Title

Ensemble deep learning of embeddings for clustering multimodal single-cell omics data.

Permalink

https://escholarship.org/uc/item/3c77k765

Journal

Bioinformatics, 39(6)

Authors

Yu, Lijia
Liu, Chunlei
Yang, Jean
et al.

Publication Date

2023-06-01

DOI

10.1093/bioinformatics/btad382

Peer reviewed

Bioinformatics, 2023, **39(6)**, btad382
https://doi.org/10.1093/bioinformatics/btad382
Advance access publication 14 June 2023
Original Paper

OXFORD

Gene expression

Ensemble deep learning of embeddings for clustering multimodal single-cell omics data

Lijia Yu[1,2,3], Chunlei Liu[1,3], Jean Yee Hwa Yang[2,3,4,5], Pengyi Yang[1,2,3,4,5,*]

[1]Computational Systems Biology Group, Children's Medical Research Institute, Faculty of Medicine and Health, The University of Sydney, Westmead, NSW 2145, Australia
[2]School of Mathematics and Statistics, Faculty of Science, University of Sydney, NSW 2006, Australia
[3]Sydney Precision Data Science Centre, University of Sydney, NSW 2006, Australia
[4]Charles Perkins Centre, The University of Sydney, Sydney, NSW 2006, Australia
[5]Laboratory of Data Discovery for Health Limited (D^{2}4H), Hong Kong Science Park, Hong Kong SAR, China

*Corresponding author. E-mail: pengyi.yang@sydney.edu.au (P.Y.)

Associate Editor: Macha Nikolski

Abstract

Motivation: Recent advances in multimodal single-cell omics technologies enable multiple modalities of molecular attributes, such as gene expression, chromatin accessibility, and protein abundance, to be profiled simultaneously at a global level in individual cells. While the increasing availability of multiple data modalities is expected to provide a more accurate clustering and characterization of cells, the development of computational methods that are capable of extracting information embedded across data modalities is still in its infancy.

Results: We propose SnapCCESS for clustering cells by integrating data modalities in multimodal single-cell omics data using an unsupervised ensemble deep learning framework. By creating snapshots of embeddings of multimodality using variational autoencoders, SnapCCESS can be coupled with various clustering algorithms for generating consensus clustering of cells. We applied SnapCCESS with several clustering algorithms to various datasets generated from popular multimodal single-cell omics technologies. Our results demonstrate that SnapCCESS is effective and more efficient than conventional ensemble deep learning-based clustering methods and outperforms other state-of-the-art multimodal embedding generation methods in integrating data modalities for clustering cells. The improved clustering of cells from SnapCCESS will pave the way for more accurate characterization of cell identity and types, an essential step for various downstream analyses of multimodal single-cell omics data.

Availability and implementation: SnapCCESS is implemented as a Python package and is freely available from https://github.com/PYangLab/SnapCCESS under the open-source license of GPL-3. The data used in this study are publicly available (see section 'Data availability').

1 Introduction

The development of novel single-cell technologies, such as cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) (Stoeckius *et al.* 2017), shared single-cell profiling of RNA and chromatin (SHARE-seq) (Ma *et al.* 2020), and trimodal single-cell profiling by TEA-seq (Swanson *et al.* 2021), enables the profiling of gene expression, protein abundance, and/or chromatin accessibility in the same cell. The availability of multiple data modalities in individual cells promises more precise characterization of cells such as clustering cells into distinctive cell types (Zhu *et al.* 2020). To analyze such multimodal single-cell omics data, however, require effective computational methods that are capable of integrating data modalities for extracting the underlying biological signals.

While various methods exist for enabling the clustering of multimodal single-cell omics data, such as by simple feature concatenation or more sophisticated methods that integrate clustering output from each modality (Adossa *et al.* 2021, Miao *et al.* 2021) or construct similarity graphs for joint clustering (Wu *et al.* 2022), a popular approach is to integrate data modalities through learning an embedding that encodes multiple data modalities into a shared latent space, from which any clustering algorithm that accepts the embedding as input could be used for clustering cells (Lin *et al.* 2022, Liu *et al.* 2023). For example, Jvis-learn performs joint dimension reduction of data modalities for generating embeddings of multimodal single-cell omics data (Do and Canzar 2021). It automatically determines the relative importance of each data modality that emphasizes distinguishing characteristics while reducing noise. MOFA+ implements a Bayesian group factor analysis framework to infer a low-dimensional embedding that captures shared variation across multiple modalities (Argelaguet *et al.* 2020). Recently, deep learning-based methods such as totalVI (Gayoso *et al.* 2021) use a variational autoencoder (VAE) to learn an embedding for integrating RNA and ADT modalities such as in CITE-seq data, and MultiVI (Ashuach *et al.* 2021) uses a VAE to integrate RNA and ATAC modalities such as in SHARE-seq data. The multimodality integrated embeddings generated from these methods can be subsequently applied for cell clustering using any clustering algorithms that accept embeddings as input. Thus, the utility and quality of the multimodality-integrated

Received: February 13, 2023. **Revised:** April 16, 2023. **Editorial Decision:** June 7, 2023. **Accepted:** June 12, 2023

embeddings will have a large impact on clustering multimodal single-cell omics data.

Here we aim to improve the clustering of multimodal single-cell omics data by developing an ensemble deep learning-based framework that generates a multi-view of multimodality-integrated embeddings. This is motivated by our previous work on autoencoder-based ensemble clustering of scRNA-seq data that demonstrates consensus derived from multiple embeddings each generated from perturbing the input data can lead to significantly better clustering results (Geddes *et al.* 2019). Such a single-cell consensus clusters of encoded subspaces (scCCESS) approach benefits from the multi-view of the input data (Cao *et al.* 2020) and is generic to clustering algorithms. Built on this concept, we propose SnapCCESS, an ensemble clustering framework that uses VAE and the snapshot ensemble learning technique (Huang *et al.* 2017) to learn multiple embeddings each encoding multiple data modalities, and subsequently generate consensus clusters for multimodal single-cell omics data by combining clusters from each embedding. The innovation in SnapCCESS includes (i) implementing an ensemble deep learning framework for creating a multi-view of latent spaces from which multimodality embeddings of multimodal single-cell omics data can be generated, and (ii) designing a snapshot ensemble learning approach to significantly improve the computational efficiency of the proposed framework.

By applying SnapCCESS to multimodal single-cell omics datasets generated by various biotechnologies and protocols, we show that SnapCCESS is effective and computationally efficient in learning multiple embeddings compared to conventional ensemble deep learning methods. We found that multimodal data generally offer more information than single modality alone and SnapCCESS leverages such information to improve cell clustering. We also show that embeddings learned from SnapCCESS are generalizable and can be coupled with various clustering algorithms for improving consensus clustering of multimodal single-cell omics data. Lastly, we demonstrate the competitive performance of SnapCCESS with the other state-of-the-art methods for generating embeddings of data modalities for clustering cells. Together, our work showcases the effectiveness of a novel unsupervised ensemble deep learning framework for performing clustering analysis of multimodal single-cell omics data.

2 Materials and methods

2.1 SnapCCESS framework for generating embeddings of multimodal single-cell data

To integrate the high-dimensional feature space in each modality of multimodal single-cell omics data, SnapCCESS encodes features from multiple data modalities into a latent space using the VAE component of our recently published Matilda framework (Liu *et al.* 2023) to jointly learn to reconstruct each data modality (Fig. 1a).

Specifically, SnapCCESS first learns different data modalities using modality-specific encoders and decoders (denoted using different colors in Fig. 1a). The encoders in the VAE component include one learnable point-wise parameters layer and one fully connected layer to the input layer. Because surface protein modality has significantly fewer features than RNA and ATAC modalities, we empirically set the numbers of neurons for encoders of RNA, surface proteins, and ATAC modalities to be 185, 30, and 185, respectively. To learn a latent space that integrates the information across modalities, SnapCCESS concatenates the output from the encoder trained from each data modality to perform joint learning using a fully connected layer with 100 neurons. The embeddings of input data were obtained from the latent space of VAE by minimizing the loss function L defined as follows:

$$L = \sum_{m=1}^{M} \|(X_m - \hat{X}_m)\|^2 + KL[N(\mu_{x_m}, \sigma_{x_m}), N(0,1)]$$

where X represents the original input, $\hat{X}$ represents the reconstructed data, i is the ith modality, and M is the number of modalities in a multimodal data. $N(\cdot)$ presents a normal distribution, which is learnable in VAE.

SnapCCESS employs the snapshot technique for ensemble learning (Huang *et al.* 2017). The key idea behind snapshot ensemble learning is to train multiple versions of a single model by using a cyclic learning rate scheduler. Since the ensemble is formed from a single training process, snapshot ensemble learning is significantly faster while maintaining a similar performance when compared to conventional ensemble methods that train individual models from multiple training processes [see Huang *et al.* (2017) for more comparative analyses]. In SnapCCESS, the learning rate of the training model was set up to be the shifted cosine function, which could help the training model converge to multiple local minima and then get multiple lower-dimensional embeddings (Fig. 1b). This is defined as follows:

$$S(t) = \frac{S_0}{2}\left(\cos\left(\frac{\pi \bmod\left(t-1, \mathrm{ceil}\left(\frac{T}{E}\right)\right)}{\mathrm{ceil}\left(\frac{T}{E}\right)}\right) + 1\right)$$

where S_0 is the initial learning rate, t is the iteration number, T is the total number of training iterations, E is the number of learning rate cycles, and $\bmod(\cdot)$ refers to the modulo operator.

2.2 Clustering algorithms for ensemble clustering of embeddings

To create consensus clustering, embeddings generated from SnapCESS and conventional VAE ensemble method were used for generating clustering results and then combined for deriving consensus results. Since the embeddings generated by these methods can be coupled with various clustering algorithms for creating consensus clustering results, we have included three different clustering algorithms for testing their effectiveness. These include a simple k-means clustering algorithm, a more sophisticated spectral clustering method, and SIMLR, a kernel-based clustering method designed for scRNA-seq data analysis (Wang *et al.* 2017). In particular, for the simple k-means clustering, we utilized the `kmeans` function in the stats package with the default settings. For the spectral clustering algorithm, we employed the `spectralClustering` function from the CiteFuse package (Kim *et al.* 2020) with the default parameters to perform spectral clustering. Lastly, we used the `SIMLR_Large_Scale` function in the SIMLR package with the number of principal components set to 20 as recommended. For all clustering algorithms, the number of clusters was set to be the same as the number of cell types in each dataset based on the cell-type annotation from the original study. After obtaining individual

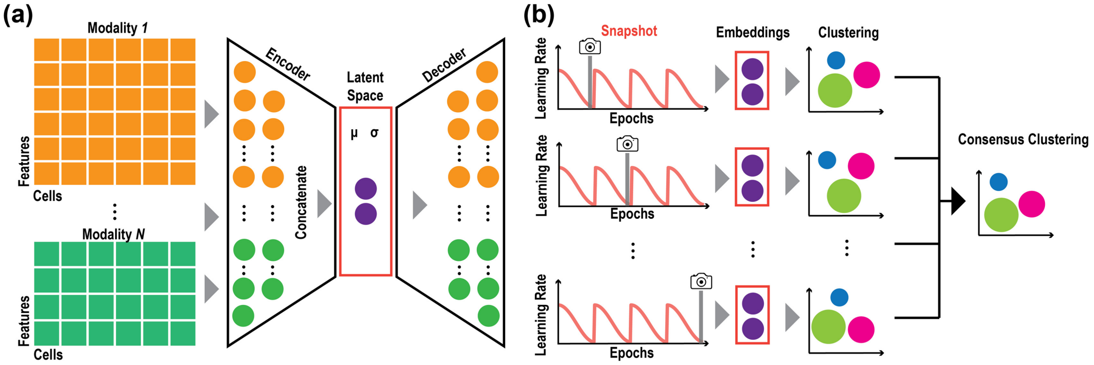

Figure 1 The proposed SnapCCESS framework of ensemble deep learning of embeddings for multimodal single-cell data clustering. (a) A VAE is used to encode the high-dimensional features from multimodal data to a low-dimensional latent space. (b) The training process of SnapCCESS is based on the snapshot ensemble deep-learning model using learning rate annealing cycles where the model converges to and then escapes from multiple local minima, and multiple snapshots were taken at these minima for creating a multi-view of embeddings. The schematic illustrates using epoch of 1 for generating snapshots. The consensus clustering is derived from combining individual clustering results each from a snapshot embedding.

clustering output from each clustering method, a fixed-point iteration algorithm $g(\cdot)$ for obtaining hard least squares Euclidean consensus partitions was applied to compute the consensus clusters of individual partitions using clue package `cl_consensus` function (Hornik 2005):

$$R = g(f(\alpha_1), f(\alpha_2), \ldots, f(\alpha_n))$$

where R is the clustering result and $f(\cdot)$ represents performing a given clustering approach using an embedding α_i. All clustering analyses were carried out in the R programming environment.

2.3 Evaluation settings and datasets

2.3.1 Conventional VAE ensemble

Conventional ensemble deep learning typically relies on perturbing the initialization and/or input data to train individual base models that can be used for creating the final ensemble model (Cao *et al.* 2020). To compare the performance of SnapCCESS with the conventional VAE ensemble method, we implemented a conventional VAE ensemble learning model by using random initialization of the VAE neural networks. For a fair comparison, we used the same VAE model as in SnapCCESS to learn the latent space for integrating the data modalities in multimodal single-cell omics data. The same numbers of neurons for encoders and decoders were used as in SnapCCESS and the learning rate was fixed at 0.02. To obtain multiple embeddings, multiple VAEs were trained each with a different set of network initialization weights on the input dataset.

2.3.2 Settings for other multimodal embedding generation methods

MOFA+: MOFA2 (v1.6.0) which implements MOFA+ (Argelaguet *et al.* 2020) was used to generate multimodality embeddings of the six datasets. Following the author's tutorial (https://raw.githack.com/bioFAM/MOFA2_tutorials/master/R_tutorials/getting_started_R.html), pre-normalized datasets were first used to create the mofa object using `create_mofa` function. The data were then used as input for the modal training using `run_mofa` function with default parameters. We use the "get factors" function with factors = "all" to obtain the embeddings for each input dataset.

Jvis-learn: Jvis-learn (v0.0.12) (Do and Canzar 2021) was employed for generating multi-modality embeddings of the six datasets. The pre-normalized datasets were used as input for creating the j-SNE embeddings that joint multimodal omics data via the "JTSNE" function.

totalVI: totalVI is designed for anayzing CITE-seq data (Gayoso *et al.* 2021). In this study, the totalVI procedure implemented in the scvi-tools package (v0.17.3) was used for generating multimodality embeddings of the three CITE-seq datasets. Following the author's tutorial (https://docs.scvi-tools.org/en/stable/tutorials/notebooks/totalVI.html), the raw count matrices of RNA and ADT were first normalized using the `normalize_total` and `log1p` functions and then top 4000 most variable genes were selected using the `highly_variable_genes` function. The data were input for model training using `scvi.model.TOTALVI.setup_anndata`, `scvi.model.TOTALVI`, and `train` functions in scvi-tools. The latent space of RNA and ADT modalities was generated using the `get_latent_representation` function.

MultiVI: MultiVI is a sibling of totalVI and specifically designed for anayzing data with RNA and ATAC modalites (Ashuach *et al.* 2021). Here, the MultiVI procedure implemented in the scvi-tools (v0.17.3) was used for multimodality integration of the SHARE-seq and SNARE-seq datasets. In accordance with the author's tutorial (https://docs.scvi-tools.org/en/stable/tutorials/notebooks/MultiVI_tutorial.html), the raw count matrices of RNA and gene activity score matrices from ATAC and the paired matrix of RNA and ATAC were utilized as input. These data were first concatenated using the `organize_multiome_anndatas` function in scvi-tools and then used for model training using `scvi.model.MULTIVI.setup_anndata`, `scvi.model.MULTIVI` and `train` functions in scvi-tools. The latent space of RNA and ATAC modalities was generated using the `get_latent_representation` function.

2.3.3 Datasets and pre-processing

Ramaswamy CITE-seq dataset (Ramaswamy *et al.* 2021): The raw RNA and ADT matrices of PBMC from three healthy donors were downloaded from NCBI GEO using the accession number GSE166489. We used the healthy donor (GSM5073072) in our analysis. After filtering RNA and ADT expressed in less than 1% of the cells and genes, discarding cell types with fewer than 50 cells, we obtained 9745 cells and 21 cell types, with 11 039 RNA and 189 ADT features.

Stephenson CITE-seq dataset (Stephenson *et al.* 2021): The PBMC CITE-seq data of healthy individuals sequenced by NCL medical center was used in this study. The raw matrices of RNA and ADT and the annotation of cells to their respective cell types from the original study were downloaded from the EMBL-EBI ArrayExpress database under the accession number E-MTAB-10026. RNA and ADT in this dataset were filtered by removing those that expressed in less than 1% of the cells and genes, cell types were filtered by removing those that have less than 50 cells. After filtering, 64 197 cells from 15 cell types (4999 RNA, 192 ADT) were kept for analysis.

Hao CITE-seq dataset (Hao *et al.* 2021): The raw RNA and ADT matrices from this CITE-seq dataset were downloaded from NCBI GEO under the accession number GSE164378. As the above, RNA and ADT in this dataset were filtered by removing those that expressed in less than 1% of the cells and genes, and cell types were filtered by removing those that have less than 50 cells. In total, 67 035 cells (11 451 RNA, 228 ADT) and 29 cell types in batch 1 of the dataset were used in the analysis.

Chen SNARE-seq dataset (Chen *et al.* 2019): The SNARE-seq data that measures RNA and ATAC from matched cells in the adult mouse brain cortex sample (AdBrainCortex) was downloaded from NCBI GEO under the accession number GSE126074. The cell-type information was obtained from the authors. For ATAC data, peaks with no expression across cells were removed. We then summarized the ATAC data from peak level into gene activity scores using the `CreateGeneActivityMatrix` function in Seurat. We filtered out RNA and ATAC quantified in fewer than 1% of the cells and genes, and removed cell types that have less than 50 cells, resulting in a dataset with 9930 cells (11 011 RNA, 16 443 ATAC peak features) and 20 cell types for the subsequent analyses.

Ma SHARE-seq dataset (Ma *et al.* 2020): The SHARE-seq data that measures RNA and ATAC from matched cells in mouse skin samples were downloaded from NCBI Gene Expression Omnibus (GEO) under the accession number GSE140203. Similar to the above, we first removed peaks with no expression across cells, and then summarized the ATAC data from peak level into gene activity scores using the `CreateGeneActivityMatrix` function in Seurat. We filtered out RNA and ATAC quantified in fewer than 1% of the cells and genes, and remove cell types that have less than 50 cells, resulting in a dataset with 32 968 cells (8765 RNA, 17 413 ATAC peak features) and 23 cell types for the subsequent analyses.

Swanson TEA-seq dataset (Swanson *et al.* 2021): TEA-seq enables simultaneous single-cell profiling of transcripts, epitopes, and chromatin accessibility. The processed matrices of TEA-seq data from measuring PBMC were downloaded from the NCBI GEO under the accession number GSE158013, with raw RNA expression, ADT expression, and peak accessibility (ATAC) measured for the same cells in four data batches. Due to the low batch effect presented in the four datasets, we merged the four data batches. We summarized the matrix of ATAC from peak level to gene activity scores using the `CreateGeneActivityMatrix` function in the Seurat package. Genes with fewer than 1% quantifications across all cells and all genes in the three modalities are removed. This resulted in a dataset with 25 286 cells and nine cell types, including 9772 RNA, 46 ADT, and 16 520 ATAC peak features.

Each data modality was first normalized by a z-score transformation and then fed into each of their modality-specific encoders in SnapCCESS. The integrated embeddings were jointly learned across all modalities using outputs of modality-specific encoders (see Section 2.1).

2.4 Performance evaluation criteria

2.4.1 Clustering concordance performance evaluation

For evaluating clustering performance, we used adjusted Rand index (ARI) and normalized mutual information (NMI) to evaluate the clustering concordance with respect to predefined cell-type annotations from their original studies (Kim *et al.* 2019). Let S be a set of N cells, then a clustering U on S is a way of partitioning S into non-overlap subset $U_1, U_2, \ldots, U_R$. Here, we define $U = U_1, U_2, \ldots, U_R$ as the real cell-type labels with R cell types, $V = V_1, V_2, \ldots, V_c$ is a partition with C clusters generated by a clustering. Pair counting-based measures can be used for counting pairs of items on which the partition U and V agree or disagree. Specifically, the $\binom{N}{2}$ item pairs in S can be classified into one of the four types: (i) N_{11}: the number of pairs that are in the same partition in both U and V; (ii) N_{00}: the number of pairs that are in different partitions in both U and V; (iii) N_{01}: the number of pairs that are in the same partition in U but in different partitions in V; (iv) N_{10}: the number of pairs that are in different partitions in U but in the same partition in V. Following this, ARIand NMI can be defined as follows:

$$\mathrm{ARI}(U,V) = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{(N_{00}+N_{01})(N_{01}+N_{11}) + (N_{00}+N_{10})(N_{10}+N_{11})}$$

$$\mathrm{NMI}(U,V) = \frac{I(U;V)}{H(U)+U(V)}$$

where $I(U;V)$ is the mutual information between U and V, defined as

$$I(U;V) = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{|U_i \cap V_j|}{N} \log_2 \frac{N|U_i \cap V_j|}{|U_i||V_j|}$$

and $H(.)$ is the entropy of partitions, in which

$$H(U) = -\sum_{i=1}^{R} \frac{|U_i|}{N} \log_2 \frac{|U_i|}{N}; \quad H(V) = -\sum_{j=1}^{C} \frac{|V_j|}{N} \log_2 \frac{|V_j|}{N}$$

To investigate the performance of SnapCCESS in terms of major and minor cell-type identification, we split the cell types in each dataset into two sets. The first set included major cell types, defined as those with a number of cells greater than or equal to the median value of the number of cells per cell type. The second set included minor cell types, defined as those with a number of cells less than the median value of the number of cells per cell type.

2.4.2 Assessment of the run time usage

To evaluate the computation speed of SnapCCESS and the conventional VAE ensemble clustering method, all benchmark tasks were allocated to a research server with an NVIDIA GPU GeForce RTX 2080 Ti. The elapsed run time was calculated by using the Python function `time.perf_counter()`. Time for each method only takes into account the deep learning network building and model training steps.

3 Results

3.1 SnapCCESS is an effective and efficient ensemble deep learning method for clustering multimodal single-cell omics data

We first evaluated the performance of SnapCCESS and compared its performance with a conventional VAE ensemble clustering method on the six multimodal single-cell omics datasets generated from different biotechnological platforms. The concordance of the *k*-means clustering output from each method with respect to the cell-type annotation from the original study of each dataset was quantified using ARI (Fig. 2a) and NMI (Supplementary Fig. S1) and the procedure was repeated 20 times to account for the variability in the clustering results. Notably, we found that ensemble learning improves the clustering performance of both methods and the clustering concordance increases while the variability reduces with the ensemble size (i.e. the number of base clusters). We also found that in general the improvement plateau at around the ensemble size of 50.

A key advantage of SnapCCESS is its ability to generate informative embeddings from multiple local minima and therefore requires much fewer epochs during the ensemble learning process (Fig. 1b). To validate this, we tested using epochs of 1, 2, and 10 in SnapCCESS and the conventional VAE ensemble clustering. As expected, in most cases the conventional VAE ensemble clustering requires a high number of epochs to achieve high performance (Fig. 2a and Supplementary Fig. S1). In comparison, SnapCCESS achieves comparable performance using only one epoch in training the ensemble model at the sizes of 50 and 100. Since the epoch is a key parameter that defines the number of times the VAE will work through an input dataset, in general training on more epochs requires more computing time. To evaluate this, we recorded the computation time of SnapCCESS and the conventional VAE ensemble on each dataset. Indeed, we found that, for both methods, fewer epochs resulted in significantly faster computation, especially with large ensemble sizes (Fig. 2b). Under the same number of epochs, the computation time of both methods is very similar. Nevertheless, from our above analyses, only SnapCCESS could achieve high performance in clustering cells with a low training epoch and significantly outperforms the conventional VAE ensemble with an epoch of 1. Taken together, these findings demonstrate that combining individual clustering results derived from multiple embeddings does lead to more accurate and reproducible consensus clustering of multimodal single-cell omics data, and the SnapCCESS framework for ensemble deep learning of embeddings can achieve high-performance cell clustering using significantly less computational time.

3.2 Diagnostic analysis of SnapCCESS reveals its ability to improve major and minor cell-type identification

We evaluated the performance of SnapCCESS using different initial learning rates and varying numbers of cells in the dataset. Using the Swanson TEA-seq dataset, we found that the default learning rate of 0.02 resulted in high cell clustering performance whereas a larger learning rate of 0.2 led to a significant reduction in performance (Supplementary Fig. S2a). Next, we subsampled the cells from the Swanson TEA-seq dataset to test the impact of the number of cells on the performance of SnapCCESS. We found that datasets with small numbers of cells have a significant impact on the performance of SnapCCESS when no ensemble was used (base of 1)

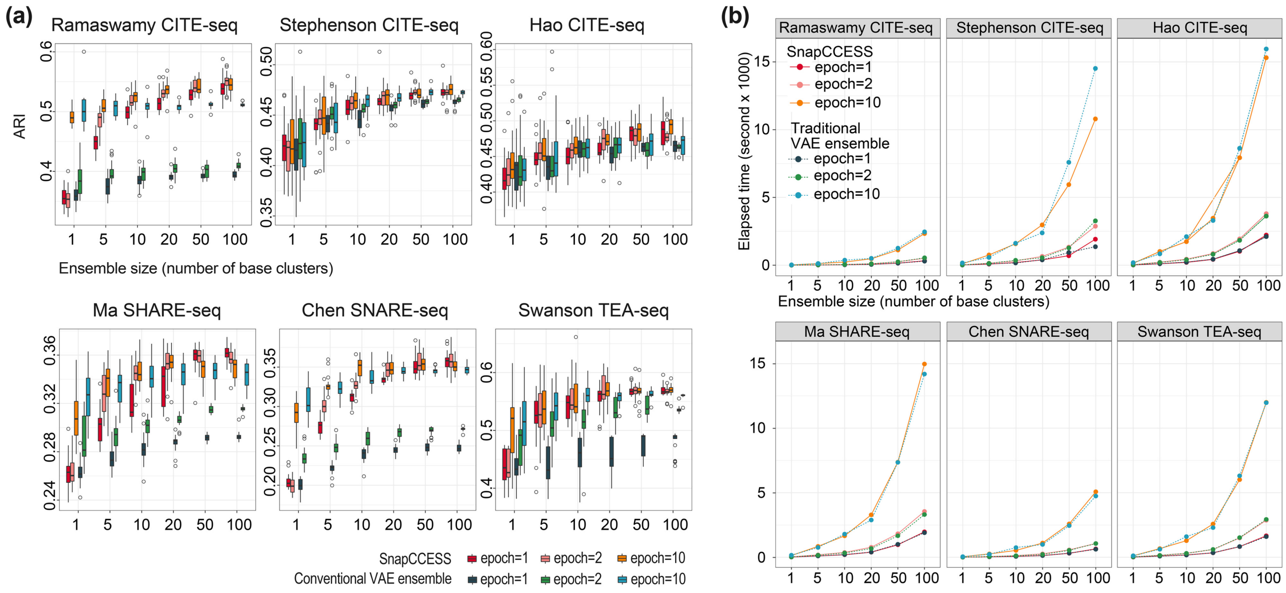

Figure 2 Clustering performance of SnapCCESS and conventional cluster ensembles trained by different numbers of epochs. (a) Concordance of cell-type clustering on six multimodal single-cell omics data. The *x*-axis is the number of base clusters for the ensemble and the *y*-axis is the concordance of the clustering output and the cell-type annotation in the original studies quantified by ARI. The *k*-means clustering algorithm was used for clustering the embeddings generated from each method. The entire procedure was repeated 20 times for capturing the performance variability. (b) Comparison of computation time for SnapCCESS and conventional VAE cluster ensembles. The *x*-axis is the number of base clusters included in the cluster ensembles and the *y*-axis is the elapsed time in the unit second. Results from SnapCCESS are presented as solid lines and those from conventional VAE cluster ensembles are presented as dashed lines. Epochs of 1, 2, and 10 were tested and denoted using different colours.

(Supplementary Fig. S2b). In comparison, SnapCCESS with an ensemble size of 50 base clusters are much more robust to data with small cell sizes, highlighting a key advantage of the ensemble technique.

To further investigate the performance of SnapCCESS on minor cell-type identification, we split the cell types in each dataset into two sets with one containing major cell types and the other containing minor cell types (see Section 2.4.1). We then assessed each set for its concordance of clustering output and the pre-defined cell-type labels. We found in all six datasets a higher clustering performance of SnapCCESS (epoch = 1) with an ensemble of 50 compared to those without using the ensemble in both the major and minor partitions of the cell types (Fig. 3a and Supplementary Fig. S3a). Lastly, as a case study, we visualized the Ramaswamy CITE-seq dataset and highlighted the MAIT-NKT cells using either the ADT modality (Fig. 3a) or the RNA modality (Supplementary Fig. S3b). We found that SnapCCESS with a base of 50 led to much better identification of cells from this cell type than those from using a base of 1 (Fig. 3c and Supplementary Fig. S3c). Taken together, these results suggest that the use of the snapshot ensemble technique in SnapCCESS can improve the identification of both the major and minor cell types.

3.3 Integrated embedding of multimodality generally leads to more precise cell clustering

One of the key motivations in conducting multimodal single-cell omics experiments is the anticipation that the availability of multiple molecular features in individual cells will lead to more precise characterization of cell identity and heterogeneity in complex multicellular organisms and biological systems (Zhu et al. 2020). To investigate this, for each dataset, we trained SnapCCESS (epoch = 1) using either all available data modalities or each unimodality independently, and then performed cell clustering using the k-means clustering algorithm on either the integrated embedding of multimodality or embeddings from each unimodality. We found that in general clustering of cells using integrated embeddings of multimodality leads to significantly better results than from using any unimodality alone (Fig. 4a and Supplementary Fig. S4). Among the clustering results using unimodal embeddings, those generated from RNA modality generally performed similarly or better compared to those from ADT modality. The clustering performance of ATAC modality appears to be lower compared to other modalities, which may be due to the higher dimensionality and data sparsity in ATAC data modality (Xiong et al. 2019). Together, these results support the expectation that taking into consideration of multiple molecular features of cells can lead to a more precise downstream characterization of the biological systems, and further highlight the utility of modality integration methods for analyzing multimodal single-cell omics data.

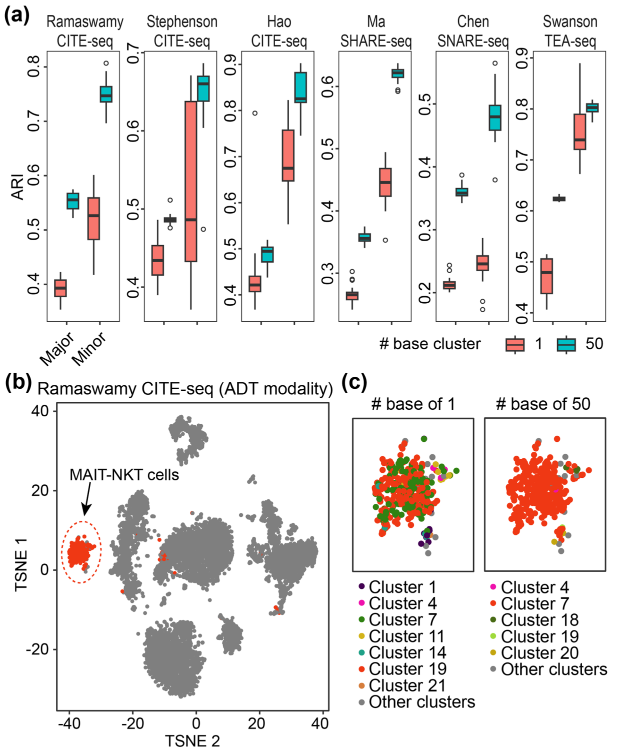

Figure 3 Concordance of clustering output and the pre-defined cell-type labels on (a) major and minor cell type quantified by ARI in each of the six datasets. (b) TSNE visualization of Ramaswamy CITE-seq data using ADT modality. MAIT-NKT cells are highlighted in red. (c) Zoom in of MAIT-NKT cell clustering using SnapCCESS with 1 or 50 base clusters.

3.4 SnapCCESS improves various clustering algorithms

Since the ensemble deep learning of embeddings in SnapCCESS is independent of clustering algorithms, we next tested the performance of the SnapCCESS framework (epoch = 1) by coupling it with a spectral clustering algorithm and SIMLR, a kernel-based clustering algorithm. Note that the two large CITE-seq datasets, Stephenson CITE-seq and Hao CITE-seq, were excluded due to the exponential growth of computational complexity with the number of cells in a dataset for the spectral clustering algorithm. Overall, we observed a clear increase in clustering performance with the increasing ensemble size, regardless of the types of clustering algorithms and concordance evaluation metrics (Fig. 4b and Supplementary Fig. S5). Nonetheless, compared to the simple k-means clustering algorithm, the application of more advanced SIMLR clustering and spectral clustering algorithms generally led to improved cell clustering as measured by their concordance to the cell-type annotation. These findings are of particular interest to SIMLR, which was originally designed for analyzing unimodal scRNA-seq data, as they demonstrate that embeddings learned by SnapCCESS from multimodal single-cell omics data can be used for a clustering algorithm designed for unimodal single-cell omics data.

Similar to the results from k-means clustering, the improvement from SIMLR and spectral clustering also peaked around the ensemble size of 50 (Fig. 4b and Supplementary Fig. S5). This is also consistent with the results from scCCESS for scRNA-seq data analysis (Geddes et al. 2019), suggesting an ensemble size of 50 may be a suitable choice for the ensemble deep learning component in the SnapCCESS framework.

3.5 SnapCCESS performs competitively to the state-of-the-art embedding generating methods for multimodal single-cell clustering

Various methods exist for generating embeddings from multiple data modalities in single-cell multimodal omics data. Some of the state-of-the-art examples include totalVI designed for combining RNA and ADT modalities in CITE-seq data and its sibling MultiVI for combining RNA and ATAC modalities such as in SHARE-seq and SNARE-seq, and Jvis-learn

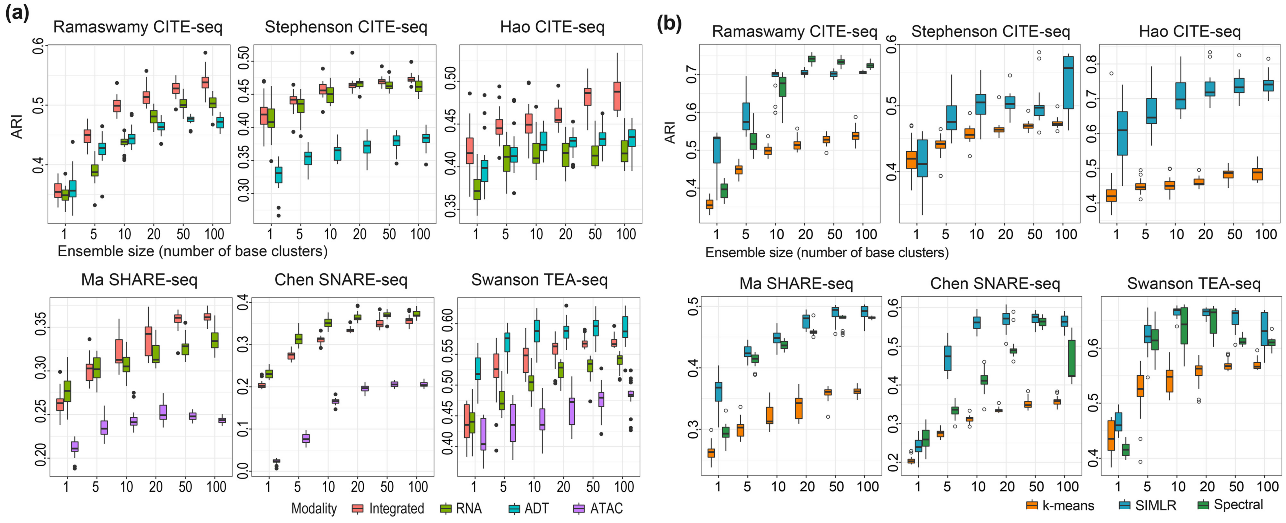

Figure 4 Comparison of embeddings learned from multimodality and unimodality, and evaluation of the SnapCCESS framework with alternative clustering algorithms on multimodal single-cell omics data. (a) Concordance of cell-type clustering quantified by ARI on six multimodal single-cell omics data using SnapCCESS generated embeddings from either all modalities in a dataset or each data modality alone. The entire procedure was repeated 20 times for capturing the performance variability. The *k*-means clustering algorithm was used. (b) Concordance of the cell-type annotations and cell clustering output from each clustering algorithm. The *x*-axis is the number of base clusters in the ensemble and the *y*-axis is the concordance quantification by either ARI. Both (a) and (b) were repeated 20 times for capturing the performance variability.

and MOFA+ which are generic and can be applied to all data modality combinations. Given the applicability of each method, we compared SnapCCESS (epoch = 1 and ensemble size of 50) with TotalVI on the CITE-seq datasets, MultiVI on SHARE-seq and SNARE-seq datasets, and MOFA+ and Jvis-learn on all six datasets with two or three modalities. In particular, we used the multimodality-integrated embeddings generated from each of these methods as input to *k*-means, SMILR, and spectral clustering algorithms and examined the concordance of clustering output with cell-type annotation using ARI (Fig. 5a) and NMI (Fig. 5b).

We found that SnapCCESS performed competitively to other multimodality embedding generation methods. In most cases, its performance is significantly better than totalVI when applied to CITE-seq datasets and MultiVI when applied to SHARE-seq and SNARE-seq datasets (Fig. 5). While MOFA+ also performed well especially on CITE-seq datasets, SnapCCESS appears to be slightly better than MOFA+ when used for generating embeddings and performing cell clustering on the SHARE-seq and the trimodal TEA-seq datasets. Interestingly, while clustering of the embeddings from Jvis-learn using *k*-means and spectral clustering algorithms performed reasonably well, the use of SIMLR on Jvis-learn generated embeddings leads to poor results in many cases. These results may indicate the varying degree of generalizability of the multimodality embedding generation methods on different clustering algorithms. To this end, multimodality embeddings generated from the other four methods (i.e. SnapCCESS, totalVI, MultiVI, and MOFA+) worked well regardless of the used clustering algorithm.

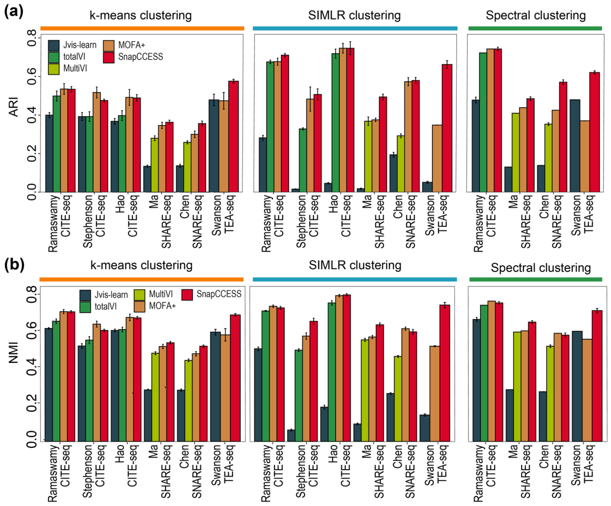

Figure 5 Comparison of SnapCCESS framework (epoch = 1 and ensemble size of 50) with other multimodality embedding generation methods on cell clustering using *k*-means, SIMLR, and spectral clustering algorithm. (a) Quantification of clustering concordance to cell-type annotation using ARI. (b) Similar to (a) but quantifying clustering concordance using NMI. Each bar indicates the average performance across datasets with 20 repeats, and error bars represent the standard deviation.

4 Discussion and conclusion

A main challenge in single-cell omics data analysis is in handling the high-dimensionality of the feature space (Yang *et al.* 2021). Embedding learning is a popular approach for reducing feature dimension for subsequent analysis such as clustering of cells. While many methods have been designed to generate embeddings for unimodality scRNA-seq data, only a few methods are specifically designed for and can be applied to integrate data modalities in multimodal single-cell omics data. Our comparison to these methods highlights the utility of forming ensembles of embeddings for dimension reduction of multimodal single-cell omics data and their subsequent application in cell clustering.

A key innovation in SnapCCESS is the adaptation of the snapshot ensemble learning technique (Huang *et al.* 2017) which significantly reduces the computation time and resources for multiple embeddings compared to conventional VAE ensembles. The underlying idea of the snapshot ensemble is to save multiple versions of a single model during the training process by using a cyclic learning rate scheduler. In SnapCCESS, this allows us to capture the multi-view of the latent space which leads to better clustering performance (Cao *et al.* 2020). Nevertheless, the clustering of each embedding generated from SnapCCESS is still performed individually. While parallelization can be implemented to speed up the process, designing clustering algorithms that can cluster cells via multiple embeddings simultaneously could further improve computational efficiency. Related to this, there are various ensemble deep learning methods that aim to reduce computation time by using techniques such as model branching and neuron deactivation (Cao *et al.* 2020). The effectiveness of these alternative approaches for learning embeddings from multimodal single-cell omics data remains to be tested.

The utility of the embeddings generated from multimodal single-cell omics data is much wider than cell clustering. While clustering is one application that can make use of embeddings learned from such data, other tasks such as supervised cell-type classification (Abdelaal *et al.* 2019) and unsupervised number of cell-type estimation (Yu *et al.* 2022) that take the embeddings as input can also be applied for analyzing multimodal single-cell omics data. Therefore, designing methods that can generate better embeddings will impact various downstream analyses and applications of multimodal single-cell omics data. To this end, the utility of ensemble deep learning methods for these applications (e.g. cell-type classification, number of cell-type estimation) should be investigated in future studies.

While the current study evaluates the SnapCCESS framework for clustering cells into discrete groups, the cell-type structures from many biological systems are hierarchical, with subpopulations of cells existing in each major cell type (Lin *et al.* 2020). The development of multimodal single-cell omics technologies facilitates the characterization of such hierarchical cell-type relationships. Therefore, developing methods that are capable of multi-resolution clustering of cells on datasets with multimodal molecular attributes is a direction of future research. Another recent expansion in the single-cell omics field is the increasing availability of spatial single-cell omics data produced by an array of new spatial profiling technologies (Larsson *et al.* 2021). Combining spatial data with other omics data types produced from the same cells and samples has the potential to uncover a wealth of information, including spatial-related cell-type structure, and will help us gain a deeper understanding of cell and tissue development and disease progression. Developing clustering algorithms for integrating spatial data with other omics datasets is challenging and requires further methodological innovation.

In summary, our previous work demonstrated that ensemble learning of embeddings provides an effective approach for improving downstream clustering analyses by providing a multi-view of the input data (Geddes *et al.* 2019). Here we extend this idea for single-cell multimodal omics data analysis by introducing SnapCCESS, an efficient ensemble deep learning framework using VAE and snapshot techniques, gaining high performance in cell clustering while alleviating the limitation on computation efficiency in conventional ensemble learning methods. Since the clustering of individual embeddings can be performed independently from each other, the proposed framework can benefit from further speed up by parallelization of embedding clustering. We expect SnapCCESS to serve as a useful tool and spark the future development of ensemble deep learning methods for multimodal single-cell omics data analysis.

Acknowledgements

The authors thank all their colleagues, particularly those at the Sydney Precision Data Science Centre, The University of Sydney, for their intellectual engagement and constructive feedback.

Author contributions

P.Y. conceived the study. L.Y. led the data analysis with input from P.Y., C.L., and J.Y.H.Y.; L.Y. and P.Y. wrote the manuscript with input from C.L. and J.Y.H.Y.; all authors read and approved the final manuscript.

Supplementary data

Supplementary data is available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work was supported by a National Health and Medical Research Council (NHMRC) Investigator [1173469 to P.Y.]; a Postgraduate Research Excellence Award (PREA) Tuition Fee Scholarship [to L.Y.]; and the AIR@innoHK programme of the Innovation and Technology Commission of Hong Kong [to P.Y. and J.Y.].

Data availability

All data used in this study are publicly available. Details of each dataset are reported in Section 2. The accession links are summarized here including the Ramaswamy CITE-seq dataset [GSE166489], Stephenson CITE-seq dataset [E-MTAB-10026], Hao CITE-seq dataset [GSE164378], Chen SNARE-seq dataset [GSE126074], Ma SHARE-seq dataset [GSE140203], Swanson TEA-seq [GSE158013].

Code availability

SnapCCESS is implemented as a Python package and is freely available from https://github.com/PYangLab/SnapCCESS.

References

Abdelaal T, Michielsen L, Cats D *et al.* A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol* 2019;**20**:1–19.

Adossa N, Khan S, Rytkönen K *et al.* Computational strategies for single-cell multi-omics integration. *Comput Struct Biotechnol J* 2021;**19**:2588–96.

Argelaguet R, Arnol D, Bredikhin D *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* 2020;**21**:1–17.

Ashuach T, Gabitto M, Jordan M *et al.* MultiVI: deep generative model for the integration of multi-modal data. bioRxiv, August 2021, preprint: not peer reviewed.

Cao Y, Geddes T, Yang JYH *et al.* Ensemble deep learning in bioinformatics. *Nat Mach Intell* 2020;**2**:500–8.

Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* 2019;**37**:1452–7.

Do VH, Canzar S. A generalization of t-SNE and UMAP to single-cell multimodal omics. *Genome Biol* 2021;**22**:1–9.

Gayoso A, Steier Z, Lopez R *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat Methods* 2021;**18**:272–82.

Geddes TA, Kim T, Nan L *et al.* Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis. *BMC Bioinformatics* 2019;**20**: 1–11.

Hao Y, Hao S, Andersen-Nissen E *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573–87.e29.

Hornik K. A clue for cluster ensembles. *J Stat Softw* 2005;**14**:1–25.

Huang G, Li Y, Pleiss G *et al.* Snapshot ensembles: train 1, get m for free. arXiv, arXiv:1704.00109, 2017, preprint: not peer reviewed.

Kim HJ, Lin Y, Geddes T *et al.* CiteFuse enables multi-modal analysis of CITE-seq data. *Bioinformatics* 2020;**36**:4137–43.

Kim T, Chen IR, Lin Y *et al.* Impact of similarity metrics on single-cell RNA-seq data clustering. *Brief Bioinformatics* 2019;**20**: 2316–26.

Larsson L, Frisén J, Lundeberg J. Spatially resolved transcriptomics adds a new dimension to genomics. *Nat Methods* 2021;**18**:15–8.

Lin X, Tian T, Wei Z *et al.* Clustering of single-cell multi-omics data with a multimodal deep learning method. *Nat Commun* 2022;**13**: 7705.

Lin Y, Cao Y, Kim HJ *et al.* scClassify: sample size estimation and multiscale classification of cells using single and multiple reference. *Molecular Systems Biology* 2020;**16**:e9389.

Liu C, Huang H, Yang P. Multi-task learning from multimodal single-cell omics with Matilda. *Nucleic Acids Res* 2023;**51**:e45.

Ma S, Zhang B, LaFave LM *et al.* Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* 2020;**183**: 1103–16.e20.

Miao Z, Humphreys B, McMahon A *et al.* Multi-omics integration in the age of million single-cell data. *Nat Rev Nephrol* 2021;**17**:710–24.

Ramaswamy A, Brodsky NN, Sumida TS *et al.* Immune dysregulation and autoreactivity correlate with disease severity in SARS-Cov-2-associated multisystem inflammatory syndrome in children. *Immunity* 2021;**54**:1083–95.e7.

Stephenson E, Reynolds G, Botting R *et al.*; Cambridge Institute of Therapeutic Immunology and Infectious Disease-National Institute of Health Research (CITIID-NIHR) COVID-19 BioResource Collaboration. Single-cell multi-omics analysis of the immune response in COVID-19. *Nat Med* 2021;**27**:904–16.

Stoeckius M, Hafemeister C, Stephenson W *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 2017; **14**:865–8.

Swanson E, Lord C, Reading J *et al.* Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife* 2021;**10**:e63632.

Wang B, Zhu J, Pierson E *et al.* Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat Methods* 2017;**14**:414–6.

Wu W, Zhang W, Ma X. Network-based integrative analysis of single-cell transcriptomic and epigenomic data for cell types. *Brief Bioinformatics* 2022;**23**:bbab546.

Xiong L, Xu K, Tian K *et al.* Scale method for single-cell ATAC-seq analysis via latent feature extraction. *Nat Commun* 2019;**10**:4576.

Yang P, Huang H, Liu C. Feature selection revisited in the single-cell era. *Genome Biol* 2021;**22**:17.

Yu L, Cao Y, Yang JY *et al.* Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. *Genome Biol* 2022;**23**:1–21.

Zhu C, Preissl S, Ren B. Single-cell multimodal omics: the power of many. *Nat Methods* 2020;**17**:11–4.