BIBS SEMINAR
Thursday, January 04th, 2024

NDAGIJIMANA, Frank Aimee Rodrigue

Seoul National University

Interdisciplinary Program in Bioinformatics.

TODAY'S ARTICLE

Bioinformatics, 39(4), 2023, btad133
https://doi.org/10.1093/bioinformatics/btad133
Advance Access Publication Date: 17 March 2023
Original Paper

OXFORD

Data and text mining

scMCs: a framework for single-cell multi-omics data integration and multiple clusterings

Liangrui Ren[1,2], Jun Wang [2]*, Zhao Li[3], Qingzhong Li[1,2], Guoxian Yu [1,2]

[1]School of Software, Shandong University, Jinan 250101, Shandong, China
[2]Joint SDU-NTU Centre for Artificial Intelligence Research, Shandong University, Jinan 250101, China
[3]College of Computer Science, Zhejiang University, Hangzhou 310058, China

*Corresponding author. Joint SDU-NTU Centre for Artificial Intelligence Research, Shandong University, No. 1500, Shunhua Road, Lixia District, Jinan, Shandong 250101, China. E-mail: kingjun@sdu.edu
Associate Editor: Jonathan Wren

Received 1 January 2023; revised 20 February 2023; accepted 13 March 2023

TABLE OF CONTENTS

01.

INTRODUCTION

1.1. Single-Cell Multi-omics

- The advancement of single-cell sequencing techniques assists researchers to simultaneously obtain multiple omics data.
  - **Single- cell RNA-sequencing (scRNA-seq)** quantifies the mRNA abundance of genes in each cell.
  - **Single-cell Assay for Transposase- Accessible Chromatin using sequencing (scATAC)** characterizes the openness of cis-regulatory elements in nearby genes
- The joint analysis of **scRNA-seq** and **scATAC** data can strength key genetic information of different omics, and decipher gene regulatory relationships related with cellular heterogeneity

1.2. Challenges in Single-Cell Multi-Omics Integration

- Inherent characteristics of single cell data bring great computational and analytical challenges.
  - **High Sparsity**: Only a small fraction of molecular features are detected in each cell. This may be due to technical limitations, biological variability, and not all genes are active or expressed in every cell.
  - **Noise**: Random variations or errors introduced during experimental and measurement processes. This may be due to sample preparation, amplification of genetic material, and actual measurement of omics data.
  - **Dimensionality Mismatch**: Different omics data types have varying dimensionalities, representing information differently.

1.3. Approaches to Overcome Integration Challenges

- Some methods build on **non-negative matrix factorization** or **principal component analysis(PCA)** to integrate single-cell multi-omics data.
  - ➢ **Limitation:** these methods **ignore omics-specific information** and **disregard non-linear geometries of multi- omics data**.
- **Manifold alignment methods** aim to align embedded low-dimensional manifolds of different omics data and characterize intrinsic cellular structures.
  - ➢ **Limitation:**Although these alignment-based methods can capture non-linear geometries across multi-omics data, they suffer a **high time complexity.**

1.4. Deep Learning Approaches for Single-Cell Multi-omics Integration

- **Current Deep Learning-Based Approaches**
  - Single-cell Multimodal variational AutoEncoder (scMVAE)
  - Deep Cross-omics Cycle Attention (DCCA)
- **Limitations:**
  - These methods focus on a shared representation, but disregard the omics individuality, and cannot integrate different levels of biological features.
  - Available single-cell clustering methods only focus on the cell type clustering, which cannot mine the alternative clustering to comprehensively analyze cells.

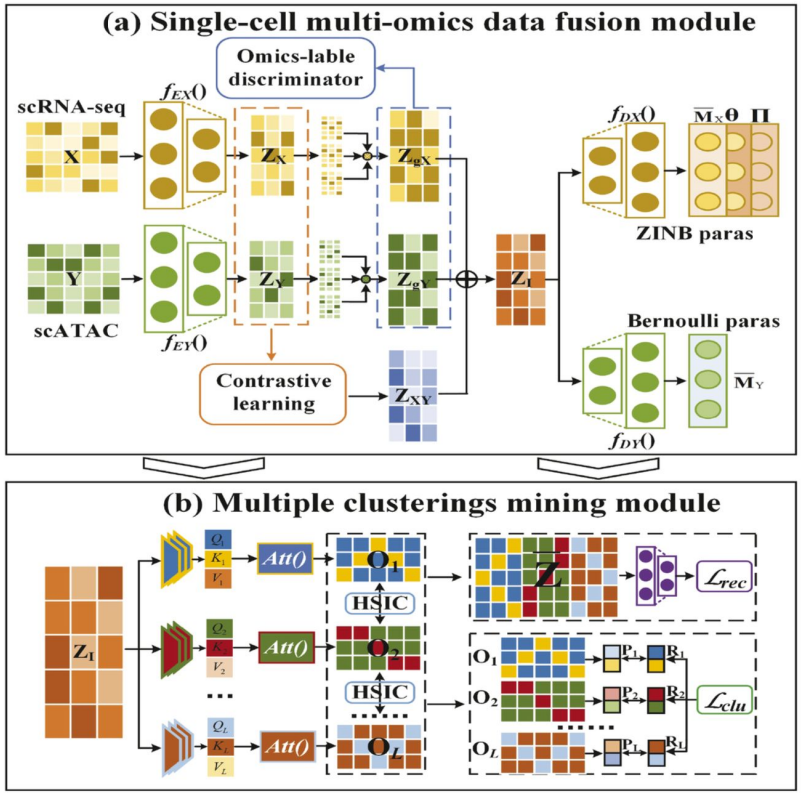
1.5. scMCs, a Solution for Single-Cell Multi-omics Data Integration

- The proposed method, scMCs, aims **to process individuality and commonality from heterogeneous omics**, constructing a comprehensive representation for single-cell multi-omics data fusion, clustering, and multiple clustering.
- It uses omics-independent deep autoencoders, attention mechanism, omics-label discriminator, contrastive learning strategy, multi-head attention mechanism, and Kullback–Leibler divergence-based clustering loss to generate multiple salient subspaces and **generate high-quality clusterings** in an end-to-end framework.

02.

MATERIALS AND METHODS

2.1. Framework Overview



- The Figure shows the overall framework of the proposed method.
- Part (a) aims at multi-omics data fusion and cell clustering
- Part (b) targets to explore multiple clusterings with quality and diversity embedded in multi-omics data

2.2. Multi-omics Data Encoder for Individuality

- Let $X \in \mathbb{R}^{N \times D_X}$ and $Y \in \mathbb{R}^{N \times D_Y}$ be the normalized scRNA-seq data and scATAC data, where N is the number of samples, D_X and D_Y are the number of features.
- scMCs firstly employs **two independent encoders** f_{EX} and f_{EY} to learn respective *d*-dimensional feature representations $\{\mathbf{Z_X}, \mathbf{Z_Y}\} \in \mathbb{R}^{\mathbf{N \times d}}$:

$$\mathbf{Zx} = \boldsymbol{f}_{\mathbf{EX}}(\mathbf{X}), \ \mathbf{Zy} = \boldsymbol{f}_{\mathbf{EY}}(\mathbf{Y}), \ \text{where:}$$

d: is the dimension of embedding space;

$\mathbf{Z_X}$**:** is the latent low- dimensional representation of cells and genes in scRNA-seq data,

$\mathbf{Z_Y}$**:** encodes the latent patterns between cells and peaks in scATAC data.

2.2. Multi-omics Data Encoder for Individuality

- To extract the individuality and explore the complementary information among different omics, this approach incorporates the attention mechanism and omics-label discriminator into the encoder module.
- Concretely, scMCs defines two normalized attention score matrices as:

$$\mathbf{A}_X = softmax\left(\frac{\mathbf{Z}_X(\mathbf{Z}_X)^T}{\sqrt{d}}\right), \mathbf{A}_Y = softmax\left(\frac{\mathbf{Z}_Y(\mathbf{Z}_Y)^T}{\sqrt{d}}\right)$$

where:

- The elements in A_X and A_Y quantify the similarity of a pair of cells for different omics.
- Softmax$(\cdot)$ normalizes the weight to [0, 1] to avoid modeling negative correlations.

2.2. Multi-omics Data Encoder for Individuality

- With the normalized attention scores, this study reorganizes the low-dimensional representations by considering the similarity among cells:

$$\mathbf{Z}_{gX} = \mathbf{A}_X \mathbf{Z}_X,\ \mathbf{Z}_{gY} = \mathbf{A}_Y \mathbf{Z}_Y.$$

- The attention mechanism plays important roles in the encoding module.
  - On one hand, it measures the importance of biological signals in the intrinsic feature spaces of different omics, and extracts omics individuality.
  - On the other hand, it explores the similarity between cells and enables to explore the representation relationship between cells and features from a global perspective.

2.2. Multi-omics Data Encoder for Individuality

- In supervised learning tasks, **labels** can indicate the class or identity of the samples.
- Given that, omics labels can be used as the supervised signals to extract individual features of each omics, here the method explicitly defines the omics labels, i.e. **cells from the same omics are labeled as one type**.
- Next, an omics-label discriminator is designed to further enhance the quality of individuality in Z_{gX} and Z_{gY}.
- The discriminator loss is defined as:

$$\mathcal{L}_{dis} = CE(\mathbf{P}, f_{dis}(\mathbf{Z}_{gX}, \mathbf{Z}_{gY})),$$

CE: the cross-entropy loss , $f_{dis}()$: the omics- label predictor.

$\mathbf{P} \in \{0,1\}^{2N \times K}$ is the true omics- label matrix, where **K** is the number of omics;

2.3. Cross-omics contrastive learning for commonality

- To extract the compact commonality features between different omics, the authors introduce the cross-omics contrastive learning strategy **to extract shared knowledge from scRNA-seq and scATAC data for fusion**.
- The core theory of contrastive learning is to maximize the consistency by maximizing the mutual information between different views.
- In this way, one can obtain more informative embedded features by maximizing the information entropy, and avoid the simple solution of assigning all samples to the same cluster.

2.4. Multi-omics Data Fusion and Imputation for Clustering

- scMCs can learn two latent representations Z_{gX} and Z_{gY} to encode omics individuality, and a latent representation Z_{XY} to encode commonality, which are key factors for clustering and imputing single-cell multi-omics data.
- Here, the authors performed an element- wise sum operation with scale parameters **λx** and **λy** to aggregate them, and generate a more discriminative co-embedding representation ZI :

$$\mathbf{Z}_I = \mathbf{Z}_{XY} + \lambda_x \mathbf{Z}_{gX} + \lambda_y \mathbf{Z}_{gY}$$

2.4. Multi-omics Data Fusion and Imputation for Clustering

- A Zero Inflated Negative Binomial(ZINB) model-based decoder network is proposed to explore the global probabilistic structure of scRNA-seq data, incorporating the mean and dispersion parameters of the negative binomial distribution.
- Mathematically, ZINB is defined with the mean (μ_x) and dispersion (θ) parameters of the negative binomial distribution and a coefficient (π) that describes the probability of dropout events:

$$NB(\mathbf{x};\mu_{\mathbf{x}},\theta)=\frac{\Gamma(\mathbf{x}+\theta)}{\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu_{\mathbf{x}}}\right)^{\theta}\left(\frac{\mu_{\mathbf{x}}}{\theta+\mu_{\mathbf{x}}}\right)^{\mathbf{x}},$$

$$ZINB(\mathbf{x};\pi,\mu_{\mathbf{x}},\theta)=\pi\zeta_0(\mathbf{x})+(1-\pi)NB(\mathbf{x};\mu_{\mathbf{x}},\theta)$$

where :
x is a vector from the original scRNA-seq data.

2.4. Multi-omics Data Fusion and Imputation for Clustering

- the ZINB-based decoder estimates the parameters $\{\pi,\mu_x,\theta\}$ based on Z_I through three different fully connected layers as follows:

$$\mathbf{\Pi} = sigmoid(f_{DX}(\mathbf{Z}_I, \mathbf{W}_\pi)),$$

$$\overline{\mathbf{M}}_X = exp(f_{DX}(\mathbf{Z}_I, \mathbf{W}_{\mu_x})),\ \mathbf{\Theta} = exp(f_{DX}(\mathbf{Z}_I, \mathbf{W}_\theta))$$

- where $\{\prod;\overline{\mathbf{M}}_X;\Theta\}$ is the matrix form of $\{\pi,\mu,\theta\}$; f_{DX} is a decoder with fully connected layer; W_π, W_{μ_x}, and W_θ are three learnable parameter matrices.
- The activation function of $\prod$ is sigmoid () because the dropout probability is between 0 and 1.
- In addition, since the mean and dispersion parameters are non-negative, the exponential function exp() is selected as the activation function for $\overline{\mathbf{M}}_X$ and Θ.

2.4. Multi-omics Data Fusion and Imputation for Clustering

- Different from the traditional mean squared error loss-based autoencoder, the loss function of ZINB-based decoder network is the negative log of the ZINB likelihood:

$$\mathcal{L}_{ZINB} = -log(ZINB(\mathbf{X}|\mathbf{\Pi}, \overline{\mathbf{M}}_X, \mathbf{\Theta})).$$

2.4. Multi-omics Data Fusion and Imputation for Clustering

- Considering the extremely sparse and nearly binary nature of scATAC data, a **Bernoulli distribution (Ber)-based decoder network** was used to model scATAC data:

$$Ber(\mathbf{y};\mu_{\mathbf{y}}) = \mathbf{y}\,log(\mu_{\mathbf{y}}) + (1-\mathbf{y})log(1-\mu_{\mathbf{y}})$$

y: vector from the original scATAC data

μ_y is the mean parameters of Ber.

- The Bernoulli-based decoder estimates μ_y based on Z_I through a fully connected layer with sigmoid() as activation function:

$$\overline{\mathbf{M}}_Y = sigmoid(f_{DY}(\mathbf{Z}_I, \mathbf{W}_{\mu_{\mathbf{y}}}))$$

$\overline{\mathbf{M}}_Y$: the matrix form of μy,

$\mathbf{W}_{\mu y}$: the weight parameter matrix.

- Finally, the Bernoulli-based autoencoder can be optimized by the cross-entropy loss:

$$\mathcal{L}_{Ber} = CE(\mathbf{Y}, \overline{\mathbf{M}}_Y)$$

2.4. Multi-omics Data Fusion and Imputation for Clustering

- To pursue a more discriminative and informative co-embedding representation that incorporates individuality and commonality of multi-omics data,the authors unify the objective of imputing the scRNA-seq data and scATAC data, predicting the omics labels, and cross-omics contrastive learning loss as follows:

$$\mathcal{L}_1 = \text{argmin}_{\Phi_1}((-log(ZINB(\mathbf{X}|\mathbf{\Pi}, \overline{\mathbf{M}}_X, \mathbf{\Theta}))) + \alpha_1 CE(\mathbf{Y}, \overline{\mathbf{M}}_Y)$$
$$+\alpha_2 CE(\mathbf{P}, f_{dis}(\mathbf{Z}_{gX}, \mathbf{Z}_{gY}))$$
$$+\alpha_3(-(I(\mathbf{Q}_X, \mathbf{Q}_Y) + \epsilon(H(\mathbf{Q}_X) + H(\mathbf{Q}_Y))))),$$

- By optimizing this equation the individual and shared feature representations can be learned from multi-omics data, and they can be merged into an informative co-embedded representation for clustering and multiple clustering.

2.5. Multiple Clusterings Mining Module

- Contemporary single-cell multi-omics analysis methods focus on integrating cross-omics shared features to find optimal cell division patterns, neglecting other important patterns.
- Multi-view multiple clustering, unlike traditional multi-view methods, incorporates consistent and specific features, generating multiple meaningful and non-redundant clusterings.
- This helps divide cells from different perspectives and explain cell heterogeneity.
- scMCs introduces another module to more comprehensively mine single-cell multi-omics data, **utilizing omics individuality and commonality** to explore alternative clusterings embedded in the multi-omics data.

2.5. Multiple Clusterings Mining Module

- The module utilizes **multi-head attention** to generate different salient subspaces, ensuring diversity.
- To enhance the quality and reduce redundancy between clusterings, **Hilbert Schmidt Independence Criterion (HSIC)** is employed.
- The optimization process involves learning sets of cluster centers in each subspace using KL divergence loss and an auxiliary target distribution.
- The overall objective combines reconstruction loss, redundancy reduction, and clustering loss, providing a comprehensive framework to mine single-cell multi-omics data effectively.

03.

RESULTS

- Experimental Setup
- Cell Clustering and Visualization
- Evaluation of Data Imputation
- Evaluation of Multiple Clustering
- Ablation Study and Parameter Sensitivity Analysis

3.1. Experimental Setup

3.1.1. Datasets

- In the experiments, the performance of scMCs is evaluated by jointly modeling the scRNA-seq data and scATAC data.
- Four preprocessed single-cell multi-omics data with paired profiles were collected from a previous study (Zuo et al. 2021).

Data set	# Cells	Details
Cell MIx	1047	• Downloaded from GEO (D1, GSE126074) • The chromatin accessibility and gene expression in each single-cell are simultaneously co-assayed using the SNARE-seq
PBMC 3K (D2)	3012	Downloaded from 10X Genomics
Mouse skin	34774	Derived from adult mouse skin by SHARE-seq.
AdBrain	10 309	• Downloaded from GEO (D4, GSE126074) • the chromatin accessibility and gene expression in each single-cell are derived from the adult mouse cerebral cortex.

3.1. Experimental Setup

3.1.2. Evaluation Protocols

- **K-means** is applied for single clustering to cluster cells based on the low-dimensional co-embedding representation Z_I.
- To evaluate the clustering performance, the current study used **Normalized Mutual Information(NMI)** and **Adjusted Rand Index(ARI)**.
- To evaluate multiple clustering, **NMI** and **Jaccard Index(JI)** were used to measure the overlap between different clusterings, and **Silhouette coefficient** and **Dunn Index(DI)** to evaluate the quality of each clustering.

3.1. Experimental Setup

3.1.3. Comparing Baselines

This study implements scMCs with the MindSpore deep learning framework and compare it against four competitive single-cell multi-omics data fusion methods:

- JSNMF (Ma et al. 2022)
- UnionCom (Cao et al. 2020)
- scMVAE(Zuo and Chen 2021)
- DCCA (Zuo et al. 2021)

3.2. Cell Clustering and Visualization

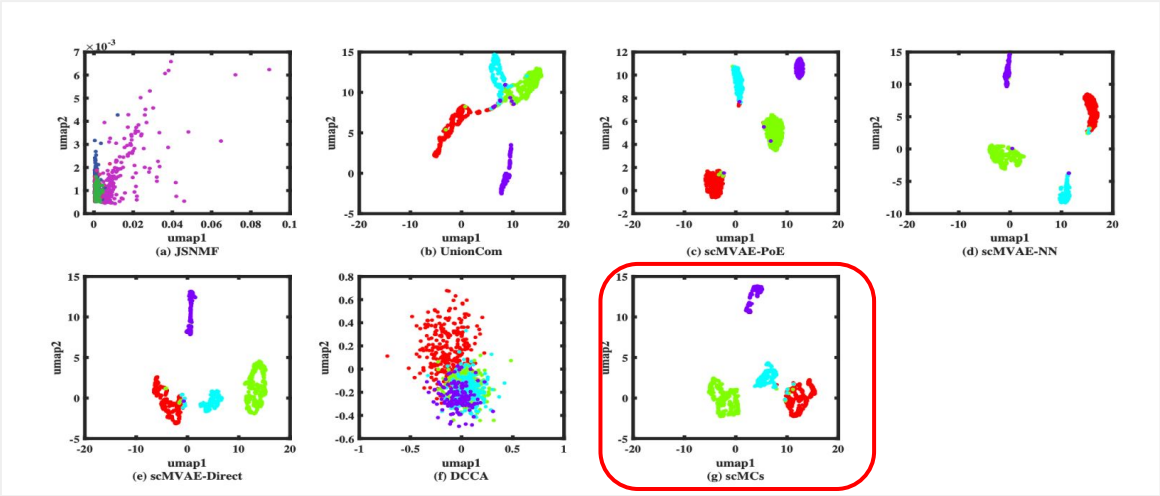
- Each method repeats five times to take the average and variance
- UnionCom is too time-consuming on large datasets, so its results on Mouse skin are not reported.
- scMCs performs well on the four data- sets in terms of NMI and ARI, and the clustering results are statistically better than other methods in most cases.

Table 1. Performance of single clustering of compared methods on different datasets.[a]

		JSNMF	UnionCom	scMVAE-PoE	scMVAE-NN	scMVAE-Direct	DCCA	scMCs
D1	NMI	0.262 ± 0.003•	0.704 ± 0.004•	0.852 ± 0.002•	0.817 ± 0.001•	0.811 ± 0.000•	0.619 ± 0.000•	**0.907 ± 0.000**
	ARI	0.196 ± 0.003•	0.670 ± 0.005•	0.839 ± 0.001•	0.819 ± 0.000•	0.811 ± 0.001•	0.513 ± 0.001•	**0.939 ± 0.000**
D2	NMI	0.416 ± 0.000•	0.606 ± 0.000°	0.603 ± 0.002°	**0.611 ± 0.001°**	0.505 ± 0.002•	0.414 ± 0.000•	0.534 ± 0.001
	ARI	0.284 ±0.004•	0.400 ± 0.001•	0.452 ± 0.007•	0.447 ± 0.003•	0.441 ± 0.004•	0.404 ± 0.000•	**0.596 ± 0.000**
D3	NMI	0.140 ± 0.000•		0.334 ± 0.000•	0.331 ± 0.000•	0.294±0.001•	0.265 ± 0.000•	**0.433 ± 0.000**
	ARI	0.087 ± 0.000•		0.250 ± 0.000•	0.260 ± 0.000°	0.232 ± 0.002•	0.250 ± 0.001•	**0.260 ± 0.000**
D4	NMI	0.269 ± 0.000•	0.305 ± 0.001•	0.325 ± 0.001•	0.287 ± 0.005•	0.273±0.003•	0.296 ± 0.003•	**0.510 ± 0.000**
	ARI	0.194 ± 0.001•	0.248 ± 0.005•	0.268 ± 0.001•	0.164 ± 0.002•	0.125 ± 0.002•	0.197 ± 0.001•	**0.554 ± 0.001**

3.2. Cell Clustering and Visualization (Continued)

- To illustrate the quality of Z_I, UMAP was applied to visualize cell clustering points of scMCs and other baselines on each benchmark dataset.
- scMCs has the **clearest division boundaries** and the **lowest misclassification rate**.
- These results also explain why scMCs achieves a better clustering performance.



3.3. Evaluation of Data Imputation

- Besides accurate cell clustering, scMCs also realizes data imputation based on Z_I using two independent deep generative decoder net- works.
- To evaluate the quality of imputed scRNA-seq data and scATAC data, this study visualizes the raw data and the imputed data generated by scMCs,and other deep learning methods: scMVAE-PoE, scMVAE-Direct, scMVAE-NN, and DCCA.
- Specifically, the raw data and imputed data were projected into different 2D spaces via UMAP, and cell clusterings were explored.
- NMI and ARI were considered to evaluate the clustering given by each method.

3.3. Evaluation of Data Imputation(Continued)

- We see the NMI and ARI scores of scMCs are significantly higher than those of other baselines.
- The visualization results also confirm the cell clustering found by scMCs is more separated between different clusters and more compact within clusters.
- All these confirm that scMCs can generate an informative embedding representation ZI, which can be used for data imputation.

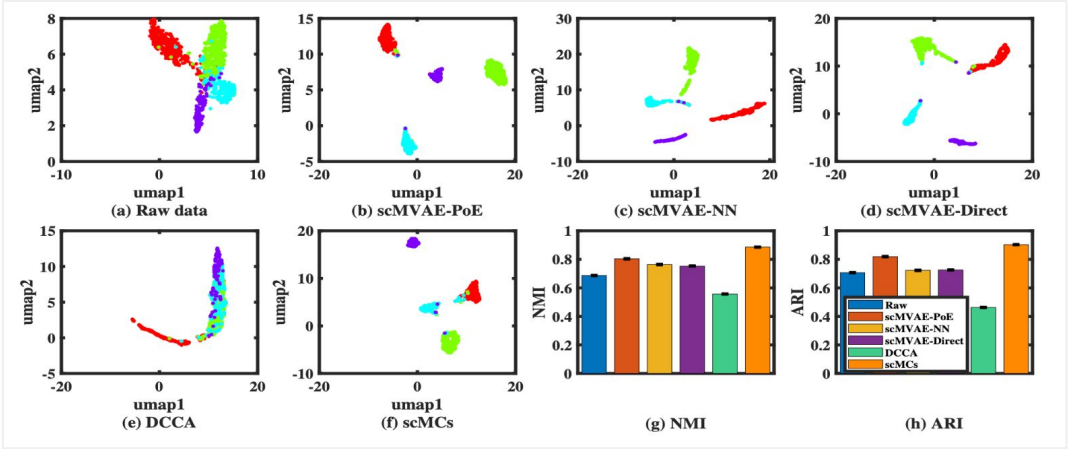


Figure S5:
Cell clustering visualization of each method on raw and imputed CellMix scRNA-seq data. (a) Raw data; (b) scMVAE-PoE; (c) scMVAE-NN;
(d) scMVAE-Direct; (e) DCCA; (f) scMCs; (g) NMI values; (h) ARI values.

3.4. Evaluation of Multiple Clusterings

- Existing single-cell data clustering methods can only find one clustering pattern of cell types.
- However, with increased single-cell data, there are alternative and meaningful clusterings that can uncover new patterns of cells more comprehensively.
- scMCs can project co-embedding representations into different subspaces and find different clusterings.
- Users can specify the number of clusterings and clusters based on datasets or user expectations.
- In experiments, scMCs project Z_I into subspaces, generate two clusterings, and measure their overall quality.

3.4. Evaluation of Multiple Clusterings

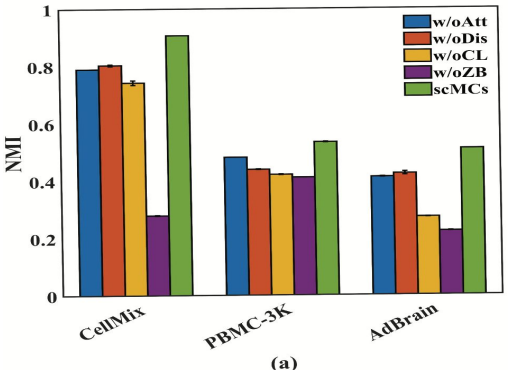
- C1 has a high similarity with the ground truth Ct , while the smaller NMI and JI values indicate that C2 is not similar to Ct.
- In addition, the high SC and DI values suggest that C2 is a potential alternative clustering with high quality.

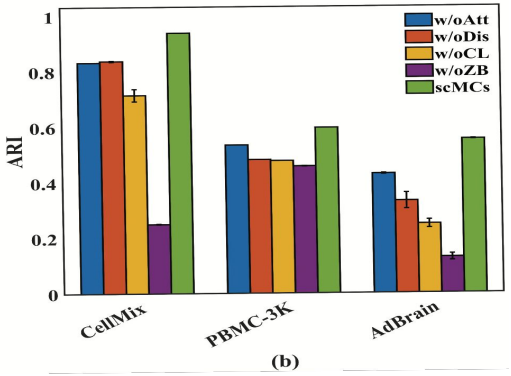
Table 2. Diversity and quality of multiple clusterings generated by scMCs on benchmark datasets.[a]

		CellMix $\mathcal{C}_t$	PBMC_3K $\mathcal{C}_t$	AdBrain $\mathcal{C}_t$
NMI↑	$\mathcal{C}_1$	0.845	0.695	0.513
	$\mathcal{C}_2$	0.365	0.204	0.289
JI↑	$\mathcal{C}_1$	0.860	0.378	0.364
	$\mathcal{C}_2$	0.355	0.197	0.291
SC↑	$\mathcal{C}_1$	0.666	0.644	0.268
	$\mathcal{C}_2$	0.599	0.826	0.579
DI↑	$\mathcal{C}_1$	0.076	0.071	0.048
	$\mathcal{C}_2$	0.054	0.040	0.053

3.5. Ablation Study

- To study the contribution factors of scMCs, four variants are introduced :**w/oAtt**, **w/oDiscriminator**, **w/oCL**, and **w/oZB**, which separately disregard the attention layer, omics-label discriminator, contrastive learning, and ZINB loss and Bernoulli loss.
- scMCs **outperforms** its variants by a clear margin, which confirms that attention layer, omics-label, contrastive learning mechanism, and generative decoder indeed contribute to the quality of cell clustering.





04.

CONCLUSION

- This article Proposes Single-Cell Multi-omics Clustering (scMCs) for single-cell multi-omics data integration.
- scMCs extract individual and shared features of multi-omics data and fuse them into informative co-embedding representation.
- scMCs can comprehensively mine multi-omics data by projecting the co-embedding representation into different subspaces.

- Experimental results show superior and competitive performance in cell clustering and data implementation.
- scMCs find multiple clustering structures with diversity and quality, providing insights into diverse cellular roles.
- Future pursuits include combining data fusion and multiple clustering mining into a unified method and simplifying scMCs with fewer parameters.

THANK YOU
FOR YOUR ATTENTION !