MultiGAI: Global Attention-Based Integration of Single-Cell Multi-Omics with Batch Effects Correction

Jinyu Zhang

Abstract

Single-cell multi-omics technologies can measure multiple molecular layers within the same cell, such as gene expression and chromatin accessibility (scRNA-seq + scATAC-seq) or gene expression and cell-surface protein abundances (scRNA-seq + ADT), providing a multidimensional view of cellular states and regulatory mechanisms. However, these modalities often differ substantially in their distributions and noise levels, and they are further affected by technical biases introduced during experimental batches and sample handling, which may obscure true biological signals. To address these challenges, we propose MultiGAI, a variational autoencoder (VAE) framework that incorporates a global attention mechanism. MultiGAI incorporates global information from the dataset during encoding and uses specially designed components to restrict the propagation of batch-specific information. This design enables effective batch effects correction while preserving essential biological signals, resulting in high-quality latent representations of cells. Overall, MultiGAI offers a new strategy that jointly preserves biological information and corrects batch effects, providing fresh insights for future methods in single-cell multi-omics integration.

Keywords: single-cell multi-omics integration, batch effects correction, global attention

1 Introduction

The development of single-cell multi-omics technologies has significantly advanced our understanding of cellular states and their regulatory mechanisms. Compared with traditional single-omics sequencing, these technologies enable the simultaneous capture of multiple layers of molecular information within the same cell, such as scRNA-seq combined with scATAC-seq (e.g., 10x Multiome [1]) or scRNA-seq combined with ADT (e.g., CITE-seq [2]), providing complementary perspectives on cellular function and identity. Leveraging these data, researchers can not only dissect the molecular characteristics of different cell types within heterogeneous tissues but also further uncover cross-modality regulatory mechanisms and explore dynamic cellular changes in physiological and pathological processes [3, 4].

Nevertheless, integrating single-cell multi-omics data remains challenging. On one hand, different modalities exhibit substantial differences in dimensionality, noise characteristics, and measurement ranges; on the other hand, technical biases arising from experimental preparation, sequencing platforms, and batch effects can obscure true biological signals and compromise the accuracy of downstream analyses [3]. Therefore, effectively correcting for batch effects while obtaining high-quality low-dimensional representations of cells during integration remains a critical problem.

To address this challenge, various methods have been proposed to correct for batch effects and reduce their interference in single-cell multi-omics integration. Batch effects correction strategies can be conceptually likened to the architecture of a variational autoencoder (VAE) [5]. Depending on where the correction is

applied, these methods can be categorized into three types: encoder-level, latent space, and decoder-level batch effects correction. Among them, latent space and decoder-level batch effects correction are the most commonly used approaches. For instance, Seurat [6] employs latent space batch effects correction by first identifying biologically similar cell pairs (anchors) across different batches in the low-dimensional space, and then using these anchors to align cells from all batches into a shared space. Meanwhile, MultiVI [7] and Multigrate [8] adopt decoder-level batch effects correction: they inject batch information into the decoder and constrain it through the loss function to achieve batch effects correction. More recently, MIMA [9] adopts an encoder-level batch effects correction strategy by partitioning the latent space into three subspaces: shared, modality-specific, and batch-related. This design achieves batch effects correction by disentangling the latent representations.

Nevertheless, these methods still have certain limitations. For latent space batch effects correction strategies, no additional mechanism is applied during dimensionality reduction. As a result, batch information may be entangled with biological signals in the latent representations, which makes it challenging to ensure that biological information is fully preserved during correction. For decoder-level batch effects correction strategies, which are commonly used in current deep learning methods, some batch information often remains in the latent space due to the reconstruction loss, potentially weakening the effectiveness of batch effects correction. In contrast, encoder-level batch effects correction strategies hold greater potential. Although information loss is inevitable during dimensionality reduction, if this loss mainly involves batch information, more biological signals can be retained in the latent space, resulting in higher-quality latent representations. Taking MIMA [9] as an example, although it employs an encoder-level batch effects correction strategy, the encoder lacks additional mechanisms, which allows biological information to spill over into the batch subspace, resulting in some loss of biological signals. Moreover, non-deep-learning methods typically focus more on inter-sample relationships and the global structure of the dataset, thereby better preserving overall information, which is crucial for effective single-cell multi-omics integration. In contrast, existing deep learning integration methods often treat each sample independently, leveraging such information less. Taken together, incorporating an encoder-based dimensionality reduction strategy that suppresses the flow of batch information, while considering the global information of the dataset, can effectively correct batch effects and preserve more biological signals in the latent space, producing high-quality latent representations.

Based on the above analysis, we propose MultiGAI, a variational autoencoder framework incorporating a global attention mechanism. The core innovation lies in the encoder: by leveraging global attention, the encoder can capture global information during dimensionality reduction and use this structure to suppress the propagation of batch-specific signals, thereby better preserving biological information in the latent representations. In addition, the model incorporates mechanisms to handle technical noise, such as sequencing depth [10], mitigating the impact of non-biological factors on integration performance. Experimental results demonstrate that MultiGAI achieves a balanced trade-off between batch effects correction and the preservation of biological signals in single-cell multi-omics data integration.

2 Methods

2.1 Global Attention Mechanism

Currently, the attention mechanism in Transformers [11] is widely used in tasks that require understanding global relationships due to its strong capability for contextual modeling. This mechanism dynamically weights information through interactions among queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$), thereby enhancing the model's expressive power. In single-cell multi-omics data, we aim for the model to capture the contextual information of each cell. However, if all cells are treated as the key-value set $\{\mathbf{K}, \mathbf{V}\}$, both computational

complexity and memory usage increase dramatically with the number of cells, making training difficult.

To address this issue, our method draws on the core idea of Perceiver [12] by introducing a global, trainable, and fixed-size key-value set $\{\mathbf{K}, \mathbf{V}\}$, which serves as a compressed representation of the entire dataset.

From a mechanistic perspective, in conventional attention mechanisms, the query vector $\mathbf{q}$ of each token interacts with the key-value set $\{\mathbf{K}, \mathbf{V}\}$ to generate the representation of that token. Here, the key-value set is composed of the key vectors $\mathbf{k}$ and value vectors $\mathbf{v}$ of all tokens. However, in biological contexts, cells exhibit significant correlations. Therefore, a fixed-size key-value set $\{\mathbf{K}, \mathbf{V}\}$ can serve as a compact approximation of the full key-value set containing all cells.

This approximation is essentially a lossy representation. However, during training, through the reconstruction loss and the explicit injection of batch information in the decoder, the model does not need to encode batch-related information in $\{\mathbf{K}, \mathbf{V}\}$. Consequently, it is forced to prioritize biologically relevant information within the limited capacity, thereby forming an information bottleneck [13, 14] that suppresses the flow of batch information.

From a biological perspective, the observed state of each cell can be assumed to be composed of several fundamental cellular states. We aim to enable the key-value set $\{\mathbf{K}, \mathbf{V}\}$ to learn these shared states and construct latent representations for each cell accordingly, thereby obtaining a latent space with biological interpretability.

Specifically, MultiGAI learns a query vector $\mathbf{q}^{(m)}$ for the input data $\mathbf{x}^{(m)}$ of each modality m, which interacts with the key-value set of that modality via attention to produce a modality-specific intermediate representation $\mathbf{e}^{(m)}$. Subsequently, the intermediate representations of all modalities are averaged element-wise to obtain a joint query vector $\mathbf{q}$. This joint query vector then interacts with the joint key-value set through attention to generate the joint representation $\mathbf{e}$, which is further used to parameterize the distribution of the joint latent variable $\mathbf{z}$ via separate neural networks for its mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$.

The advantages of this method include:

1. **Efficient global context integration**: The fixed global key-value set serves as a shared memory for the dataset, allowing each sample's query to access a set of condensed cross-sample features. This effectively enhances the cross-sample consistency and generalization of the latent representations, while avoiding the high computational cost of conventional attention mechanisms on large datasets.

2. **Dynamic information selection and information bottleneck**: Through the interaction between the query vector $\mathbf{q}$ and the global key set $\{\mathbf{K}\}$, the model can dynamically weight and extract the most relevant information from the global value set $\{\mathbf{V}\}$ for the current sample. Meanwhile, the global attention mechanism forms an information bottleneck [13, 14] during dimensionality reduction, effectively suppressing irrelevant batch information. As a result, the latent variable $\mathbf{z}$ flexibly integrates global information while retaining richer biologically relevant signals, thereby improving reconstruction quality and downstream analysis performance.

2.2 Model Architecture

MultiGAI is a single-cell multi-omics integration method based on the variational autoencoder (VAE) framework [5], incorporating a global attention mechanism and using a zero-inflated negative binomial (ZINB) distribution [10] as the observation model.

The overall workflow is as follows: MultiGAI first learns a query vector $\mathbf{q}^{(m)}$ for each modality's single-cell data $\mathbf{x}^{(m)}$, which interacts with the modality-specific key-value pairs $\{\mathbf{K}^{(m)}, \mathbf{V}^{(m)}\}$ via attention [11] to produce a modality-specific intermediate representation $\mathbf{e}^{(m)}$. The intermediate representations from all

modalities are then averaged element-wise to obtain a joint query vector $\mathbf{q}$, which attends to the joint key-value pairs $\{\mathbf{K}, \mathbf{V}\}$ to generate a unified representation $\mathbf{e}$. Based on this unified representation, separate neural networks learn the parameters (mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$) of the latent variable's distribution, and the latent variable $\mathbf{z}$ is then sampled from this distribution using the reparameterization trick [5]. Finally, the latent variable $\mathbf{z}$ is concatenated with the one-hot encoded batch labels and fed into the decoder to generate the ZINB distribution parameters for each modality [10], which are then used to reconstruct the original data while accounting for sequencing depth or total counts for each modality. The model is trained in an end-to-end manner by maximizing the log-likelihood of the observations for each modality while simultaneously minimizing the KL divergence between the latent distribution and the prior. In addition, to enhance consistency across intermediate representations of different modalities, a cosine similarity-based inter-modality consistency constraint is introduced, further improving the quality of the joint latent representation. Figure 1 illustrates the overall architecture of the MultiGAI model. The detailed mathematical formulations of the model, including the encoder, latent variable sampling, decoder, ZINB [10] likelihood function, and loss function, are provided in Section B.

3 Results

To comprehensively evaluate the performance of MultiGAI, we assessed it on two single-cell multi-omics datasets, focusing on latent space quality, query mapping, and feature prediction.

The datasets are described as follows:

- **Dataset 1** [15]: Multiome (scRNA-seq + scATAC-seq) data derived from bone marrow mononuclear cells (BMMCs) of healthy human donors. This dataset contains 69,249 cells from 13 batches, 13,431 genes, and 116,490 peaks, covering 22 distinct cell types.
- **Dataset 2** [15]: CITE-seq (scRNA-seq + ADT) data, also derived from BMMCs of healthy donors. This dataset contains 90,261 cells from 12 batches, 13,953 genes, and 134 ADT features, covering 45 distinct cell types.

3.1 Latent Variable Quality

To comprehensively evaluate the quality of the latent variables learned by MultiGAI, we employed the scIB [16] framework for integration quality assessment. This framework includes multiple metrics. We selected six core metrics covering batch effects correction and biological conservation (see Appendix C for detailed formulas).

As baseline methods, we selected MultiVI [7] (and its CITE-seq version totalVI [17]), GLUE [18], MOFA+ [19], MIMA [9], and Seurat [6]. For MIMA [9], the experimental setup followed the default configuration in the official tutorial, with the random seed set to 42 and the latent dimension set to 50. For GLUE [18], we evaluated each metric based on the integrated scRNA-seq and scATAC-seq latent variables separately. For the remaining methods, aside from fixing the random seed at 42, all configurations followed the recommendations in Multigrate [8].

The configurations of MultiGAI and its variants are as follows: the latent dimension is set to 30, and the network consists of a single hidden layer structure with a linear layer, LayerNorm, ReLU activation, and Dropout, with a hidden layer width of 128. The number of key-value pairs is 128, and both the query vectors and value vectors have a dimension of 128. The model is trained for 200 epochs, with α representing the weight of the cosine similarity-based inter-modality consistency loss and β representing the weight of the

KL divergence loss. For the first 100 epochs, $\alpha = 1$ and $\beta = 0$, and for the remaining 100 epochs, β is set to 0.1. The random seed is fixed at 42, and the input consists of raw counts from each modality.

In this experiment, data preprocessing for GLUE [18] was performed according to its official tutorial. For all other methods, we followed the strategy of Multigrate [8], selecting 4,000 highly variable genes and 20,000 highly variable peaks from the data.

Based on the experimental results, MultiGAI achieved the highest overall scores on both datasets, demonstrating its strong ability to correct batch effects while preserving biological variation (See Table 1 and Figures 2–3).Specifically, in Dataset1, except for MIMA [9], MultiGAI achieves near-top scores in biological preservation while clearly outperforming other methods in batch effect correction. For MIMA [9], although it attains the highest batch effect correction score, its biological preservation score is slightly lower. In Dataset2, MultiGAI leads in biological preservation scores while maintaining a relatively high level of batch effect correction. Overall, MultiGAI is able to generate high-quality latent representations in single-cell multi-omics integration tasks with batch effects.

3.2 Query Mapping

To evaluate the query mapping capability of MultiGAI, we designed three experiments on each of the two datasets mentioned above, and two experiments on the combined dataset obtained by integrating these two datasets.

3.2.1 Mapping of Single-Modality Data

For Dataset1 and Dataset2, we performed stratified sampling based on batch labels and cell type labels to obtain three subsets corresponding to data containing only scRNA-seq, only scATAC-seq (or only ADT), and both modalities, with relative proportions of 3:3:4, respectively. The data containing both modalities were used as the training set, and the trained model was then applied to infer the entire dataset. The UMAP [20] visualization was generated as follows: UMAP [20] parameters were first trained using the latent variables from the training set, and then the entire dataset was projected onto this UMAP [20] space. This experiment aims to evaluate the model's ability to map data when only a single modality is provided.

Since the joint query vector in MultiGAI is the mean of the intermediate representations from all modalities, directly using the intermediate representation of each modality as the joint query vector is still inappropriate, even though modality consistency was enforced during training via a cosine similarity loss. Therefore, we added a simple neural network for each modality in MultiGAI to learn the mapping from that modality's intermediate representation to the joint query vector. Specifically, after training the model on the training set, we collected the intermediate representations of each modality along with their corresponding joint query vectors, trained each neural network using mean squared error (MSE) loss, and then used these networks to map the intermediate representations of each modality to the joint query vectors.

From the UMAP visualizations [20] (see Figures 4 and 5), it can be observed that cells containing only a single modality are evenly mixed in the latent space with cells that have multi-omic data, without any noticeable structural shifts. These results indicate that MultiGAI is able to effectively map cells into a latent space consistent with multi-omic data, even when only a single modality is provided as input.

3.2.2 Mapping of Unseen Cell Types

For Dataset1 and Dataset2, we removed a cell type located at the edge of the UMAP [20] space and another located near the center, resulting in two datasets per original dataset, each lacking a specific cell type. These datasets with missing cell types were used as the training sets, and the trained models were then applied to

infer the entire dataset. UMAP [20] visualizations were generated by projecting the latent variables of the full dataset. This experiment aims to evaluate the model's ability to map unseen cell types.

From the UMAP [20] visualizations (See Figures 6 and 7), we observe that the held-out cell types do not mix with other distinct cell types; instead, they form separate clusters whose relative positions are largely consistent with those obtained using the full data. These results indicate that MultiGAI exhibits strong generalization ability and can correctly identify cell types that are similar to, but not present in, the training set.

3.2.3 Integration and Mapping of Multiome and CITE-seq

To further evaluate the integration capability of MultiGAI, we first constructed a tri-modal dataset in which each sample contains only two modalities, based on the two datasets described above. Specifically, we first integrated the shared modality (scRNA-seq) from the two datasets, selected the common genes, and identified 4,000 highly variable genes using the same procedure as before. Next, for each sample, the missing modality was filled with zeros. This resulted in a tri-modal dataset containing 4,000 genes, 20,000 peaks (selected from Dataset1), and 134 ADT features. For different multi-omics datasets, we added modality-type labels to the existing batch labels in order to distinguish samples originating from different omics sources during integration.

After completing the integration experiment, we further performed a query mapping experiment. Following the stratified sampling strategy used in the *Mapping of Single-Modality Data* subsubsection, we partitioned the integrated dataset into subsets with a 1:1:1:2 ratio, corresponding to samples containing only scRNA-seq, only scATAC-seq, only ADT, and Multiome or CITE-seq, respectively. The proportion of Multiome and CITE-seq samples was kept consistent with their proportion in the integrated dataset. We then visualized the entire dataset using the same training procedure and UMAP [20] workflow as in the *Mapping of Single-Modality Data* subsubsection.

From the UMAP [20] visualizations (see Figure 8, where the top subfigures correspond to the integration task and the bottom subfigures correspond to the query mapping task), we observe that in both the integration and query mapping experiments, cells from different modalities are well mixed and cluster according to cell type. This demonstrates that MultiGAI possesses strong integration and mapping capabilities for single-cell multi-omics datasets that share at least one common modality.

3.3 Feature Imputation

To evaluate the feature imputation capability of MultiGAI, we conducted experiments on the integrated dataset used for query mapping. We selected several sets of canonical immune cell marker genes, their corresponding surface protein markers, and representative chromatin accessible regions to systematically assess MultiGAI's imputation performance across scRNA-seq, ADT, and scATAC-seq modalities. Specifically, we focused on the following features:

- T cell–related features: the marker gene *CD3G*, the corresponding surface marker *CD3* [21], and a chromatin accessible region associated with *CD3G* expression, `chr11:118,343,914-118,344,801` [7];
- NK cell–related features: the marker gene *FCGR3A*, the corresponding surface marker *CD16*, which is primarily expressed in NK cells but can also be detected in CD16+ monocytes (non-classical/intermediate) [22];
- B cell–related features: the marker gene *MS4A1* and the corresponding surface marker *CD20* [23].

As shown in the figures (see Figures 9 and 10), the marker gene, surface protein, and chromatin accessible region for T cells all clearly delineate the T cell population; similarly, the marker gene and surface protein for B cells effectively identify the B cell population. For NK cells, both the marker gene *FCGR3A* and the surface marker *CD16* label NK cells as well as CD16+ monocytes, consistent with known biological observations [22].

Notably, even when many cells contain only a single modality (scRNA-seq, scATAC-seq, or ADT), the gene expression, surface protein abundance, and chromatin accessibility imputed by MultiGAI exhibit consistent and stable patterns within the corresponding cell populations.

3.4 Ablation Study

To evaluate the effectiveness of the global attention mechanism in MultiGAI, we designed an ablation model, MultiGAI_N, which has the same architecture and loss functions as MultiGAI but does not utilize the global attention mechanism. The metrics and datasets used in the ablation study are consistent with those in the latent variable quality experiments.

From the metric plots and tables (See Figure 11 and Table 2), we observe that while MultiGAI and MultiGAI_N achieve similar scores in biological conservation, MultiGAI significantly outperforms MultiGAI_N in batch effects correction. This indicates that incorporating the global attention mechanism can substantially improve batch effects correction while maintaining biological information, thereby demonstrating its effectiveness.

To evaluate the ability of the global attention mechanism to suppress batch effects while preserving the overall structure of the dataset, we extracted the joint query vectors, joint representations, and the set of joint global value vectors obtained from the integration experiments described in the *Integration and Mapping of Multiome and CITE-seq* subsection and performed visualization analysis. The UMAP plot (see Fig. 12) shows that the joint query vectors, when colored by cell type, still exhibit pronounced batch effects, whereas the joint representations processed by the global attention module effectively eliminate these batch effects, indicating that the global attention mechanism can effectively suppress the propagation of batch information. Furthermore, the UMAP plot (see Fig. 13) demonstrates that the joint global value vectors can effectively capture the overall structure of the dataset's joint representations, suggesting that the global attention mechanism preserves the dataset's global information.

4 Conclusion and Outlook

To address the limitations of current single-cell multi-omics integration methods, we proposed a variational autoencoder (VAE) [5] framework incorporating a global attention mechanism—MultiGAI. MultiGAI effectively removes batch effects while preserving biological information, thereby enhancing the integration of single-cell multi-omics data and providing strong integration and latent space mapping capabilities.

Currently, MultiGAI has been successfully applied to the integration of scRNA-seq, scATAC-seq, and ADT data, but it has not yet been extended to other omics types such as metabolomics [24, 25] and spatial omics [26,27]. Future studies could focus on how to leverage the biological characteristics of different omics and design corresponding mechanisms to incorporate them into the MultiGAI framework.

In addition, the current model still involves a considerable number of hyperparameters. Clarifying the specific impact of each hyperparameter on model performance and exploring more efficient hyperparameter selection strategies (e.g., adaptive tuning based on dataset characteristics) would help improve the model's generalizability and application efficiency.

It should also be noted that the MultiGAI decoder primarily reconstructs the proportional structure of each modality rather than raw counts. In future work, we plan to further optimize the decoder to explore more accurate reconstruction methods, thereby broadening the applicability of the model.

Data Availability

The datasets used in this study are publicly available and can be accessed from the NCBI GEO database under accession number GSE194122.

Code Availability

The source code for this project is available on our GitHub: `https://github.com/Recklesszjy/Multigai`.

References

[1] C. Zhu, S. Preissl, and B. Ren. Integrated single-cell analysis of transcriptional and epigenetic states in human hematopoiesis. *Nature Biotechnology*, 38:1601–1610, 2020.

[2] M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14:865–868, 2017.

[3] C. Liu, S. Ding, H.J. Kim, and et al. Multitask benchmarking of single-cell multimodal omics integration methods. *Nature Methods*, 22:2449–2460, 2025.

[4] S. Fu, S. Wang, D. Si, and et al. Benchmarking single-cell multi-modal data integrations. *Nature Methods*, 22:2437–2448, 2025.

[5] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2014. Preprint on arXiv.

[6] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck III, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zagar, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar B. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 2021.

[7] Tom Ashuach et al. MultiVI: Deep Generative Model for the Integration of Multimodal Data. *Nature Methods*, 18(7):787–795, 2021.

[8] Anastasia Litinetskaya, Maiia Schulman, Fabiola Curion, Artur Szalata, Alireza Omidi, Mohammad Lotfollahi, and Fabian Theis. Integration and querying of multimodal single-cell data with poe-vae. In S. Sankararaman, editor, *Research in Computational Molecular Biology. RECOMB 2025. Lecture Notes in Computer Science*, volume 15647. Springer, Cham, 2025.

[9] Jose Ignacio Alvira Larizgoitia, Gabriele Partel, Lorenzo Venturelli, Wanqiu Zhang, Xander Spotbeen, Sebastiaan Vanuytven, Sam Kint, Katy Vandereyken, David Wouters, Anis Ismail, Regis Scarceriaux, Jakub Idkowiak, Tassiani Sarretto, Shane R. Ellis, Massimo Loda, Fabio Socciarelli, Thomas Gevaert, Steven Joniau, Marc Claesen, Nico Verbeeck, Thierry Voet, Johannes V. Swinnen, Jelle Jacobs, and Alejandro Sifrim. Multi-omics integration and batch correction using a modality-agnostic deep learning framework. *bioRxiv*, 2025. Preprint.

[10] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef. Deep Generative Modeling for Single-Cell Transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[12] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, Demis Hassabis, and Koray Kavukcuoglu. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning (ICML)*, 2021.

[13] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377, 1999.

[14] Naftali Tishby and Naftali Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015.

[15] Daniel B. Burkhardt, Malte D. Luecken, Christopher Lance, et al. Single-Cell Multiomics Data of Human Bone Marrow Mononuclear Cells. GEO Accession: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE194122`, 2022.

[16] Malte D. Luecken and Fabian J. Theis. Benchmarking Atlas-Level Data Integration in Single-Cell Genomics. *Nature Methods*, 16(12):1394–1401, 2019.

[17] Antonio Gayoso, Zachary Steier, Romain Lopez, et al. Joint probabilistic modeling of single-cell multiomic data with totalvi. *Nature Methods*, 18:272–282, 2021.

[18] C. Wang et al. Multi-Omics Single-Cell Data Integration and Regulatory Inference with Graph-Linked Embedding. *Nature Communications*, 13:448, 2022.

[19] Ricard Argelaguet et al. MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1):111, 2020.

[20] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. In *arXiv preprint arXiv:1802.03426*, 2018.

[21] Abul K Abbas, Andrew H Lichtman, and Shiv Pillai. *Immunology: Basic and Clinical Principles*. Elsevier, 2020.

[22] Lewis L Lanier. Nk cell recognition. *Annual Review of Immunology*, 23:225–274, 2005.

[23] Thomas F Tedder. B lymphocyte development and function. *Annual Review of Immunology*, 21:551–592, 2003.

[24] Gary J. Patti, Oscar Yanes, and Gary Siuzdak. Metabolomics: The Apogee of the Omics Trilogy. *Nature Reviews Molecular Cell Biology*, 13(4):263–269, 2012.

[25] C. H. Johnson, J. Ivanisevic, and G. Siuzdak. Metabolomics: Beyond Biomarkers and Towards Mechanisms. *Nature Reviews Molecular Cell Biology*, 17(7):451–459, 2016.

[26] Patrik L Ståhl, Fredrik Salmén, Siniša Vickovic, Anders Lundmark, Johan F Navarro, Jakob Magnusson, Stefania Giacomello, Mikael Asp, Jakob O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

[27] Samuel G Rodriques, Robert R Stickels, Alexandro Goeva, Carson A Martin, Eli Murray, Charles R Vanderburg, Justin Welch, Leo M Chen, Fei Chen, and Evan Z Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.

A Model Architecture

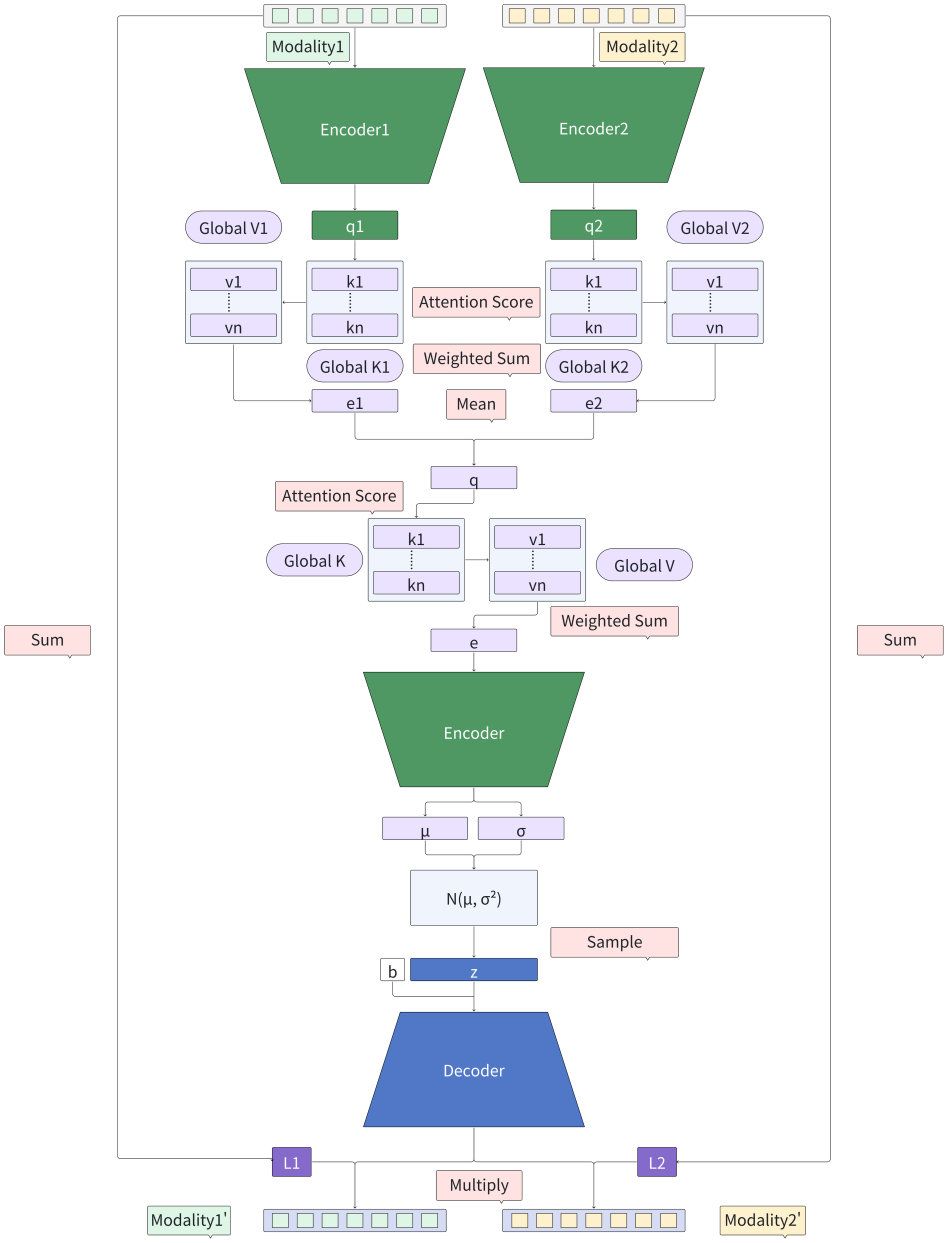


Figure 1: Overall architecture of the MultiGAI model

B Detailed Model Formulation

Below we provide the detailed mathematical formulation of the MultiGAI model, including the encoder, latent variable sampling, decoder, ZINB likelihood, and loss function.

B.1 Encoder

The encoder maps single-cell multi-omics inputs into a unified latent space using a hierarchical global attention mechanism. For a single sample with multi-omics inputs $\{\mathbf{x}^{(m)}\}_{m=1}^{M}$, where M denotes the number of modalities, the encoder first maps each modality to a query vector:

$$\mathbf{q}^{(m)} = f_{\phi}^{(m)}(\mathbf{x}^{(m)}), \qquad \mathbf{q}^{(m)} \in \mathbb{R}^{1\times d_q}. \tag{1}$$

A global trainable key-value set is defined for each modality:

$$\{\mathbf{K}^{(m)}, \mathbf{V}^{(m)}\}, \quad \mathbf{K}^{(m)} \in \mathbb{R}^{n\times d_q}, \quad \mathbf{V}^{(m)} \in \mathbb{R}^{n\times d_v}, \tag{2}$$

where n is the number of key-value pairs, d_q is the query dimension, and d_v is the value dimension.

The modality-specific intermediate representation is computed via attention:

$$\mathbf{e}^{(m)} = \text{softmax}\left(\frac{\mathbf{q}^{(m)}{\mathbf{K}^{(m)}}^{\top}}{\sqrt{d_q}}\right)\mathbf{V}^{(m)} \in \mathbb{R}^{1\times d_v}. \tag{3}$$

Intermediate representations from all modalities are averaged to obtain a joint query vector:

$$\mathbf{q} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{e}^{(m)} \in \mathbb{R}^{1\times d_v}. \tag{4}$$

A joint global key-value set is defined:

$$\{\mathbf{K}, \mathbf{V}\}, \quad \mathbf{K} \in \mathbb{R}^{n\times d_v}, \quad \mathbf{V} \in \mathbb{R}^{n\times d_v}. \tag{5}$$

The unified representation is computed by attention between the joint query vector $\mathbf{q}$ and the joint key-value set:

$$\mathbf{e} = \text{softmax}\left(\frac{\mathbf{q}\mathbf{K}^{\top}}{\sqrt{d_v}}\right)\mathbf{V} \in \mathbb{R}^{1\times d_v}. \tag{6}$$

Finally, the joint representation $\mathbf{e}$ is fed into two separate linear layers to predict the mean and standard deviation of the latent variable:

$$\boldsymbol{\mu} = h_{\mu}(\mathbf{e}), \qquad \boldsymbol{\sigma} = h_{\sigma}(\mathbf{e}), \tag{7}$$

where $h_{\mu}, h_{\sigma} : \mathbb{R}^{d_v} \to \mathbb{R}^{d}$ and d is the latent dimension.

B.2 Latent Variable Sampling

We employ the reparameterization trick to sample latent variables from a Gaussian distribution:

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{8}$$

where $\mathbf{z} \in \mathbb{R}^{1\times d}$ denotes the sampled latent vector.

B.3 Decoder

The decoder takes as input the concatenation of the latent variable $\mathbf{z}$ and the batch label $\mathbf{b}$:

$$\mathbf{h} = [\mathbf{z}, \mathbf{b}] \in \mathbb{R}^{1\times(d+B)}, \tag{9}$$

where $\mathbf{b} \in \{0,1\}^B$ is a one-hot encoding of the batch label and B is the total number of batches.

For each modality m, the decoder predicts the parameters of a zero-inflated negative binomial (ZINB) distribution:

$$(\boldsymbol{\pi}^{(m)}, \boldsymbol{\mu}^{(m)}, \mathbf{r}^{(m)}) = g_\theta^{(m)}(\mathbf{h}), \tag{10}$$

where $\boldsymbol{\mu}^{(m)}$ is normalized via softmax to represent relative expression proportions.

The sequencing depth for modality m is:

$$l^{(m)} = \sum_{i=1}^{D_m} x_i^{(m)}, \tag{11}$$

and the expected expression is scaled:

$$\boldsymbol{\mu}'^{(m)} = l^{(m)} \cdot \boldsymbol{\mu}^{(m)}. \tag{12}$$

B.4 ZINB Log-Likelihood

The log-likelihood of a single feature under the ZINB distribution:

$$\log p(x_i) = \begin{cases} \log\left[\pi_i + (1-\pi_i)\left(\frac{r_i}{r_i+\mu_i'}\right)^{r_i}\right], & x_i = 0, \\ \log(1-\pi_i) + \log\Gamma(x_i+r_i) - \log\Gamma(r_i) - \log\Gamma(x_i+1) & \\ \quad + x_i \log\mu_i' - x_i\log(r_i+\mu_i') + r_i \log\left(\frac{r_i}{r_i+\mu_i'}\right), & x_i > 0, \end{cases} \tag{13}$$

Total log-likelihood for all features in modality m:

$$\log p(\mathbf{x}^{(m)}) = \sum_{i=1}^{D_m} \log p(x_i), \tag{14}$$

where $\mathbf{x}^{(m)} = [x_1, \ldots, x_{D_m}]^\top$.

B.5 Loss Function

The training objective minimizes a composite loss:

$$\mathcal{L}(\theta,\phi) = \underbrace{\sum_m -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}^{(m)} \mid \mathbf{h})\right]}_{\text{ZINB reconstruction loss}} + \underbrace{\beta\,\mathrm{KL}\Big(\mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2)) \,\|\, \mathcal{N}(\mathbf{0},\mathbf{I})\Big)}_{\text{latent variable KL divergence}} + \underbrace{\gamma\,\mathcal{L}_{\text{cos}}}_{\text{cross-modality consistency loss}}, \tag{15}$$

where $\mathbf{h} = [\mathbf{z}, \mathbf{b}]$, $\beta, \gamma > 0$, and $\mathcal{L}_{\text{cos}}$ is the cosine similarity loss between intermediate representations of different modalities.

C Metrics

To comprehensively evaluate the quality of the latent variables learned by MultiGAI, we employed the scIB [16] framework for integration quality assessment. We selected six core metrics covering two key dimensions:

- **Batch-effect Correction Metrics** (weight: 40%):
  - `Batch ASW`: Batch mixing silhouette width
  - `iLISI`: Local inverse Simpson' s index for batch mixing
  - `kBET`: k-nearest neighbor batch effect test
- **Biological Conservation Metrics** (weight: 60%):
  - `Cell-type ASW`: Cell-type silhouette width
  - `NMI`: Normalized mutual information (between clustering and true labels)
  - `ARI`: Adjusted Rand index (between clustering and true labels)

The overall score is calculated as:

$$\text{Overall Score} = 0.4 \times \underbrace{\left(\frac{\text{Batch ASW} + \text{iLISI} + \text{kBET}}{3}\right)}_{\text{Batch-effect Correction}} + 0.6 \times \underbrace{\left(\frac{\text{Cell-type ASW} + \text{NMI} + \text{ARI}}{3}\right)}_{\text{Biological Conservation}} \quad (16)$$

D Latent Variable Quality

Table 1: Latent variable quality metrics

Dataset	Method	Score	Bat	Bio	ASW_b	iLISI	kBET	NMI	ARI	ASW
Multiome	MultiGAI	**0.6342**	0.5777	0.6719	0.9201	0.3428	0.4703	0.7679	**0.6965**	0.5513
Multiome	GLUE_ATAC	0.5990	0.5102	0.6582	0.8854	0.3114	0.3339	0.7456	0.6534	0.5755
Multiome	GLUE_RNA	0.5960	0.4704	**0.6797**	0.8789	0.2836	0.2486	**0.7808**	0.6781	0.5801
Multiome	MIMA	0.6171	**0.6018**	0.6274	**0.9663**	0.3409	**0.4980**	0.7313	0.6378	0.5132
Multiome	MOFA+	0.4392	0.2967	0.5343	0.8190	0.0246	0.0464	0.6424	0.4158	0.5447
Multiome	MultiVI	0.5897	0.4663	0.6719	0.8637	0.2827	0.2527	0.7730	0.6572	**0.5856**
Multiome	Seurat	0.5869	0.5068	0.6403	0.8051	**0.3499**	0.3655	0.7343	0.6013	0.5854
CITE-seq	MultiGAI	**0.6273**	0.5118	**0.7044**	0.9154	0.2914	0.3284	**0.7946**	**0.7624**	0.5562
CITE-seq	MIMA	0.5962	0.5015	0.6593	**0.9524**	0.2725	0.2798	0.7633	0.6883	0.5264
CITE-seq	MOFA+	0.4206	0.2988	0.5018	0.8040	0.0135	0.0789	0.6328	0.3025	0.5701
CITE-seq	Seurat	0.5846	**0.5352**	0.6176	0.8014	**0.3722**	**0.4319**	0.7341	0.5288	**0.5898**
CITE-seq	totalVI	0.6001	0.5154	0.6566	0.9013	0.3063	0.3385	0.7865	0.6246	0.5586

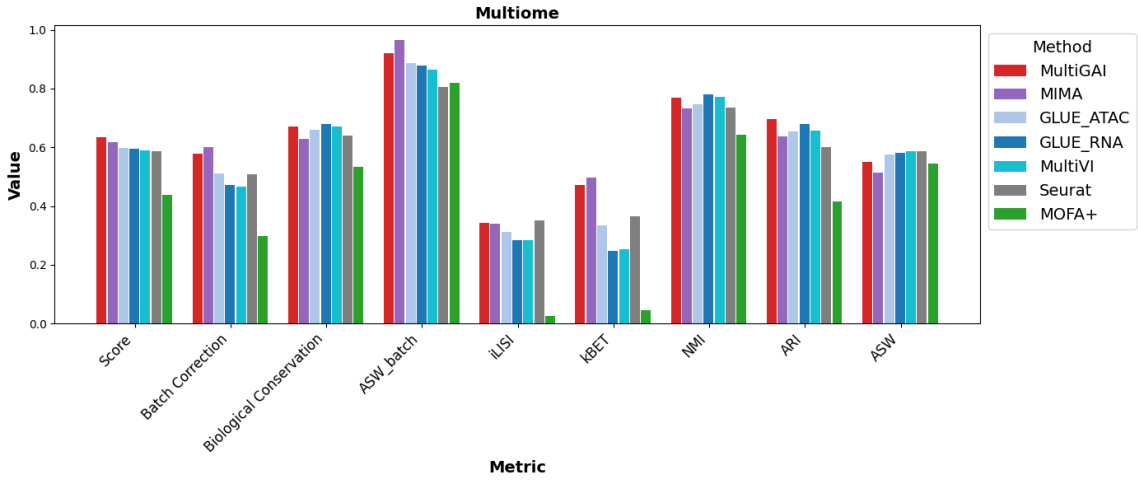


Figure 2: Latent variable quality metrics for Multiome

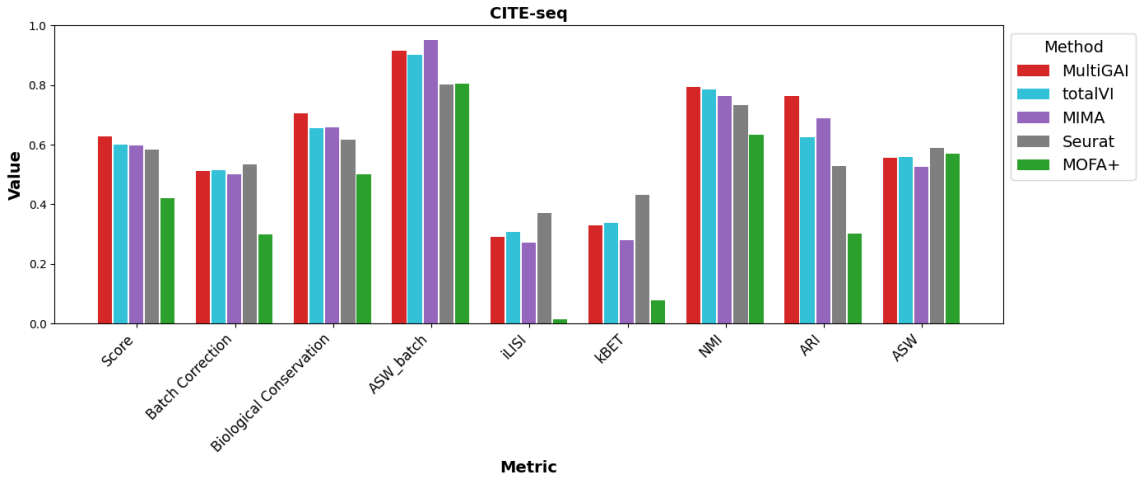
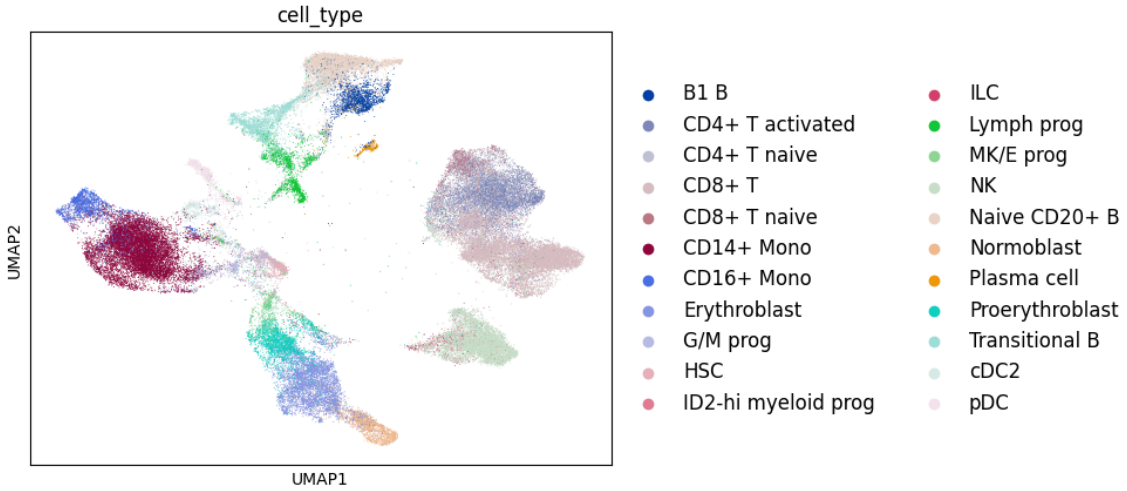


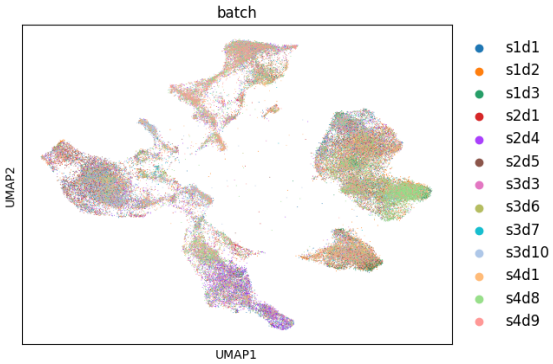
Figure 3: Latent variable quality metrics for CITE-seq

E Query Mapping

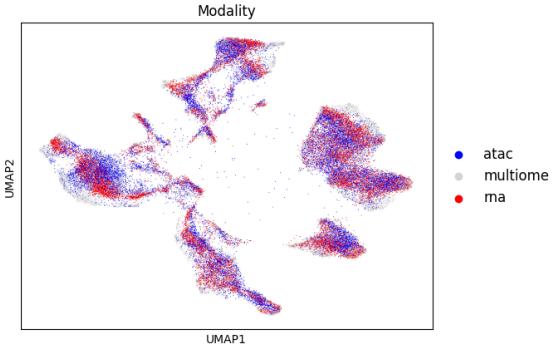
E.1 Mapping of Single-Modality Data



(a) UMAP plot of cell labels for Multiome

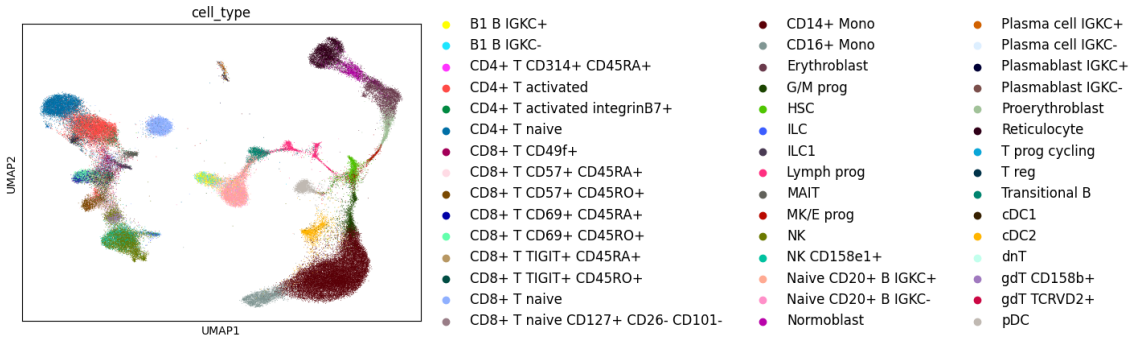


(b) UMAP plot of batch labels for Multiome

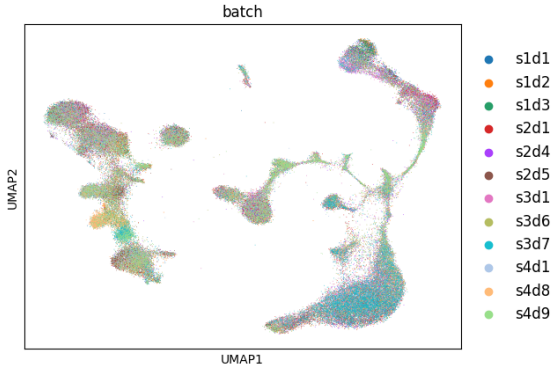


(c) UMAP plot of modality labels for Multiome

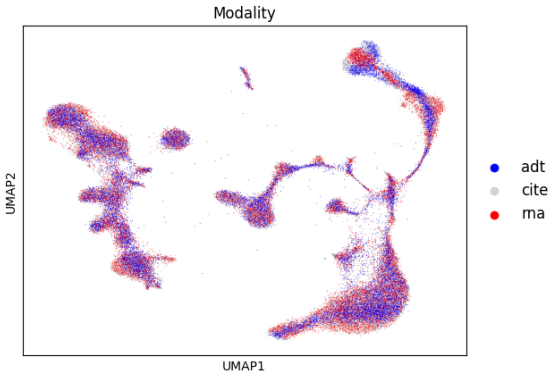
Figure 4: UMAP plots for Multiome dataset: cell labels (top), batch and modality labels (bottom).



(a) UMAP plot of cell labels for CITE-seq



(b) UMAP plot of batch labels for CITE-seq

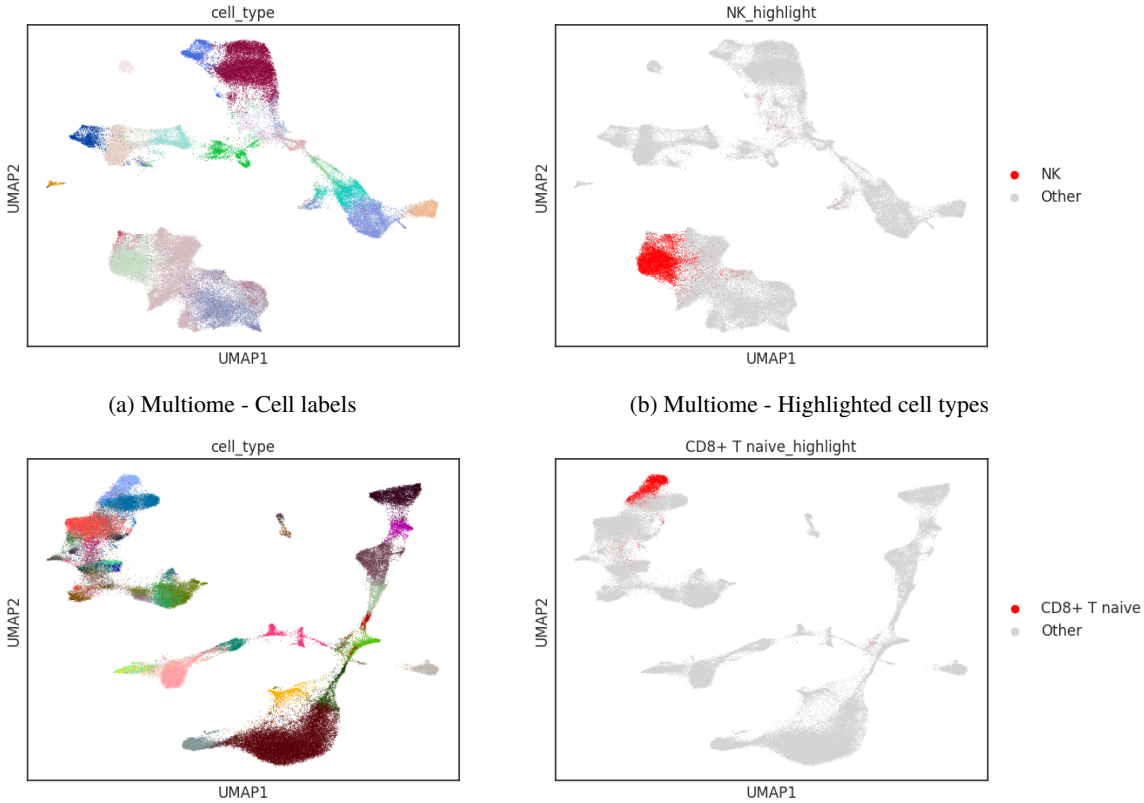


(c) UMAP plot of modality labels for CITE-seq

Figure 5: UMAP plots for CITE-seq dataset: cell labels (top), batch and modality labels (bottom).

E.2 Mapping of Unseen Cell Types

E.2.1 Edge Cells



(a) Multiome - Cell labels

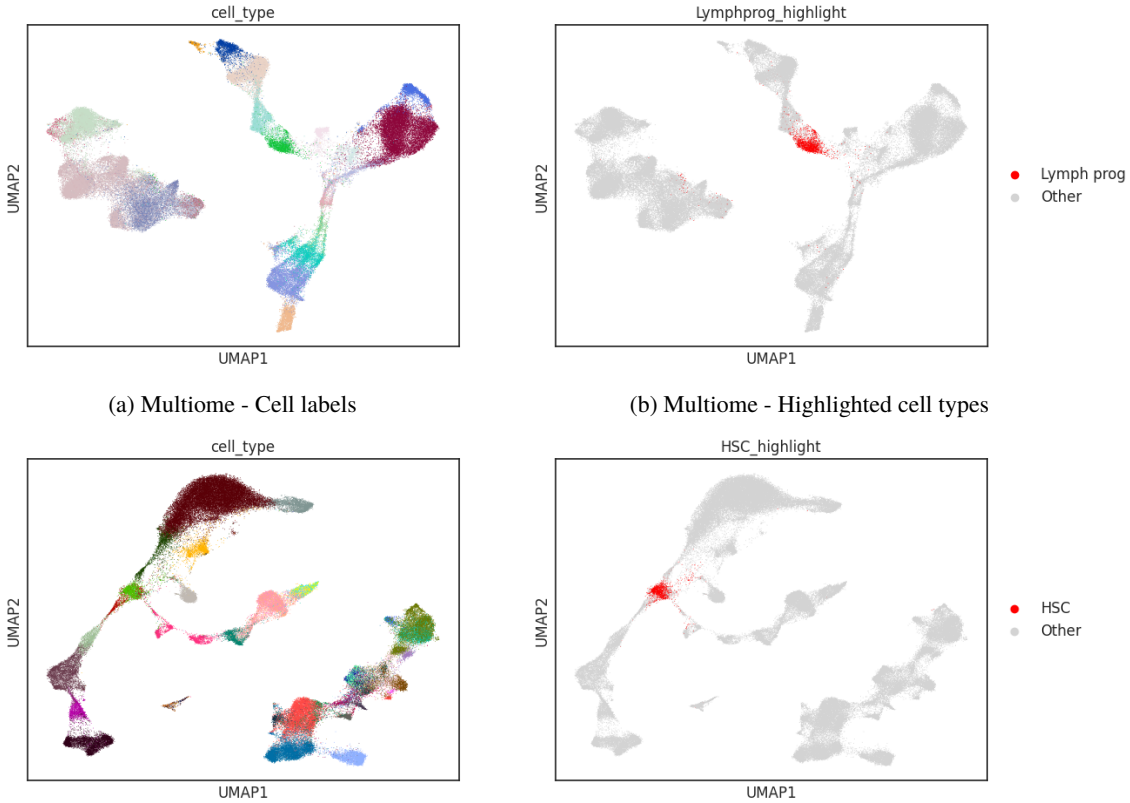
(b) Multiome - Highlighted cell types

(c) CITE-seq - Cell labels

(d) CITE-seq - Highlighted cell types

Figure 6: UMAP plots of edge cell labels and highlighted cell types for Multiome and CITE-seq datasets.

E.2.2 Center Cells



(a) Multiome - Cell labels

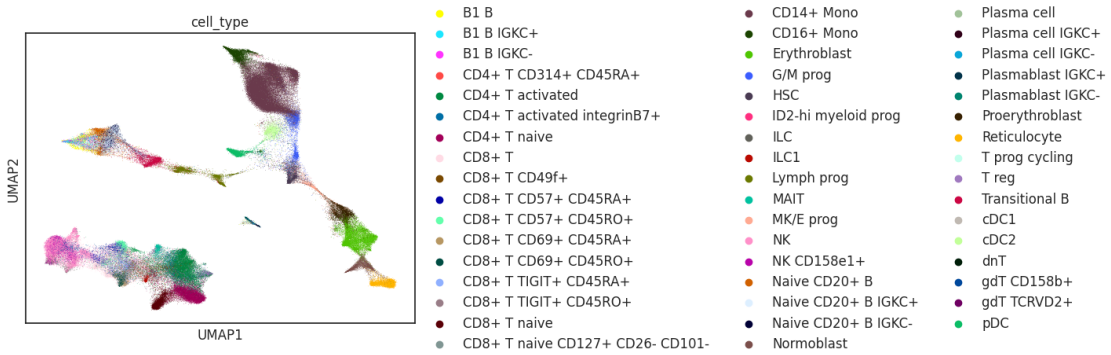
(b) Multiome - Highlighted cell types

(c) CITE-seq - Cell labels

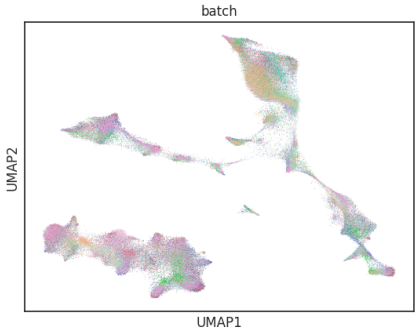
(d) CITE-seq - Highlighted cell types

Figure 7: UMAP plots of center cell labels and highlighted cell types for Multiome and CITE-seq datasets.

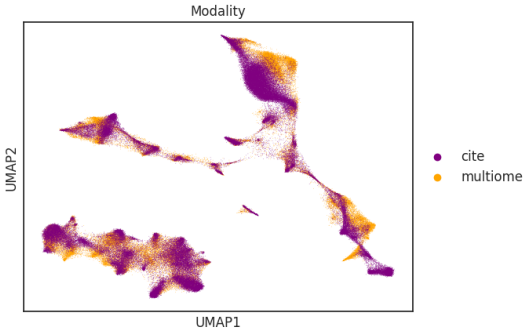
E.3 Integration and Mapping of Multiome and CITE-seq



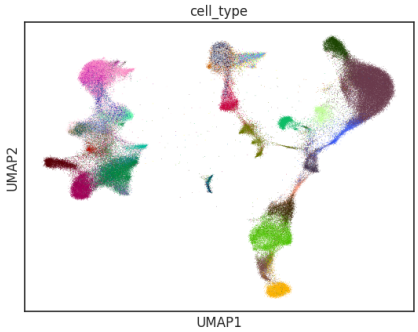
(a) UMAP plot of cell labels for integration



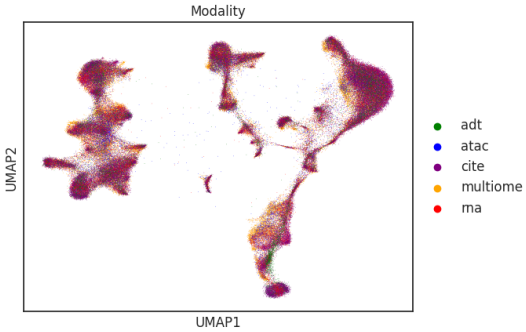
(b) UMAP plot of batch labels for integration



(c) UMAP plot of modality labels for integration



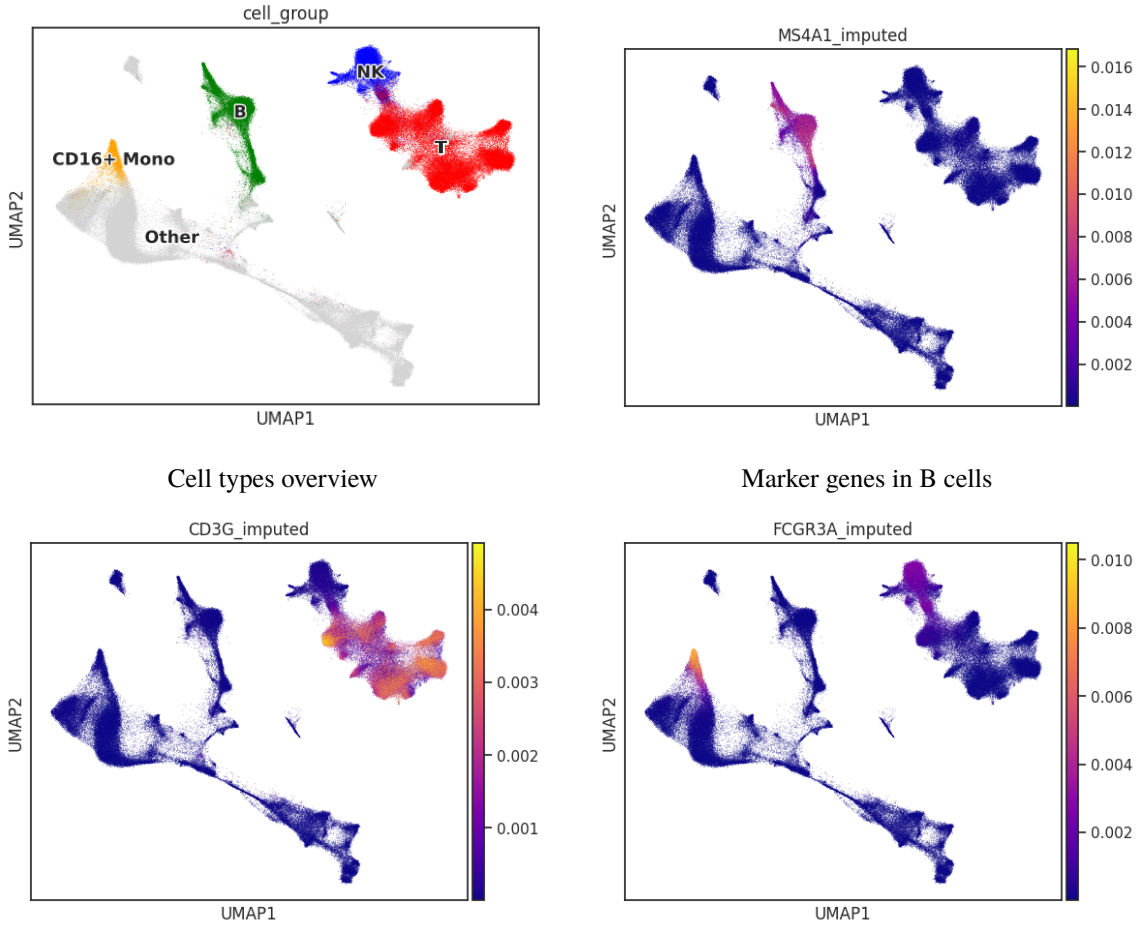
(d) UMAP plot of cell labels for query mapping



(e) UMAP plot of modality labels for query mapping

Figure 8: Integration and query mapping of Multiome and CITE-seq datasets.

F Feature Imputation



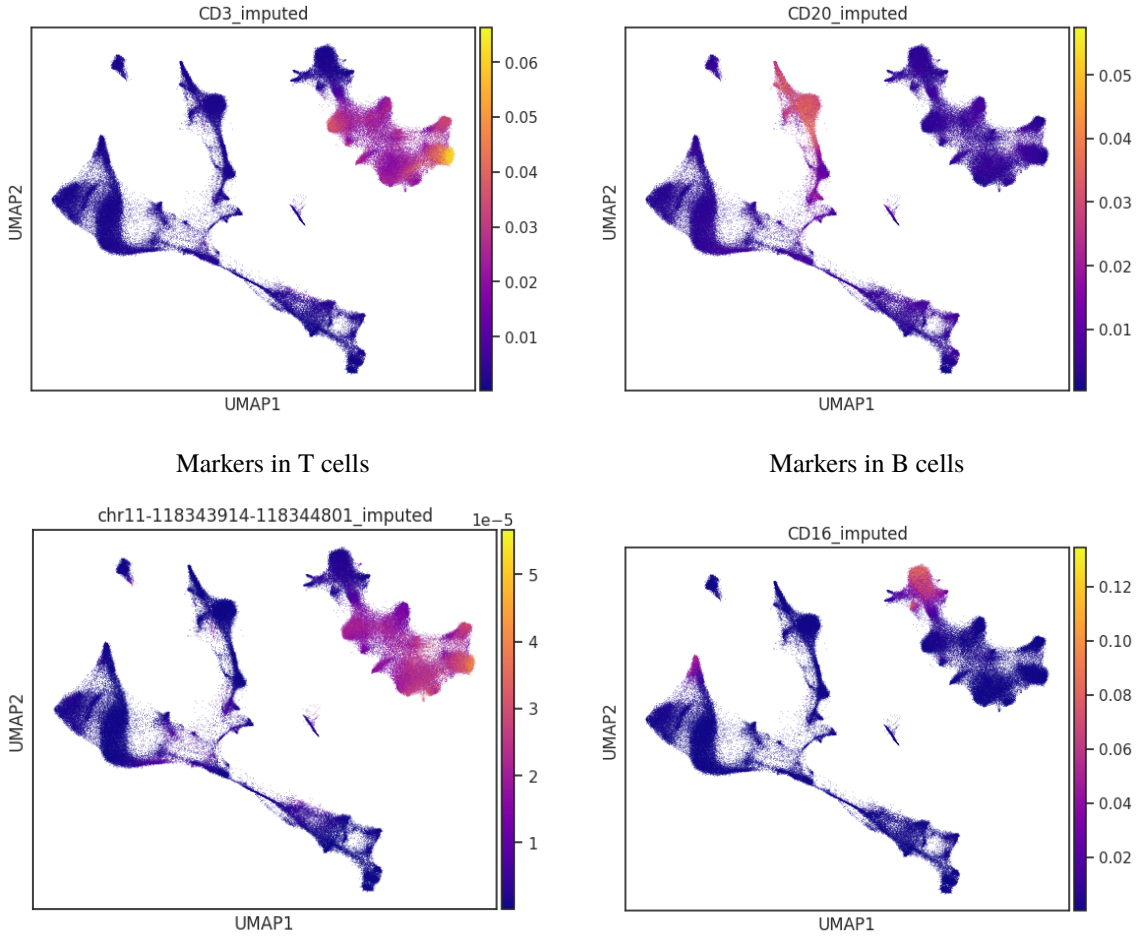
Cell types overview

Marker genes in B cells

Marker genes in T cells

Marker genes in NK cells

Figure 9: Marker genes across different cell types.



Markers in T cells

Markers in B cells

Specific open regions in T cells

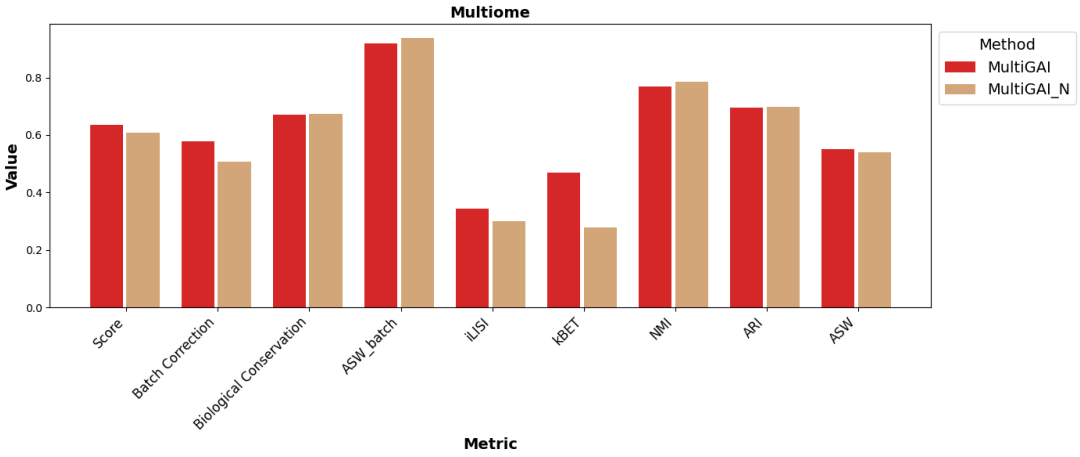
Markers in NK cells

Figure 10: Markers and specific open regions across different cell types.

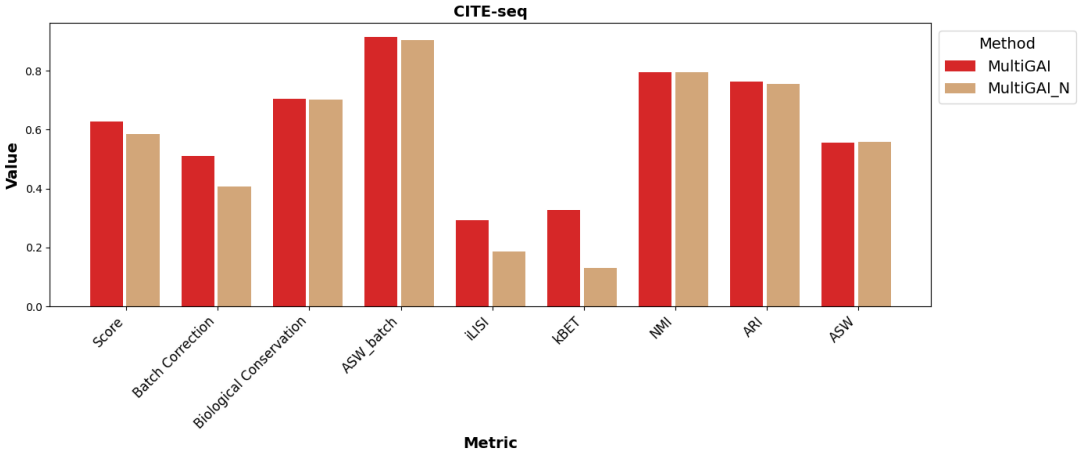
G Ablation Study

Table 2: Ablation study metrics

Dataset	Method	Score	Bat	Bio	ASW_b	iLISI	kBET	NMI	ARI	ASW
Multiome	MultiGAI	**0.6342**	**0.5777**	0.6719	0.9201	**0.3428**	**0.4703**	0.7679	0.6965	**0.5513**
Multiome	MultiGAI_N	0.6068	0.5061	**0.6740**	**0.9390**	0.3006	0.2786	**0.7856**	**0.6970**	0.5392
CITE-seq	MultiGAI	**0.6273**	**0.5118**	**0.7044**	**0.9154**	**0.2914**	**0.3284**	**0.7946**	**0.7624**	0.5562
CITE-seq	MultiGAI_N	0.5844	0.4068	0.7028	0.9039	0.1853	0.1313	**0.7946**	0.7554	**0.5584**

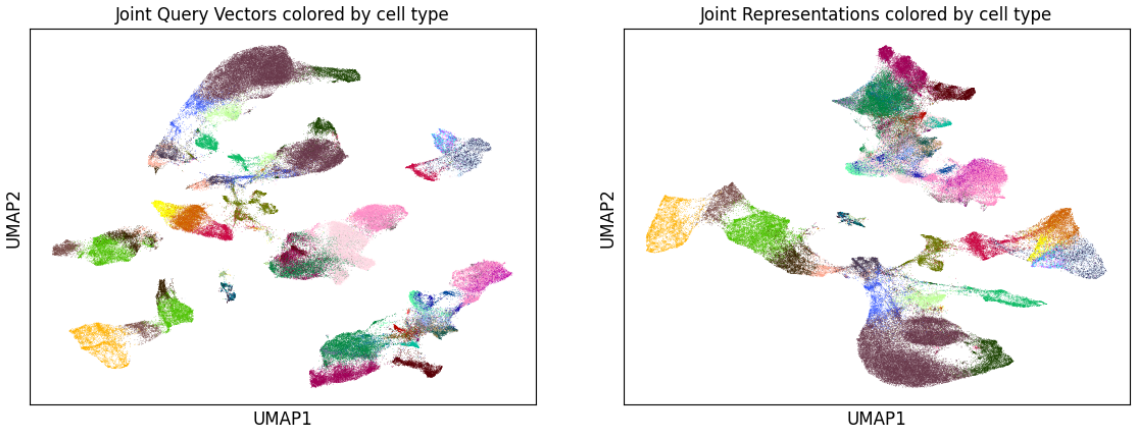


(a) Metrics of the ablation study for Multiome



(b) Metrics of the ablation study for CITE-seq

Figure 11: Metrics of the ablation studies for Multiome and CITE-seq datasets.



Joint Query Vectors colored by cell type

Joint Representations colored by cell type

Figure 12: UMAP plots of joint query vectors and joint representations colored by cell type

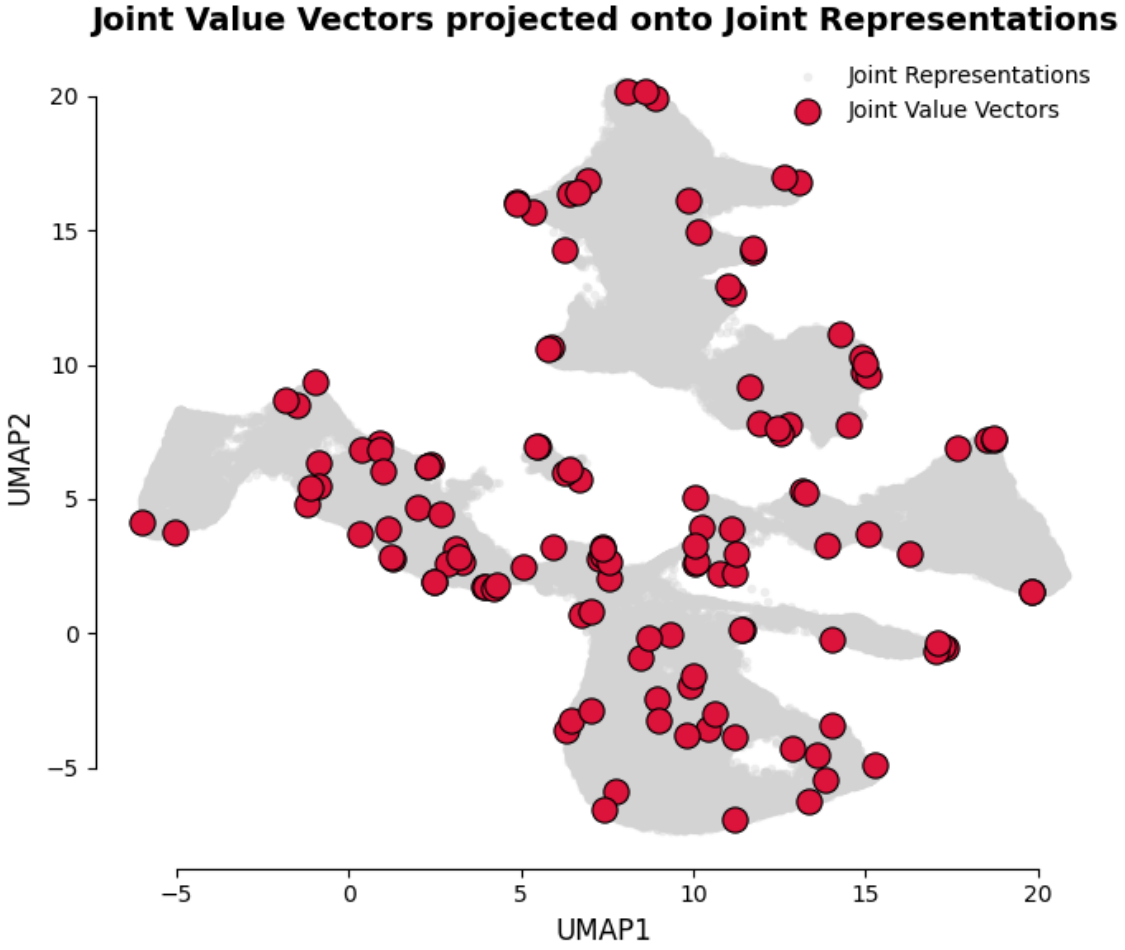


Figure 13: UMAP plots of joint value vectors and joint representations