JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

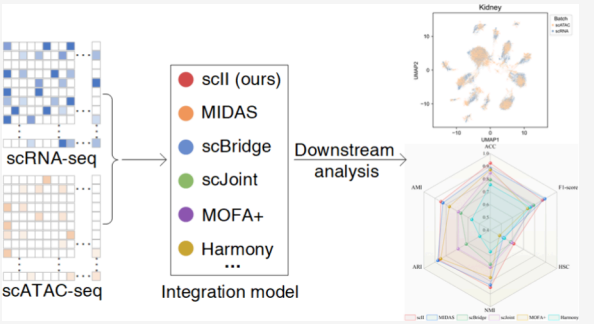
pubs.acs.org/jcim
Article

scII: Dual-Threshold Adaptive Integration of Single-Cell Multiomics Data Driven by Imputation

Yi Zhang, Yuru Li,* Zhicheng Jin, Ye Tian, and Chen Su

Cite This: J. Chem. Inf. Model. 2026, 66, 1445−1456

ABSTRACT: Single-cell multiomics technologies provide unprecedented opportunities to dissect cellular heterogeneity by capturing multidimensional information on complex cellular states and regulatory networks. However, challenges such as high dimensionality, extreme data sparsity, and modality-specific discrepancies hinder the accuracy, interpretability, and scalability of the existing integration methods. Existing integration paradigms, including horizontal, vertical, and diagonal strategies, are further limited by their inability to fully capture nonlinear biological relationships, their reliance on high-quality data, and their substantial computational demands. Here, we present scII (Dual-Threshold Adaptive Integration of Single-Cell Multiomics Data Driven by Imputation), an adaptive framework designed to integrate gene expression (scRNA-seq) and chromatin accessibility (scATAC-seq) data. Our approach is built on several key conceptual innovations: (i) scRNA-seq−guided signal imputation to enhance information integrity in scATAC-seq; (ii) a multilayer perceptron with the Maxout activation function to improve the modeling of complex nonlinear relationships and mitigate the vanishing gradient problem; (iii) a dynamic dual-threshold adaptive selection mechanism that jointly evaluates cross-modality feature similarity and classification reliability to select high-quality cells; and (iv) Bayesian Information Criterion (BIC)-based optimization to dynamically determine the number of Gaussian Mixture Model components according to data distribution, thereby eliminating reliance on manually preset parameters. Extensive experiments on multiple real-world and simulated data sets demonstrate that scII not only enables efficient integration of unpaired scRNA-seq and scATAC-seq data but also achieves accurate transfer of cell-type annotations, allowing high-precision cell-type prediction for scATAC-seq.



1. INTRODUCTION

Single-cell multiomics data integration aims to jointly analyze multiple layers of molecular information derived from the same cell or a cell population, encompassing the genome,[1] transcriptome,[2] epigenome,[3] proteome,[4] as well as data generated by various hybrid-modality techniques (e.g., scMT-seq,[5] 10× Multiome). Compared with conventional single-omics analysis, multiomics integration enables the elucidation of cross-layer regulatory interactions, thereby providing a more comprehensive characterization of cellular states, resolving functional heterogeneity, and revealing the underlying dynamic regulatory mechanisms. In recent years, advances in technologies such as SHARE-seq,[6] SNARE-seq,[7] and Paired-seq[8] have made it possible to simultaneously profile gene expression and chromatin accessibility within the same cell. Meanwhile, CITE-seq[9] and REAP-seq[10] allow the joint measurement of RNA expression and surface protein abundance. These technological developments provide researchers with powerful tools to integrate multiple layers of genomic information and to gain a more comprehensive understanding of how different cellular mechanisms interact.[11,12] Nevertheless, single-cell multiomics data are inherently challenged by high dimensionality, extreme sparsity, and substantial heterogeneity,[13,14] which significantly increase the computational complexity and analytical difficulty of data integration. To systematically address these challenges, existing studies generally categorize integration methods into three main classes based on their data pairing relationships and shared structures[15]:

- **Horizontal integration.** Exemplified by methods such as MOFA,[16] this strategy employs linear factor models to project multiple omics data sets into a shared latent space. It is computationally efficient and offers high interpretability; however, its effectiveness depends strongly on the number of shared features across modalities. Consequently, its utility is constrained by missing features and sparse signals, and it often fails to capture complex nonlinear biological relationships.

Received: October 5, 2025
Revised: December 23, 2025
Accepted: December 23, 2025
Published: January 15, 2026

https://doi.org/10.1021/acs.jcim.5c02396
J. Chem. Inf. Model. 2026, 66, 1445−1456

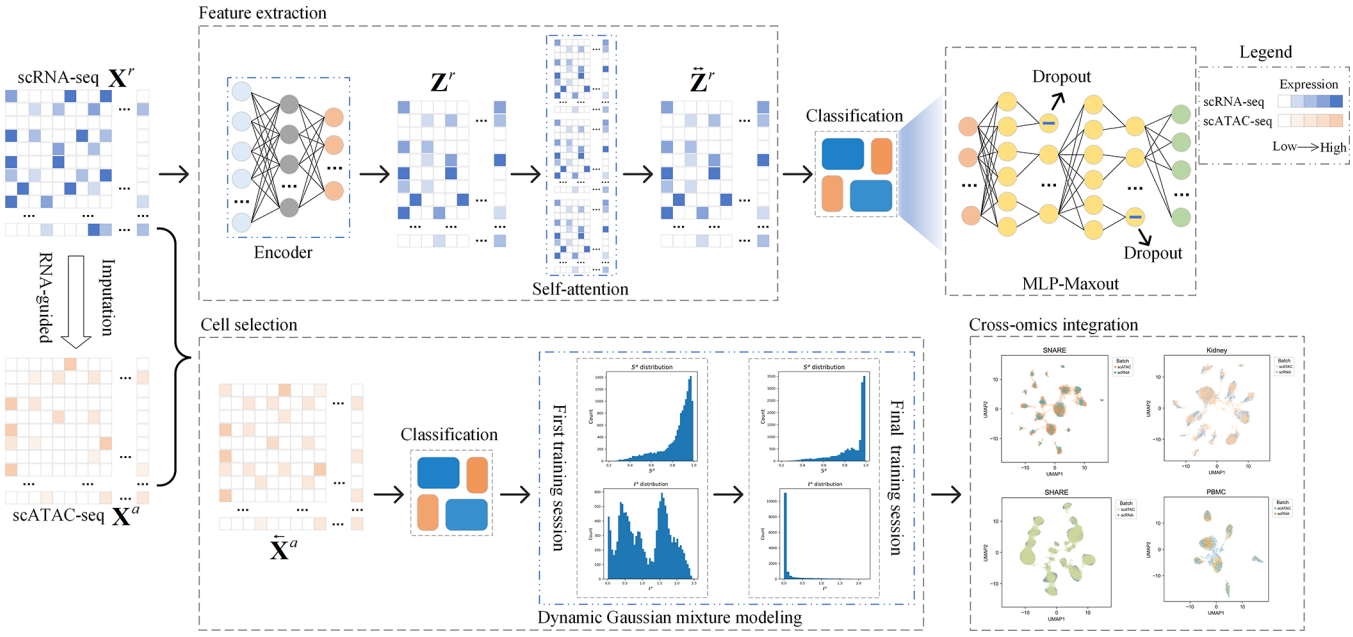


Figure 1. scII workflow. The proposed model incorporates data imputation, a Maxout-activated MLP classifier, and a dynamic dual-threshold selection mechanism to achieve precise cross-modality cell type annotation transfer and integration. Here, $\mathbf{X}^r$ denotes the scRNA-seq expression matrix, and $\mathbf{X}^a$ denotes the scATAC-seq gene activity matrix.

- **Vertical integration.** Designed for paired multiomics data from the same cell (e.g., scRNA-seq and scATAC-seq),[17−19] this approach leverages one-to-one cellular correspondence to model nonlinear cross-modal relationships and enable cross-modality prediction. Despite these advantages, vertical integration requires high-quality paired data, incurs substantial experimental costs, and demands significant computational resources, which limit its scalability to large data sets.
- **Diagonal integration.** Unlike the other two categories, this strategy does not rely on shared features or cell-level pairing. Instead, it employs dimensionality reduction or deep learning to construct a common latent space, thereby enabling integrative analysis across distinct omics modalities.[19−21] While highly flexible and suitable for unpaired data sets, diagonal integration remains challenged by modeling complexity, inconsistent evaluation standards, and reduced interpretability and robustness due to high-dimensional noise and data sparsity.

Despite recent advances in multiomics integration and regulatory network reconstruction, key limitations remain in nonlinear modeling, sparsity compensation, and adaptive parameter selection. These challenges are particularly pronounced in scATAC-seq, where the number of detected open chromatin regions is only about one-third of the genes typically captured by scRNA-seq, substantially limiting cell-type identification.[22,23] In contrast, scRNA-seq benefits from mature library construction platforms (e.g., 10× Genomics) and standardized annotation pipelines (e.g., CellTypist[24]), which provide high-quality reference atlases. This disparity highlights the potential of cross-omics annotation transfer, where well-annotated scRNA-seq data sets can guide scATAC-seq cell-type prediction in the absence of direct annotations, thereby enhancing classification accuracy and biological interpretability.

To address these challenges, we propose scII (Dual-Threshold Adaptive Integration of Single-Cell Multiomics Data Driven by Imputation), a robust framework for integrating unpaired single-cell multiomics data. scII introduces four key innovations:

1. RNA-guided imputation: Leveraging the high signal-to-noise ratio of scRNA-seq data to guide the imputation of highly sparse scATAC-seq data, thereby improving information integrity.
2. Nonlinear modeling: Employing a multilayer perceptron with Maxout activation to better capture nonlinear dependencies and mitigate vanishing gradient problems.
3. Dynamic dual-threshold selection: Jointly evaluating cross-omics feature similarity and classification reliability to identify high-quality cells for model training.
4. Adaptive clustering: Optimizing Gaussian Mixture Model (GMM) parameters using the Bayesian Information Criterion (BIC), enabling data-driven inference of clustering structures without manual parameter tuning.

By integrating these strategies, scII achieves effective alignment of scRNA-seq and scATAC-seq data and enables accurate transfer of cell-type annotations, providing a scalable and reliable solution for single-cell multiomics integration.

2. RESEARCH METHODS

2.1. Overview of scII

scII is an integrative framework consisting of four key modules: the RNA-guided imputation module, the feature extraction and classification module, the cell selection module, and the cross-omics integration module. A detailed schematic of the model's architecture is provided in Figure 1.

2.2. Data Preprocessing

2.2.1. Data Set Selection. To comprehensively evaluate the performance of our model scII in integrating scRNA-seq and scATAC-seq data, four publicly available single-cell multiomics data sets together with one adaptively simulated data set were employed for testing. The real-world data sets were carefully selected to cover diverse biological contexts, including various tissue types, species, and sequencing platforms, thereby ensuring broad biological relevance and methodo-

https://doi.org/10.1021/acs.jcim.5c02396
J. Chem. Inf. Model. 2026, 66, 1445−1456

Table 1. Overview of the Experimental Data sets

data set	cell number (scRNA-seq/scATAC-seq)	gene number	data source	species	modality-paired[a]
PBMC	4644/4502	17441	GSE156478	human	no
SNARE	9134/9134	16750	GSE126074	mouse	yes
Kidney	14527/14527	20105	Datasets | 10× Genomics	mouse	yes
SHARE	55520/55520	17701	GSE207308	human	yes
SimData[b]	26968/26648	17701			no

[a]"yes" indicates that the numbers of scRNA-seq and scATAC-seq cells are equal. [b]"SimData" indicates that simulated data sets.

logical generalizability. An overview of these data sets is provided in Table 1.

2.2.2. Preprocessing Procedure. The preprocessing pipeline was tailored to each data modality. For scRNA-seq data, the established preprocessing workflow implemented within the Seurat package[25,26] was utilized. This standard workflow encompasses sequential steps of normalization, log transformation, and subsequent Z-score standardization. For scATAC-seq data, peak counts were first converted into gene activity matrices,[27] followed by a Term Frequency-Inverse Document Frequency (TF-IDF) transformation and Z-score standardization. After preprocessing, the data sets are mathematically represented with the following notation:

- m: Number of cells in the scRNA-seq data set.
- n: Number of cells in the scATAC-seq data set.
- g: Number of shared genes between scRNA-seq and scATAC-seq data.
- C: Total number of cell categories in the scRNA-seq data set.
- $\mathbf{X}^r \in \mathbb{R}^{m\times g}$: Gene expression matrix for scRNA-seq data.
- $\mathbf{X}^a \in \mathbb{R}^{n\times g}$: Gene activity matrix for scATAC-seq data.
- $\mathbf{Y}^r \in \mathbb{R}^{m\times 1} = \{\mathbf{Y}_1^r, \cdots, \mathbf{Y}_i^r, \cdots, \mathbf{Y}_m^r\}$: Cell label set for scRNA-seq data, where $\mathbf{Y}_i^r \in \{1, 2, \cdots, C\}$.

2.3. RNA-Guided Imputation

2.3.1. Dimensionality Reduction. Principal component analysis (PCA)[28] was applied to the original matrix $\mathbf{X}^r$ for dimensionality reduction, projecting it from a high-dimensional space into a lower-dimensional subspace. The reduced matrix $\widetilde{\mathbf{X}}^r \in \mathbb{R}^{m\times d}$ was obtained as follows (eq 1):

$$\widetilde{\mathbf{X}}^r = \mathbf{X}^r \cdot \mathbf{W} \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{g\times d}$ denotes the PCA projection matrix. The variable d ($d \ll g$) represents the number of selected principal components, which corresponds to the dimensionality of the reduced feature space. In this specific analysis,[29,30] the value of d was set to 15.

2.3.2. Similarity Calculation. In the low-dimensional embedding space of scRNA-seq data, the similarity between cell i and cell j was quantified using cosine similarity,[31] yielding a RNA−RNA similarity matrix $\mathbf{M} \in \mathbb{R}^{m\times m}$, where each matrix element $\mathbf{M}_{ij}$ was computed as (eq 2):

$$\mathbf{M}_{ij} = \frac{\widetilde{\mathbf{X}}_{i*}^r \cdot \widetilde{\mathbf{X}}_{j*}^r}{\left\|\widetilde{\mathbf{X}}_{i*}^r\right\| \cdot \left\|\widetilde{\mathbf{X}}_{j*}^r\right\|} \tag{2}$$

where $\widetilde{\mathbf{X}}_{i*}^r$ and $\widetilde{\mathbf{X}}_{j*}^r \in \mathbb{R}^{1\times d}$ denote the feature vectors of the ith and jth cells in the reduced space, $\|\cdot\|$ denotes the Euclidean norm, and $\cdot$ indicates the dot product.

2.3.3. Cross-Omics Imputation. Cross-omics imputation was performed by leveraging similarity matrix $\mathbf{M}$ to infer missing scATAC-seq features. For an arbitrary scATAC cell i, its gene activity vector $\mathbf{X}_{i*}^a$ was mapped into the same gene feature space as RNA, and its corresponding similarity vector $\mathbf{M}_{i*}$ was sorted to select the top k most similar RNA neighbors, forming the neighbor set N. For each neighbor cell $j \in N$, its feature vector $\mathbf{X}_{j*}^a \in \mathbb{R}^{1\times g}$ from matrix $\mathbf{X}^a$ was then aggregated through a similarity-weighted averaging scheme, resulting in the imputed feature $\overleftarrow{\mathbf{X}}_{i*}^a \in \mathbb{R}^{1\times g}$ for cell i, it was calculated as shown in eq 3:

$$\overleftarrow{\mathbf{X}}_{i*}^a = \sum_{j\in N} \frac{\mathbf{M}_{ij}}{\sum_{p\in N} \mathbf{M}_{ip}} \mathbf{X}_{j*}^a \tag{3}$$

This aggregation strategy ensures that highly similar neighbors contribute more strongly to the imputed feature representation. The collection of all imputed feature vectors $\overleftarrow{\mathbf{X}}_{i*}^a$ forms the imputed feature matrix $\overleftarrow{\mathbf{X}}^a \in \mathbb{R}^{n\times g}$.

2.4. Feature Extraction

2.4.1. Autoencoder Encoding. The low-dimensional embedding $\mathbf{Z}_{i*}^r \in \mathbb{R}^{1\times h}$ was derived using the encoder of a two-layer fully connected autoencoder, where h denotes the final dimensionality of the scRNA-seq data after this second stage of dimensionality reduction via the autoencoder.[32] The embedding was computed as (eq 4):

$$\mathbf{Z}_{i*}^r = f_E(\mathbf{X}_{i*}^r) \tag{4}$$

where $f_E(\cdot)$ represents the encoder function of the autoencoder. The set of low-dimensional embeddings collectively form the embedding matrix $\mathbf{Z}^r \in \mathbb{R}^{m\times h}$.

2.4.2. Embedding Reconstruction. Using a self-attention module,[33] the embedding matrix $\mathbf{Z}^r$ was weighted to obtain the reconstructed matrix $\overleftrightarrow{\mathbf{Z}}^r \in \mathbb{R}^{m\times h}$, computed as follows (eq 5):

$$\overleftrightarrow{\mathbf{Z}}^r = soft\max\left(\frac{\mathbf{Z}^r(\mathbf{Z}^r)^{\mathrm{T}}}{\sqrt{d}}\right)\cdot \mathbf{Z}^r \tag{5}$$

where $soft$ max $(\cdot)$ denotes the activation function and $\sqrt{d}$ is a scaling factor, typically set to $d = h$. Similarly, the imputed scATAC-seq matrix $\overleftarrow{\mathbf{X}}^a$ was encoded via an autoencoder and reconstructed through low-dimensional embeddings to obtain the reconstructed scATAC-seq embedding matrix $\overleftrightarrow{\mathbf{Z}}^a \in \mathbb{R}^{n\times h}$.

2.5. Feature Classification

To achieve accurate cell type classification under high-dimensional and sparse conditions, we developed a classifier based on a multilayer perceptron (MLP).[34] This classifier employs two Maxout[35] layers as the encoder and a linear layer as the decoder, enabling nonlinear feature learning from scRNA-seq and scATAC-seq data.

2.5.1. MLP-Based Encoding. *2.5.1.1. Linear Transformation.* The input feature matrix $\overleftrightarrow{\mathbf{Z}}^r$ was first processed by a two-layer encoder. For each cell i, the corresponding feature vector $\dot{\mathbf{Z}}_{i*}^r \in \mathbb{R}^{1\times d}$ underwent a linear transformation. This transformation is represented by eq 6:

$$\dot{\mathbf{Z}}_{i*}^r = \overleftrightarrow{\mathbf{Z}}_{i*}^r \mathbf{W}^{(1)} + b^{(1)} \tag{6}$$

where $\mathbf{W}^{(1)} \in \mathbb{R}^{h\times d}$ denotes the weight matrix of the first MLP layer, and $b^{(1)} \in \mathbb{R}^{1\times d}$ represents the bias vector of the first MLP layer. The output dimensionality of this linear layer, d, is determined by the Maxout activation function[35] and was set to 128.

2.5.1.2. Maxout Activation. To enhance nonlinear representational capacity and reduce the influence of noise and other interfering factors,

https://doi.org/10.1021/acs.jcim.5c02396
J. Chem. Inf. Model. 2026, 66, 1445−1456

a Maxout activation function[36] was applied. It selected the maximum value from every two consecutive elements as follows (eq 7):

$$\overline{\mathbf{Z}}^{r}_{i\theta} = Maxout(\dot{\mathbf{Z}}^{r}_{i*}) = \max\{\dot{\mathbf{Z}}^{r}_{i(2\theta-1)}, \dot{\mathbf{Z}}^{r}_{i(2\theta)}\},\ \theta = 1, \cdots, d/2 \tag{7}$$

where $\overline{\mathbf{Z}}^{r}_{i\theta} \in \mathbb{R}^{1\times(d/2)}$ represents the vector of activation matrix $\overline{\mathbf{Z}}^{r} \in \mathbb{R}^{m\times(d/2)}$. This process was repeated in the second MLP layer, yielding intermediate matrices $\ddot{\mathbf{Z}}^{r} \in \mathbb{R}^{m\times d}$ and $\overline{\overline{\mathbf{Z}}}^{r} \in \mathbb{R}^{m\times(d/2)}$.

2.5.2. MLP-Based Decoding. The high-order feature $\overline{\overline{\mathbf{Z}}}^{r}_{i*} \in \mathbb{R}^{1\times(d/2)}$ extracted by the two-layer encoder, was fed into a final linear layer to compute class scores. The score for cell i belonging to class c was calculated as (eq 8):

$$g(\overleftrightarrow{\mathbf{Z}}^{r}_{i*})[c] = \sum_{j=1}^{d/2} \mathbf{W}^{(3)}_{jc}\overline{\overline{\mathbf{Z}}}^{r}_{ij} + b^{(3)}_{c} \tag{8}$$

where $\mathbf{W}^{(3)}_{jc}$ represents the matrix element of the weight matrix $\mathbf{W}^{(3)} \in \mathbb{R}^{(d/2)\times c}$, and $b^{(3)} \in \mathbb{R}^{1\times C} = \{b^{(3)}_{1}, \cdots, b^{(3)}_{c}, \cdots, b^{(3)}_{C}\}$ represents the bias vector of the final layer, respectively.

Finally, the obtained scores were normalized by the Softmax function to yield the probability of assigning cell to class c, denoted as $t^{r}_{i*}[c]$, its value was calculated using eq 9:

$$t^{r}_{i*}[c] = \frac{\exp(g(\overleftrightarrow{z}^{r}_{i*}[c]))}{\sum_{t=1}^{C}\exp(g(\overleftrightarrow{z}^{r}_{i*})[t])} \tag{9}$$

The same procedure was applied to scATAC-seq cells to obtain the cell assignment probability $t^{a}_{i*}[c]$.

2.5.3. Classification Loss Calculation. To effectively train the classifier and address the issue of cell-type imbalance in real single-cell data sets, we employed a weighted cross-entropy loss function. This strategy assigned higher weights to under-represented rare cell types and lower weights to over-represented common cell types. The procedure consisted of two stages: weight precomputation and iterative computation of the weighted loss.

2.5.3.1. Weight Precomputation. The weight computation was based on the total number of cells belonging to each class in the scRNA data set. For a given cell class c, the number of cells n_c was calculated as (eq 10):

$$n_c = \sum_{i=1}^{m}[\mathbf{1}_{\{c=\mathbf{Y}^{r}_{i}\}}] \tag{10}$$

where $\mathbf{1}_{\{c=\mathbf{Y}^{r}_{i}\}}$ denotes the indicator function, equal to 1 if $c = \mathbf{Y}^{r}_{i}$ and 0 otherwise.

The class weight w_c (eq 11) was then determined by using inverse frequency weighting:

$$w_c = \frac{C}{n_c\cdot\sum_{t=1}^{C}1/n_t} \tag{11}$$

2.5.3.2. Weighted Loss Computation. During model training, a mini-batch gradient descent strategy was adopted. In each iteration, a batch of B cells was randomly sampled, and the loss was computed on this batch. The weighted classification loss l_{ce} for the batch was calculated as (eq 12):

$$l_{ce} = -\frac{1}{B}\sum_{i=1}^{B} w_c \log\left(\frac{\exp(g(\overleftrightarrow{\mathbf{Z}}^{r}_{i*})[\mathbf{Y}^{r}_{i}])}{\sum_{c=\mathbf{Y}^{r}_{1}}^{\mathbf{Y}^{r}_{m}}\exp(g(\overleftrightarrow{\mathbf{Z}}^{r}_{i*})[c])}\right) \tag{12}$$

where w_c denotes the precomputed weight corresponding to the true class label $\mathbf{Y}^{r}_{i}$ of cell i. By minimizing l_{ce}, the classifier effectively learned the discriminative features of all cell types while enhancing the recognition accuracy of rare cell populations.

2.6. Cell Selection

Due to inherent differences between scRNA-seq and scATAC-seq, directly applying an scRNA-seq-trained MLP classifier to scATAC-seq data often reduced accuracy and reliability. To overcome this, we developed a Dynamic Gaussian Mixture Model (GMM)-based strategy[37−39] that leveraged the correlation between gene expression and chromatin accessibility to adaptively identify "high-quality" scATAC-seq cells. These cells showed minimal transcriptomic−epigenomic discrepancies and high classification confidence. Selection was based on two criteria: feature similarity and classification reliability.

2.6.1. Feature Similarity Calculation. To assess the degree of alignment between each scATAC-seq cell and the scRNA-seq data in the feature space, a similarity score was calculated.

- scRNA-seq data: In the reconstruction scRNA-seq matrix $\overleftrightarrow{\mathbf{Z}}^{r}$, cells belonging to class c were identified, and their mean feature vector of was calculated and denoted as $\mathbf{V}^{r}_{c} \in \mathbb{R}^{1\times d}$.
- scATAC-seq data: For each cell i in the reconstructed scATAC-seq matrix $\overleftrightarrow{\mathbf{Z}}^{a}$, the cosine similarity between its feature vectors $\overleftrightarrow{\mathbf{Z}}^{a}_{i*}$ and $\mathbf{V}^{r}_{c}$ was computed. The maximum similarity across all classes is then defined as the cross-omics similarity score, $S^{a}_{i}(S^{a}_{i} \in [0,1])$, its value was calculated using eq 13:

$$S^{a}_{i} = \max_{c\in\{1,\cdots,C\}}\frac{\overleftrightarrow{\mathbf{Z}}^{a}_{i*}\cdot\mathbf{V}^{r}_{c}}{\left\|\overleftrightarrow{\mathbf{Z}}^{a}_{i*}\right\|\cdot\left\|\mathbf{V}^{r}_{c}\right\|},\ \mathbf{V}^{r}_{c} = \frac{1}{|I^{r}_{c}|}\sum_{i\in I^{r}_{c}}\overleftrightarrow{\mathbf{Z}}^{r}_{i*} \tag{13}$$

where $I^{r}_{c} = \{i|c = Y^{r}_{i}\}$ denotes the index set of scRNA-seq cells labeled as class c, and $|I^{r}_{c}|$ represents the number of elements in this set.

2.6.2. Classification Reliability Loss Calculation. To evaluate the accuracy and confidence of the classifier in predicting scATAC-seq cells, a classification reliability loss l^{a}_{i} (eq 14) for each cell i was defined, with higher values indicating greater uncertainty:

$$l^{a}_{i} = -\log(\max_{c\in\{1,\cdots,C\}}(t^{a}_{i*}[c])) \tag{14}$$

where $t^{a}_{i*}[c]$ is the Softmax function predicted probability for class c of cell i. Since $l^{a}_{i} \geq 0$, the loss is strictly positive.

2.6.3. Dynamic Gaussian Mixture Modeling. scATAC-seq cells with minimal cross-omics discrepancies relative to scRNA-seq data was identified through two steps: normalization and component selection.

2.6.3.1. Normalization. The cross-omics similarity score S^{a}_{i} (eq 15) and the classification loss l^{a}_{i} (eq 16) were normalized as

$$\tilde{S}^{a}_{i} = \frac{S^{a}_{i} - \min_{t} S^{a}_{t}}{\max_{t} S^{a}_{t} - \min_{t} S^{a}_{t} + \varepsilon} \tag{15}$$

$$\tilde{l}^{a}_{i} = \frac{l^{a}_{i} - \min_{t} l^{a}_{t}}{\max_{t} l^{a}_{t} - \min_{t} l^{a}_{t} + \varepsilon},\ t \in \{1, \cdots, n\} \tag{16}$$

where $\varepsilon > 0$ is a small constant for numerical stability. The normalized values constituted the sets $\tilde{S}^{a} = \{\tilde{S}^{a}_{i}\}^{n}_{i=1}$ and $\tilde{l}^{a} = \{\tilde{l}^{a}_{i}\}^{n}_{i=1}$, respectively.

2.6.3.2. Component Selection. (a) Number of Gaussian Components.

For $\tilde{S}^{a}$ and $\tilde{l}^{a}$, the likelihood-based model was defined as (eqs 17 and 18):

$$p(X|\theta) = \sum_{k=1}^{K}\pi_k\mathcal{N}(X|\mu_k, \sigma^{2}_{k}),\ X = \tilde{S}^{a} \text{ or } \tilde{l}^{a} \tag{17}$$

$$L(\theta) = \sum_{i=1}^{n}\ln p(x_i|\theta) = \sum_{i=1}^{n}\ln\left(\sum_{k=1}^{K}\pi_k|\mu_k, \sigma^{2}_{k}\right),\ x_i = \tilde{S}^{a}_{i} \text{ or } \tilde{l}^{a}_{i} \tag{18}$$

where $\pi_k(\pi_k > 0,\ \sum_{k=1}^{K}\pi_k = 1)$, μ_k, σ^{2}_{k} and $\mathcal{N}(\cdot)$ denote the mixture weight, mean, variance, and probability density function of the kth Gaussian component. $\theta = \{(\pi_k, \mu_k, \sigma^{2}_{k})\}^{K}_{k=1}$ represents the full set of GMM parameters, and K denotes the number of components.

(b) Optimal Number of Gaussian Components.

https://doi.org/10.1021/acs.jcim.5c02396
J. Chem. Inf. Model. 2026, 66, 1445−1456

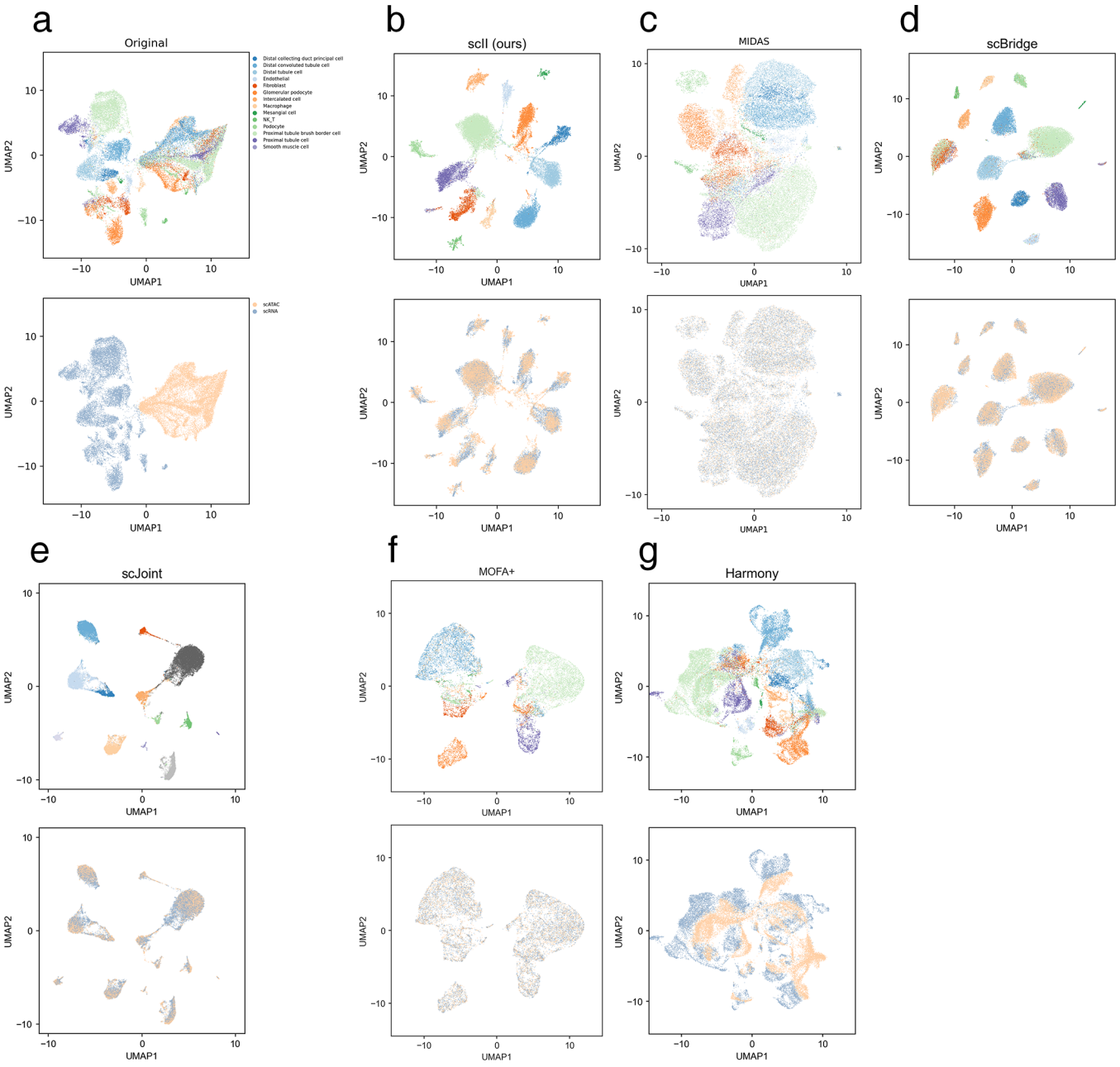


Figure 2. Integration performance on the Kidney data set.

Traditional GMM requires prespecifying the number of components K, which often leads to underfitting, overfitting, or high computational cost due to heuristic selection. To address this, we adopted a data-driven approach using the Bayesian Information Criterion (BIC)[40]: for each candidate K, a GMM was fitted using the Expectation-Maximization (EM) algorithm[41] and the corresponding BIC value (BIC_K) was computed using eq 19:

$$\mathrm{BIC}_K = -2\hat{L}_K + (3K - 1)\ln(n) \tag{19}$$

where $\hat{L}_K$ denotes the maximum log-likelihood under K components. The optimal number of components, K^*, was selected as the K value that minimized the BIC_K, found through a grid search. This procedure was applied independently to $\tilde{S}^a$ and $\tilde{I}^a$, yielding optimal component counts K_s^* and K_I^*, respectively.

2.6.3.3. High-Quality Cell Selection. Cell quality was assessed based on the concurrent achievement of two metrics: a high cross-omics similarity score and a low cross-omics classification loss. This quantitative standard directly dictates the reliability of cell-type identification.

- For the K_s^* Gaussian distributions of the set $\tilde{S}^a$, the Gaussian distribution with the largest mean was selected from the set $\{\mu_1, \cdots, \mu_{K_s^*}\}$ as the highest cross-omics similarity score.
- For the K_I^* Gaussian distributions of the set $\tilde{I}^a$, the Gaussian distribution with the smallest mean was selected from the set $\{\mu_1, \cdots, \mu_{K_I^*}\}$ as the low classification loss.

For each cell i, posterior probabilities of belonging to the high-similarity and low-loss component were computed as (eqs 20 and 21):

$$P_s(\tilde{S}_i^a) = \frac{\pi_{\hat{k}}\mathcal{N}(\tilde{S}_i^a|\mu_{\hat{k}}, \sigma_{\hat{k}}^2)}{\sum_{k=1}^{K_s^*}\pi_k\mathcal{N}(\tilde{S}_i^a|\mu_k, \sigma_k^2)} \tag{20}$$

https://doi.org/10.1021/acs.jcim.5c02396
J. Chem. Inf. Model. 2026, 66, 1445−1456

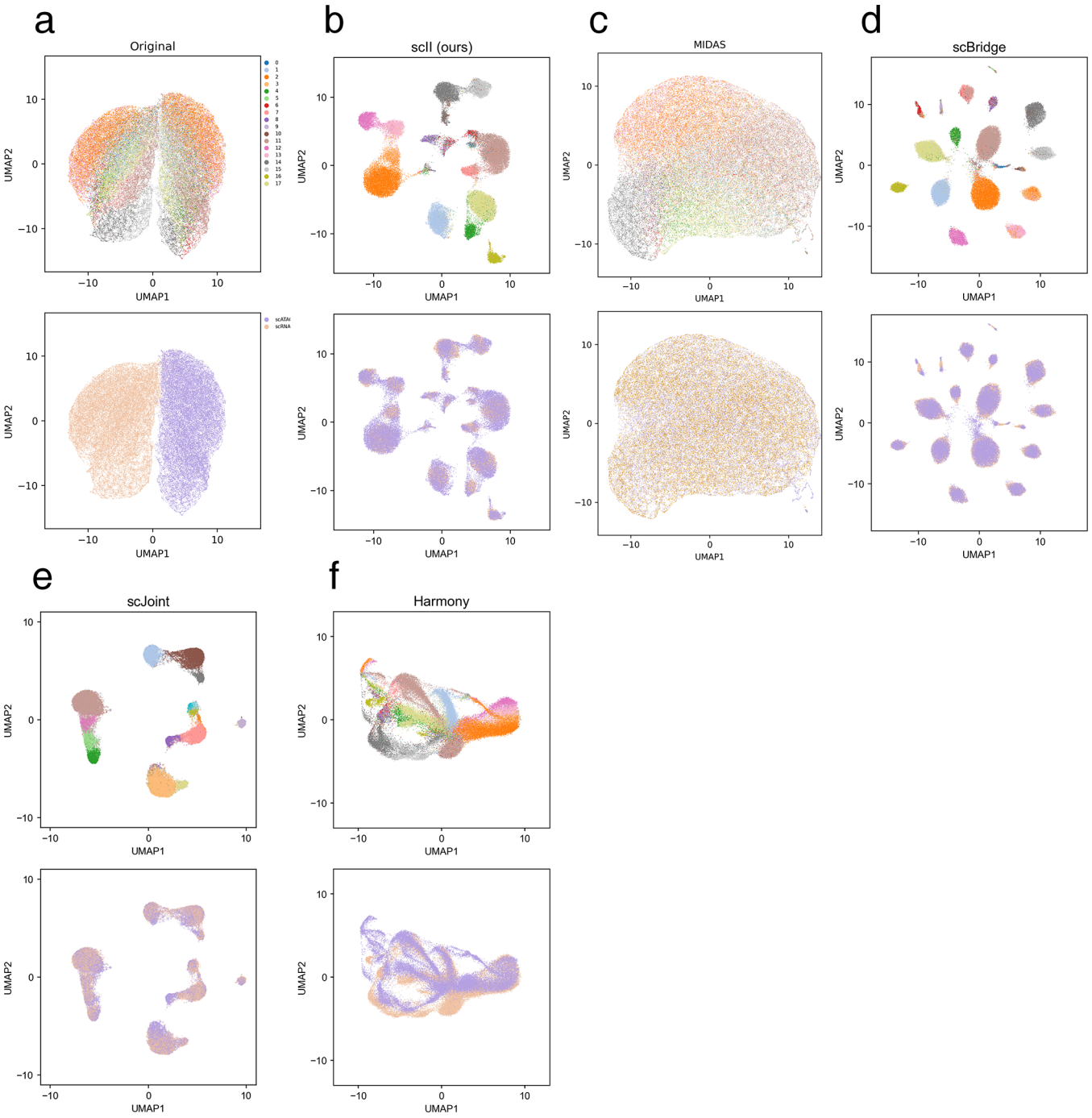


Figure 3. Integration performance on the Simulated data set.

$$P_l(\tilde{l}_i^a) = \frac{\pi_{\check{k}} \mathcal{N}(\tilde{l}_i^a | \mu_{\check{k}}, \sigma_{\check{k}}^2)}{\sum_{k=1}^{K_l^*} \pi_k \mathcal{N}(\tilde{l}_i^a | \mu_k, \sigma_k^2)} \tag{21}$$

where $P_s(\tilde{S}_i^a)$ and $P_l(\tilde{l}_i^a)$ denote the probabilities of the cross-omics similarity score $\tilde{S}_i^a$ and the cross-omics classification loss $\tilde{l}_i^a$ for cell i, respectively. The mixture coefficient, mean, and variance of the Gaussian distribution with the largest mean in the sets $\{\mu_1, \cdots, \mu_{K_s^*}\}$ are represented by $(\pi_{\hat{k}}, \mu_{\hat{k}}, \sigma_{\hat{k}}^2)$; similarly, those of the Gaussian distribution with the smallest mean in the sets $\{\mu_1, \cdots, \mu_{K_l^*}\}$ are represented by $(\pi_{\check{k}}, \mu_{\check{k}}, \sigma_{\check{k}}^2)$. Finally, high-quality cells are selected based on whether $P_s(\tilde{S}_i^a)$ and $P_l(\tilde{l}_i^a)$ meet the predefined thresholds.

2.7. Cross-Omics Integration

2.7.1. Omics Discrepancy Reduction. To reduce cross-omics discrepancies in selected high-quality cells, the mean square error between the scRNA-seq and scATAC-seq modalities was minimized. The alignment loss l_{mse} was defined as eq 22:

$$l_{mse} = \sum_{c=1}^{C} \left\| \mathbf{v}_c^r - \mathbf{v}_c^a \right\|_2^2 \tag{22}$$

Here, $\mathbf{v}_c^a \in \mathbb{R}^{1\times d}$ is the weighted mean feature vector of all scATAC-seq cells belonging to class c, computed as eq 23:

https://doi.org/10.1021/acs.jcim.5c02396
J. Chem. Inf. Model. 2026, 66, 1445−1456

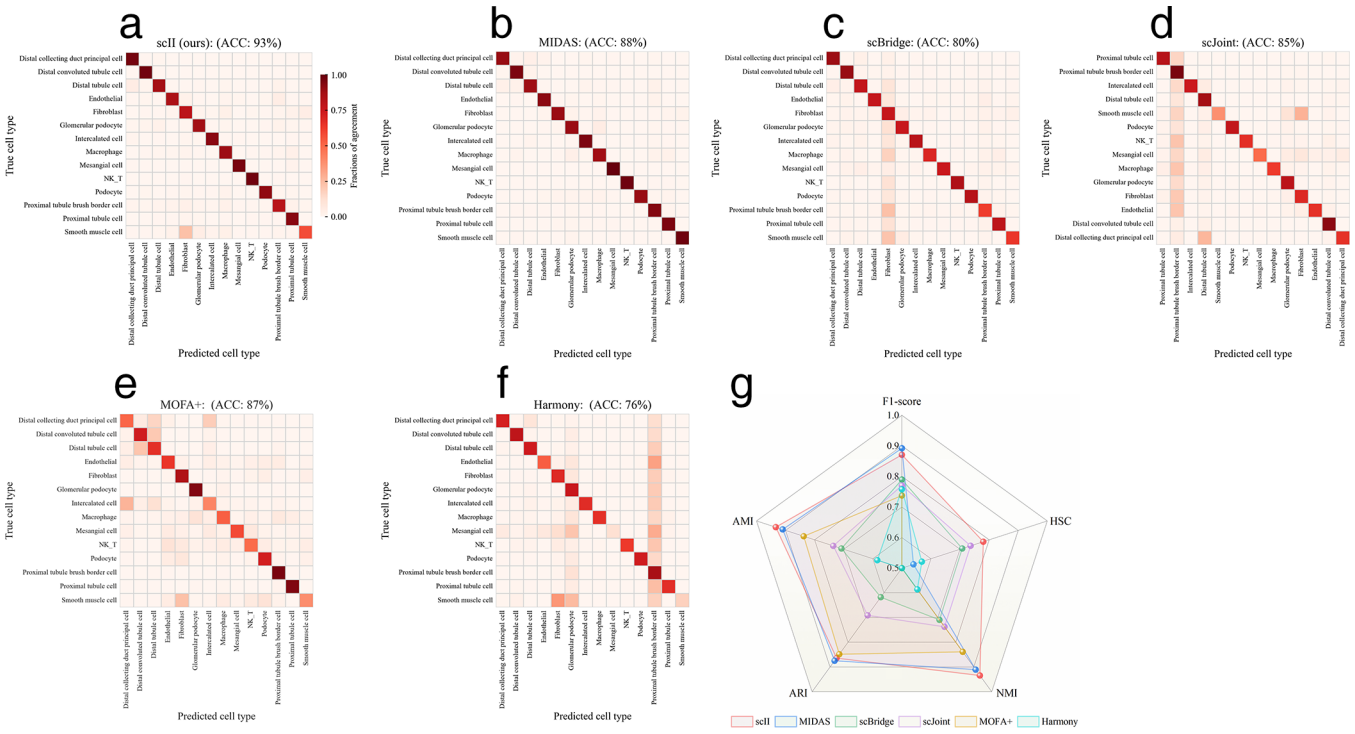


Figure 4. Multimetric clustering performance on the Kidney data set.

$$\mathbf{V}_c^a = \frac{\sum_{i \in I_c^a} (q_i \cdot \overleftrightarrow{\mathbf{Z}}_{i*}^a)}{\sum_{i \in I_c^a} (q_i)} \tag{23}$$

where $I_c^a = \{i | \hat{Y}_i^a = c\}$ denotes the index set of scATAC-seq cells labeled as class c. $\hat{\mathbf{Y}}_i^a = \arg\max_{c \in \{1,\cdots,C\}} (t_{i*}^a[c])$ is the predicted class label, and $\arg\max(\cdot)$ selects the class with the maximum predicted score.

The weight q_i was defined by the product of two probabilities (eq 24):

$$q_i = P_s(\tilde{S}_i^a) \cdot P_l(\tilde{l}_i^a) \tag{24}$$

2.7.2. Self-Training Strategy. To leverage high-quality scATAC-seq cells and iteratively refine the model, a self-training framework was employed. In each iteration, previously identified high-quality cells are removed from the scATAC-seq data set $\mathbf{X}^a$ and merged into the scRNA-seq data set $\mathbf{X}^r$ as pseudolabeled samples, thereby augmenting the training set with high-confidence predictions. The updated $\mathbf{X}^r$ was then used to retrain the entire model (autoencoder and MLP classifier) end-to-end, minimizing a composite loss by combining classification and cross-modal alignment objectives. This process was repeated until no additional high-quality cells were identified.

3. MATERIALS

3.1. Performance Assessment

Model performance was assessed using a multidimensional framework comprising harmonized silhouette coefficient (HSC), accuracy (ACC), weighted F1-score, normalized mutual information (NMI), adjusted Rand index (ARI), and adjusted mutual information (AMI). These metrics were applied across data sets of varying scales and qualities to comprehensively evaluate integration effectiveness, feature representation, and biological relevance. Detailed calculation formulas and implementation details for each metric are provided in Supporting Material Section 1.

3.2. Benchmark Methods

Five representative methods, including MIDAS,[21] scBridge,[42] scJoint,[19] MOFA+,[16] and Harmony,[43] were selected as benchmarks based on their methodological relevance, technical innovation, and complementary strategies in single-cell multiomics integration. Detailed descriptions of these methods are provided in Supporting Information Section 2.

4. RESULTS

For clarity and space considerations, we present results on the real Kidney data set and the simulated data set in the main text. Analyses of the remaining data sets are provided in Supporting Material Figures S1−S6.

4.1. Integration Effect Visualization

UMAP was applied visualization to compare integration performance on different data sets, focusing on cellular space alignment, cross-modality consistency, and the preservation of cell cluster structures.

4.1.1. Kidney Data Set. To assess robustness across tissue types and platforms, we compared scII with five benchmark methods on the Kidney data set, as shown in Figure 2.

The unintegrated original data (Figure 2a), cells colored by type, revealed 14 distinct populations, while coloring by modality revealed strong modality effects between scRNA-seq and scATAC-seq. In contrast, our scII method (Figure 2b) demonstrated superior performance by successfully mixing cells from both modalities, while forming well-defined clusters for each cell type. Other methods (Figure 2c−g) showed varying limitations. Specifically, MIDAS (Figure 2c) and scJoint (Figure 2e) performed poorly in clustering, with MIDAS failing to fully eliminate modality separation and scJoint unable to effectively distinguish between cell types. Although scBridge (Figure 2d), MOFA+ (Figure 2f), and Harmony (Figure 2g), while partially mixing modalities, failed to effectively distinguish cell types,

https://doi.org/10.1021/acs.jcim.5c02396
J. Chem. Inf. Model. 2026, 66, 1445−1456

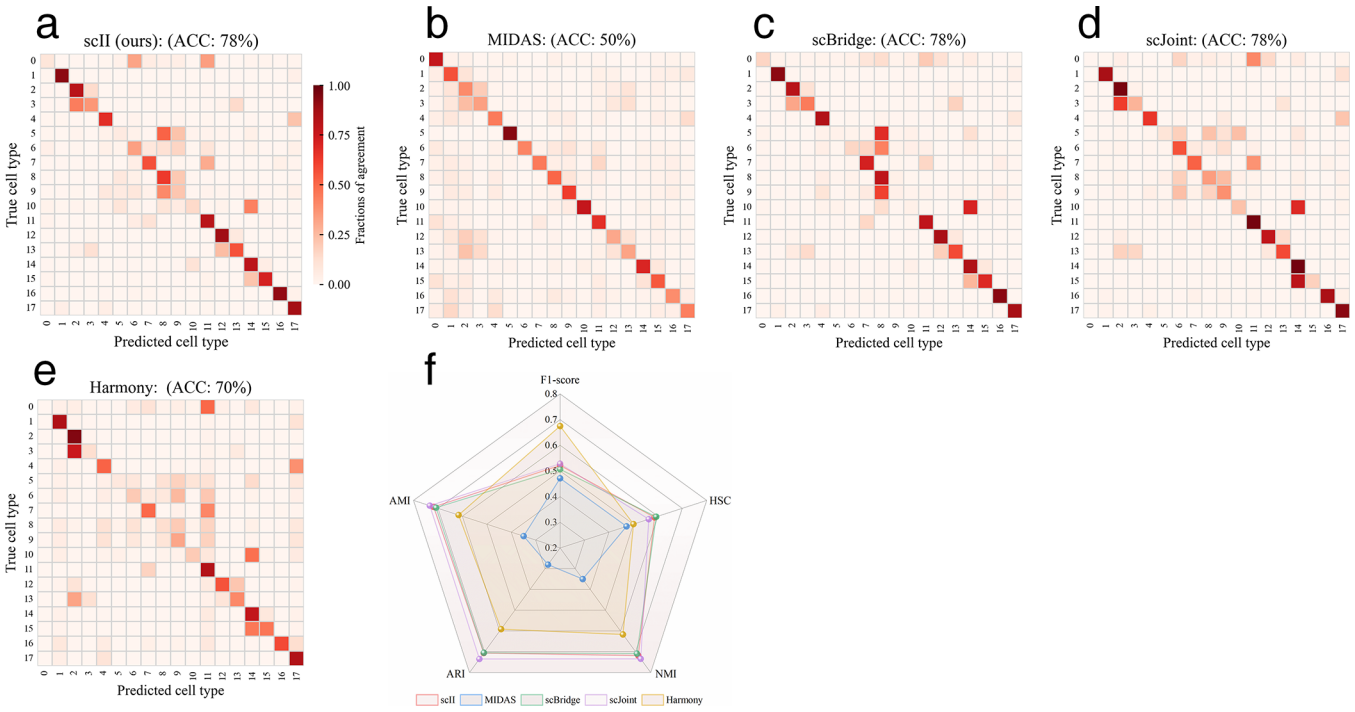


Figure 5. Multimetric clustering performance on the Simulated data set.

leading to poor biological clustering and a reduced significance of their integration results.

4.1.2. Simulated Data Set. To evaluate performance under high complexity and strong modality effects, we benchmarked scII against four methods on the simulated data set (Figure 3).

In the unintegrated data (Figure 3a), cells were heavily mixed by type with no discernible structure, while modality coloring revealed a complete separation between scRNA-seq and scATAC-seq, indicating strong modality effects. Our scII method (Figure 3b) achieved optimal integration, resolving 18 compact, well-separated clusters and fully eliminating modality stratification. In contrast, MIDAS (Figure 3c) only partially mitigated modality effects, producing low-resolution embeddings. ScBridge (Figure 3d) generated discrete clusters but showed oversegmentation and persistent modality bias. ScJoint (Figure 3e) failed to resolve specific cell types, erroneously merging them, and retained localized modality artifacts. Harmony (Figure 3f) produced elongated, blurry structures, where modality effects remained dominant.

4.2. Multimetric Clustering Performance

We evaluated cross-omics label transfer performance using six state-of-the-art methods: MIDAS, scBridge, scJoint, MOFA+, Harmony, and our proposed scII. These methods were tested across five data sets: SNARE, Kidney, SHARE, PBMC, and SimData.

4.2.1. Kidney Data Set. To assess generalization in a complex tissue environment, we evaluated six integration methods on the Kidney data set comprising 14 annotated renal cell types (Figure 4).

Our scII (Figure 4a) and MIDAS (Figure 4b) achieved high classification accuracy (93% and 88%), effectively transferring labels from scRNA-seq to scATAC-seq and reconstructing renal cell identities with strong discriminative power. scBridge (Figure 4c) reached 80% accuracy but showed substantial confusion between proximal tubule brush border cells and fibroblasts, reflecting the limited ability to resolve similar cell types. MOFA+ (Figure 4e) achieved 87% accuracy yet misclassified distal collecting duct principal cells and intercalated cells, indicating difficulty in distinguishing functionally related subpopulations. scJoint (Figure 4d) obtained 85% accuracy with errors mainly between smooth muscle cells and fibroblasts, likely due to feature loss during alignment. The accuracy of the Harmony (Figure 4f) model is relatively low because it mistakenly classifies a large number of brush-border cells in the proximal tubules as other cell types.

Clustering-based metrics (Figure 4g) further confirmed the superiority of our scII and MIDAS, both achieving near-optimal AMI, NMI, ARI, and F1-scores. However, scII exhibited a higher HSC (0.612), compared to 0.516 for MIDAS, highlighting its stronger ability to preserve the global structure and ensure cross-modality consistency.

4.2.2. Simulated Data Set. Figure 5 evaluates cross-omics integration under controlled conditions with known ground-truth labels.

Our scII achieved 78% accuracy across 18 cell types (Figure 5a), demonstrating robust label transfer and strong subtype discrimination. As shown in Figure 5e, Harmony (70% accuracy) showed basic classification ability but was clearly inferior. MIDAS (50% accuracy, Figure 5b) failed markedly with extensive confusion among multiple cell types, indicating poor adaptation to designed modality gaps. While scBridge (Figure 5c) and scJoint (Figure 5d) matched scII in accuracy (78%), they misclassified closely related subtypes (e.g., types 5−6, 8−10, 14−15), revealing weaker fine-grained resolution.

Radar plots (Figure 5f) further distinguish the integration quality. scII and scJoint not only achieved high accuracy and F1-score but also maintained superior clustering metrics (AMI, NMI, ARI) and HSC, indicating strong discriminative representations that preserve true population structures. MIDAS showed uniformly low scores, especially in HSC and AMI, confirming its inability to model nonlinear cross-modal relationships. Harmony performed worst overall, with lowest

https://doi.org/10.1021/acs.jcim.5c02396
J. Chem. Inf. Model. 2026, 66, 1445−1456

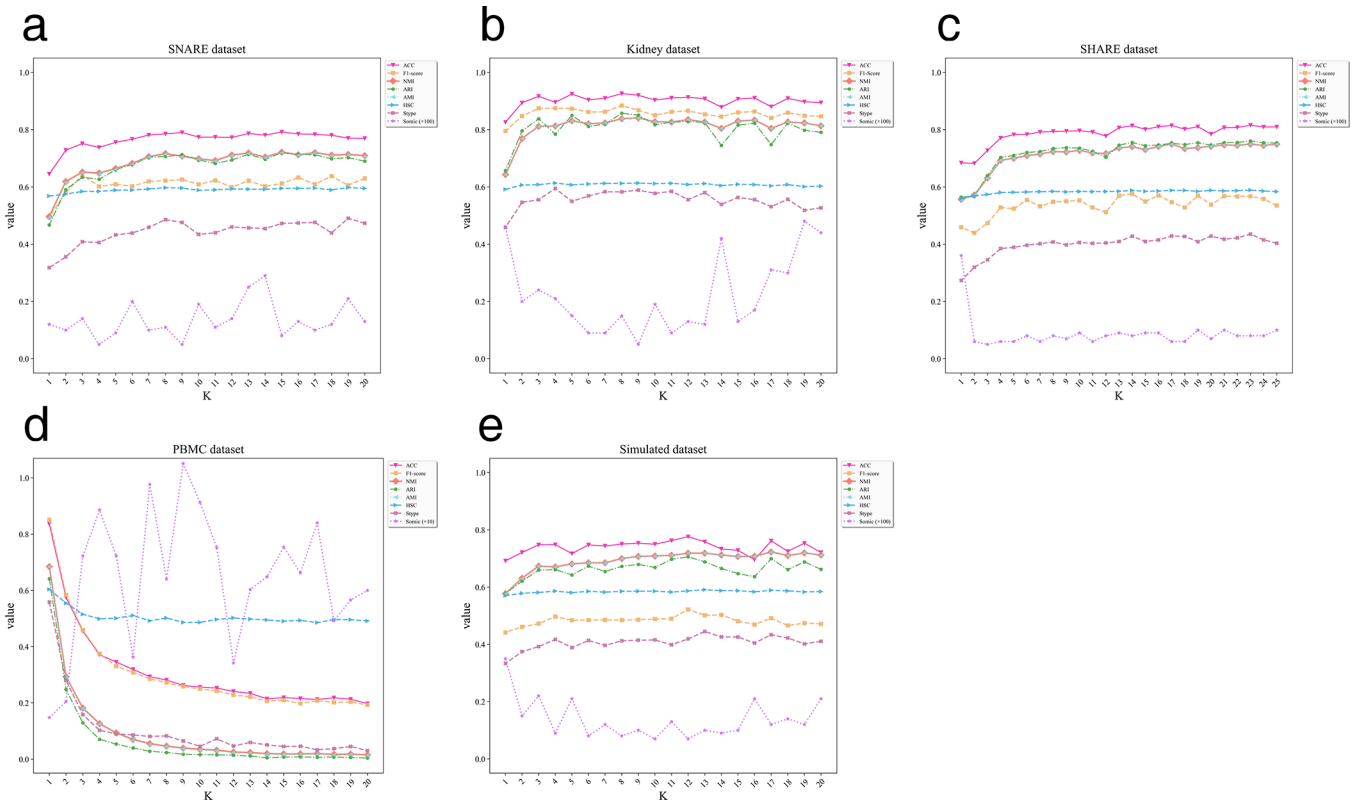


Figure 6. Parameter K optimization.

Table 2. Performance of Model Variants on Five Benchmark Data sets

data set	variant	ACC	F1-score	HSC	NMI	ARI	AMI
PBMC	S^a-alone	0.8499	0.8623	0.6112	0.7264	0.6806	0.7257
	MLP-alone	0.86	0.8706	0.6128	0.7193	0.6917	0.7187
	scII	0.8359	0.8517	0.6036	0.6846	0.6410	0.6839
SNARE	S^a-alone	0.7536	0.606	0.5946	0.7046	0.6598	0.7022
	MLP-alone	0.7945	0.6199	0.6007	0.7278	0.7225	0.7257
	scII	0.7901	0.6258	0.5959	0.7074	0.7119	0.7051
Kidney	S^a-alone	0.8983	0.8455	0.6107	0.8164	0.8009	0.8159
	MLP-alone	0.9391	0.8916	0.6171	0.8607	0.8902	0.8603
	scII	0.9256	0.8708	0.612	0.8474	0.8638	0.847
SHARE	S^a-alone	0.8034	0.579	0.5838	0.7301	0.7432	0.7298
	MLP-alone	0.7476	0.5385	0.5868	0.7094	0.6568	0.7091
	scII	0.8201	0.5762	0.5874	0.7491	0.7647	0.7488
SimData	S^a-alone	0.7035	0.4903	0.5814	0.7226	0.6868	0.722
	MLP-alone	0.7538	0.4963	0.5926	0.7089	0.6689	0.7083
	scII	0.7759	0.5219	0.5864	0.7182	0.7057	0.7176

accuracy and compressed radar profile, suggesting suitability for batch correction but inadequate for deep cross-omics alignment.

4.3. Parameter Optimization

In scRNA-seq guided scATAC-seq imputation, the parameter K (number of nearest neighbors) directly affects the imputation accuracy and computational efficiency. To determine the optimal setting, we evaluated $K \in \{1, \cdots, 20\}$ (extended to 25 for SHARE) across the SNARE, Kidney, SHARE, PBMC, and SimData, using eight metrics (ACC, F1-score, Somic, Stype, HSC, NMI, ARI, and AMI) (Figure 6).

As shown in Figure 6, optimal K varied by data sets, reflecting differences in intrinsic structure and data quality. Performance plateaued at $K = 9$ for SNARE (Figure 6a), 8 for Kidney (Figure 6b), 14 for SHARE (Figure 6c), and 12 for SimData (Figure 6e). These results suggest moderately larger neighborhoods improve robustness in complex or heterogeneous data by leveraging complementary signals. In contrast, PBMC peaked at $K = 1$ (Figure 6d) and declined thereafter, likely due to its high quality, sharp lineage boundaries, and strong modality specificity, where additional neighbors introduce noise and blur distinctions. In summary, K should be adaptively selected based on data set scale, heterogeneity, and quality, rather than applying a fixed uniform value, to achieve optimal cross-omics integration.

4.4. Ablation Study

An ablation analysis was conducted to evaluate the contribution of the dual-threshold mechanism, specifically its use of the cross-omics similarity score (S^a) and MLP classification reliability loss (l^a) for high-quality cells selection. The performance of the

https://doi.org/10.1021/acs.jcim.5c02396
J. Chem. Inf. Model. 2026, 66, 1445−1456

complete scII framework was contrasted with that of two single-criterion variants:

- S^a-alone: Selects high-quality scATAC-seq cells solely based on the cross-omics similarity score S^a.
- MLP-alone: Selects high-quality scATAC-seq cells solely based on the MLP classification loss l^a.
- scII: The proposed model, which combines the dual-threshold mechanism (S^a and l^a).

This evaluation was performed across five benchmark data sets (PBMC, Kidney, SNARE, Simulated, and SHARE). To ensure a fair comparison, the training pipeline and all hyperparameters were kept strictly identical across all variants with results detailed in Table 2.

The data in Table 2 reveal distinct performance trends based on data set size:

- Small data sets (such as PBMC): The single-criterion variants (S^a-alone or MLP-alone) achieve performance comparable to, or slightly exceeding, that of scII. This suggests that simpler selection criteria may suffice when the cell population size is limited and the data structure is less complex.
- Complex data sets (such as SHARE, SimData): As data set size and complexity increase, scII generally achieves the best or near-best performance metrics. Notably, scII attains a superior trade-off between the degree of cross-omics integration (measured by HSC) and clustering quality (measured by NMI, ARI, and AMI) alongside classification accuracy (ACC, F1-score).

By maintaining high integration scores while improving clustering accuracy, scII demonstrates that the benefits of integrating scRNA-seq and scATAC-seq data via the dual-threshold mechanism become increasingly pronounced as the data complexity grows.

5. CONCLUSION AND FUTURE DIRECTION

We present scII, a cross-omics integration framework that utilizes scRNA-seq to guide scATAC-seq imputation. By employing cosine similarity for cell comparisons and weighted imputation, scII effectively mitigates technical noise and bridges modality gaps. The framework incorporates dual reliability assessment, integrating feature similarity with classification loss, and leverages progressive knowledge transfer to enhance robustness against low-quality or outlier samples. Benchmarking on real and simulated data sets demonstrates scII's superior performance, particularly in complex data sets, achieving over 78% annotation accuracy with high discriminative power, structural fidelity, compact intracluster cohesion, and distinct intercluster boundaries. These attributes highlight scII's biological interpretability and analytical reliability.

A current limitation of the scATAC-seq preprocessing pipeline is its reliance on gene activity matrices constructed solely from proximal regulatory regions (gene bodies), which exclude information from distal accessible regions.

As sequencing technologies advance toward triand quad-omics profiling (encompassing transcriptome, epigenome, proteome, and metabolome), scII will be extended to accommodate higher-dimensional data and to explicitly map distal accessible regions to their target genes. This direction aims to exploit the full regulatory landscape, capture more intricate cross-modal interactions, and further enhance the resolution of transitional and rare cell states.

■ ASSOCIATED CONTENT

Data Availability Statement

The scRNA-seq and scATAC-seq data used in this manuscript are all publicly available. The SNARE, SHARE and PBMC data are available in the Gene Expression Omnibus (GEO) under accession codes GSE126074, GSE207308, and GSE156478, respectively. The Kidney data set was sourced from 10× Genomics (Datasets | 10× Genomics). The source code of scII is available at https://github.com/yufatang123la/scII.git.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c02396.

Definitions of evaluation metrics, benchmark methods, visualizations of integration effects, and visualizations of clustering performance (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Yuru Li − *School of Computer Science and Engineering and Guangxi key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin 541004, China;* orcid.org/0009-0003-6563-0337; Email: Liyuru@glut.edu.cn

Authors

Yi Zhang − *School of Computer Science and Engineering and Guangxi key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin 541004, China;* orcid.org/0000-0001-6167-7945

Zhicheng Jin − *School of Computer Science and Engineering and Guangxi key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin 541004, China*

Ye Tian − *School of Computer Science and Engineering and Guangxi key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin 541004, China*

Chen Su − *School of Computer Science and Engineering and Guangxi key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin 541004, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c02396

Author Contributions

Y.L., Y.Z.: Conceptualization, formal analysis, writing—original draft, writing—review and editing. Y.L., Z.J., Y.T., C.S.: Data curation. Y.L., Z.J.: Visualization. Funding acquisition: Y.Z.

Funding

This research was funded by the National Natural Science Foundation of China, grant number 62166014, and the Natural Science Foundation of Guangxi Zhuang Autonomous Region, grant number 2025GXNSFAA069627.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for suggestions that helped improve the paper substantially.

REFERENCES

(1) Liu, J.; Adhav, R.; Xu, X. Current Progresses of Single Cell DNA Sequencing in Breast Cancer Research. *International journal of biological sciences* **2017**, *13* (8), 949.

(2) Adil, A.; Kumar, V.; Jan, A. T.; Asger, M. Single-Cell Transcriptomics: Current Methods and Challenges in Data Acquisition and Analysis. *Frontiers in Neuroscience* **2021**, *15*, No. 591122.

(3) Stein, C. M.; Weiskirchen, R.; Damm, F.; Strzelecka, P. M. Single-cell Omics: Overview, Analysis, and Application in Biomedical Science. *J. of Cellular Biochemistry* **2021**, *122* (11), 1571−1578.

(4) Vistain, L. F.; Tay, S. Single-Cell Proteomics. *Trends in biochemical sciences* **2021**, *46* (8), 661−672.

(5) Hu, Y.; An, Q.; Guo, Y.; Zhong, J.; Fan, S.; Rao, P.; Liu, X.; Liu, Y.; Fan, G. Simultaneous Profiling of mRNA Transcriptome and DNA Methylome from a Single Cell. In *Single Cell Methods*; Proserpio, V., Ed.; Methods in Molecular Biology; Springer New York: New York, NY, 2019; Vol. *1979*, pp 363−377.

(6) Ma, S.; Zhang, B.; LaFave, L. M.; Earl, A. S.; Chiang, Z.; Hu, Y.; Ding, J.; Brack, A.; Kartha, V. K.; Tay, T.; Law, T.; Lareau, C.; Hsu, Y.-C.; Regev, A.; Buenrostro, J. D. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **2020**, *183* (4), 1103−1116.e20.

(7) Chen, S.; Lake, B. B.; Zhang, K. High-Throughput Sequencing of the Transcriptome and Chromatin Accessibility in the Same Cell. *Nature biotechnology* **2019**, *37* (12), 1452−1457.

(8) Zhu, C.; Yu, M.; Huang, H.; Juric, I.; Abnousi, A.; Hu, R.; Lucero, J.; Behrens, M. M.; Hu, M.; Ren, B. An Ultra High-Throughput Method for Single-Cell Joint Analysis of Open Chromatin and Transcriptome. *Nature structural & molecular biology* **2019**, *26* (11), 1063−1070.

(9) Stoeckius, M.; Hafemeister, C.; Stephenson, W.; Houck-Loomis, B.; Chattopadhyay, P. K.; Swerdlow, H.; Satija, R.; Smibert, P. Simultaneous Epitope and Transcriptome Measurement in Single Cells. *Nat. Methods* **2017**, *14* (9), 865−868.

(10) Peterson, V. M.; Zhang, K. X.; Kumar, N.; Wong, J.; Li, L.; Wilson, D. C.; Moore, R.; McClanahan, T. K.; Sadekova, S.; Klappenbach, J. A. Multiplexed Quantification of Proteins and Transcripts in Single Cells. *Nature biotechnology* **2017**, *35* (10), 936−939.

(11) Cao, J.; Cusanovich, D. A.; Ramani, V.; Aghamirzaie, D.; Pliner, H. A.; Hill, A. J.; Daza, R. M.; McFaline-Figueroa, J. L.; Packer, J. S.; Christiansen, L.; Steemers, F. J.; Adey, A. C.; Trapnell, C.; Shendure, J. Joint Profiling of Chromatin Accessibility and Gene Expression in Thousands of Single Cells. *Science* **2018**, *361* (6409), 1380−1385.

(12) Ma, A.; McDermaid, A.; Xu, J.; Chang, Y.; Ma, Q. Integrative Methods and Practical Challenges for Single-Cell Multi-Omics. *Trends Biotechnol.* **2020**, *38* (9), 1007−1022.

(13) Kang, M.; Ko, E.; Mersha, T. B. A Roadmap for Multi-Omics Data Integration Using Deep Learning. *Briefings Bioinf.* **2022**, *23* (1), bbab454.

(14) Song, M.; Greenbaum, J.; Luttrell, J., IV; Zhou, W.; Wu, C.; Shen, H.; Gong, P.; Zhang, C.; Deng, H.-W. A Review of Integrative Imputation for Multi-Omics Datasets. *Frontiers in genetics* **2020**, *11*, No. 570255.

(15) Athaya, T.; Ripan, R. C.; Li, X.; Hu, H. Multimodal Deep Learning Approaches for Single-Cell Multi-Omics Data Integration. *Brief Bioinform* **2023**, *24* (5), bbad313.

(16) Argelaguet, R.; Arnol, D.; Bredikhin, D.; Deloro, Y.; Velten, B.; Marioni, J. C.; Stegle, O. MOFA+: A Statistical Framework for Comprehensive Integration of Multi-Modal Single-Cell Data. *Genome Biol.* **2020**, *21* (1), 111.

(17) Zuo, C.; Chen, L. Deep-Joint-Learning Analysis Model of Single Cell Transcriptome and Open Chromatin Accessibility Data. *Briefings Bioinf.* **2021**, *22* (4), bbaa287.

(18) Gong, B.; Zhou, Y.; Purdom, E. Cobolt: Integrative Analysis of Multimodal Single-Cell Sequencing Data. *Genome Biol.* **2021**, *22* (1), 351.

(19) Lin, Y.; Wu, T.-Y.; Wan, S.; Yang, J. Y.; Wong, W. H.; Wang, Y. R. scJoint Integrates Atlas-Scale Single-Cell RNA-Seq and ATAC-Seq Data with Transfer Learning. *Nature biotechnology* **2022**, *40* (5), 703−710.

(20) Cao, Z.-J.; Gao, G. Multi-Omics Single-Cell Data Integration and Regulatory Inference with Graph-Linked Embedding. *Nat. Biotechnol.* **2022**, *40* (10), 1458−1466.

(21) He, Z.; Hu, S.; Chen, Y.; An, S.; Zhou, J.; Liu, R.; Shi, J.; Wang, J.; Dong, G.; Shi, J.; et al. Mosaic Integration and Knowledge Transfer of Single-Cell Multimodal Data with MIDAS. *Nat. Biotechnol.* **2024**, *42* (10), 1594−1605.

(22) Chen, H.; Lareau, C.; Andreani, T.; Vinyard, M. E.; Garcia, S. P.; Clement, K.; Andrade-Navarro, M. A.; Buenrostro, J. D.; Pinello, L. Assessment of Computational Methods for the Analysis of Single-Cell ATAC-Seq Data. *Genome Biol.* **2019**, *20* (1), 241.

(23) Stuart, T.; Satija, R. Integrative Single-Cell Analysis. *Nat. Rev. Genet.* **2019**, *20* (5), 257−272.

(24) Domínguez Conde, C.; Xu, C.; Jarvis, L. B.; Rainbow, D. B.; Wells, S. B.; Gomes, T.; Howlett, S. K.; Suchanek, O.; Polanski, K.; King, H. W.; Mamanova, L.; Huang, N.; Szabo, P. A.; Richardson, L.; Bolt, L.; Fasouli, E. S.; Mahbubani, K. T.; Prete, M.; Tuck, L.; Richoz, N.; Tuong, Z. K.; Campos, L.; Mousa, H. S.; Needham, E. J.; Pritchard, S.; Li, T.; Elmentaite, R.; Park, J.; Rahmani, E.; Chen, D.; Menon, D. K.; Bayraktar, O. A.; James, L. K.; Meyer, K. B.; Yosef, N.; Clatworthy, M. R.; Sims, P. A.; Farber, D. L.; Saeb-Parsy, K.; Jones, J. L.; Teichmann, S. A. Cross-Tissue Immune Cell Analysis Reveals Tissue-Specific Features in Humans. *Science* **2022**, *376* (6594), No. eabl5197.

(25) Stuart, T.; Butler, A.; Hoffman, P.; Hafemeister, C.; Papalexi, E.; Mauck, W. M.; Hao, Y.; Stoeckius, M.; Smibert, P.; Satija, R. Comprehensive Integration of Single-Cell Data. *Cell* **2019**, *177* (7), 1888−1902.e21.

(26) *Integrated analysis of multimodal single-cell data: Cell.* https://www.cell.com/cell/fulltext/S0092-8674(21)00583-3?s=09 (accessed Dec 01, 2025).

(27) Stuart, T.; Srivastava, A.; Madad, S.; Lareau, C. A.; Satija, R. Single-Cell Chromatin State Analysis with Signac. *Nat. Methods* **2021**, *18* (11), 1333−1341.

(28) Ghojogh, B.; Crowley, M. Unsupervised and Supervised Principal Component Analysis: Tutorial. arXiv August 1, 2022. .

(29) Cui, S.; Nassiri, S.; Zakeri, I. Mcadet: A Feature Selection Method for Fine-Resolution Single-Cell RNA-Seq Data Based on Multiple Correspondence Analysis and Community Detection. *PLOS Computational Biology* **2024**, *20* (10), No. e1012560.

(30) Zhu, Q.; Zhao, X.; Zhang, Y.; Li, Y.; Liu, S.; Han, J.; Sun, Z.; Wang, C.; Deng, D.; Wang, S.; et al. Single Cell Multi-Omics Reveal Intra-Cell-Line Heterogeneity across Human Cancer Cell Lines. *Nat. Commun.* **2023**, *14* (1), 8170.

(31) Alake, R. Understanding Cosine Similarity and Its Applications. *builtin. com* **2023**.

(32) Yu, S.; Principe, J. C. Understanding Autoencoders with Information Theoretic Concepts. *Neural Networks* **2019**, *117*, 104−123.

(33) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in neural information processing systems*; **2017**; Vol. *30*.

(34) Popescu, M.-C.; Balas, V. E.; Perescu-Popescu, L.; Mastorakis, N. Multilayer Perceptron and Neural Networks. *WSEAS Trans. Circ. Syst.* **2009**, *8* (7), 579−588.

(35) Goodfellow, I.; Warde-Farley, D.; Mirza, M.; Courville, A.; Bengio, Y. Maxout Networks. In *International conference on machine learning*; PMLR: 2013; pp 1319−1327.

(36) Sun, W.; Su, F.; Wang, L. Improving Deep Neural Networks with Multi-Layer Maxout Networks and a Novel Initialization Method. *Neurocomputing* **2018**, *278*, 34−40.

(37) Wan, H.; Wang, H.; Scotney, B.; Liu, J. A Novel Gaussian Mixture Model for Classification. In *2019 IEEE international conference on systems, man and cybernetics (SMC)*; IEEE: 2019; pp 3298−3303.

(38) Reynolds, D. Gaussian Mixture Models. In *Encyclopedia of biometrics*; Springer: 2015; pp 827−832.

(39) Chen, Z.; Ellis, T. A Self-Adaptive Gaussian Mixture Model. *Computer Vision and Image Understanding* **2014**, *122*, 35−46.

https://doi.org/10.1021/acs.jcim.5c02396
J. Chem. Inf. Model. 2026, 66, 1445−1456

(40) Watanabe, S. A Widely Applicable Bayesian Information Criterion. *J. Mach. Learn. Res.* **2013**, *14* (1), 867−897.
(41) Jannah, W.; Saputro, D. R. Parameter Estimation of Gaussian Mixture Models (GMM) with Expectation Maximization (EM) Algorithm. In *AIP Conference Proceedings*; AIP Publishing LLC: 2022; Vol. *2566*, p 040002.
(42) Li, Y.; Zhang, D.; Yang, M.; Peng, D.; Yu, J.; Liu, Y.; Lv, J.; Chen, L.; Peng, X. scBridge Embraces Cell Heterogeneity in Single-Cell RNA-Seq and ATAC-Seq Data Integration. *Nat. Commun.* **2023**, *14* (1), 6045.
(43) Korsunsky, I.; Millard, N.; Fan, J.; Slowikowski, K.; Zhang, F.; Wei, K.; Baglaenko, Y.; Brenner, M.; Loh, P.; Raychaudhuri, S. Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony. *Nat. Methods* **2019**, *16* (12), 1289−1296.

https://doi.org/10.1021/acs.jcim.5c02396
J. Chem. Inf. Model. 2026, 66, 1445−1456