Briefings in Bioinformatics, 2025, **26(2)**, bbaf157
https://doi.org/10.1093/bib/bbaf157
Problem Solving Protocol

FactVAE: a factorized variational autoencoder for single-cell multi-omics data integration analysis

Linjie Wang[1], Huixia Zhang[1], Bo Yi[1], Weidong Xie[1], Kun Yu[2], Wei Li[3,4,*], Keqin Li[5], Dazhe Zhao[1,*]

[1]School of Computer Science and Engineering, Northeastern University, 110819, Shenyang, China
[2]College of Medicine and Bioinformation Engineering, Northeastern University, 110819, Shenyang, China
[3]Key Laboratory of Intelligent Computing in Medical Image (MIIC), Northeastern University, 110000, Shenyang, China
[4]National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Northeastern University, 110819, Shenyang, China
[5]Department of Computer Science, State University of New York, Albany, NY 12561, United States

*Corresponding authors. Wei Li, E-mail: liwei@cse.neu.edu.cn; Dazhe Zhao, E-mail: zhaodazhe@mail.neu.edu.cn

Abstract

Single-cell multi-omics technologies have revolutionized the study of cell states and functions by simultaneously profiling multiple molecular layers within individual cells. However, existing methods for integrating these data struggle to preserve critical feature information and fail to exploit known regulatory knowledge, which is essential for understanding cell functions. This limitation hinders their ability to provide comprehensive and accurate insights into cells. Here, we propose FactVAE, an innovative factorized variational autoencoder designed for the robust and accurate understanding of single-cell multi-omics data. FactVAE integrates the factorization principle into the variational autoencoder framework, ensuring the preservation of feature information while leveraging the non-linear capture of sample information by neural networks. Additionally, known regulatory knowledge is incorporated during model training, and a knowledge transfer strategy is employed for cell embedding optimization and data augmentation. Comparative analyses of single-cell multi-omics datasets from different protocols and the spatial multi-omics dataset demonstrate that FactVAE not only outperforms benchmark methods in clustering performance but also generates augmented data that reveals the clearest cell-type-specific motif expression. Moreover, the feature embeddings captured by FactVAE enable the inference of potential and reliable gene regulatory relationships. Overall, FactVAE's superior performance and strong scalability make it a promising new solution for single-cell multi-omics data analysis.

Keywords: single-cell multi-omics data; factorization; variational autoencoder (VAE)

Introduction

The emergence of single-cell multi-omics sequencing technology has offered a ground-breaking approach to researchers, allowing them to simultaneously acquire diverse information such as gene expression and chromatin states of individual cells [1]. The consistent and complementary information derived from multi-omics data provides a more comprehensive and insightful understanding of cellular states, paving the way for resolving complex tissues at the single-cell resolution. The most noteworthy methods are the widely adopted ones that are capable of simultaneously detecting epigenomic and transcriptomic information of individual cells, such as SNARE-seq [2], sci-CAR [3], SHARE-seq [4], and 10X Genomics. As the assay for transposase-accessible chromatin (ATAC) data reveals the regulatory landscape of cells, while RNA data provides insights into gene expression, researchers can more precisely delineate distinct cell types and states by harnessing the complementary information from these two modalities. Single-cell multi-omics data integration analysis has gradually become a pivotal direction as data volumes increase and analytical techniques advance. However, this integration analysis faces several challenges, primarily because of the high dimensionality and diversity of data [1, 5]. Effective methods are required to accurately characterize and interpret various data types and feature layers.

Various methods have been proposed for single-cell multi-omics data analysis, which can be categorized as matrix decomposition-, alignment-, and neural network (NN)-based [6]. Matrix decomposition-based methods such as Non-negative Matrix Factorization (NMF) [7] and Canonical Correlation Analysis (CCA) [8] transform high-dimensional data into more comprehensible, low-dimensional factor matrices. These methods explore shared information within data by mapping single-cell multi-omics data into a shared low-dimensional latent space to integrate data. However, while simple and intuitive, matrix decomposition-based methods strictly assume linear relationships in the data, limiting their ability to capture nonlinear relationships common in single-cell multi-omics data. Alignment-based methods aim to identify shared sample-to-sample relationships among different datasets. Since the graph structure of cells is typical in single-cell genomic data processing, some studies have integrated data through graph alignment. Joint diffusion [9] constructs graph structures for different omics data and performs

Received: January 15, 2025. **Revised:** March 02, 2025. **Accepted:** March 21, 2025

joint diffusion through a random walk process to integrate data from different omics. In addition, the distance matrix is a common way to express cell-to-cell relationships. CiteFUSE [10] employs similarity network fusion algorithms to process the transcriptomic and proteomic data. UnionCom [11] constructs a geometric distance matrix through shortest distances and maps features from different omics to a unified latent space through geometric matching. These methods are excellent at handling complex relationships within data, capturing nonlinear relationships, and representing the data. However, their high time complexity, $O(N^3)$ (where N is the number of samples), makes it difficult to handle large-scale data [12].

Recently, NN-based methods have been high-profile due to their powerful learning capabilities, which are primarily based on variational autoencoders (VAEs) with dedicated encoders and decoders for omics data. Various training strategies are employed to project data from heterogeneous omics into a common latent space, including omics data transformation [13, 14], joint learning [15, 16], adversarial training [17, 18], and contrastive learning [19, 20]. In particular, scCobra [21] prospectively leverages both contrast learning and adversarial training, and treats the integration of multi-omics data as a batch correction problem through a modality conversion strategy. In addition, autoencoder-based approaches, such as DCCA [22], employ coupled autoencoders for different omics data and introduce an attention mechanism to constrain the encoders to learn information between omics data. The success of these methods underscores the flexibility and robust nonlinear modeling capabilities of NN-based approaches, ensuring they capture complex relationships and patterns within data, even in highly noisy data.

Despite these advancements, NN-based methods for single-cell multi-omics data analysis still encounter significant challenges, particularly in extracting feature information and exploiting regulatory knowledge. These methods primarily focus on capturing cell-level information, while neglecting feature-level details [13–16, 18, 19, 22]. This oversight hinders the exploration of rich feature-level correlations across omics data, which is essential for a comprehensive understanding of cell states. Many studies have insufficiently leveraged pre-existing regulatory relationships between genes and peaks to inform the learning process of the model. This knowledge is critical for gene-peak relationship inference, the absence of which limits the reliability of models in revealing the gene's underlying regulatory mechanism [23]. Moreover, some studies simplify ATAC data by converting peak matrices to gene activity matrices, which improves computational efficiency but sacrifices data integrity [8, 15]. However, such conversion relying on fragment files is not always feasible, as some ATAC datasets lack the necessary files, limiting the applicability of these methods.

To address these issues, we propose a novel factorized variational autoencoder (FactVAE) to comprehensively analyse ATAC and RNA data derived from the same cells. Specifically, FactVAE combines factorization with VAEs to extract feature information and capture the nonlinear characteristics of the data through NNs. Furthermore, FactVAE leverages known gene-peak to guide model training, ensuring the reliability of the derived information, and employs a knowledge transfer strategy to mitigate the inherent sparsity of ATAC data [24]. Through unsupervised analysis of single-cell multi-omics data, FactVAE can accomplish the following goals: (i) generating embeddings for each omics data, facilitating cell visualization and clustering; (ii) augmenting ATAC data with insights from RNA data; and (iii) inferring potential gene-peak regulatory relationships based on the feature information captured by decoders.

Materials and methods

Single-cell ATAC data $X^a \in \mathbb{R}^{N\times D_a}$ and RNA data $X^r \in \mathbb{R}^{N\times D_r}$ are simultaneously derived from cells. Here, N represents the number of cells, D_a and D_r represent the detected number of genes and peaks, respectively. The objective is to perform an integrated analysis to obtain cell embeddings $Z^a \in \mathbb{R}^{N\times d}$ and $Z^r \in \mathbb{R}^{N\times d}$, where d denotes the dimension of the latent space, with $d \ll D_a$ and $d \ll D_r$. This approach allows for a more precise cell identification than analysing each individual type of omics data. Additionally, augmenting scATAC-seq data is expected to facilitate downstream analysis by mitigating the challenges posed by its sparsity.

FactVAE, depicted in Fig. 1(A), consists of two specialized VAEs: RNA-VAE for the RNA data and ATAC-VAE for the ATAC data. Unlike classical VAEs, FactVAE introduces factorization into VAE. The encoders project raw data to a latent space, capturing the sample (cell) information. The decoders capture the feature (gene or peak) information and reconstruct the data distribution by multiplying feature embeddings with cell embeddings. FactVAE's training involves two iterative training processes: a co-optimization of RNA-VAE and ATAC-VAE and an independent optimization of the ATAC-VAE's encoder. Co-optimization guides the model to capture reliable feature information under the constraint of established gene-peak regulatory relationships. The independent optimization, facilitated by a similarity structure alignment, encourages the ATAC-VAE's encoder, as the student model, to learn as much knowledge as possible from RNA-VAE as the teacher model, thereby optimizing the cell embeddings and augmenting the ATAC data. As shown in Fig. 1(B), the cell embeddings derived from FactVAE facilitate downstream cell-level analyses such as cell clustering, while the feature embeddings reveal insights into potential gene-peak regulatory relationships. Moreover, the augmented ATAC data generated by FactVAE is suitable for downstream analysis, such as differential motif detection.

Datasets

In this study, we collect eight publicly available multi-omics datasets (see Supplementary Table S1 for details), spanning human and mouse species from five major sequencing platforms to evaluate the performance of FactVAE. Each single-cell dataset provides cell-type annotation as ground truth. The details of data preprocessing can be found in Supplementary S1.

The specialized VAE with factorization principle

The key idea of factorization is to decompose a high-dimensional data matrix into the product of two low-dimensional matrices, capturing both sample and feature information. By integrating this principle into the VAE framework, FactVAE aims to address the common issue in NN-based methods, in which feature information is ignored. Specifically, the encoder employs a multilayer perceptron (MLP) to extract sample embeddings, whereas the decoder reconstructs the data using trainable matrices that represent feature embeddings. This structure enables FactVAE to capture the nonlinear characteristics of samples while preserving the feature details.

To accommodate the unique characteristics of different omics data, FactVAE trains independent models for each modality, effectively capturing omics-specific information. The encoders focus

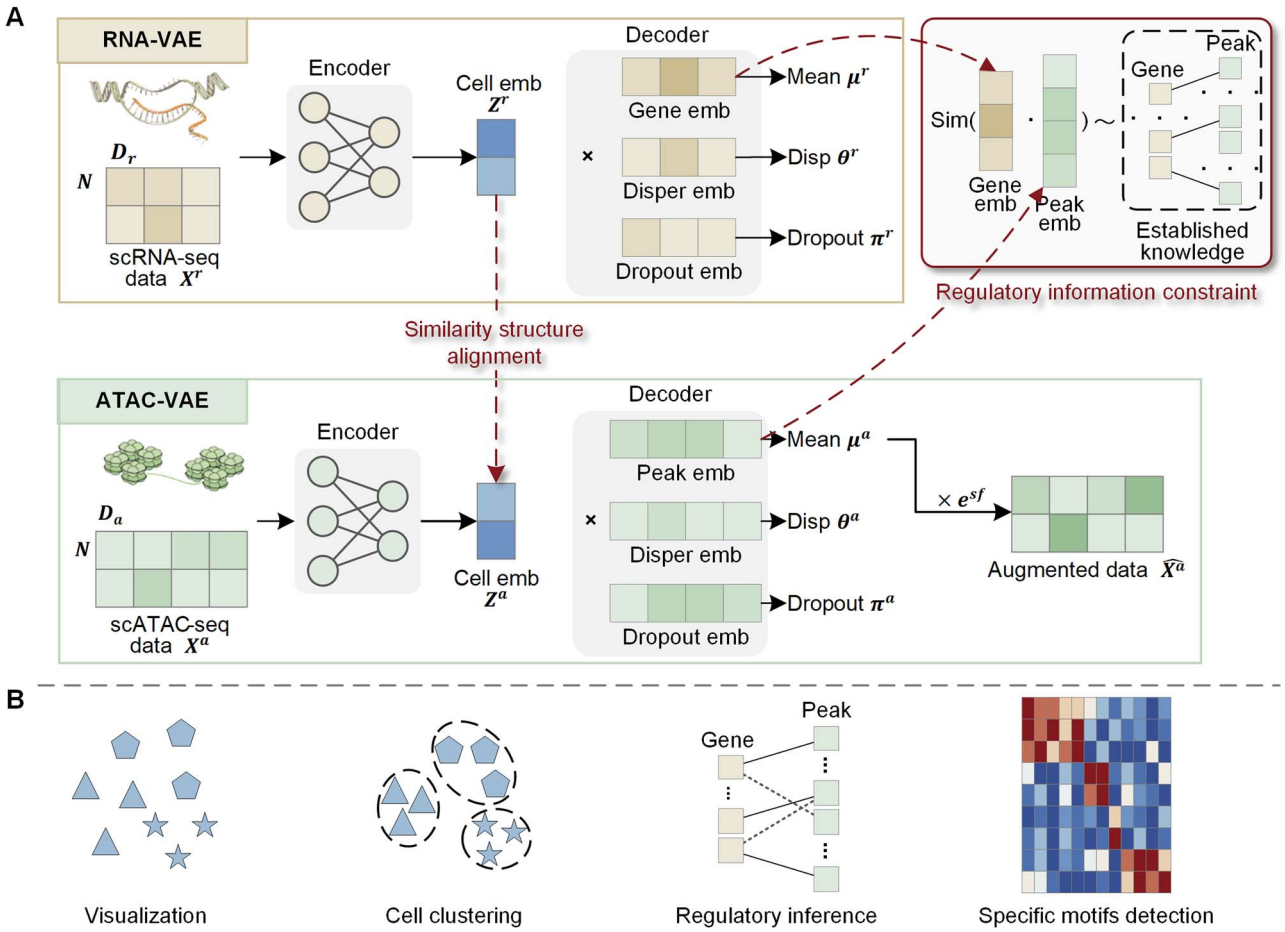

Figure 1. Overview of FactVAE. (A) Preprocessed RNA and ATAC data are fed into their respective VAEs, RNA-VAE and ATAC-VAE. Each encoder extracts cell embeddings from its corresponding omics data. The decoding process utilizes matrix multiplication to generate parameters of the zero-inflated negative binomial (ZINB) distribution, including mean μ, dispersion θ, and zero-inflation parameter π, reconstructing the data's distribution from the latent space. During training, known gene regulatory information is leveraged to constrain the gene and peak embeddings within the decoders. Furthermore, a similarity structure alignment (SSA) loss term is employed to facilitate knowledge transfer from RNA-VAE to ATAC-VAE, guiding ATAC-VAE's encoder. (B) The downstream analyses supported by FactVAE include cell embeddings for visualization and clustering of cells, gene and peak embeddings for regulation inference, and augmented ATAC data for differential motifs detection.

on deriving shared cell embeddings across omics data, while the decoders capture unique feature embeddings. For illustrative purposes, we explain the operation using ATAC-VAE as an example, the same principles apply to RNA-VAE by replacing the superscript a with r.

The loss function for ATAC-VAE, as shown in Equation 1, consists of a reconstruction loss and a Kullback-Leibler (KL) divergence term. The former ensures the model captures essential information, while the latter encourages latent variables to follow a Gaussian distribution.

$$L_{\text{atac}} = -\mathbb{E}_q[\log p\left(\mathrm{x}^a \mid \mathrm{z}^a\right)] + D_{\text{KL}}\left(q\left(\mathrm{z}^a \mid \mathrm{x}^a\right) \parallel p\left(\mathrm{z}^a\right)\right). \tag{1}$$

Here, $-\mathbb{E}_q[\log p\left(\mathrm{x}^a \mid \mathrm{z}^a\right)]$ denotes the negative log-likelihood of the reconstruction loss, where $p\left(\mathrm{x}^a \mid \mathrm{z}^a\right)$ is the generative distribution from the decoder. And $q\left(\mathrm{z}^a \mid \mathrm{x}^a\right)$ represents the posterior distribution of the latent variable z^a from the encoder. The prior distribution $p\left(\mathrm{z}^a\right)$ is defined as a standard normal distribution, $N(0, I)$.

The encoder processes the gene expression matrix x^a using MLP layers to generate the mean μ and variance σ of the latent variable z^a through Equations 2 and 3, respectively. The latent variable z^a is sampled using the reparameterization trick as shown in Equation 4, enabling back-propagation through sampling. This allows the encoder to project raw data from a high-dimensional space to a low-dimensional latent space, generating cell embeddings z^a.

$$\mu = w_\mu\left(ELU\left(w_0 \mathrm{x}^a + b_0\right)\right) + b_\mu \tag{2}$$

$$\sigma = w_\sigma\left(ELU\left(w_0 \mathrm{x}^a + b_0\right)\right) + b_\sigma \tag{3}$$

$$q_\phi\left(\mathrm{z}^a \mid \mathrm{x}^a\right) = \mu + \epsilon \odot \sigma, \epsilon \sim \mathrm{N}(0, I) \tag{4}$$

where ϵ is noise sampled from a standard normal distribution.

Since both single-cell ATAC and RNA data adhere to the ZINB distribution [25, 26], the decoders in FactVAE compute the reconstruction loss by fitting this distribution:

$$\begin{aligned} p_\theta(\mathrm{x}|\mathrm{z}^a) &= \mathrm{ZINB}(\mathrm{x}; \pi^a, \mu^a, \theta^a) \\ &= \pi^r \delta(X) + (1-\pi) NB(\mathrm{x}; \mu^a, \theta^a) \end{aligned} \tag{5}$$

where the negative binomial (NB) component is given by:

$$NB(\mathrm{x}; \mu^a, \theta^a) = \frac{\Gamma(\mathrm{x} + \theta^a)}{\Gamma(\theta^a)\Gamma(\mathrm{x}+1)}\left(\frac{\mu^a}{\mu^a + \theta^a}\right)^{\mathrm{x}}\left(\frac{\theta^a}{\theta^a + \mu^a}\right)^{\theta^a} \tag{6}$$

where $\delta(X)$ is the Dirac function. Parameters π^a, μ^a, and θ^a represent the zero-inflation term, mean, and dispersion, respectively. These are generated through matrix multiplication with specific matrices within the decoder:

$$\begin{aligned}\mu^a &= \mathbf{Softmax}(z^a \times peak^T)\\ \theta^a &= \mathbf{exp}(Softmax(z^a \times disp^T))\\ \pi^a &= z^a \times dropout^T.\end{aligned} \tag{7}$$

These trainable parameter matrices are initialized by Xavier normal distribution and encapsulate distinct aspects of the data's characteristics.

- Peak Matrix (*peak*): This matrix encodes the primary feature information. Matrix multiplication with cell embeddings (z^a) generates the mean (μ^a) of the ZINB distribution, reflecting the expected reconstruction data.
- Dispersion Matrix (*disp*): This matrix captures the dispersion of counts in the data, modeling the inherent variability within single-cell data.
- Dropout Matrix (*dropout*): This matrix determines the probability of dropout events in single-cell data by modeling the zero-inflation property.

By applying factorization principles, FactVAE explicitly learns both sample and feature embeddings. This approach ensures that the unique characteristics of each data type are preserved in multi-omics data analysis, which is essential for exploring the interactions between different omic layers.

Regulatory information constraint

FactVAE introduces a regulatory information constraint to bridge ATAC and RNA data and capture gene regulatory relationships. This constraint supervises the decoders of the RNA-VAE and ATAC-VAE, ensuring biologically meaningful relationships between gene and peak embeddings derived from the respective omics data.

The regulatory knowledge matrix $G2P \in \{0,1\}^{D_r \times D_a}$ is constructed based on the known locations between genes and peaks. Specifically, if gene locus i covers peak j, then $G2P_{i,j} = 1$, indicating a potential regulatory association; otherwise $G2P_{i,j} = 0$.

A similarity matrix $S^{g2p} \in \mathbb{R}^{D_r \times D_a}$ is calculated to represent the association strength between the gene and peak embeddings generated by RNA-VAE and ATAC-VAE respectively:

$$S^{g2p} = \frac{gene \cdot peak^T}{\|gene\| \cdot \|peak\|}. \tag{8}$$

FactVAE minimizes the mean squared error (MSE) loss function to align S^{g2p} and $G2P$:

$$L_{reg} = \frac{1}{|pair|} \sum_{(i,j)\in pair} \left(S^{g2p}_{ij} - 1\right)^2 \tag{9}$$

where $pair = \{(i,j) \mid G2P_{ij} = 1\}$. This loss function is selectively applied to known location proximal gene-peak pairs, concentrating on the potential gene-peak relationships inference. By exploiting this constraint, FactVAE incorporates relationships supported by location proximal, which indicates cis-regulatory interactions. This not only improves the model's interpretability but also directs the discovery of potential regulatory relationships.

Similarity structure alignment constraint

In single-cell multi-omics data, cells across different omics layers share similar distributions and are coupled [17]. Additionally, models trained on RNA data generally outperform those trained on ATAC data in representing cells [22]. By treating RNA-VAE as a teacher model and ATAC-VAE as a student model, FactVAE leverages the knowledge captured by RNA-VAE to enhance the ATAC-VAE's cell representation ability and augment the data, addressing the challenges caused by data sparsity in downstream analysis.

FactVAE employs the SSA constraint to fine-tune the encoder in ATAC-VAE by aligning the similarity structures of cell embeddings between RNA and ATAC data. These similarity structures, represented by matrices S^r and S^a, reflect inter-cell distribution patterns. They are calculated from RNA and ATAC cell embeddings, z^r and z^a, respectively:

$$S^r = \frac{z^r \cdot z^r}{\|z^r\| \cdot \|z^r\|},\ S^a = \frac{z^a \cdot z^a}{\|z^a\| \cdot \|z^a\|}. \tag{10}$$

The SSA loss function aligns these structures by maximizing their cosine similarity:

$$L_{ssa} = \frac{1}{N} \sum_i 1 - \frac{s^r_i \cdot s^a_i}{\|s^r_i\| \cdot \|s^a_i\|}. \tag{11}$$

Through the application of SSA constraint, FactVAE allows the encoder in ATAC-VAE to assimilate meaningful cell context information inferred from RNA-VAE, enriching its representation capability. By leveraging the enriched representation, FactVAE mitigates the data inherent sparsity in ATAC data and provides more informative and comprehensive information for downstream analysis. High-quality augmented data are reconstructed from the decoder:

$$\hat{x^a} = e^{sf} \cdot \mu^a \tag{12}$$

where sf represents the size factor of each cell in X^a. This augmented data supports more accurate cell clustering and enhances the detection of differential motifs in downstream tasks.

Training strategy

FactVAE's training process consists of two iterative phases, each with a specific objective to ensure the effective co-optimization of RNA-VAE and ATAC-VAE while emphasizing the independent optimization of ATAC-VAE's encoder. The Adam optimizer trains the model in each phase. This two-phase strategy not only captures cell and feature information but also mitigates the inherent sparsity of ATAC data.

The former phase aims at the co-optimization of RNA-VAE and ATAC-VAE to extract cell and feature information from their respective omics data, under the guidance of known gene-peak regulatory relationships. The objective function for this co-optimization phase is:

$$L_{joint} = L_{rna} + L_{atac} + \alpha e^{L_{reg}} \tag{13}$$

Table 1. ARI scores for various methods on six single-cell multi-omics datasets

	Method	hCellLine	hPBMC3k	mBrainSc	mKidney	mCortex	hPBMC10k
Single-omic scRNA	scDCCA	0.751(±0.143)	0.506(±0.053)	0.514(±0.057)	0.408(±0.046)	0.357(±0.036)	0.455(±0.070)
	scGNN	0.372(±0.006)	**0.627(±0.015)**	0.212(±0.048)	0.177(±0.007)	0.085(±0.020)	0.501(±0.033)
	scziDesk	0.923(±0.012)	0.590(±0.025)	0.543(±0.042)	0.391(±0.022)	0.355(±0.020)	0.469(±0.050)
	scVAEBGM	0.801(±0.032)	0.542(±0.026)	0.533(±0.035)	0.366(±0.004)	0.304(±0.013)	0.477(±0.036)
Single-omic scATAC	scDCCA	0.064(±0.047)	0.329(±0.031)	0.029(±0.010)	0.033(±0.009)	0.047(±0.017)	0.431(±0.030)
	scGNN	-0.012(±0.007)	0.231(±0.002)	0.020(±0.003)	0.008(±0.000)	0.001(±0.000)	0.215(±0.010)
	scziDesk	0.007(±0.002)	0.260(±0.028)	0.065(±0.011)	0.012(±0.002)	0.091(±0.007)	0.483(±0.039)
	scVAEBGM	0.423(±0.022)	0.380(±0.014)	0.000(±0.001)	0.000(±0.000)	0.007(±0.016)	0.453(±0.008)
Multi-omics	MOFA+	0.559(±0.000)	0.319(±0.004)	0.494(±0.032)	0.002(±0.000)	0.31(±0.011)	–
	seuratV3	0.623(±0.024)	0.437(±0.031)	0.134(±0.008)	0.267(±0.030)	0.300(±0.012)	0.535(±0.014)
	Unicom	0.916(±0.009)	0.477(±0.035)	0.521(±0.044)	0.414(±0.028)	0.388(±0.032)	0.444(±0.063)
	scMVAE	0.870(±0.070)	0.384(±0.041)	0.405(±0.070)	0.368(±0.041)	0.265(±0.020)	0.377(±0.043)
	DCCA	0.728(±0.037)	0.262(±0.050)	0.191(±0.024)	0.086(±0.027)	0.322(±0.030)	0.469(±0.017)
	scMM	0.883(±0.024)	0.379(±0.060)	0.385(±0.008)	0.301(±0.029)	0.231(±0.018)	0.414(±0.052)
	MatchCLOT	0.812(±0.020)	0.453(±0.006)	0.395(±0.014)	**0.501(±0.026)**	0.421(±0.009)	0.428(±0.012)
	scGLUE	0.859(±0.106)	0.539(±0.039)	0.516(±0.016)	–	0.423(±0.019)	0.450(±0.013)
	FactVAE	**0.942(±0.005)**	0.593(±0.013)	**0.613(±0.041)**	0.452(±0.020)	**0.435(±0.014)**	**0.539(±0.040)**

where L_{rna} and L_{atac} denote loss terms for RNA-VAE and ATAC-VAE, respectively. The term $\alpha e^{L_{reg}}$ accounts for regulatory information loss, with the hyper-parameter α defaulting to 1.0. This phase ensures that both VAEs are trained to learn relevant cell and feature representations while following known relationships.

The latter phase focuses on the independent fine-tuning of ATAC-VAE's encoder, encouraging it to refine its cell representations by leveraging knowledge transferred from the teacher model, RNA-VAE, through the SSA constraint. So, the objective function for this phase is L_{ssa}. The targeted focus on the ATAC-VAE's encoder contributes to mitigating the influence of data sparsity and enhances its ability to accurately represent cell states.

Upon completion of the training process, integrated cell embeddings are constructed by concatenating the embeddings from both models:

$$z = concat(z^a, z^r). \tag{14}$$

This integrated representation encapsulates a more comprehensive view of cell states by combining information from both omics layers. The ablation study and parameter analysis of FactVAE can be found in Supplementary S2.

Results

Benchmark methods and evaluation metrics

We compare FactVAE against four single-omic data analysis methods and eight well-established multi-omics data analysis methods. Single-omic methods, including scDCCA [27], scGNN [28], scziDesk [29], and scVAEBGM [30], applied to individual omic data. Multi-omics methods comprise factorization-based methods like MOFA+ [31] and SeuratV3 [8], alignment-based methods like UnionCom [11], and NN-based methods like scMVAE [15], DCCA [22], scMM [16], MatchCLOT [19], and scGLUE [17].

All methods are executed on the same NVIDIA GeForce RTX 3090 server, using default parameters as specified in their respective papers or GitHub repositories. Both SeuratV3 and scMVAE require the gene activity matrix. We follow the transformation strategy utilized by scGLUE to obtain the gene activity matrix from the peak matrix. Specifically, if a peak overlaps with the genome or the proximal promoter region (defined as 2 kb upstream of the transcription start site (TSS)), the peak is linked to the gene. The counts of peaks associated with specific genes are aggregated to create gene activity scores.

The quality of cell embeddings generated by different models reflects their ability to extract meaningful cell information. Clustering, a vital downstream analysis, assesses these methods' ability to accurately represent cells [32]. The clustering performance is quantified by adjusted rand index (ARI) and normalized mutual information (NMI), where $\mathbf{ARI} \in [-1, 1]$ and $\mathbf{NMI} \in [0, 1]$. Higher values for these metrics indicate better clustering performance. The performance of the inferred gene-peak regulatory relationships is assessed by analysing the receiver operating characteristic (ROC) curve. The area under the ROC curve (AUROC) is calculated using the sklearn package [33] for quantitative assessment.

Clustering performance of FactVAE

We apply each method combined with K-means on six single-cell multi-omics datasets, with k set to the known number of cell types. Each method is executed five times, and the mean and standard deviation of the results are reported. As shown in Tables 1 and 2, clustering results for scRNA-seq data outperform those for scATAC-seq data when analysed individually, which may be attributed to the reliability of the scRNA-seq data. In contrast, integrated multi-omics analyses yield more precise clustering results. FactVAE achieves the best performance across most datasets, with the exception of the mKidney dataset, where it ranks slightly behind MatchCLOT (with differences of 0.051 in ARI and 0.027 in NMI). Notably, some benchmark methods encounter specific challenges in this experiment: scGLUE fails to process the mKidney dataset due to the presence of decimal values, MOFA+ encounters out-of-memory issues when handling the hPBMC10k dataset. Among NN-based multi-omics methods, both MatchCLOT and FactVAE exhibit low variance across all datasets, highlighting robustness and stable performance. We visualize cell embeddings by UMAP and quantify them by silhouette scores based on ground truth. FactVAE exhibits a distinct structure with concentrated cells of the same type and dispersed cells of different types, indicating that FactVAE outperforms in generating clear and interpretable visualizations (see Supplementary Fig. S1-S6).

Table 2. NMI scores for various methods on six single-cell multi-omics datasets

	Method	hCellLine	hPBMC3k	mBrainSc	mKidney	mCortex	hPBMC10k
Single-omic scRNA	scDCCA	0.715(±0.128)	0.675(±0.010)	0.655(±0.025)	0.554(±0.023)	0.550(±0.013)	0.685(±0.025)
	scGNN	0.438(±0.005)	0.626(±0.006)	0.491(±0.027)	0.338(±0.008)	0.262(±0.024)	0.665(±0.007)
	scziDesk	0.884(±0.013)	0.713(±0.009)	0.677(±0.022)	0.542(±0.014)	0.571(±0.013)	0.689(±0.016)
	scVAEBGM	0.759(±0.030)	0.724(±0.019)	0.687(±0.015)	0.542(±0.010)	0.548(±0.009)	0.709(±0.012)
Single-omic scATAC	scDCCA	0.072(±0.035)	0.493(±0.019)	0.089(±0.013)	0.029(±0.003)	0.125(±0.022)	0.620(±0.010)
	scGNN	0.042(±0.005)	0.382(±0.008)	0.05(±0.002)	0.010(±0.000)	0.016(±0.000)	0.428(±0.011)
	scziDesk	0.069(±0.007)	0.428(±0.017)	0.164(±0.018)	0.024(±0.003)	0.212(±0.009)	0.637(±0.017)
	scVAEBGM	0.371(±0.014)	0.564(±0.005)	0.017(±0.002)	0.004(±0.000)	0.037(±0.061)	0.687(±0.005)
Multi-omics	MOFA+	0.650(±0.000)	0.561(±0.002)	0.640(±0.012)	0.009(±0.000)	0.498(±0.005)	–
	seuratV3	0.524(±0.017)	0.602(±0.015)	0.333(±0.015)	0.381(±0.032)	0.525(±0.008)	0.734(±0.005)
	Unicom	0.863(±0.012)	0.709(±0.021)	0.659(±0.020)	0.551(±0.021)	0.629(±0.014)	0.690(±0.027)
	scMVAE	0.844(±0.046)	0.604(±0.025)	0.577(±0.050)	0.510(±0.031)	0.503(±0.021)	0.630(±0.021)
	DCCA	0.688(±0.048)	0.436(±0.049)	0.375(±0.028)	0.180(±0.036)	0.551(±0.038)	0.673(±0.015)
	scMM	0.823(±0.032)	0.590(±0.008)	0.588(±0.021)	0.439(±0.032)	0.379(±0.015)	0.628(±0.026)
	MatchCLOT	0.744(±0.022)	0.687(±0.006)	0.589(±0.005)	**0.632(±0.012)**	0.667(±0.004)	0.704(±0.006)
	scGLUE	0.817(±0.089)	0.708(±0.014)	0.634(±0.020)	–	0.656(±0.006)	0.691(±0.007)
	FactVAE	**0.905(±0.007)**	**0.758(±0.005)**	**0.715(±0.013)**	0.606(±0.007)	**0.679(±0.010)**	**0.761(±0.014)**

These findings underscore FactVAE's outstanding capability to export reliable cell embeddings, facilitating high-accuracy clustering.

Augmentation performance of ATAC-VAE in FactVAE

FactVAE leverages insights from RNA-VAE to guide the training of ATAC-VAE's encoder, thereby optimizing cell embeddings derived from ATAC data. Additionally, the ATAC-VAE's decoder, designed with a ZINB distribution, enables data augmentation, enhancing biologically meaningful expression patterns of transcription factor (TF) motifs.

To assess the impact of the SSA constraint on ATAC-VAE, we compare the clustering results of cell embeddings from ATAC data extracted by ATAC-VAE with and without this constraint, evaluating the significance of the differences by T-test. As depicted in Fig. 2(A), incorporating SSA consistently improves the clustering performance across all datasets, with the most pronounced enhancement observed in the hCellLine dataset (ARI increases by 0.06 and NMI by 0.08). T-test results confirm that most improvements are statistically significant ($p \leq 0.05$).

To further assess FactVAE's augmentation capability, we benchmark it against three other augmentation-capable methods: scMVAE, DAAC, and scMM. ChromVAR [34] determines the chromatin accessibility deviations for each cell, and the Wilcoxon test identifies cell-type-specific TF motifs ($p<0.05$). We report the overlap of cell-type-specific TF motifs and the log2 fold change (log2FC) values of these shared TF motifs in the raw and augmented data generated from each method on three datasets with varying cell counts: hCellLine, hPBMC3k, and hPBMC10k. As shown in Fig. 2(B), the top panel demonstrates the overlap counts, while the bottom scatter plots compare log2FC values of shared motifs in raw (X-axis) with augmented (Y-axis) data. FactVAE outperforms other methods in preserving cell-type-specific TF motifs from raw data and exhibits higher log2FC values post-augmentation. This indicates that FactVAE effectively retains and enhances the cell-type-specific expression patterns inherent in raw data.

Collectively, these results demonstrate that FactVAE effectively integrates RNA-VAE knowledge through the SSA constraint to guide ATAC-VAE training, enhancing its ability to extract information and data augmentation for ATAC data, thereby alleviating challenges posed by data sparsity in downstream analyses, like differential motif detection.

Regulatory inference performance of FactVAE

FactVAE and scGLUE are advanced multi-omics data integration analysis methods that generate embeddings for gene-peak regulatory inference. While scGLUE employs a graph neural network (GNN) framework to reconstruct known gene-peak relationships directly, FactVAE adopts a more flexible approach by treating these relationships as constraints rather than reconstruction targets, facilitating potential relationship discovery while maintaining biological validity. For regulatory inference from feature embeddings derived by FactVAE's decoders, we apply a strategy consistent with scGLUE: randomly shuffled gene-peak pairs are considered as background, and the significance of the target pairs is quantified by cosine similarity comparison with the background distribution.

To validate the inferred regulatory relationships, we refer to two external databases: promoter capture Hi-C (pcHi-C) interaction maps for 17 human hematopoietic cell types [35] and expression quantitative trait loci (eQTL) identified by the Genotype-Tissue Expression (GTEx) project [36]. As shown in Fig. 3(A) and (B), FactVAE achieves superior performance in both the area under the receiver operating characteristic curve (AUROC) and the average precision (AP) score, outperforming commonly used methods such as Spearman correlation, LASSO, Cicero [37], and scGLUE.

FactVAE successfully identifies reliable long-range regulatory relationships, despite relying solely on gene-peak overlap relationships. As illustrated in Fig. 3(C), FactVAE accurately infers regulatory peaks for TESK2 ~130 kb upstream and ~235 kb downstream of TSS. TESK2, as a major kinase of cofilin, plays a critical role in the regulation of the actin cytoskeleton [38]. These predictions are not only well supported by the eQTL and pcHi-C databases, but also align with the potential regulator CTCF, as predicted by hTFtarget [39]. CTCF is a key regulator of chromatin conformation and has been studied for its association with the actin cytoskeleton [40]. Similarly, the inferred regulatory peaks for PHACTR4 are confirmed. POLR2A, the largest subunit of RNA polymerase II, directly contributes to transcription initiation and elongation. PHACTR4 restricts cancer cell proliferation and tumor

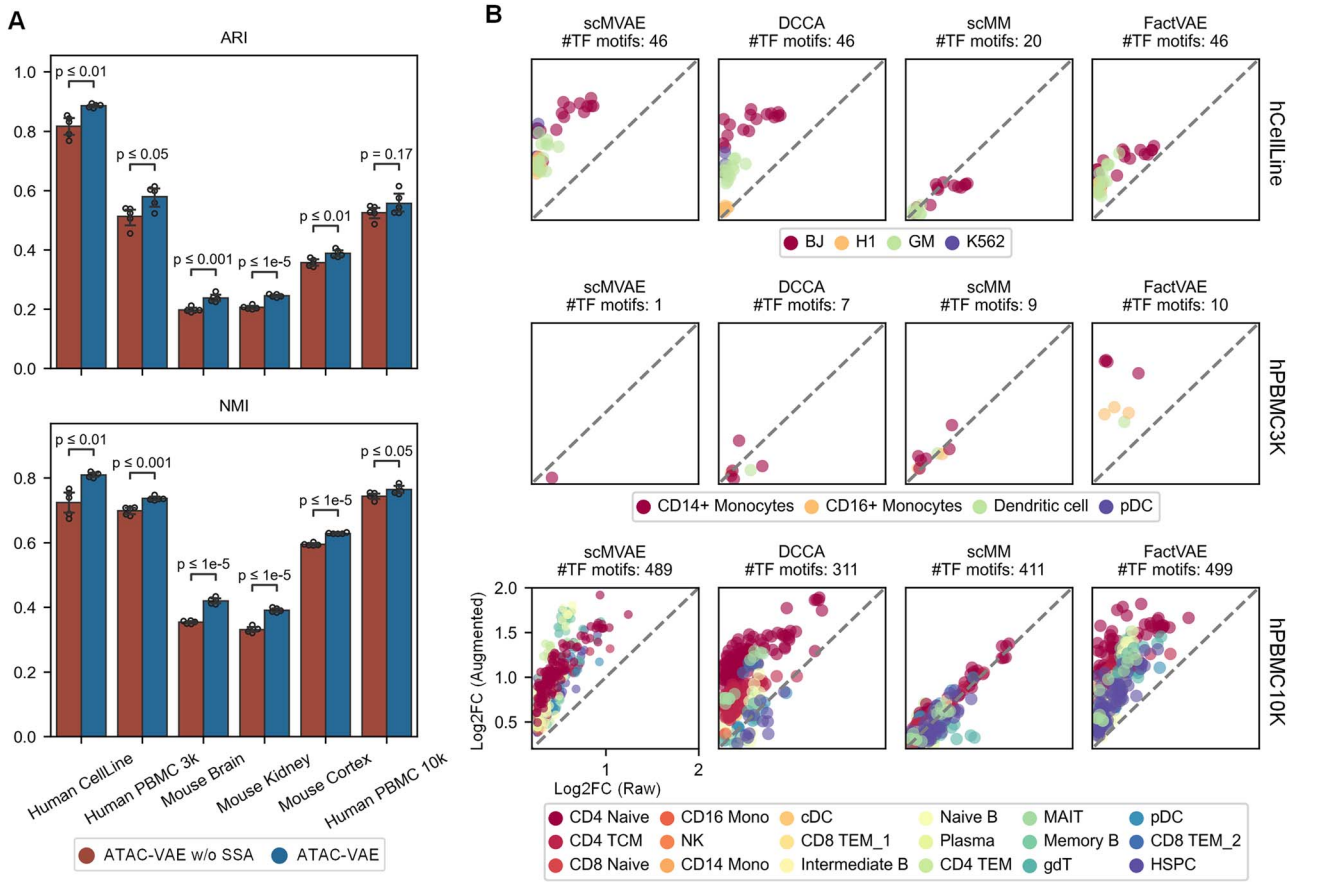

Figure 2. The SSA constraint strengthens the information extraction and data augmentation capabilities of ATAC-VAE. (A) Impact of SSA constraint on clustering performance of ATAC-VAE across six single-cell multi-omics datasets. (B) Comparison of cell type-specific TF motifs before and after augmentation by different methods on three single-cell multi-omics datasets.

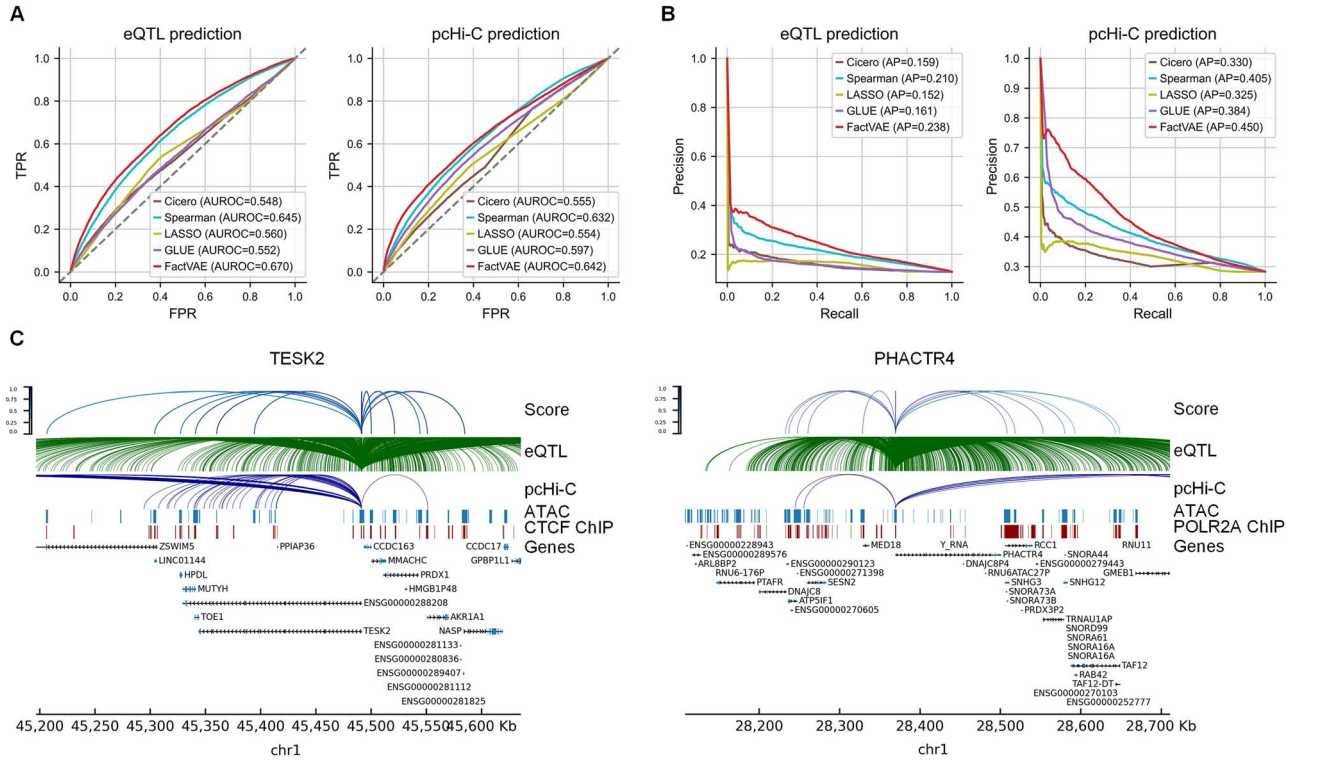

Figure 3. Comparison of FactVAE with benchmark methods in gene regulation inference performance. (A) Comparison of ROC curves for gene regulatory scores predicted by different methods using eQLT and pcHi-C datasets as ground truth. (B) Comparison of PR curves for gene regulatory scores inferred by different methods using eQLT and pcHi-C datasets as ground truth. (C) Highly significant regulatory relationships inferred by FactVAE for PHACTR4 and TESK2 genes.

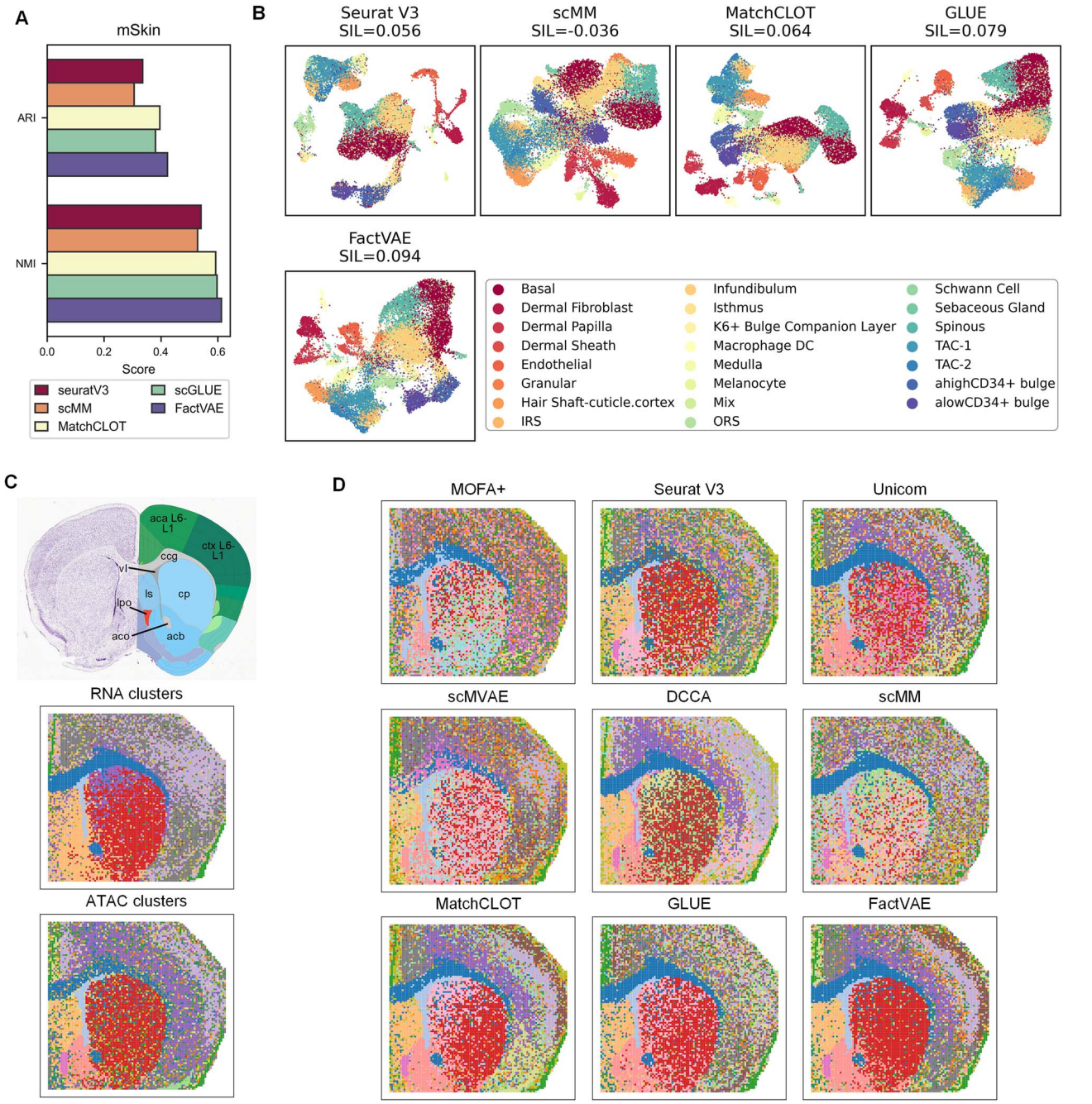

Figure 4. Clustering performance on the large-scale single-cell dataset (mskin) and spatial dataset (mBrainSp). (A) A barplot of ARI and NMI of various multi-omics methods on the mSkin dataset. (B) The UMAP visualizations of cell embeddings result from various multi-omics methods on the mSkin dataset. (C) Reference annotations for the coronal section of the mouse brain from the Allen Mouse Brain Atlas and the original analysis by Zhang *et al.* (D) Domain detection performance of various multi-omics methods on the mBrainSp dataset.

growth [41], and precise targeting POLR2A has been considered as a therapeutic strategy for triple-negative breast cancer (TNBC) [42], indicating the inferred regulatory relationships may be significant in tumorigenesis understanding and cancer therapeutic target discovery. These findings consistently reflect the reliability of FactVAE in discovering biologically meaningful regulatory relationships.

Scalability of FactVAE on large-scale and spatial datasets

With the advancement of technology, performance on large-scale data has become increasingly important due to the rising production of single-cell multi-omics data. So we utilize the mSkin dataset consisting of 34,774 cells to assess the performance of FactVAE. Due to resource limitations, such as insufficient memory or GPU capacity, MOFA+, UnionCom, scMVAE, and DCCA are excluded from this analysis as these methods could not handle this dataset on the same server. K-means is applied to cluster the cell embeddings derived from each method. As depicted in Fig. 4(A) and (B), FactVAE outperforms benchmark methods in both ARI and NMI scores. Moreover, its highest silhouette score indicates superior visualization quality.

Beyond the rapid increase in throughput of single-cell multi-omics sequencing technologies, spatial multi-omics technologies

have received attention for preserving spatial context. To evaluate the performance of FactVAE on spatial multi-omics data, we utilize the mBrainSp dataset, which is derived from the coronal section of a mouse brain on the 22nd day after birth. All methods are applied to generate spot embeddings, followed by spatial domain detection using Leiden [43]. The original independent analysis of the two omics layers by Zhang *et al.* [44] is presented in Fig. 4(C). The results indicate that both omics effectively identify the lateral ventricle (vl) and anterior commissure olfactory limb (aco) regions. RNA data can clearly identify the lateral ventricle (vl) and anterior commissure olfactory limb (aco) regions, with less clarity in the cortex (ctx) layers. Conversely, ATAC data distinguish the caudoputamen (cp) and parts of the ctx layers but have limited delineation of the corpus callosum (ccg). When integrating both data types (Fig. 4(D)), most methods effectively detect the vl and aco regions. Compared to benchmark methods, FactVAE uniquely identifies critical regions such as the lateral septal nucleus (ls), lateral preoptic area (lpo), nucleus accumbens (acb) and ccg regions, while revealing distinct hierarchical structures in the aca and ctx regions (L6 in purple, L5 in yellow, L4 in light purple, L2/3 in brown, and L1 in green).

In summary, FactVAE FactVAE efficiently processes large-scale single-cell data and accurately separates spatial domains, advancing biological research by enabling precise cell identification and localization.

Discussion and conclusion

The proposed FactVAE demonstrates significant advantages in the integration analysis of single-cell multi-omics data. By incorporating the factorization principle into VAE, FactVAE effectively captures both cell and feature information, facilitating a comprehensive multi-omics data integration analysis. The outstanding performance on six single-cell multi-omics datasets evidences its robust capability for extracting cell information. Leveraging known regulatory information, FactVAE ensures the reliability of the extracted feature information, as verified by its superior performance in regulatory relationship inference. Meanwhile, the SSA constraint optimizes the training process of ATAC-VAE, enhancing the model's ability to handle sparse ATAC data. The ablation study highlights the importance of each component, as the integration of regulatory knowledge and SSA constraints improves model performance. Furthermore, FactVAE's excellent performance on large-scale single-cell and spatial multi-omics datasets confirms its broad applicability.

Despite its strengths, FactVAE has room for further improvement. While FactVAE focuses on integrating ATAC and RNA data, the framework can be extended to other omics data, such as proteomics and epigenomics, offering the potential to construct a more comprehensive depiction of cellular states. While FactVAE achieves superior results on the spatial multi-omics dataset, it does not exploit specific spatial information. Moreover, the high cost of single-cell multi-omics technologies causes independent profiling of different omics layers in many studies. Future work should aim to incorporate more diverse datasets, address the sample mismatch issue in multi-omics data, and explore multi-omics integration within spatial contexts, providing more comprehensive solutions for future research.

In summary, FactVAE, as a novel factorized variational autoencoder, successfully addresses the challenges of insufficient cross-modal knowledge utilization and missing feature information by integrating factorization principles and leveraging known regulatory information. It provides a powerful tool for the in-depth resolution of cellular states and functions from single-cell multi-omics data, with promising advances in precision medicine and beyond.

Key Points

- Innovatively enhancing VAE with factorization to simultaneously capture nonlinear sample and feature information, providing a new solution for multi-omics data analysis.
- Leveraging known regulatory relationships to guide model training, enabling the inference of potential regulatory relationships while ensuring reliable feature information captured by the model.
- Employing a knowledge transfer strategy during training to guide the model in capturing more comprehensive characteristics of ATAC data and augmenting the data to effectively alleviate the inherent sparsity of ATAC data burdening the downstream tasks.
- Extensive experiments validate that the proposed method outperforms well-established benchmark methods and exhibits high scalability in large-scale single-cell and spatial multi-omics data.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work is supported by the research project funded by National key research and development program, China (2023YFC2413500), and the National Frontiers Science Center for Industrial Intelligence and Systems Optimization the 111 Project (B16009).

Data and software availability

All of the datasets analysed in this manuscript are publicly available. Please see the section 2.1 for details. The source code could be accessed and downloaded from https://github.com/WangDaMiao97/FactVAE.

References

1. Baysoy A, Bai Z, Satija R. *et al.* The technological landscape and applications of single-cell multi-omics. *Nat Rev Mol Cell Biol* 2023;**24**:695–713. https://doi.org/10.1038/s41580-023-00615-w
2. 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. https://doi.org/10.1038/s41587-019-0290-0
3. Cao J, Cusanovich DA, Ramani V. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* 2018;**361**:1380–5.

4. Ma S, Zhang B, LaFave LM. *et al.* Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell* 2020;**183**:1103–1116.e20. https://doi.org/10.1016/j.cell.2020.09.056
5. Lee J, Hyeon DY, Hwang D. Single-cell multiomics: technologies and data analysis methods. *Exp Mol Med* 2020;**52**:1428–42. https://doi.org/10.1038/s12276-020-0420-2
6. Stanojevic S, Li Y, Ristivojevic A. *et al.* Computational methods for single-cell multi-omics integration and alignment. *Genom Proteom Bioinform* 2022;**20**:836–49. https://doi.org/10.1016/j.gpb.2022.11.013
7. Liu J, Gao C, Sodicoff J. *et al.* Jointly defining cell types from multiple single-cell datasets using liger. *Nat Protoc* 2020;**15**:3632–62. https://doi.org/10.1038/s41596-020-0391-8
8. Stuart T, Butler A, Hoffman P. *et al.* Comprehensive integration of single-cell data cell. 2019;**177**:1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031
9. Kuchroo M, Godavarthi A, Tong A. *et al.* Multimodal data visualization and denoising with integrated diffusion. In: *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. Gold Coast, Australia: IEEE, pp. 1–6, 2021. https://doi.org/10.1109/MLSP52302.2021.9596214
10. Kim HJ, Lin Y, Geddes TA. *et al.* Citefuse enables multi-modal analysis of cite-seq data. *Bioinformatics* 2020;**36**:4137–43. https://doi.org/10.1093/bioinformatics/btaa282
11. Cao K, Bai X, Hong Y. *et al.* Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics* 2020;**36**:i48–56.
12. Ren L, Wang J, Li Z. *et al.* Scmcs: a framework for single-cell multi-omics data integration and multiple clusterings. *Bioinformatics* 2023;**39**:btad133.
13. Squair JW, Gautier M, Kathe C. *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* 2021;**12**:5692. https://doi.org/10.1038/s41467-021-25960-2
14. Wu KE, Yost KE, Chang HY. *et al.* Babel enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc Natl Acad Sci* 2021;**118**:e2023070118. https://doi.org/10.1073/pnas.2023070118
15. Zuo C, Chen L. Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data. *Brief Bioinform* 2021;**22**:bbaa287. https://doi.org/10.1093/bib/bbaa287
16. Minoura K, Abe K, Nam H. *et al.* A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell Rep Methods* 2021;**1**:100071.
17. Cao Z-J, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol* 2022;**40**:1458–66. https://doi.org/10.1038/s41587-022-01284-4
18. Stark SG, Ficek J, Locatello F. *et al.* Scim: universal single-cell matching with unpaired feature sets. *Bioinformatics* 2020;**36**:i919–27.
19. Gossi F, Pati P, Chouvardas P. *et al.* Matching single cells across modalities with contrastive learning and optimal transport. *Brief Bioinform* 2023;**24**:bbad130. https://doi.org/10.1093/bib/bbad130
20. Zhao J, Zhao B, Song X. *et al.* Subtype-dcc: decoupled contrastive clustering method for cancer subtype identification based on multi-omics data. *Brief Bioinform* 2023;**24**:bbad025.
21. Zhao B, Song K, Wei D-Q. *et al.* Sccobra allows contrastive cell embedding learning with domain adaptation for single cell data integration and harmonization. *Commun Biol* 2025;**8**:233.
22. Zuo C, Dai H, Chen L. Deep cross-omics cycle attention model for joint analysis of single-cell multi-omics data. *Bioinformatics* 2021;**37**:4091–9. https://doi.org/10.1093/bioinformatics/btab403
23. Wang J-C, Chen Y-J, Zou Q. Grace: unveiling gene regulatory networks with causal mechanistic graph neural networks in single-cell rna-sequencing data. *IEEE Trans Neural Netw Learn Syst* 2024. https://doi.org/10.1109/TNNLS.2024.3412753
24. Chen H, Lareau C, Andreani T. *et al.* Assessment of computational methods for the analysis of single-cell atac-seq data. *Genome Biol* 2019;**20**:1–25. https://doi.org/10.1186/s13059-019-1854-5
25. Dayu H, Liang K, Dong Z. *et al.* Effective multi-modal clustering method via skip aggregation network for parallel scrna-seq and scatac-seq data. *Brief Bioinform* 2024;**25**:bbae102.
26. 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. https://doi.org/10.1038/s41467-022-35031-9
27. 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**:bbac625.
28. Wang J, Ma A, Chang Y. *et al.* Scgnn is a novel graph neural network framework for single-cell rna-seq analyses. *Nat Commun* 2021;**12**:1–11.
29. Chen L, Wang W, Zhai Y. *et al.* Deep soft k-means clustering with self-training for single-cell rna sequence data. *NAR Genom Bioinform* 2020;**2**:lqaa039.
30. Duan H, Li F, Shang J. *et al.* Scvaebgm: clustering analysis of single-cell atac-seq data using a deep generative model. *Interdiscip Sci Comput Life Sci* 2022;**14**:917–28. https://doi.org/10.1007/s12539-022-00536-w
31. 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.
32. Xie Y, Wang X, Wang P. *et al.* A pseudo-label supervised graph fusion attention network for drug–target interaction prediction. *Exp Syst Appl* 2025;**259**:125264. https://doi.org/10.1016/j.eswa.2024.125264
33. Pedregosa F, Varoquaux G, Gramfort A. *et al.* Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.
34. Schep AN, Beijing W, Buenrostro JD. *et al.* Chromvar: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* 2017;**14**:975–8. https://doi.org/10.1038/nmeth.4401
35. Javierre BM, Burren OS, Wilder SP. *et al.* Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* 2016;**167**:1369–1384.e19. https://doi.org/10.1016/j.cell.2016.09.037
36. Lonsdale J, Thomas J, Salvatore M. *et al.* The genotype-tissue expression (gtex) project. *Nat Genet* 2013;**45**:580–5. https://doi.org/10.1038/ng.2653
37. Pliner HA, Packer JS, McFaline-Figueroa JL. *et al.* Cicero predicts cis-regulatory dna interactions from single-cell chromatin accessibility data. *Mol Cell* 2018;**71**:858–871.e8. https://doi.org/10.1016/j.molcel.2018.06.044
38. Ohashi K. Roles of cofilin in development and its mechanisms of regulation. *Dev Growth Differ* 2015;**57**:275–90.
39. Zhang Q, Liu W, Zhang H-M. *et al.* Htftarget: a comprehensive database for regulations of human transcription factors and their targets. *Genom Proteom Bioinform* 2020;**18**:120–8.
40. Lin L, Dong J, Shun X. *et al.* Autoinhibition and relief mechanisms for mical monooxygenases in f-actin disassembly. *Nat Commun* 2024;**15**:6824. https://doi.org/10.1038/s41467-024-50940-7

41. Solimini NL, Liang AC, Xu C. *et al.* Stop gene phactr4 is a tumor suppressor. *Proc Natl Acad Sci* 2013;**110**:E407–14.
42. Jiangsheng X, Liu Y, Li Y. *et al.* Precise targeting of polr2a as a therapeutic strategy for human triple negative breast cancer. *Nat Nanotechnol* 2019;**14**:388–97.
43. Traag VA, Waltman L, Van Eck NJ. From louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 2019;**9**:1–12.
44. Zhang D, Deng Y, Kukanja P. *et al.* Spatial epigenome–transcriptome co-profiling of mammalian tissues. *Nature* 2023;**616**:113–22. https://doi.org/10.1038/s41586-023-05795-1

Briefings in Bioinformatics, 2025, **26(2)**, bbaf157
https://doi.org/10.1093/bib/bbaf157
Problem Solving Protocol