# scCMIA: Self-supervised Dual Model for Mitigating Information Loss in Single-cell Cross-Modal Alignment

**Anonymous authors**
Paper under double-blind review

## Abstract

Recent technological advances in single-cell sequencing have enabled simultaneous profiling of multiple omics modalities within individual cells. Despite these advancements, challenges such as high noise levels and information loss during computational integration persist. While existing methods align different modalities, they often struggle to balance alignment accuracy with the preservation of modality-specific information needed for downstream biological discovery. In this paper, we introduce scCMIA, a novel framework guided by Mutual Information (MI) principles that leverages a VQ-VAE architecture. scCMIA achieves robust cross-modal alignment in a unified discrete latent space while enabling high-fidelity reconstruction of the original data modalities. Crucially, our framework transforms the learned discrete representations into a tool for tangible biological discovery, allowing for the quantification of regulatory programs and cross-modal relationships. Our extensive experiments demonstrate that scCMIA achieves state-of-the-art performance across multiple datasets. Our code is available at: https://anonymous.4open.science/r/scCMIA-77E3.

## 1 Introduction

Multimodal learning is becoming increasingly crucial in the field of biology. Biological processes within cells involve multiple regulatory levels, including DNA, RNA, and proteins Tang et al. (2023); Li et al. (2024). The intricate interactions and influences between these levels necessitate an integrated multimodal understanding to fully comprehend these biological processes Tu et al. (2022). In recent years, technological advancements that enable the analysis of multimodal information at single-cell resolution have been pivotal in cataloging cell types and states. For instance, single-cell RNA sequencing (scRNA-seq) Picelli et al. (2013) is used to profile the transcriptomes of individual cells, offering deep insights into cellular heterogeneity and gene expression patterns. Similarly, the single-cell assay for transposase-accessible chromatin with high throughput sequencing (scATAC-seq) Cusanovich et al. (2015) profiles the chromatin accessibility of individual cells, providing valuable information about gene regulatory networks and chromatin structure.

Although data from individual modalities are readily available and easy to analyze, they offer limited fundamental information on how different layers of genomic regulation interact within a single cell Wu et al. (2021). With the further development of technology, various multimodal single-cell protocols Chen et al. (2019b); Ma et al. (2020); Xu et al. (2022) have been proposed to obtain a more comprehensive view of individual cells and simultaneously profile gene expression and chromatin accessibility. Despite the potential benefits, integrating multimodal data often faces significant challenges due to large differences in the modal feature spaces Chen et al. (2019a). For instance, accessible chromatin regions in scATAC-seq and genes in scRNA-seq exhibit substantial discrepancies Argelaguet et al. (2021), making it difficult to effectively perform a joint analysis of data from both modalities.

Intuitively, a direct strategy for managing multimodal data involves embedding diverse modalities into a unified representation space. Certain methodologies accomplish this by transforming multimodal inputs directly into a shared feature space, often utilizing prior knowledge Duren et al. (2018); Zeng et al. (2019). Conversely, alternative approaches, such as those described by Minoura et al. (2021);

Gong et al. (2021) , integrate multiple modalities without aligning them in a shared space. This approach, however, fails to facilitate the interaction between different modalities and consequently cannot fully exploit potentially complementary information across modalities. Furthermore, some techniques Cao & Gao (2022); Ashuach et al. (2023) attempt to enhance model performance by combining cross-modal alignment with modality reconstruction tasks. Despite these efforts, there remains significant scope for improvement in both the efficacy of cross-modal alignment and the accuracy of modality reconstruction.

To address these challenges, we developed a novel VQ-VAE-based cell-level alignment framework, called single-cell cross-modal mutual information (MI) alignment (scCMIA). This framework effectively achieves cross-modal alignment at the single-cell level and reconstructs data in their original modality spaces. scCMIA utilizes the RNA to ATAC (RtA) module to initially align scRNA and scATAC sequences in a continuous shared feature space. Subsequently, it constructs a cross-modal unified codebook in discrete space, facilitating enhanced cross-modal interaction and significantly improving the robustness of both alignment and reconstruction processes. This approach not only enhances the alignment accuracy but also effectively reconstruct multi-modal data, thereby addressing the issue of unimodal information deficiency and providing a comprehensive solution for multi-modal data integration and analysis. Our main contributions are summarized as follows:

- We propose scCMIA, a single-cell multi-omics integration framework centered on mutual information (MI) theory. This framework achieves alignment by maximizing cross-modal MI while minimizing intra-modal MI for feature decoupling, providing theoretical assurance for high-precision alignment and high-fidelity data generation. To this end, we designed a robust dual-space alignment strategy that first performs alignment in a continuous space and then refines it in a discrete space using a unified discrete codebook. This approach significantly enhances the model's overall performance and robustness.
- We demonstrate that the designed unified codebook can learn structures with high biological significance, and propose methods to quantify the conservation of regulatory programs across cell lineages and reveal differences in regulatory coupling among distinct cell types, providing powerful new tools for downstream biological exploration.
- Through extensive experiments across multiple benchmark datasets, we demonstrate that scCMIA achieves state-of-the-art performance across a range of key tasks. Compared to existing methods, our model exhibits significant advantages in cross-modal alignment, data reconstruction, and data interpolation tasks, comprehensively validating the effectiveness and superiority of our proposed framework.

## 2 Related Work

Multimodal alignment is rapidly advancing in fields such as text, vision, and speech. Methods like CLIP Radford et al. (2021), ALBEF Li et al. (2021), and GLIP Li et al. (2022) have played significant roles in their respective domains. Concurrently, in the field of biology, multimodal alignment and reconstruction methods are also making important contributions. In the field of biology, we can categorize multimodal integration strategies into three types: (1). multimodal alignment, (2). multimodal reconstruction, and (3). multimodal alignment and reconstruction.

**Multimodal Alignment**. Techniques such as Pamona Cao et al. (2022), UnionCon Cao et al. (2020), Seurat V3 Stuart et al. (2019), MMD_MA Singh et al. (2020), SCOT Demetci et al. (2020) and scGCL Xiong et al. (2023) align cells from different omics layers through nonlinear flows. These methods eliminate the need for prior knowledge and minimize information loss between modalities. However, they suffer from poor alignment robustness when handling noisy and difficult to apply on large scale data processing. **Multimodal Reconstruction**. Techniques such as scMM Minoura et al. (2021), Cobolt Gong et al. (2021) and scButterfly Cao et al. (2024) focus primarily on reconstructing missing or incomplete data across different modalities. However, as these methods do not explicitly align modalities into a shared latent space, their utility for tasks requiring direct cross-modal comparison, querying, and label transfer can be limited. By not creating a unified representation, they may also not fully leverage complementary information across modalities for certain downstream analyses.

**Multimodal Alignment and Reconstruction**. Most methods in this category are based on autoencoders. These approaches not only align data from different modalities but also reconstruct

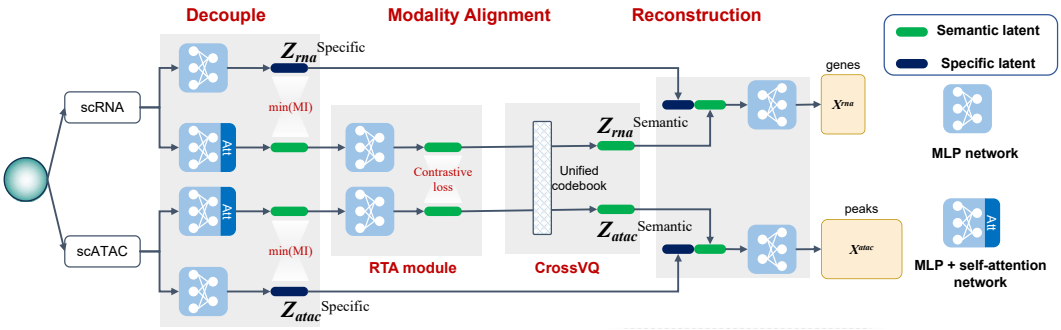



Figure 1: The pipeline of scCMIA. The scCMIA framework is designed to perform intra-modal decoupling and cross-modal alignment, thereby enabling dynamic interaction between modalities while reconstructing individual modalities to capture their intrinsic semantic information.

the original input data from the aligned representations. GLUE Cao & Gao (2022) utilizes graph variational autoencoders (VAE) to model known regulatory relationships between open chromatin regions and genes, enabling efficient cross-modal feature translation. However, it can only reconstruct the original spatial data of the scATAC that is related to the scRNA. MutiVI constructs multiple VAE models employing a joint latent representation to integration embedding spaces Ashuach et al. (2023). The performance on alignment and refactoring tasks still has room for further improvement.

**Feature Decoupling**. Feature decoupling plays an important role in many fields, such as Peng et al. (2019) proposed decoupled representation learning framework for multigraphs to capture complete and clean common information. The decoupling of class-independent features is proposed Mo et al. (2023) and the alignment of source domain and target domain is realized. Uni-code Xia et al. (2024) proposed dual cross-modal information uncoupling and multimodal EMA, which unified expression in audio and video, audio text, and even audio - video - text three modes, and realized cross-modal generalization of various tasks in the downstream. Our approach has similar ideas to Uni-code, but we have adopted a completely different strategy in terms of framework and unified codebook design, making it more suitable for the single-cell multi-omics field.

# 3 METHOD

## 3.1 PRELIMINARY

In this section, we first introduce preliminary work on the design of the scCMIA framework (Fig. 1). Our objective is to balance the model in terms of alignment and reconstruction performance. However, unimodal data contains both modality-specific features and semantic characteristics, where modality-specific features may hinder cross-modal alignment. To address this challenge, we propose decoupling and alignment methods that leverage the bounds of MI. This approach facilitates achieving our goal by effectively managing these modality-specific features while preserving semantic consistency across different modalities.

MI is a measure of mutual dependence between two random variables. Intuitively, MI represents the amount of information contained in one random variable about another. Given random variables $X$ and $Y$, the MI is defined as the Kullback-Leibler (KL) divergence between its joint distribution $p(x, y)$ and the product $p(x)p(y)$ of the marginal distributions:

$$\begin{aligned} \mathrm{I}(X;Y) &= D_{\mathrm{KL}}(p(x,y)\|p(x)p(y)) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right] \\ &= \mathrm{H}(Y) - \mathrm{H}(Y \mid X) = \mathrm{H}(X) - \mathrm{H}(X \mid Y). \end{aligned} \tag{1}$$

where $\mathrm{H}(X)$ and $\mathrm{H}(Y)$ are marginal entropies, and $\mathrm{H}(X|Y)$ and $\mathrm{H}(Y|X)$ are conditional entropies. When $X$ and $Y$ correspond to each other, $\mathrm{I}(X;Y) = \mathrm{H}(X) = \mathrm{H}(Y)$; when $X$ and $Y$ are independent of each other, $\mathrm{I}(X;Y) = 0$.

In single-cell scRNA and scATAC, let's define random variables $x$ and $\hat{x} \in X_{\text{RNA}}$, and $y$ and $\hat{y} \in Y_{\text{ATAC}}$ representing the semantic and modality-specific features of scRNA and scATAC, respectively, the upper bound of the intra-modal mutual information can be expressed as $\mathcal{L}_{\text{intra}} = \mathrm{I}(x;\hat{x}) + \mathrm{I}(y;\hat{y})$.

To achieve the decoupling of modality-specific and semantic features, our objective is to minimize this upper bound. However, relying solely on minimizing the upper bound of MI within modalities is insufficient for effectively decoupling modality-specific and semantic information in a directed manner. Therefore, to efficiently decouple modality-specific and semantic features, it is necessary to leverage the correlations between cross-modal semantics as guidance. Our objective is to maximize the lower bound of cross-modal MI $\mathcal{L}_{\text{inter}} = \mathrm{I}(x;y)$ to bring the cross-modal semantic relationships closer together.

Therefore, our overall optimization objective combines minimizing the upper bound of MI within modalities to achieve intra-modal decoupling, and maximizing the lower bound of MI across modalities to achieve cross-modal semantic alignment ($\psi$ is the decoupling module parameter).

$$\min_{\psi} \mathcal{L}_{\text{intra}} - \mathcal{L}_{\text{inter}}. \tag{2}$$

### 3.2 Intra-modal Decoupling Learning

When two modes are very different, forcing the two modes to align directly will lead to semantic information loss in the shared space of each mode Niu et al. (2024). To mitigate this tradeoff between alignment and modal information loss, we introduce the following assumption:

Given sequencing data of two modalities from the same single cell, scRNA, and scATAC, represented by $X$ and $Y$ respectively, we can decouple them into modality-specific $\hat{x}$ and modality-agnostic semantic representations $x$. $Y$ is decoupled into modality-specific representation $\hat{y}$ and modality-agnostic semantic representation $y$. The distance between the decoupled scRNA semantic representations $x$ and $y$ should satisfy the following condition:

$$\max\left(\mathrm{I}(x,\hat{x}), \mathrm{I}(y,\hat{y})\right) < \mathrm{I}(x,y). \tag{3}$$

On the basis of this assumption, we uses Contrastive Log-ratio Upper Bound (CLUB) Cheng et al. (2020) to estimate the upper bound of the MI between $\hat{x}$ and $x$ within the scRNA modality: $\mathrm{I}(x;\hat{x}) \leq \mathrm{I}_{\mathrm{CLUB}}(x;\hat{x})$. Similarly, the upper bound on the MI of modality-specific and modality-agnostic representation in scATAC is estimated by $\mathrm{I}(y;\hat{y}) \leq \mathrm{I}_{\mathrm{CLUB}}(y;\hat{y})$. Therefore, in the context of a single modality, CLUB is defined as:

$$\mathrm{I}_{\mathrm{vCLUB}}(x;\hat{x}) := \mathbb{E}_{p(x,\hat{x})}\left[\log q_\theta(\hat{x} \mid x)\right] - \mathbb{E}_{p(x)}\mathbb{E}_{p(y)}\left[\log q_\theta(\hat{x} \mid x)\right]. \tag{4}$$

where $q_\theta\left(\hat{x}|x\right)$ is a variational distribution with parameter $\theta$ to approximate $p\left(\hat{x}|x\right)$. To achieve the decoupling of scRNA and scATAC modality-specific representations and modality-agnostic semantic representations, we constructed two dual modality-specific and semantic encoders (as shown in Fig. 1) and then optimized the MI upper bound between the modality feature representation $\hat{m}$ and the semantic representation $m$ ($m \in (\mathrm{RNA}, \mathrm{ATAC})$:

$$\hat{\mathrm{I}}_{\mathrm{vCLUB}} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\log q_\theta(\hat{m}_i|m_i) - \log q_\theta(\hat{m}_j|m_i)\right]. \tag{5}$$

However, it is difficult to obtain meaningful semantic representations for the semantic encoder by relying solely on this module. Therefore, according to the previous assumption, cross-modal mutual information (MI) should be maximized, as higher cross-modal MI helps the CLUB module achieve better decoupling.

### 3.3 Inter-modal Contrastive Learning

The **RtA** module maximizes a lower bound on the MI between different "views" of the semantic representation of an scRNA-scATAC pair. This is achieved by symmetrically calculating a contrastive

loss from the RNA-to-ATAC direction and the ATAC-to-RNA direction. Formally, we defined $x$ and $y$ to be the semantic representations of the outputs of the scRNA and scATAC semantic encoders, respectively. We aimed to maximize the MI between $x$ and $y$. In practice, we maximized the MI between $x$ and $y$ by minimizing the InfoNCE loss, which is defined as

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}_{p(x,y)}\left[\log \frac{\exp(x \cdot y)}{\sum_{y \in \hat{Y}} \exp(x \cdot y)}\right]. \tag{6}$$

where $Y$ contains the positive sample $y$ and $||y|| - 1$ negative samples drawn from a proposal distribution. Next, we reformulate the loss term for the RtA module based on InfoNCE as follows:

$$\mathcal{L}_{\text{RtA}} = -\frac{1}{2}\mathbb{E}_{p(R,A)}\left[\log \frac{\exp(R \cdot A/\tau)}{\sum_{m=1}^{M} \exp\left(R \cdot A_m/\tau\right)} + \log \frac{\exp(A \cdot R/\tau)}{\sum_{m=1}^{M} \exp\left(A \cdot R_m/\tau\right)}\right]. \tag{7}$$

where $\tau$ is the temperature coefficient. Minimizing $\mathcal{L}_{\text{RtA}}$ is, by contrast, equivalent to maximizing $\mathcal{L}_{\text{inter}}$. Thus, the RtA module can be treated as two perspectives on the semantic representation of scRNA-scATAC pairs, which can be maximized by training the RtA module to maximize $I(x; y)$.

### 3.4 Cross-Modality Unified Codebook

Properties such as scRNA and scATAC, which are inherently discrete data types and possess characteristics somewhat inconsistent with Gaussian assumptions, present a challenge when it comes to understanding cellular heterogeneity quantitatively. This is especially true given that the potential embeddings generated by existing methods are often continuous and may lack direct biological significance Cui et al. (2024). Therefore, to improve the performance of the model in different downstream tasks, we constructed a cross-modal discrete unified codebook. Inspired by SimVQ Zhu et al. (2024), it only adds a simple and efficient linear transformation to the codebook of VQ, which can achieve accelerated convergence and improve the coding utilization rate. We designed Cross-modal VQ (CrossVQ). CrossVQ first initializes a cross-modal shared codebook $e$ where $e \in \{e_1, e_2, \ldots, e_k\}$, along with a randomly initialized learnable weight matrix $W$. For an individual scRNA, its representation after VQ is formulated as:

$$\begin{aligned} z_q^{\text{RNA}} &= z^{\text{RNA}} + \text{sg}[q^{\text{RNA}}W - z^{\text{RNA}}], \\ q^{\text{RNA}} &= \underset{e \in \{e_1, e_2, \cdots, e_K\}}{\text{argmin}} \|z^{\text{RNA}} - eW\|. \end{aligned} \tag{8}$$

Here, $z^{\text{RNA}}$ represents the encoding obtained from the scRNA data. To enable the update of the codebook for better cross-modal alignment, we further designed Cross-modal VQ to update the learnable weight matrix $W$. Focusing on a single modality, the process can be described as follows:

$$\mathcal{L}_{\text{VQ}}^{\text{RNA}} = \|q^{\text{RNA}}W - \text{sg}[z^{\text{RNA}}]\|^2 + \beta\|q^{\text{RNA}}W - \text{sg}[z^{\text{ATAC}}]\|^2. \tag{9}$$

Additionally, the loss function for updating the scRNA encoder is formulated as follows:

$$\mathcal{L}_{\text{encoder}}^{\text{RNA}} = \frac{\beta}{2}\left\|z^{\text{RNA}} - \text{sg}[q^{\text{RNA}}W]\right\|^2. \tag{10}$$

Here, $\beta$ is used to weight the codebook and encoder loss terms, respectively. For scATAC in cross-modal settings, the loss function is similar. Therefore, the overall loss term for CrossVQ can be expressed as follows:

$$\mathcal{L}_{\text{CrossVQ}} = \mathcal{L}_{\text{encoder}}^{m} + \mathcal{L}_{\text{VQ}}^{m}, m \in \{\text{RNA}, \text{ATAC}\}. \tag{11}$$

### 3.5 Overall Training Objection

scCMIA is divided into three main components: decoupling, cross-modal alignment, and reconstruction of the original space data. The scCMIA's framework is shown in Fig. 1. First, intra-modal maximization of mutual information upper bound (Eq. 5) is performed to achieve effective decoupling. Next, RtA and CrossVQ are used for cross-modal alignment in continuous (Eq. 7) and discrete spaces (Eq. 11), facilitating efficient interaction between modalities. Finally, each modality is efficiently reconstructed in the original space.

$$\mathcal{L}_{scCMIA} = \hat{\mathrm{I}}_{\mathrm{vCLUB}} + \mathcal{L}_{\mathrm{RtA}} + \mathcal{L}_{\mathrm{CrossVQ}} + \mathcal{L}_{\mathrm{rec}}. \tag{12}$$

Here $\mathcal{L}_{\mathrm{rec}}$ represents the reconstruction loss for each modality.

## 4 Experiments

In this section, we systematically evaluate the performance of scCMIA. We begin by describing the datasets and evaluation metrics. We then demonstrate the superiority of our framework on the primary tasks of cross-modal alignment and data reconstruction, and verify the efficacy of its key components through ablation studies. To provide a more comprehensive validation, we present additional experiments in the Appendix, which cover reconstruction quality, imputation accuracy, biological discovery, and model convergence.

### 4.1 Datasets and Metrics

**Datasets** Single-cell multi-omics data are often hindered by complex and sophisticated techniques, low throughput, and high noise levels. Therefore, in this paper, we use well-studied single-cell multimodal data from the community for testing purposes. Including 10x Multiome PBMC Genomics (2020), SHARE-seq Ma et al. (2020), SNARE-seq Chen et al. (2019b), and ISSAAC-seq Xu et al. (2022). Detailed information on these datasets is shown in Appendix Table 6.

**Evaluation Metrics** The fraction of samples closer than the true match (**FOSCTTM**) Singh et al. (2020) was used to assess the accuracy of the alignment of the single cell level. A lower FOSCTTM value indicates a higher accuracy in correctly identifying that two modalities originate from the same cell. Additionally, we use Root Mean Square Error (**RMSE**) and Mean Absolute Error (**MAE**) to evaluate the reconstruction performance of the model, of which the lower value indicates a better reconstruction performance. Finally, we also verify that the representations obtained from the latent space contain cell identity information using several clustering metrics including Adjusted Rand Index (**ARI**), Normalised Mutual Information (**NMI**), Adjusted Mutual Information (**AMI**) and the **Homogeneity (HOM)** metric items.

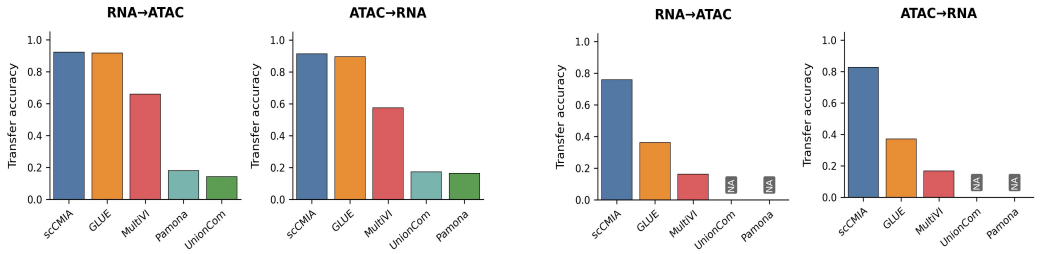
### 4.2 Alignment Experiments

The core objective of multi-modal alignment is to precisely match corresponding cells across modalities like scRNA-seq and scATAC-seq, a task challenged by their inherently heterogeneous data distributions. To systematically evaluate how different alignment strategies preserve this biological correspondence, we conducted extensive comparative experiments, with quantitative results detailed in Table 1.

The results of the multimodal alignment experiment show that our method achieved the best performance in most datasets, and on average, it reduced the error of the best competing alignment method by 26.60%. To evaluate cross-modal label transferability between scRNA-seq and scATAC-seq data, we performed bidirectional cell type annotation transfer experiments. Specifically, we trained a kNN classifier to transfer cell type labels from scRNA-seq to scATAC-seq data (RNA-to-ATAC transfer) and vice versa (ATAC-to-RNA transfer). The comparative results, as summarized in Fig. 2, demonstrate that our scCMIA method significantly outperformed existing methods in both transfer directions. Notably, the high concordance observed between transferred labels across modalities provides strong evidence for successful cross-modal alignment. These findings highlight scCMIA's exceptional capability to preserve biological consistency while integrating heterogeneous single-cell omics data, thereby establishing its effectiveness for multimodal data harmonization tasks.

Table 1: Alignment performance measured by FOSCTTM (mean ± std, lower is better ↓). **Bold** = best, *italic* = second best. Average is mean across non-NA datasets.

| Method | 10X Multiome | ISSAAC-seq | SHARE-seq | SNARE-seq | Average |
|---|---|---|---|---|---|
| MultiVI | 0.2482±0.092 | 0.3679±0.01 | 0.1989±0.003 | 0.2567±0.008 | 0.2346 |
| Seurat v3 | 0.0777±0.0002 | 0.0778±0.0002 | 0.1214±0.001 | 0.2501±0.0004 | 0.1318 |
| GLUE | *0.0172±0.002* | *0.0111±0.002* | *0.0343±0.003* | **0.0127±0.006** | 0.0188 |
| MMD-MA | 0.2998±0.014 | 0.4027±0.049 | — | 0.5280±0.0138 | 0.4102 |
| Pamona | 0.4968 | 0.5007 | — | 0.5025 | 0.5000 |
| UnionCom | 0.5041±0.029 | 0.4875±0.059 | — | 0.4800±0.027 | 0.4905 |
| **scCMIA** | **0.0132±0.008** | **0.0027±0.001** | **0.0165±0.003** | *0.0227±0.006* | **0.0138** |




(a) 10× PBMC (b) SHARE-seq

Figure 2: Bidirectional label transfer accuracy across integration methods.

### 4.3 Reconstruction Experiments

For the data reconstruction evaluation, we performed systematic experiments to validate scCMIA's capability to accurately reconstruct both the scRNA-seq and scATAC-seq data spaces. As evidenced in Tables 2 and 3, scCMIA demonstrates superior multimodal reconstruction performance compared to benchmark methods, achieving the lowest error metrics across both scRNA and scATAC modalities.

Table 2: Reconstruction performance (mean ± std) on scRNA across four random seeds. Lower is better (↓). **Bold** = best, *italic* = second best.

| Method | Align | Recon | Metric | SHARE-seq | SNARE-seq | 10×PBMC | ISSAAC-seq |
|---|---|---|---|---|---|---|---|
| GLUE | ✓ | ✓ | RMSE | 0.5214±0.013 | 0.8298±0.278 | 1.8103±0.005 | 1.3244±0.001 |
| | | | MAE | 0.0978±0.001 | 0.0272±0.136 | **0.1494±0.001** | 0.4807±0.001 |
| MultiVI | ✓ | ✓ | RMSE | 0.9385±0.001 | 0.9322±0.003 | **0.9508±0.001** | 3.5434±0.154 |
| | | | MAE | 0.3018±0.02 | 0.2665±0.005 | 0.3061±0.004 | 2.4638±0.111 |
| Cobolt | | ✓ | RMSE | *0.3832±0.021* | 0.8319±0.003 | 1.8127±0.001 | 1.3261±0.001 |
| | | | MAE | **0.0300±0.001** | **0.0268±0.005** | *0.1499±0.013* | 0.4814±0.001 |
| scMM | | ✓ | RMSE | 0.4824±0.003 | *0.6309±0.008* | 1.7939±0.014 | 1.2817±0.035 |
| | | | MAE | 0.0905±0.0003 | 0.1228±0.0003 | 0.2945±0.007 | 0.2525±0.04 |
| scButterfly | | ✓ | RMSE | 2.1405±0.048 | 1.3840±0.033 | 3.7336±0.050 | *1.2452±0.027* |
| | | | MAE | 0.1464±0.041 | *0.0750±0.001* | 0.2756±0.024 | *0.1918±0.012* |
| **scCMIA** | ✓ | ✓ | RMSE | **0.3213±0.025** | **0.5490±0.003** | *1.0140±0.013* | **0.8931±0.001** |
| | | | MAE | *0.0624±0.014* | 0.0919±0.046 | 0.1591±0.080 | **0.0750±0.043** |

Notably, while scCMIA maintained top-2 ranking on the 10× PBMC dataset reconstruction task, it exhibited marginally higher error metrics compared to GLUE and MultiVI. The observed performance differences were quantitatively minimal, with MAE and RMSE discrepancies of merely 0.0097 and 0.0632 respectively against GLUE. This narrow performance gap suggests comparable reconstruction fidelity among the top-performing methods, while scCMIA maintains a distinct advantage in its capability for simultaneous multimodal reconstruction. Furthermore, the comprehensive cross-modal

Table 3: Reconstruction performance (mean ± std) on scATAC across four random seeds. Lower is better (↓). **Bold** = best, *italic* = second best.

| Method | Align | Recon | Metric | SHARE-seq | SNARE-seq | 10×PBMC | ISSAAC-seq |
|---|---|---|---|---|---|---|---|
| MultiVI | ✓ | ✓ | RMSE | 0.3148±0.001 | 0.2784±0.003 | *1.4572±0.0002* | 1.7726±0.005 |
| | | | MAE | 0.1655±0.001 | 0.2045±0.010 | 0.9447±0.004 | 0.4512±0.004 |
| Cobolt | | ✓ | RMSE | *0.2621±0.001* | *0.2460±0.001* | 1.5283±0.002 | 2.1371±0.161 |
| | | | MAE | **0.0406±0.0003** | 0.0623±0.001 | 0.7039±0.013 | 0.7646±0.005 |
| scMM | | ✓ | RMSE | 0.3730±0.0002 | 0.3451±0.0004 | 1.6689±0.0002 | 3.0510±0.103 |
| | | | MAE | 0.1202±0.0001 | 0.0998±0.0002 | 0.8064±0.002 | *0.4316±0.0001* |
| scButterfly | | ✓ | RMSE | 0.4942±0.001 | 0.5001±0.022 | 1.4746±0.016 | *1.6950±0.008* |
| | | | MAE | 0.4924±0.024 | 0.4995±0.131 | **0.4688±0.021** | 0.4931±0.011 |
| **scCMIA** | ✓ | ✓ | RMSE | **0.2607±0.003** | **0.2459±0.001** | **1.2086±0.016** | **1.1996±0.0001** |
| | | | MAE | *0.0532±0.145* | **0.0561±0.123** | *0.6713±0.019* | **0.2404±0.0167** |

Table 4: Comparison of dependency and direction consistency metrics across multi-omics datasets. Higher values indicate stronger semantic alignment.

| **Metric** | | **Dataset** | | | |
|---|---|---|---|---|---|
| Category | Type | SHARE-seq | SNARE-seq | 10× PBMC | ISSAAC-seq |
| Dependency (MI) | Sem-Sp (scRNA) | 0.0181 | 0.0127 | 0.0791 | 0.0484 |
| | Sem-Sp (scATAC) | 0.0030 | 0.0037 | 0.0048 | 0.0113 |
| | Sem-Sem (RtA) | **0.2466** | **0.1761** | **0.2751** | **0.2226** |
| Direction Consistency (Cosine Sim.) | Sem-Sp (scRNA) | 0.0264 | 0.0134 | -0.0097 | -0.0044 |
| | Sem-Sp (scATAC) | 0.0026 | 0.00001 | 0.0090 | 0.0007 |
| | Sem-Sem (Raw) | -0.0049 | -0.0017 | -0.0039 | -0.0052 |
| | Sem-Sem (RtA) | **0.6087** | **0.5070** | **0.6013** | **0.6818** |

reconstruction performance across all evaluated datasets confirms scCMIA's methodological strength in preserving data integrity during integration processes.

### 4.4 Model Validity Experiments

**Decoupling Effectiveness** The effectiveness of our model's decoupling mechanism is validated through two quantitative approaches. Our primary validation directly assesses the objective of our training strategy by measuring the MI between the resulting latent variables. This confirms that MI was successfully minimized between semantic and modality-specific representations within a modality, while being maximized between the semantic representations across modalities. As an auxiliary method, we also calculate the cosine similarity between these vectors to further verify their statistical independence.

The experimental results are shown in Table 4, which shows that the modal specificity and semantic MI in the single mode are close to 0, which also means that the more independent the two random variables are. The semantic mutual information of the two modalities after RtA has increased by 789. 19% significantly compared to the mutual information within a single modal. Furthermore, the cosine similarity calculated between modality-specific and semantic representations is also nearly zero. The decoupled semantic representation has a higher cosine similarity, which also indicates the consistency of the two representations in terms of direction. These findings strongly validate that the disentangled variables are independent and uncorrelated, which confirms the effectiveness of our decoupling approach. It is also consistent with the assumptions of Eq. 3.

Additionally, we further investigated decoupled modality-specific and semantic representations using clustering tasks. The experimental results are shown in Fig. 3. The results suggest that the semantic representations contain more information related to cell identity, whereas the modality-specific portion of the representations contain less or almost no (vs. scATAC) information related to cell identity. In particular, both the semantic representations of scRNA and scATAC are rich in cell identity

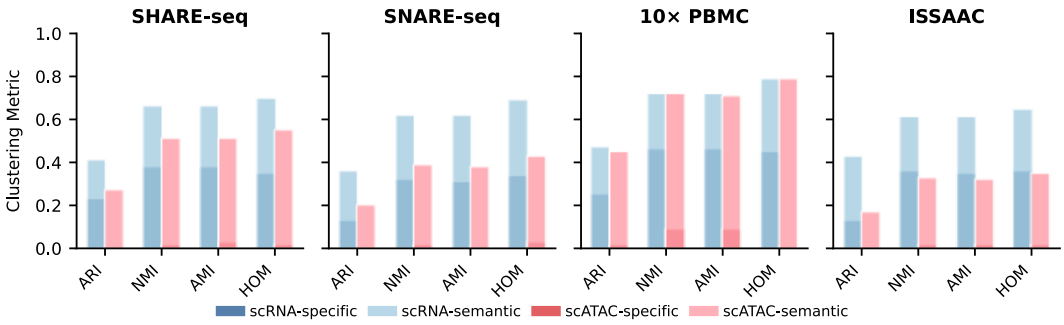



Figure 3: Comparison of clustering results of decoupled **modality-specific** and **semantic** representations across different datasets.

information. Since modality-specific data include unique information pertinent to each modality, this information is essential for subsequent reconstruction tasks to achieve better performance.

**Ablation Study** We also conducted ablation studies on each module of the scCMIA by constructing models that include different combinations of modules. Specifically, the CLUB module is used for intra-modal decoupling, while the RtA and CrossVQ modules are employed for modal alignment and to address the issue of insufficient information within individual modalities. We constructed ablation experiments with different components (in the SHARE-seq dataset) and the experimental results are shown in Table 5. The experimental results were able to observe that the inclusion of the RtA module significantly improves cross-modal alignment, while the inclusion of CrossVQ improves the performance in terms of reconstruction. In addition, the model containing all components (CrossVQ, CLUB, RtA) is able to achieve good performance in both alignment and reconstruction performance.

In addition, we also compare the performance of CLUB+RtA and Full model on the label migration and clustering tasks, and the experimental results are shown in Fig. 4, which shows that scCMIA with Full model is able to have better performance, and the experiments also validate the reasonableness of the individual modules that we have designed. Together, these modules provide a robust framework for handling multi-modal data integration and analysis.

Table 5: Ablation experiments of different modules, among which scCMIA is a full model, including CLUB+RtA+CrossVQ. (FOSCTTM↓, RMSE↓, MAE↓)

| **Components** | **FOSCTTM** | **scRNA** | | **scATAC** | |
|---|---|---|---|---|---|
| | | RMSE | MAE | RMSE | MAE |
| VQ-VAE (Baseline) | 0.4801 | 0.4557 | 0.1060 | 0.2634 | 0.1195 |
| + CrossVQ, CLUB | 0.4945 | **0.3125** | 0.0629 | 0.2677 | 0.0544 |
| + CrossVQ, RtA | 0.0231 | 0.3666 | **0.0612** | 0.2625 | 0.0539 |
| + CLUB, RtA | 0.0178 | 0.3234 | 0.0641 | 0.2612 | 0.0592 |
| scCMIA | **0.0132** | ***0.3213*** | ***0.0624*** | **0.2607** | **0.0532** |

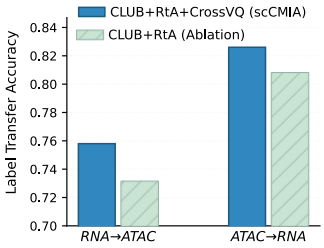



Figure 4: Compare the performance of scCMIA and CLUB + RtA modules on the label transfer task.

## 5 Conclusion

This paper introduces scCMIA, a novel self-supervised framework designed to address the challenges of integrating single-cell multi-omics data. Based on mutual information principles and a unified discrete codebook, this model not only outperforms existing methods in alignment and reconstruction tasks but also pioneers the use of its interpretable latent space for biological exploration. It successfully quantifies regulatory conservation and coupling differences across distinct cell lineages, demonstrating its immense potential as a tool for biological discovery.

**Ethics statement**. This work adheres to the ICLR Code of Ethics. Our research is based on publicly available, anonymized datasets and we foresee no direct negative societal impacts or ethical concerns.

**Reproducibility statement**. All code, model architecture details, and data preprocessing steps required to reproduce our findings are provided in the supplementary materials and detailed in the Appendix.

## REFERENCES


Ricard Argelaguet, Anna SE Cuomo, Oliver Stegle, and John C Marioni. Computational principles and challenges in single-cell data integration. *Nature biotechnology*, 39(10):1202–1215, 2021.

Tal Ashuach, Mariano I Gabitto, Rohan V Koodli, Giuseppe-Antonio Saldi, Michael I Jordan, and Nir Yosef. Multivi: deep generative model for the integration of multimodal data. *Nature Methods*, 20(8):1222–1231, 2023.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018a.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *ICML*, 2018b.

Kai Cao, Xiangqi Bai, Yiguang Hong, and Lin Wan. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics*, 36(Supplement_1):i48–i56, 2020.

Kai Cao, Yiguang Hong, and Lin Wan. Manifold alignment for heterogeneous single-cell multi-omics data integration using pamona. *Bioinformatics*, 38(1):211–219, 2022.

Yichuan Cao, Xiamiao Zhao, Songming Tang, Qun Jiang, Sijie Li, Siyu Li, and Shengquan Chen. scbutterfly: a versatile single-cell cross-modality translation method via dual-aligned variational autoencoders. *Nature Communications*, 15(1):2973, 2024.

Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, 2022.

Huidong Chen, Caleb Lareau, Tommaso Andreani, Michael E Vinyard, Sara P Garcia, Kendell Clement, Miguel A Andrade-Navarro, Jason D Buenrostro, and Luca Pinello. Assessment of computational methods for the analysis of single-cell atac-seq data. *Genome biology*, 20:1–25, 2019a.

Song Chen, Blue B Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, 37(12):1452–1457, 2019b.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pp. 1779–1788. PMLR, 2020.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Hongbo Chi, Marion Pepper, and Paul G Thomas. Principles and therapeutic applications of adaptive immunity. *Cell*, 187(9):2052–2078, 2024.

Xuejian Cui, Xiaoyang Chen, Zhen Li, Zijing Gao, Shengquan Chen, and Rui Jiang. Discrete latent embedding of single-cell chromatin accessibility sequencing data for uncovering cell heterogeneity. *Nature Computational Science*, pp. 1–14, 2024.

Darren A Cusanovich, Riza Daza, Andrew Adey, Hannah A Pliner, Lena Christiansen, Kevin L Gunderson, Frank J Steemers, Cole Trapnell, and Jay Shendure. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, 2015.

Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Gromov-wasserstein optimal transport to align single-cell multi-omics data. *bioRxiv*, 2020. doi: 10.1101/2020.04.28.066787. URL https://www.biorxiv.org/content/early/2020/11/11/2020.04.28.066787.

Zhana Duren, Xi Chen, Mahdi Zamanighomi, Wanwen Zeng, Ansuman T Satpathy, Howard Y Chang, Yong Wang, and Wing Hung Wong. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proceedings of the National Academy of Sciences*, 115(30): 7723–7728, 2018.

10X Genomics. PBMC multiome dataset, 2020. URL https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k.

Boying Gong, Yun Zhou, and Elizabeth Purdom. Cobolt: integrative analysis of multimodal single-cell sequencing data. *Genome biology*, 22:1–21, 2021.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *International Conference on Learning Representations*, 2019.

Bohan Li, Feng Bao, Yimin Hou, Fengji Li, Hongjue Li, and Yue Deng. Tissue characterization at an enhanced resolution across spatial omics platforms with deep generative mode. *Nature communications*, 15(1):6541, 2024.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022.

L Alexander Liggett and Vijay G Sankaran. Unraveling hematopoiesis through the lens of genomics. *Cell*, 182(6):1384–1400, 2020.

Sai Ma, Bing Zhang, Lindsay M LaFave, Andrew S Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K Kartha, Tristan Tay, et al. Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell*, 183(4):1103–1116, 2020.

Kodai Minoura, Ko Abe, Hyunha Nam, Hiroyoshi Nishikawa, and Teppei Shimamura. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell reports methods*, 1(5), 2021.

Yujie Mo, Yajie Lei, Jialie Shen, Xiaoshuang Shi, Heng Tao Shen, and Xiaofeng Zhu. Disentangled multiplex graph representation learning. In *International Conference on Machine Learning*, pp. 24983–25005. PMLR, 2023.

Xin Niu, Enyi Li, Jinchao Liu, Yan Wang, Margarita Osadchy, and Yongchun Fang. Mind the gap: Learning modality-agnostic representations with a cross-modality unet. *IEEE Transactions on Image Processing*, 33:655–670, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. In *International conference on machine learning*, pp. 5102–5112. PMLR, 2019.

Simone Picelli, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, 10(11):1096–1098, 2013.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pp. 5171–5180. PMLR, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ritambhara Singh, Pinar Demetci, Giancarlo Bonora, Vijay Ramani, Choli Lee, He Fang, Zhijun Duan, Xinxian Deng, Jay Shendure, Christine Disteche, et al. Unsupervised manifold alignment for single-cell multi-omics data. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 1–10, 2020.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *cell*, 177(7):1888–1902, 2019.

Xin Tang, Jiawei Zhang, Yichun He, Xinhe Zhang, Zuwan Lin, Sebastian Partarrieu, Emma Bou Hanna, Zhaolin Ren, Hao Shen, Yuhong Yang, et al. Explainable multi-task learning for multi-modality biological data analysis. *Nature communications*, 14(1):2546, 2023.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Xinming Tu, Zhi-Jie Cao, Sara Mostafavi, Ge Gao, et al. Cross-linked unified embedding for cross-modality representation learning. *Advances in Neural Information Processing Systems*, 35: 15942–15955, 2022.

Anna Vettoruzzo, Mohamed-Rafik Bouguelia, Joaquin Vanschoren, Thorsteinn Rognvaldsson, and KC Santosh. Advances and challenges in meta-learning: A technical review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Kevin E Wu, Kathryn E Yost, Howard Y Chang, and James Zou. Babel enables cross-modality translation between multiomic profiles at single-cell resolution. *Proceedings of the National Academy of Sciences*, 118(15):e2023070118, 2021.

Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. Achieving cross modal generalization with multimodal unified representation. *Advances in Neural Information Processing Systems*, 36, 2024.

Zehao Xiong, Jiawei Luo, Wanwan Shi, Ying Liu, Zhongyuan Xu, and Bo Wang. scgcl: an imputation method for scrna-seq data based on graph contrastive learning. *Bioinformatics*, 39(3):btad098, 2023.

Wei Xu, Weilong Yang, Yunlong Zhang, Yawen Chen, Ni Hong, Qian Zhang, Xuefei Wang, Yukun Hu, Kun Song, Wenfei Jin, et al. Issaac-seq enables sensitive and flexible multimodal profiling of chromatin accessibility and gene expression in single cells. *Nature Methods*, 19(10):1243–1249, 2022.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2913–2925, 2024.

Wanwen Zeng, Xi Chen, Zhana Duren, Yong Wang, Rui Jiang, and Wing Hung Wong. Dc3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data. *Nature communications*, 10(1):4613, 2019.

Yongxin Zhu, Bocheng Li, Yifei Xin, and Linli Xu. Addressing representation collapse in vector quantized models with one linear layer. *arXiv preprint arXiv:2411.02038*, 2024.

# A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

This appendix provides a comprehensive overview of the technical implementation of scCMIA. This section includes detailed diagrams of the model architecture, a complete list of training hyperparameters, a description of the data preprocessing pipeline, and the algorithm's pseudocode to ensure full reproducibility.

## A.1 MORE RELATED WORKS

**Maximize Mutual Information Lower Bound**. MI The mutual information neural estimator (MINE) Belghazi et al. (2018b) relies on kernel density estimation of random variables and estimates MI through a neural network to fit the expectation of two distributions. Deep InfoMax (DIM) Hjelm et al. (2019) utilizes autoencoders to learn latent representations of two variables, then estimates MI by maximizing the consistency between the representations. InfoNCE Oord et al. (2018) estimates a lower bound on MI by contrasting the similarity of representations between positive and negative samples. InfoNCE can be combined with deep neural networks to learn complex representations and estimate highly nonlinear MI relationships.

**Minimize Mutual Information Upper Bound**. MI minimization has found wide applications in the disentangled representation learning Von Kügelgen et al. (2021), domain adaptation Vettoruzzo et al. (2024), and information bottleneck methods Tian et al. (2020). However, these methods require closed-form density functions and tractable log-density ratios between the joint and marginal distributions, which limits the exact computation of MI to a few special cases. To address this challenge, sample-based MI estimators Belghazi et al. (2018a); Cherti et al. (2023) have been proposed. For example, L1out Poole et al. (2019) can provide more accurate MI estimates with large sample sizes. However, when applied to MI minimization models, it suffers from high numerical instability. The contrastive log-ratio upper bound (CLUB) Cheng et al. (2020) is a reliable MI estimator that can also be trained within a gradient descent framework. To obtain a tighter upper bound on the MI Yang et al. (2024), we use CLUB to evaluate the upper boundary for better intra-modal decoupling.

## A.2 SCCMIA ALGORITHM

Algorithm 1 provides a clear step-by-step outline of the scCMIA algorithm, emphasizing the key steps such as intra-modal decoupling, cross-modal alignment using contrastive learning and VQ operations, and the calculation and minimization of the total loss term $\mathcal{L}_{\mathrm{scCMIA}}$ to optimize the model parameters $\theta$.

**Algorithm 1** scCMIA

**Input:** scRNA $X$, scATAC $Y$, epoch $N$,
  scRNA modality-specific encoder $\Phi$ and semantic encoder $\hat{\Phi}$,
  scATAC modality-specific encoder $\Psi$ and semantic encoder $\hat{\Psi}$,
  model parameter $\theta$.
Initialize codebook $e \in \{e_1, e_2, \ldots, e_k\}$, learnable parameters $W$
**for** $i = 1$ **to** $N$ **do**
  $\hat{I}_{\mathrm{vCLUB}}, x, \hat{x}, y, \hat{y} \leftarrow \mathrm{CLUB}(X, Y, \Phi, \hat{\Phi}, \Psi, \hat{\Psi})$
  $\mathcal{L}_{\mathrm{RtA}}, x', y' \leftarrow \mathrm{RtA}(x, y)$
  $\mathcal{L}_{\mathrm{CrossVQ}}, x'', y'' \leftarrow \mathrm{CrossVQ}(x', y')$
  $\mathcal{L}_{\mathrm{rec}} \leftarrow \mathrm{Reconstruct\_decoder}(x'', \hat{x}, y'', \hat{y})$
**end for**
$\mathcal{L}_{\mathrm{scCMIA}} \leftarrow \hat{I}_{\mathrm{vCLUB}} + \mathcal{L}_{\mathrm{RtA}} + \mathcal{L}_{\mathrm{CrossVQ}} + \mathcal{L}_{\mathrm{rec}}$
$\theta \leftarrow \arg\min_{\theta} \mathcal{L}_{\mathrm{scCMIA}}$

### A.3 More Experimental Supplements

### A.4 Datasets

Each dataset contains variable sample sizes with different number of cell-types and the dimensions of both scRNA and scATAC are high. In Table 6, we present detailed information about the datasets used in this work, including sample size, dimensions of paired modalities, and cell types.

Table 6: Composition of the experimental datasets.

| Datasets | Sample Size | scRNA Dimension | scATAC Dimension | Cell Type |
|---|---|---|---|---|
| 10x Multiome | 9,631 | 29,095 | 107,194 | 19 |
| SHARE-seq | 32,231 | 21,478 | 340,341 | 22 |
| SNARE-seq | 9,190 | 28,930 | 241,757 | 22 |
| ISSAAC-seq | 10,361 | 32,285 | 169,180 | 23 |

Given the high dimensionality and sparsity issues prevalent in both scRNA and scATAC data, it is necessary to perform feature selection beforehand to better handle the data. For scRNA data, we select 2000 highly variable genes, while for scATAC data, we choose 30,000 high-variance regions as features.

### A.5 Evaluation Metrics

#### A.5.1 Performance Evaluation Metrics

This section introduces the calculation formulas for the Fraction of Samples Closer Than the True Match (FOSCTTM) and matching accuracy (MA).

FOSCTTM is a core alignment performance metric. It is specifically designed to evaluate whether data from two different modalities at the single-cell level has been accurately matched together. A lower FOSCTTM value indicates higher alignment precision of the model. If $N$ cells have true pairwise information, FOSCTTM is defined as

$$\begin{aligned} \text{FOSCTTM} &= \frac{1}{2N}\left(\sum_{i=1}^{N}\frac{n_1^{(i)}}{N}+\sum_{i=1}^{N}\frac{n_2^{(i)}}{N}\right), \\ n_1^{(i)} &= \left|\{j \mid d\left(\mathbf{x}_j,\mathbf{y}_i\right) < d\left(\mathbf{x}_i,\mathbf{y}_i\right)\}\right|, \\ n_2^{(i)} &= \left|\{j \mid d\left(\mathbf{x}_i,\mathbf{y}_j\right) < d\left(\mathbf{x}_i,\mathbf{y}_i\right)\}\right|. \end{aligned} \tag{13}$$

The parameters in the formula are explained as follows:

- $d(\cdot,\cdot)$: A function used to calculate the Euclidean distance.
- $n_1^{(i)}$ and $n_2^{(i)}$: Denote the number of cells that are closer to the $i$-th sample than their true match in the opposite dataset.
- The value of FOSCTTM is in the range of $[0,1]$. Smaller values of FOSCTTM indicate better performance.

Additionally, we tested the accuracy of correctly matching paired samples of another modality under given batch samples from different modal perspectives. We use the MA as a measure, which is defined by the formula:

$$\mathbf{MA} = \frac{1}{N}\frac{1}{B}\sum_{k=1}^{N}\sum_{i=1}^{B}\sum_{j=1}^{B}\mathbb{I}\left(\sum_{m\in\{\mathbf{x},\mathbf{y}\}}\text{Cos}_{\text{sim}}\left(m_i,m_i\right) > \sum_{m\in\{\mathbf{x},\mathbf{y}\}}\text{Cos}_{\text{sim}}\left(m_j,m_i\right) \text{ for all } j\neq i\right). \tag{14}$$

The parameters in the formula are explained as follows:

- $N$: Total number of samples in the dataset.

- $B$: Batch size (number of samples in each batch).
- $\mathbf{x}_i$ and $\mathbf{y}_i$: Represent the two modalities (e.g., scRNA and scATAC) of the $i$-th sample.
- $\text{Cos}_{\text{sim}}(\cdot, \cdot)$: Cosine similarity function.
- $\mathbb{I}(\cdot)$: Indicator function, which takes the value 1 if the condition is true, and 0 otherwise.
- Range of values: [0, 1].

### A.5.2 Biological Interpretability Metrics

CTSI (Cell Type Specificity Index) and Conservation Score are used to reveal whether components such as the model's latent space and encoding capture meaningful patterns consistent with biological knowledge.

Conservation Score measures the frequency with which both scRNA and scATAC modalities map to the same discrete code within the same cell. It reveals the degree of coupling in multimodal regulation within different cell types. Variations in its values constitute biological discoveries, demonstrating that the model has learned deep insights into the distinct regulatory logic of different cell types. It combines two parts: the overlap of codes used and the similarity of their usage frequency distributions. A higher score indicates that two cell types utilize the VQ codebook in a more similar or conserved manner.

Overlap Score Measures the similarity between the sets of VQ codes used by two cell types, $C_i$ and $C_j$.

$$\text{Overlap}(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|}. \tag{15}$$

Distribution similarity measures the similarity between the VQ code frequency vectors, $F_i$ and $F_j$, for two cell types.

$$\text{DistSim}(F_i, F_j) = \frac{F_i \cdot F_j}{\|F_i\|\|F_j\|} = \frac{\sum_{k=1}^{N} F_{ik} F_{jk}}{\sqrt{\sum_{k=1}^{N} F_{ik}^2}\sqrt{\sum_{k=1}^{N} F_{jk}^2}}. \tag{16}$$

Conservation Score is a weighted average of the Overlap Score and the Distribution Similarity. The $\alpha$ is a weighting factor, we set $\alpha = \frac{1}{2}$.

$$\text{Conservation}(i, j) = \alpha \cdot \text{Overlap}(C_i, C_j) + (1 - \alpha) \cdot \text{DistSim}(F_i, F_j). \tag{17}$$

Additionally, CIST measures the specificity with which different cell types utilize specific codes within a unified codebook. A high CTSI value indicates the codebook has successfully learned discrete states that distinguish distinct cellular identities, providing a biological interpretation for the model's black-box interior. It aims to quantify how specific the VQ code usage is for a particular cell type compared to all other cell types. A high CTSI score for a cell type suggests it uses at least one VQ code with a much higher frequency than any other cell type does, indicating a specific signature in the codebook space.

Let $F_i$ be the frequency vector of VQ codes for cell type $i$, and $\mathcal{T}$ be the set of all cell types. The CTSI for cell type $i$ is defined as:

$$\text{CTSI}_i = \frac{\max(F_i) - \max\left(\frac{1}{|\mathcal{T}|-1}\sum_{j \in \mathcal{T}, j \neq i} F_j\right)}{\max(F_i)}. \tag{18}$$

The parameters in the formula are explained as follows:

- $\max(F_i)$ is the maximum frequency of any single code for cell type $i$.
- $\frac{1}{|\mathcal{T}|-1}\sum_{j \in \mathcal{T}, j \neq i} F_j$ is the average frequency vector across all other cell types.

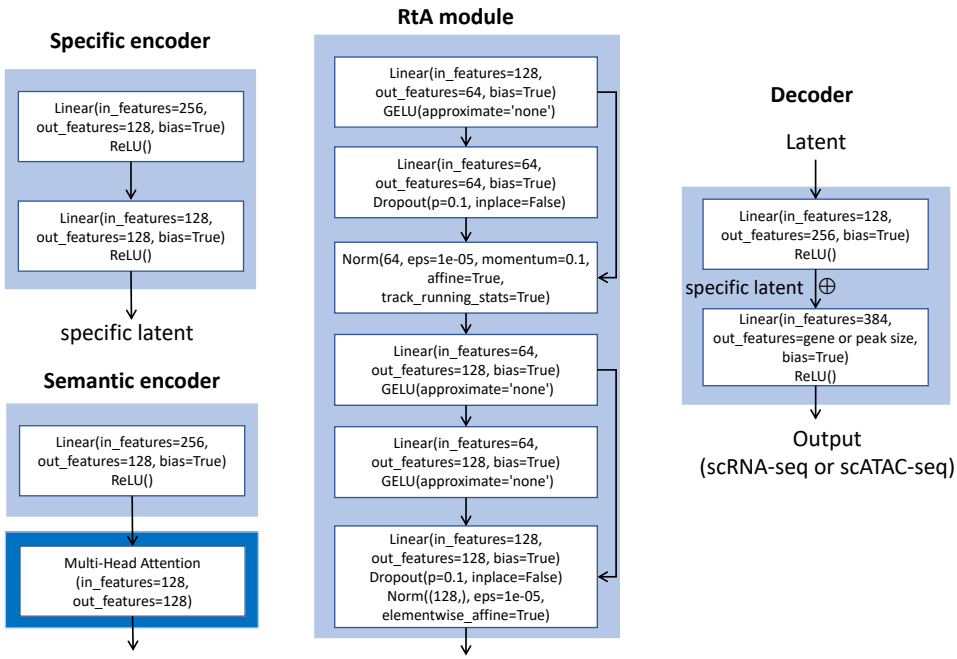



Figure 5: Model structure details. The model architectures and parameter configurations of the domain-specific encoder, semantic encoder, RtA module, and decoder in our proposed framework.

- $\max(\cdot)$ of that average vector gives the highest frequency achieved for any code on average by other cell types.

### A.6 Model architecture and implementation details

In order to improve the reproducibility of the algorithm for easy understanding, we show the architecture of the modal model and the parameter settings in detail in Fig. 5. This includes the modality-specific and semantic coders, the RtA module, and the decoders for each modality.

We implemented scCMIA on an NVIDIA RTX A6000. First, to reduce the model parameters and remove redundant information, we preprocessed scRNA (Principal Component Analysis, PCA) and scATAC (Latent Semantic Indexing, LSI) using a linear-dimensionality reduction method. The raw scRNA and scATAC data are reduced to 256 dimensions and used as input for the scCMIA model. We use Adam as the optimizer with the learning rate set to 0.00001. In the training phase, 10% of the cells were used as a validation set, the number of training iterations was set to 500, we implemented an early stopping mechanism that halts training if the loss does not decrease for 20 consecutive epochs. In addition, to validate the robustness of our method, we set up four different random seeds for the experiments.

The experimental design accounted for methodological differences among comparative frameworks: 1) GLUE exclusively relies on RNA-derived association graphs for cross-modal integration, inherently limiting its scATAC reconstruction capacity (RNA reconstruction metrics only reported); 2) The original Cobolt implementation lacked intrinsic reconstruction functionality, necessitating our implementation of a dedicated reconstruction module to enable fair performance comparison.

To assess the preservation of local structural correspondence between RNA and ATAC modalities, we conducted bidirectional label transfer experiments (RNA→ATAC and ATAC→RNA) following established protocols from reference methods (e.g., Pamona, UnionCom). A k-nearest neighbors (KNN) classifier was trained on low-dimensional embeddings of the source modality to ensure consistency with benchmark implementations. The trained model was then applied to predict cell-type labels using the target modality's embeddings, thereby enabling cross-modality prediction. Label transfer accuracy was systematically quantified across all annotated cell types to evaluate alignment fidelity between modalities. This rigorous framework not only facilitates direct comparison with prior

studies but also objectively measures the mutual translatability of cellular state representations across distinct data types, providing critical insights into multimodal integration performance.

# B Supplementary Experimental Results

This includes a quantitative analysis of how data reconstruction improves data quality, a masking experiment to verify the model's imputation accuracy, a deeper exploration of the biological insights enabled by our model's interpretable discrete codebook, and a visualization of the total loss curves to demonstrate stable model convergence across all datasets.

## B.1 Reconstruction tasks can effectively enhance data quality

To evaluate whether the reconstruction module of scCMIA can effectively improve the quality of raw single-cell data, we compared key data metrics before and after model reconstruction, with particular focus on the issue of data sparsity.

Table 7: Key metrics of scRNA and scATAC data were compared before and after scCMIA reconstruction.

| Data Type | Metric | Raw | Reconstructed | Improvement (%) |
|---|---|---|---|---|
| scRNA | Density (%) | 9.3123 | 9.4842 | +1.8 |
| | Avg. expressed genes/cell | 186 | 189 | +1.7 |
| | Avg. expressed cells/gene | 896 | 913 | +1.9 |
| scATAC | Density (%) | 20.2984 | 38.3894 | +89.1 |
| | Avg. expressed peaks/cell | 6089 | 11516 | +89.1 |
| | Avg. expressed cells/peak | 1954 | 3697.3 | +89.2 |

As shown in the Table. 7, after reconstruction by scCMIA, the data quality of both scRNA-seq and scATAC-seq modalities was significantly improved. The experimental results clearly demonstrate that the reconstruction process effectively reduces data sparsity. For scRNA-seq data, all metrics showed modest yet robust improvements. For scATAC-seq data, which suffers from more severe sparsity, both data density and feature detection rates increased by approximately 89%, demonstrating particularly significant effects. This confirms that our reconstruction module can generate a more complete and information-rich cellular landscape by filling in technologically lost information.

Although the above experiments demonstrate that reconstruction can increase data density, we must verify that this improvement stems from accurate data imputation rather than the filling of random noise. To this end, we designed a masking experiment to directly evaluate the model's imputation capability. We conducted simulations on the 10x PBMC dataset. For each cell, we randomly masked 10% to 30% of its feature values. To ensure fairness in evaluation, masked positions were strictly balanced: half were original non-zero values (to test false negatives/recall), and half were original zero values (to test false positives). The model's task is to predict these masked values. We framed this as a binary classification problem and used Recall and AUROC as evaluation metrics.

Table 8: Interpolation performance under varying masking ratios.

| Masking Ratio | scRNA Recall | scRNA AUROC | scATAC Recall | scATAC AUROC |
|---|---|---|---|---|
| 10% | 0.7361 | 0.8344 | 0.9643 | 0.8119 |
| 15% | 0.7325 | 0.8333 | 0.9643 | 0.8114 |
| 20% | 0.7330 | 0.8323 | 0.9642 | 0.8108 |
| 25% | 0.7261 | 0.8289 | 0.9641 | 0.8106 |
| 30% | 0.7244 | 0.8283 | 0.9640 | 0.8103 |

As shown in the Table. 8 , scCMIA demonstrates robust and powerful interpolation performance even under varying masking ratios. This experiment provides direct quantitative evidence that the increased data density observed earlier is not an artifact but rather a reflection of the model's strong and precise interpolation capabilities. Particularly on scATAC-seq data, the model correctly recovered

approximately 96% of genuinely open chromatin regions (peaks) that were artificially obscured. This robustly confirms that our reconstruction process provides an enhanced, biologically more accurate representation of cellular states by recovering true signals from technical noise.

### B.1.1 Biological Significance Validation and Application Exploration of Unified Codebooks

To validate the biological significance and practical value of the VQ-VAE framework and unified codebook in this study, we designed two supplementary experiments. The first experiment aimed to verify whether the discrete representations learned by the codebook itself possess interpretable biological structures. The second experiment further explored the potential of leveraging these structures for downstream biological knowledge discovery.

First, we quantitatively assessed the intrinsic properties of the unified codebook by introducing two novel metrics: CTSI and Consistency Rate. Experimental results (Fig. 9) demonstrate that high CTSI values (most $> 0.8$) confirm the codebook learned highly specialized, non-generalized discrete encodings for different cell types. Simultaneously, the substantial variation in Consistency Rate across cell types (e.g., as high as 0.98 in memory B cells versus as low as 0.33 in plasma cells) reveals the model's ability to successfully capture diverse regulatory coupling relationships between transcriptomes and chromatin accessibility across cell types, effectively avoiding excessive or forced alignment across different biological states Liggett & Sankaran (2020); Chi et al. (2024). This analysis fundamentally validates the unified codebook as a structured, interpretable layer of biological representation.

Table 9: Cell-type-specific integration (CTSI) scores and consistency rates across modalities.

| Cell Type | scRNA CTSI | scATAC CTSI | Consistency Rate |
|---|---|---|---|
| CD14 Mono | 0.7299 | 0.6116 | 0.3751 |
| CD4 Naive | 0.8448 | 0.8133 | 0.8603 |
| CD8 Naive | 0.8569 | 0.8365 | 0.0185 |
| CD8 TEM_1 | 0.7848 | 0.7578 | 0.6429 |
| HSPC | 0.8127 | 0.8473 | 0.6471 |
| Intermediate B | 0.8664 | 0.8684 | 0.8500 |
| Memory B | 0.8685 | 0.8449 | 0.9765 |
| NK | 0.8577 | 0.8378 | 0.9280 |
| Naive B | 0.8747 | 0.8716 | 0.8560 |
| Plasma | 0.7931 | 0.8632 | 0.3333 |

Second, building upon the validated unified codebook, we introduced the regulatory Conservation Score (CS) to quantify the similarity of regulatory programs across cell types, aiming to test the model's capability for biological knowledge discovery. This results ((Fig. 10)) successfully reproduced known cellular lineage relationships, cells within the same lineage (e.g., B cell subpopulations) obtained high RCS scores, while scores between different lineages (e.g., lymphoid and myeloid) were lower. More importantly, this approach reveals finer biological insights, such as quantitatively distinguishing functional differences among distinct monocyte subpopulations and capturing shared cytotoxic programs between NK cells and CD8 TEM_1 cells. This experiment demonstrates that our model transcends mere data integration, serving as a quantitative exploration tool to generate novel insights into cellular regulatory networks.

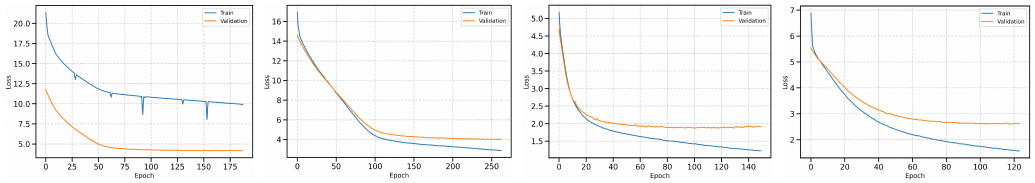
To sum up, these two complementary experiments form a complete chain of reasoning. The first experiment establishes the structural validity and interpretability of the discrete code book in our method, demonstrating it is not a complex component designed for novelty's sake. The second experiment demonstrates the functional utility of this structure, proving it can serve as a powerful tool for discovering and quantifying cellular regulatory logic. Together, they confirm that our proposed scCMIA framework not only excels at alignment and reconstruction tasks but also delivers profound biological insights, opening new analytical dimensions for single-cell multi-omics research.

Table 10: Conservation scores (CS) and biological interpretation across cell type pairs. Higher CS indicates stronger cross-modality alignment.

| Category | Cell Type Pair | RNA CS | ATAC CS | Interpretation |
|---|---|---|---|---|
| High Conservation (B cell) | Naive B vs. Plasma | 0.774 | 0.847 | Strong conservation in B-cell development |
| High Conservation (T cell) | CD8 Naive vs. CD8 TEM_1 | 0.510 | 0.551 | Conservation across T-cell subtypes |
| Low Conservation (Distant) | CD4 Naive vs. CD14 Mono | 0.180 | 0.095 | Lymphoid vs. myeloid programs |
| Nuanced Insights | CD14 Mono vs. CD16 Mono | 0.361 | 0.365 | Subtle differences between monocyte subtypes |
| | NK vs. CD8 TEM_1 | 0.410 | 0.509 | Shared cytotoxic program |

### B.1.2 Model Convergence Analysis

The overall training objective function of the cCMIA framework comprises multiple components, resulting in a complex training process. To validate the model's convergence, we visualized the evolution of the total training loss across epochs on all four benchmark datasets (10x Multiome PBMC, SHARE-seq, SNARE-seq, ISSAAC-seq).




(a) 10× PBCM (b) ISSAA-seq (c) SHARE-seq (d) SNARE-seq

Figure 6: Total loss ($\mathcal{L}_{scCMIA}$) convergence curve of scCMIA across four datasets.

As shown in Fig. 6, the model's training process remains stable even across datasets with varying complexity and feature differences. This provides robust assurance for scCMIA's broad applicability across diverse single-cell multi-omics datasets.

### B.1.3 Alignment Performance Experiment Supplement

In experiments with a fixed batch size of 56, we compared various contrastive methods based on their top1 and top5 matching accuracy for two modalities. The results in Table 11 indicate that our method, scCMIA, demonstrates robust alignment performance in both top1 and top5 across multiple datasets under this batch size. These findings further confirm the effectiveness of scCMIA.

Table 11: Performance comparison of different methods on multi-omics datasets.

| **Methods** | **Modality** | **10X Multiome** | | **ISSAAC-seq** | | **SHARE-seq** | |
|---|---|---|---|---|---|---|---|
| | | **Top1** | **Top5** | **Top1** | **Top5** | **Top1** | **Top5** |
| GLUE | RNA→ATAC | 0.6732 | 0.9662 | 0.7692 | 0.9775 | 0.1510 | 0.4120 |
| | ATAC→RNA | 0.6282 | 0.9575 | 0.7264 | 0.9705 | 0.1686 | 0.4271 |
| Pamona | RNA→ATAC | 0.0184 | 0.0881 | 0.0163 | 0.0836 | NA | NA |
| | ATAC→RNA | 0.0195 | 0.0927 | 0.0174 | 0.0847 | NA | NA |
| UnionCon | RNA→ATAC | 0.0171 | 0.0817 | 0.0148 | 0.0217 | NA | NA |
| | ATAC→RNA | 0.0080 | 0.0590 | 0.0736 | 0.0953 | NA | NA |
| MMD_MA | RNA→ATAC | 0.0430 | 0.1791 | 0.0676 | 0.2676 | NA | NA |
| | ATAC→RNA | 0.0304 | 0.1642 | 0.0633 | 0.2697 | NA | NA |
| scCMIA | RNA→ATAC | **0.7146** | **0.9855** | **0.8595** | **0.9958** | **0.6073** | **0.9552** |
| | ATAC→RNA | **0.7191** | **0.9840** | **0.8632** | **0.9967** | **0.6137** | **0.9548** |

### B.1.4 CROSS-DATASET ZERO-SHOT EXPERIMENTS

In order to apply the pre-trained model in a realistic scenario, we performed the zero-shot task. The experiment was set up as a zero-shot experiment on SNARE-seq (Cortex tissue of mouse) with data trained on SHARE-seq (Skin tissue of mouse). And we used the results obtained by Pamona, UnionCon, and MMD_MA trained on SNARE-seq data as Baseline. The experimental results are shown in Fig. 7. scCMIA demonstrates optimal matching accuracy on doing the zero-shot task, although there is a gap compared to directly on the original dataset, which may be due to the variability between datasets, tissues, and cell types resulting in the limited migration ability of the model, which requires a larger scale and diversity of data to train the model in order to effectively improve the transfer ability of the model.

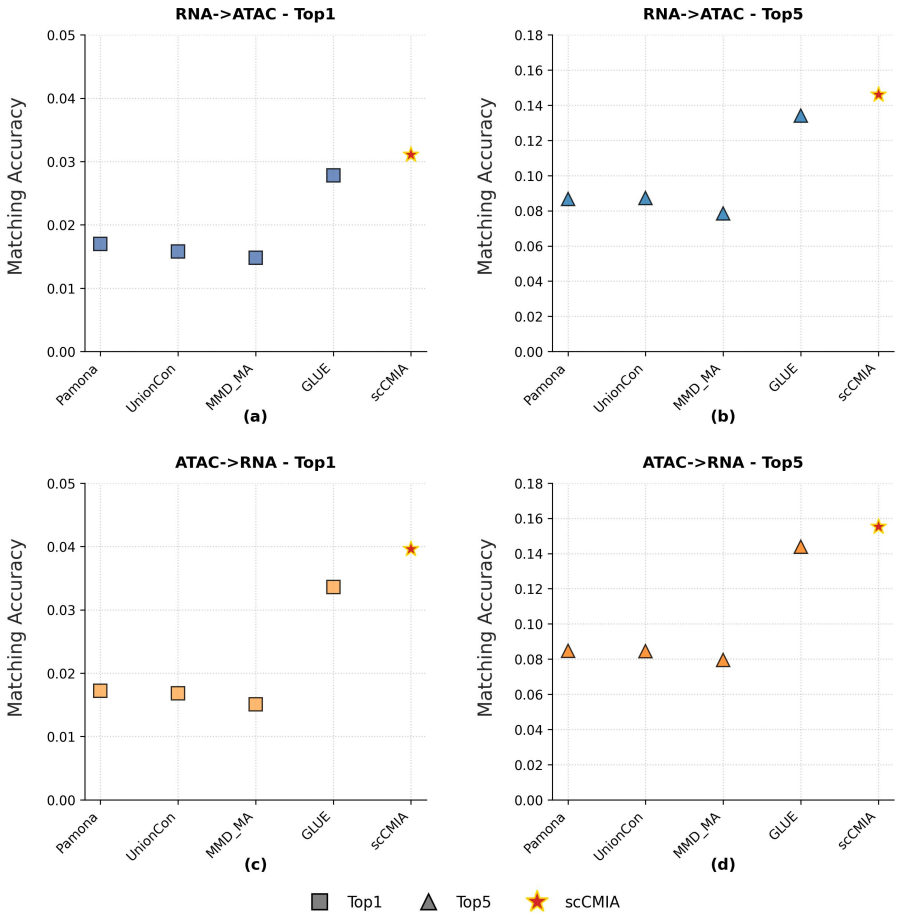



Figure 7: SHARE-seq trained models are made into pair matching zero-shot experiments on the SNARE-seq dataset.

## C USE OF LLMS

We utilized a large language model (LLM) to assist in the writing and editing of this paper. The LLM's role was strictly limited to improving grammar, phrasing, and overall readability. All scientific contributions, including the core ideas, methodology, and interpretation of results, are solely the work of the authors.