estimation of K.

The autoencoder network includes an encoder and a decoder, each consisting of two distinct fully connected layers. The encoder uses three convolutional layers followed by a fully connected layer to project X_i into a K-dimensional space, resulting in a latent representation denoted as Z_i. To comprehensively capture the features of scRNA-seq data, we incorporated a multi-dilated attention gate component into the convolutional layers.

In the latent space Z_i, soft clustering was applied by training the autoencoder with weighted reconstruction, NB parameters, and centroid loss. The decoder, comprising a fully connected layer and three deconvolutional layers, reconstructs Z_i back to the original X_i.

The autoencoder network includes an encoder and a decoder, each consisting of two distinct fully connected layers. The encoder uses three convolutional layers followed by a fully connected layer to project X_i into a K-dimensional space, resulting in a latent representation denoted as Z_i. To comprehensively capture the features of scRNA-seq data, we incorporated a multi-dilated attention gate component into the convolutional layers.

In the latent space Z_i, soft clustering was applied by training the autoencoder with weighted reconstruction, NB parameters, and centroid loss. The decoder, comprising a fully connected layer and three deconvolutional layers, reconstructs Z_i back to the original X_i.

To enhance feature capture in scRNA-seq data, we introduced a loss function grounded in the NB model, with parameters representing the mean μ and dispersion ϕ, as shown in Formula (3).

$$P_{NB}(x' \mid \mu, \phi) = \frac{\Gamma(x' + \phi)}{x'!\Gamma(\phi)} \left(\frac{\phi}{\phi + \mu}\right)^{\phi} \left(\frac{\mu}{\phi + \mu}\right)^{x'} \tag{3}$$

Within the autoencoder framework, we introduced a loss function based on the negative binomial (NB) distribution to more precisely capture the features of single-cell RNA sequencing data. This loss function enhances network performance by estimating the mean μ and dispersion ϕ of the NB model. These parameters, related to the input data D are computed via the network's weight parameter matrices W_μ and W_ϕ, as illustrated in Formula (4)Formula (5):

$$\mu = \exp\left(W\mu D\right) \tag{4}$$

$$\phi = \exp\left(W\phi D\right) \tag{5}$$

We use the exponential function to ensure that the mean and dispersion are non-negative values. The loss function based on the negative binomial (NB) is a measure of the deviation between model predictions and actual observations, as shown in Formula (6), which illustrates the NB distribution [22]:

$$L_{NB} = -\log\left(P_{NB}\left(\left[x'\right]|\mu,\phi\right)\right) \tag{6}$$

Here, x' denotes the preprocessed and rounded expression values. Formula (6) describes the probability density function of the NB distribution. Following the autoencoder's pretraining, the loss function is utilized to improve the model's performance by learning more representative low-dimensional representations. This approach not only preserves the integrity of data features but also enhances the model's adaptability to subsequent clustering tasks.

Following the clustering process, initial cluster centers are determined. The autoencoder then undergoes training K times, with each iteration corresponding to a specific cluster, thereby increasing the probability of cells associating with their correct clusters. To achieve accurate classification, we define three distinct loss functions: one for the reconstruction of the convolutional autoencoder, another for optimizing clustering centers, and a third for fitting the parameters of the NB distribution, as shown in Formula (7):

$$L_u^{(k)} = L_r^{(k)} + \alpha Lc^{(k)} + \beta L_{NB}^{(k)} \tag{7}$$

In this approach, we conduct K iterations of autoencoder training, each aligned with a specific cluster, to increase the likelihood that cell points correspond to their true cellular clusters. Additionally, three distinct loss functions are defined: the first for reconstructing the convolutional autoencoder, the second for positioning clustering centers, and the third for fitting the parameters of the Negative Binomial (NB) distribution. These are represented as three weighted sum losses: $Lr^{(k)}$, $Lc^{(k)}$, and $L_{NB}^{(k)}$, as specified in formulas (8), (9), (10), respectively. These losses address cell reconstruction, clustering center alignment, and NB fitting. Two hyperparameters, α and β, are introduced to balance these components within the loss functions.

$$L_r^{(k)} = \sum_{i=1}^{n} p_{ik}^{\lambda} \left\| x_i - \hat{x}_i \right\|_2^2 \tag{8}$$

$$L_c^{(k)} = \sum_{i=1}^{n} p_{ik}^{\lambda} \left\| u_i - u^{(k)} \right\|_2^2 \tag{9}$$

$$L_{NB}^{(k)} = -\sum_{i=1}^{n} p_{ik}^{\lambda} \sum_{j=1}^{t} \log\left(P_{NB}\left(\left[x_{ij}\right]|\mu,\theta\right)\right) \tag{10}$$

Multi-dilated attention gate

Multi-dilated convolutional layers

For a given input sequence $X = [x_1, x_2, \ldots, x_T]$, we apply a series of multi-dilated convolutional layers, each configured with a unique dilation rate. The dilation rate determines the receptive field size of the convolutional kernel, allowing the model to capture interactions across varying scales within the input sequence.

Initially, the cell gene expression matrix is processed through multiple one-dimensional convolutional layers, each utilizing distinct dilation rates. This multi-dilated convolutional structure effectively captures gene interactions and relationships across different resolutions, enabling the model to identify patterns specific to various cell types. The process can be mathematically expressed by Formula (11):

$$H_i = \text{Conv1D}(X, \text{kernel\_size}, \text{dilated\_ratio}[i]) \tag{11}$$

where H_i represents the output of the i-th convolutional layer, and Conv1D denotes the one-dimensional convolution operation.

After passing through the multi-dilated convolutional layers, the outputs are concatenated and activated using the ReLU function. These processed features are subsequently fed into a selection module for further refinement. Figure 1C illustrates the detailed architecture of the Multi-dilated Attention Gate module. This module combines and refines the outputs of the multi-dilated convolutional layers, employing attention mechanisms to selectively emphasize critical gene expression patterns. The attention mechanism ensures the model effectively prioritizes relevant features, enhancing its ability to distinguish cell types in heterogeneous populations.

Selection module (SM)

To finalize the feature processing, we introduce a selection module that automatically filters the concatenated features based on various dilation rates, selecting the most relevant ones. The Selection Module (SM) utilizes a sigmoid activation function to map each output element to the (0, 1) range, as demonstrated in formulas (12) and (13):

$$G = \text{Sigmoid}(\text{MLP}(H_{\text{att}})) \tag{12}$$

$$H_{\text{final}} = G \odot H_{\text{att}} \tag{13}$$

Here, SM denotes the selection module, and $\odot$ represents element-wise multiplication. The resulting outputH_{final}, produced through the Selection Module (SM), retains information deemed relevant by the model for the task at hand. Through the application of the Selection Module (SM), the model effectively filters out non-essential information,

preserving only the most significant features necessary for accurately differentiating cell types or states. This selective retention not only enhances model accuracy but also boosts computational efficiency. By using a sigmoid function, the Selection Module (SM) screens and prioritizes features critical to the current clustering task, marking a key step in model optimization.

By integrating these two key components with the autoencoder, the SMD model achieves cell clustering with improved accuracy and efficiency. In our experiments, we employed gene expression data from diverse cell types to train and test the model. The results demonstrate that the SMD model exhibits high accuracy and robustness in recognizing and classifying cell types. Particularly effective in handling heterogeneous cell populations, the model accurately distinguishes between different cell types, underscoring its potential in cell clustering applications.

Self-supervised learning with cellnet

The training of CellNet(Fig. 1B) is conducted in two phases. In the initial phase, CellNet is randomly initialized. After running the first module of the scSMD model, a fully connected network (CellNet) is attached to enhance cell-cell similarity. The input to CellNet is the latent representation of each cell in the autoencoder's latent space, and the output layer consists of neurons corresponding to the number of cell clusters. The Softmax function is applied in the output layer to obtain probability values, indicating the likelihood of each cell belonging to a particular cluster.

Over a set number of training iterations (epoch $= 20$), the latent space assignments P_n are used to train CellNet with a designated loss function (Formula (14)). This loss function aims to minimize intra-cluster variance while maximizing inter-cluster distances, ensuring that cells with similar features are grouped together more tightly in the latent space, while dissimilar cells are pushed apart. The hyperparameter δ plays a crucial role in defining the threshold for similarity, enabling the model to distinguish between meaningful biological relationships and noise. Specifically, δ is typically set to a value less than 1, allowing the model to focus only on the most confident cell pair similarities during training, thereby avoiding inconclusive clustering information. During each iteration, cluster centers are updated, progressively refining the latent space representation to achieve a relatively stable clustering structure.

In the second phase, CellNet undergoes fine-tuning for a specified number of epochs, utilizing the loss function defined in Formula (15). This phase employs a self-supervised approach, where pseudo-labels are generated based on the similarity between cells. Cells with high similarity are assigned the same type, while cells with significant dissimilarity are assigned to different types. This structured supervision allows CellNet to enhance its cell similarity measurements by leveraging intrinsic features, ultimately leading to improved clustering accuracy. The integration of pseudo-labels during fine-tuning helps further refine cell-type identification by encouraging clear separation between different cell types, thus improving the robustness and accuracy of the final clustering results.

As an integral part of the scSMD model, CellNet enhances the model's ability to learn and analyze cell representations with greater accuracy, contributing to improved performance in single-cell data clustering tasks. This integration allows scSMD to leverage both the strengths of autoencoder-based feature extraction and the detailed similarity

measurements provided by CellNet, ultimately yielding more precise and reliable clustering outcomes for complex single-cell datasets.

$$L_1 = \sum_{x_i, x_j \in \tilde{X}} (\mathbb{I}[p_i^T p_j \geq \delta]\left(1 - q_i^T q_j\right) + \mathbb{I}[p_i^T p_j \leq (1-\delta)](q_i^T q_j)) \tag{14}$$

$$L_2 = \sum_{x_i, x_j \in \tilde{X}} (\mathbb{I}[q_i^T q_j \geq \delta](1 - q_i^T q_j) + \mathbb{I}[q_i^T q_j \leq (1-\delta)](q_i^T q_j)) \tag{15}$$

Results

Comparison experiment

In this study, we performed a comprehensive evaluation by comparing the performance of various models across different biological datasets, with a particular emphasis on their normalized mutual information (NMI) [26] scores in cell annotation tasks. Our SMD model consistently outperformed other models across all tested datasets. These results further confirm the robust clustering precision and high adaptability of the SMD model in practical applications.

scDeepCluster demonstrated high NMI scores across all datasets, achieving a particularly notable score of 0.729 on the PBMC dataset, highlighting its effectiveness in cell population identification. scGMAI scored 0.722 on the Bhattacherjee dataset, showcasing good adaptability to this specific data type; however, its performance was slightly lower on other datasets, scoring 0.496 on the Fibroblast dataset. scziDesk excelled on the Bhattacherjee dataset with an NMI score of 0.9258 but showed relatively lower scores on other datasets, achieving 0.615 and 0.6449 on the Zeisel and Baron datasets, respectively. scCAN reached its highest score of 0.8309 on the PBMC dataset, underscoring its strong performance on this type of data, though its score on the Cancer dataset was lower at 0.502, indicating a slightly weaker performance. scDCCA also performed well on the PBMC dataset with an NMI score of 0.8165, but its performance fluctuated, as seen in its lower score of 0.5684 on the Fibroblast dataset. Similarly, DeepScena, like scziDesk, achieved a high score (0.928) on the Bhattacherjee dataset, yet displayed a significant drop to 0.567 on the Zeisel dataset, indicating notable performance variability across different datasets.

To evaluate the models, we utilized not only publicly available datasets but also validated them using osteosarcoma data from Ruijin Hospital. The validation results further confirmed the superiority of the SMD model over other models, as shown in Fig. 2. In summary, our proposed SMD model achieved the highest NMI scores across all tested datasets, with a particularly outstanding score of 0.9401 on the Bhattacherjee dataset. On the osteosarcoma data from Ruijin Hospital, Fig. 2 shows that the SMD model consistently outperformed other models, obtaining the highest ARI and NMI scores. This underscores its remarkable generalization capability and superior performance in cell annotation tasks.

Additionally, the SMD model achieved top Adjusted Rand Index (ARI) scores across multiple datasets, further highlighting its effectiveness in achieving precise clustering. To visually support these findings, we generated UMAP plots to better illustrate

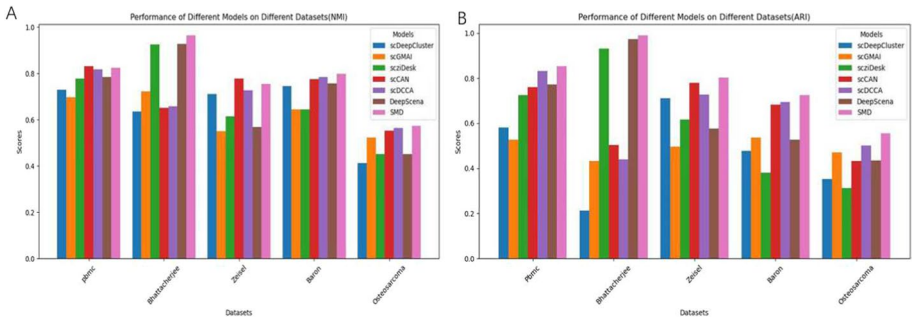

Fig. 2 ARI and NMI scores across various datasets for different models. **A** Bar chart depicting NMI scores of different models across multiple datasets. **B** Bar chart depicting ARI scores of different models across multiple datasets. Panels **C** and **D**, which provide heatmap visualizations of NMI and ARI scores, have been moved to the Supplementary Figure S2 for reference

the clustering performance of each model. Figure 3 shows UMAP visualizations of the clustering results for all models on the Osteosarcoma and Bhattacherjee datasets, while Fig. 4 compares the clustering of the scSMD model with the real cell classifications for the datasets. These plots offer a clearer perspective on the model's performance and its ability to reflect the true biological structure of the data Fig. 5

Ablation experiment

To further refine our cell clustering model, particularly in addressing gene-cell expression matrices, we introduced a Multi-Dilated Attention Gate (Multi-Dilated AG)-an innovative module designed to capture multi-scale data features using convolution operations with varied dilation rates. We experimented with multiple dilation rate combinations, including (1,3,5,7), (2,4,6,8), (4,6,8,10), (6,8,10,12), and (10,8,6,4), to

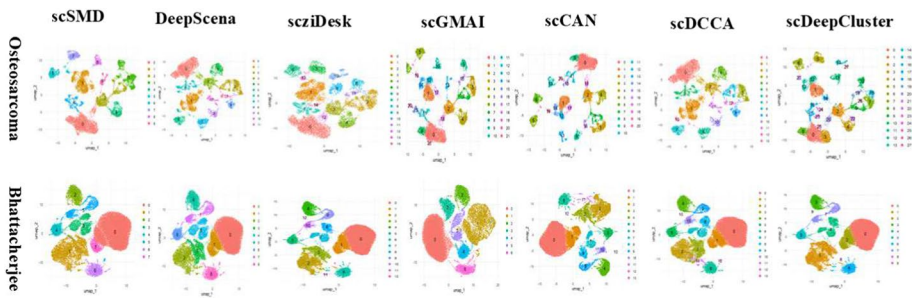

Fig. 3 Comparison of the scSMD model clustering with true cell types classifications

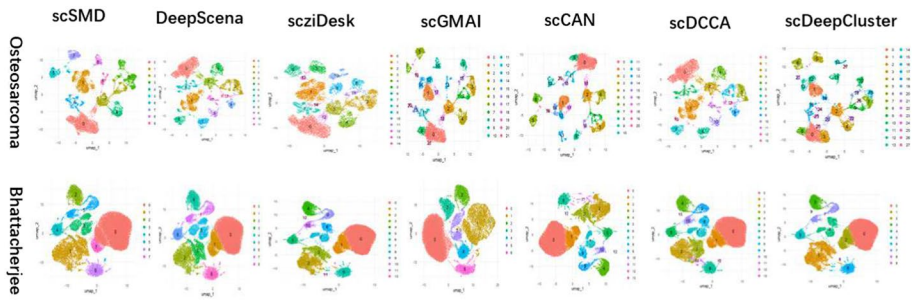

Fig. 4 UMAP visualization for all models on Osteosarcoma and Bhattacherjee datasets

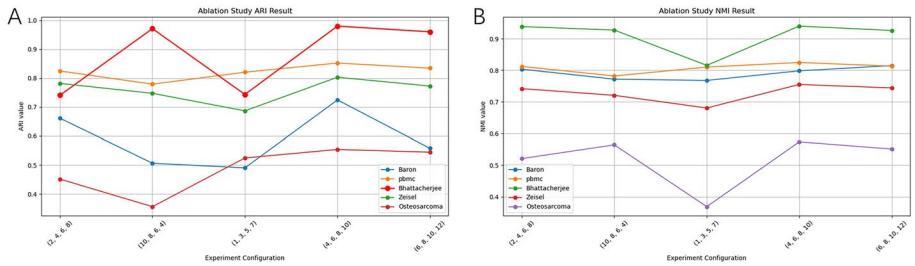

Fig. 5 ARI and NMI scores across varying expansion rates on different datasets. **A** and **B** respectively present the ARI and NMI evaluation metrics for the SMD model under different expansion rates in the ablation experiments

identify the most effective configuration. After extensive training and fine-tuning, we found that the dilation rate combination (4,6,8,10) achieved the best results. Our findings were evaluated using two widely recognized clustering metrics: Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). The results of the ablation experiments, illustrated in Fig. 2, demonstrate that across both publicly available datasets and the osteosarcoma dataset, the SMD model consistently achieved superior performance with the (4,6,8,10) configuration. The comparative outcomes are also presented in detail in Table 2 and Table 3.

This ablation experiment not only confirms the necessity of our selected dilation rate configuration but also highlights the effectiveness of the Multi-Dilated Attention Gate in improving model performance on complex datasets, particularly for gene-cell data with diverse expression patterns. These findings further validate the application potential of our SMD model for precise cell clustering, demonstrating its robustness on novel or previously unseen data.

Table 2 Normalized Mutual Information (NMI) Scores for the SMD Model Across Various Datasets and Dilation Rate Configurations

Datasets	(2,4,6,8)	(1,3,5,7)	(4,6,8,10)	(6,8,10,12)	(10,8,6,4)
Baron	0.803	0.767	0.798	0.814	0.771
pbmc	0.812	0.810	0.824	0.813	0.782
Bhattacherjee	0.815	0.810	0.940	0.926	0.927
Zeisel	0.742	0.680	0.755	0.744	0.720
Osteosarcoma	0.521	0.369	0.574	0.551	0.564

Table 3 Adjusted Rand Index (ARI) Scores for the SMD Model Across Various Datasets and Dilation Rate Configurations

Datasets	(2,4,6,8)	(1,3,5,7)	(4,6,8,10)	(6,8,10,12)	(10,8,6,4)
Baron	0.661	0.490	0.724	0.557	0.506
pbmc	0.824	0.820	0.852	0.834	0.779
Bhattacherjee	0.741	0.743	0.980	0.960	0.971
Zeisel	0.781	0.686	0.803	0.772	0.747
Osteosarcoma	0.451	0.524	0.554	0.544	0.356

Conclusion

The primary contributions of our work are centered around the development of a novel SMD model that effectively addresses the high dimensionality and inherent sparsity of single-cell RNA sequencing (scRNA-seq) data. The model incorporates Multi-Dilated Convolutional Layers, which allow it to adaptively adjust the receptive field size, enabling the capture of gene interactions at multiple scales. This adaptive approach enhances the model's ability to handle the complex relationships between genes without significantly increasing computational costs, thereby making it suitable for large-scale data analysis.

To further address the challenges posed by the sparsity of single-cell data, our model employs a loss function based on the negative binomial distribution. This loss function is well-suited to handle the excessive zero counts typically found in scRNA-seq datasets, helping to mitigate noise and improve the robustness of the clustering process. Additionally, we integrated an attention mechanism through the Multi-Dilated Attention Gate, which selectively emphasizes key gene expression features while reducing the influence of irrelevant information. This mechanism significantly improves annotation accuracy and generalization performance, making the SMD model particularly effective for analyzing complex and large-scale scRNA-seq datasets.

We benchmarked the SMD model against six other models on multiple public datasets, representing a wide range of tissue types and biological processes. The results consistently demonstrated the superior performance of our model across various metrics. The SMD model showed improved robustness in handling high-dimensional, noisy, and sparse data environments, ensuring better clustering quality and more accurate identification of cell types.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06047-x.

Supplementary Figure S1: Exploratory analyses of the osteosarcoma dataset, including PCA Loadings Plot, PCA Scores Plot, ECDF Plot, and UMAP Visualization. These figures illustrate the structure and clustering tendencies of the dataset, providing a comprehensive understanding of its characteristics.

Supplementary Figure S2: Heatmap visualizations of NMI and ARI scores across various datasets for different models. Panel (C) illustrates the NMI performance of each model, and Panel (D) illustrates the ARI performance of each model. These heatmaps provide a complementary perspective to the bar charts shown in the main text (Fig. 2), highlighting the comparative performance of the models on all tested datasets.

Acknowledgements
Not applicable.

Author Contributions
Xiaoxu Cui authored the paper and conducted specific experiments. Renkai Wu contributed conceptual ideas. Yinghao Liu, Pengchen Liang, and Qing Chang collaborated on revising and proofreading the manuscript, providing valuable insights and suggestions. Peizhan Chen provided validation using a private dataset. Changyu He has reviewed the manuscript and provided funding support.

Funding
This work was supported partly by the National Natural Science Foundation of China (Nos.82002463).

Availability of data and materials
The code is available from https://github.com/xiaoxuc/scSMD. Baron Dataset: The single-cell RNA-seq data from human pancreatic islets can be accessed from the Gene Expression Omnibus (GEO) repository under accession number GSE84133 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84133). PBMC Datasets: The PBMC datasets used in this study are available from 10x Genomics' single-cell gene expression section, which can be accessed via the following link: https://support.10xgenomics.com/single-cell-gene-expression/datasets. Bhattacherjee Dataset: The PBMC dataset from Bhattacherjee et al. (2020) is available in the GEO database under accession number GSE124952 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124952). Zeisel Dataset: The human brain cell atlas data from Zeisel

et al. (2015) can be accessed from GEO under accession number GSE60361 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60361).

Declarations

Ethics approval and consent to participate

The Osteosarcoma dataset used in this study adheres to the ethical standards of Ruijin Hospital and has obtained appropriate ethical approval and licenses.

Competing interests

The authors declare that they have no competing interest.

Received: 12 April 2024 Accepted: 13 January 2025
Published online: 29 January 2025

References

1. Villani A-C, Satija R, Reynolds G. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science. 2017;356:4573.
2. Olsen T, Baryawno N. Introduction to single-cell RNA sequencing. Curr Protoc Mol Biol. 2018;122(1):57. https://doi.org/10.1007/s12110-009-9068-2.
3. Grün D, Lyubimova A, Kester L. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015;525(7568):251–5.
4. Zhang H, Lu M, Lin G, Zheng L, Zhang W, Xu Z, Zhu F. Socube: an innovative end-to-end doublet detection algorithm for analyzing scrna-seq data. Brief Bioinform. 2023;24(3):104. https://doi.org/10.1093/bib/bbad104.
5. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. Genome Med. 2017;9(1):75.
6. Li X, Zhang S, Wong K-C. Deep embedded clustering with multiple objectives on scrna-seq data. Brief Bioinform. 2021;22(5):090. https://doi.org/10.1093/bib/bbab090.
7. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell rna-seq data. Nat Rev Genet. 2019;20(5):273–82.
8. Sun N, Yu X, Li F. Inference of differentiation time for single cell transcriptomes using cell population reference data. Nat Commun. 2017;8:1–12.
9. Papalexi E, Satija R. Single-cell rna sequencing to explore immune cell heterogeneity. Nat Rev Immunol. 2018;18:35–45.
10. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O. Scikit-learn: Machine learning in python. J Mach Learn Res. 2011;12:2825–30.
11. Gan Y, Chen Y, Xu G, Guo W, Zou G. Deep enhanced constraint clustering based on contrastive learning for scrna-seq data. Brief Bioinform. 2023;24(4):222. https://doi.org/10.1093/bib/bbad222.
12. Wang S, Shen B, Guo L, Shang M, Liu J, Sun Q, Shen B. scfed: federated learning for cell type classification with scrna-seq. Brief Bioinform. 2024;25(1):507. https://doi.org/10.1093/bib/bbad507.
13. Stuart T, Butler A, Hoffman P. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888–190221.
14. Wolf FA, Angerer P, Theis FJ. Scanpy: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15.
15. Traag VA, Waltman L, Eck NJ. From louvain to leiden: guaranteeing well-connected communities. Sci Rep. 2019;9(1):5233.
16. Subelj L, Bajec M. Unfolding communities in large complex networks: combining defensive and offensive label propagation for core extraction. Phys Rev E: Stat, Nonlin, Soft Matter Phys. 2011;83(3 Pt 2): 036103.
17. Tian T, Wan J, Song Q. Clustering single-cell rna-seq data with a model-based deep learning approach. Nat Mach Intell. 2019;1:191–8.
18. Yu B, Chen C, Qi R, Zheng R, Skillman-Lawrence PJ, Wang X, Ma A, Gu H. scgmai: a gaussian mixture model for clustering single-cell rna-seq data based on deep autoencoder. Brief Bioinform. 2021;22(4):316.
19. Eraslan G, Simon LM, Mircea M. Single-cell rna-seq denoising using a deep count autoencoder. Nat Commun. 2019. https://doi.org/10.1038/s41467-018-07931-2.
20. Wang J, Xia J, Wang H, Su Y, Zheng CH. scdcca: Deep contrastive clustering for single-cell rna-seq data based on auto-encoder network. Brief Bioinform. 2023;24(1):625.
21. Yuan M, Chen L, Deng M. scmra: A robust deep learning method to annotate scrna-seq data with multiple reference datasets. Bioinformatics. 2022;38(3):738–45.
22. Lei T, Chen R, Zhang S, Chen Y. Self-supervised deep clustering of single-cell rna-seq data to hierarchically detect rare cell populations. Brief Bioinform. 2023;24(6):335.
23. Bhattacherjee A, Djekidel MN, Chen R. Cell type-specific transcriptional programs in mouse prefrontal cortex during adolescence and addiction. Nat Commun. 2019;10(1):4169.
24. Zeisel A, Muñoz-Manchado AB, Codeluppi S. Brain structure. cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. Science (New York, NY) 347(6226), 2015;1138–1142
25. Zhou Y, Yang D, Yang Q, Lv X, Huang W. Single-cell RNA landscape of intratumoral heterogeneity and immunosuppressive microenvironment in advanced osteosarcoma. Nat Commun. 2020;11(1):6322. https://doi.org/10.1038/s41467-020-20059-6. Erratum. In: Nat Commun. 2021 Apr 30;12(1):2567.
26. Strehl A, Ghosh J. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. J Mach Learn Res. 2002;3:583–617.