scMRDR: A Scalable and Flexible Framework for Unpaired Single-Cell Multi-Omics Data Integration

Jianle Sun[1,2], **Chaoqi Liang**[1], **Ran Wei**[1], **Peng Zheng**[1], **Lei Bai**[1], **Wanli Ouyang**[1], **Hongliang Yan**[4], **Peng Ye**[1,3†]

[1] Shanghai Artificial Intelligence Laboratory, [2] Carnegie Mellon University, [3] The Chinese University of Hong Kong, [4] Guangzhou Laboratory

Abstract

Advances in single-cell sequencing have enabled high-resolution profiling of diverse molecular modalities, while integrating unpaired multi-omics single-cell data remains challenging. Existing approaches either rely on pair information or prior correspondences, or require computing a global pairwise coupling matrix, limiting their scalability and flexibility. In this paper, we introduce a scalable and flexible generative framework called single-cell Multi-omics Regularized Disentangled Representations (scMRDR) for unpaired multi-omics integration. Specifically, we disentangle each cell's latent representations into modality-shared and modality-specific components using a well-designed β-VAE architecture, which are augmented with isometric regularization to preserve intra-omics biological heterogeneity, adversarial objective to encourage cross-modal alignment, and masked reconstruction loss strategy to address the issue of missing features across modalities. Our method achieves excellent performance on benchmark datasets in terms of batch correction, modality alignment, and biological signal preservation. Crucially, it scales effectively to large-scale datasets and supports integration of more than two omics, offering a powerful and flexible solution for large-scale multi-omics data integration and downstream biological discovery.

1 Introduction

Recent advances in single-cell sequencing technologies have enabled the measurement of diverse molecular modalities at single-cell resolution, such as gene expression (scRNA), chromatin accessibility (scATAC), and protein abundance (scProtein). These complementary data sources offer a comprehensive view of cellular states and dynamics. Although a few protocols allow limited joint profiling using marker-based techniques, they still suffer from low feature coverage and reduced flexibility due to the destructive nature of single-cell assays, making it remain technically challenging to jointly measure multiple modalities within the same cell. Consequently, large-scale single-cell datasets are typically unpaired across different modalities [33]. This unpaired nature, coupled with significant technical noise such as batch effects, dropouts, and sequencing depth variation, makes data integration in a shared biologically meaningful latent space a highly nontrivial task [27, 3].

The goal of multi-omics data integration is to map single-cell data in different omics into a shared latent space, where representations across modalities are distributionally aligned while preserving biological differences between cell types and correcting for technical variations due to experimental batches (Fig.1a). Existing approaches explored joint dimension reduction (Fig.1b), like factor

†Corresponding author.

Source codes are available at `https://github.com/sjl-sjtu/scMRDR`.

Email Address: jianles@andrew.cmu.edu. Work was done during an internship at Shanghai AI Laboratory.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

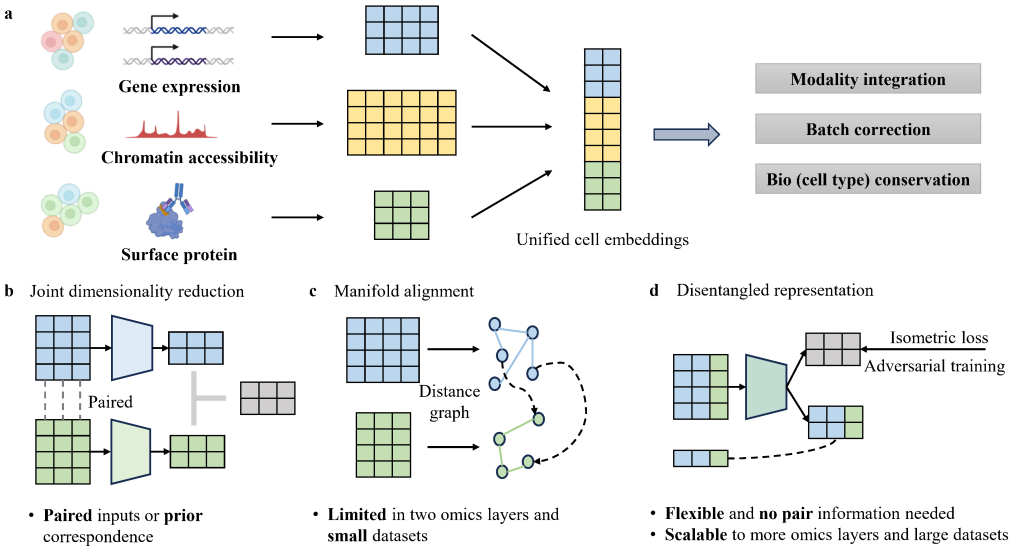


Figure 1: Method overview. (a) Multi-omics data integration. The goal is to integrate single-cell data in different modalities into an aligned latent space while preserving biological information and correcting technical noise. (b) Integration via joint dimension reduction (e.g., joint autoencoders). It typically works with paired data (measurements on different omics within the same cell). (c) Integration via manifold alignment between the geometric structures (e.g., KNN distance graphs) of different omics. It does not require paired data, but is typically limited to small-scale datasets involving only two omics modalities. (d) Our framework, based on disentangled representations, is flexible to completely unpaired data and scalable to large datasets with more than two omics.

analysis [1], or deep generative models [24]. However, they typically rely on paired or partially paired data [2, 6] to guide the integration, or require external prior knowledge [8] or pre-learned coupling matrix [11] to bridge modalities, limiting their **flexibility**. On the other hand, some approaches employ manifold alignment (Fig.1c), including optimal transport [14, 7] or unsupervised manifold transformation [5, 39]. However, these methods, relying on computing a global pairwise coupling matrix, typically restrict to integrating two modalities, and encounter serious computational issues in large datasets due to the complexity of inter-modal alignment, struggling in **scalability**.

To address the challenges, we proposed a scalable and flexible generative model named single-cell Multi-omics Regularized Disentangled Representations (scMRDR) to integrate unpaired multi-omics single-cell data into a unified latent space (Fig.1d). Unlike existing methods, scMRDR requires neither paired samples and prior information nor establishing global correspondences across different modalities. Instead, we achieve the integration based on the disentanglement of each sample's latent code into *shared* and *modality-specific* components via a well-designed β-VAE architecture, incorporating *isometric regularization* to ensure the conservation of biological information, and *adversarial training* to encourage the fusion of different modalities, with *masked loss function* to address the feature missing issue in different modalities (Fig.2).

Due to the *single unified encoder-decoder architecture*, scMRDR is flexible to completely unpaired data and able to scale to large datasets with more than two modalities naturally. Applied to real-world unpaired single-cell data, scMRDR demonstrates excellent performance in modality integration, batch correction as well as bio-conservation, surpassing a broad range of existing methods. Moreover, scMRDR scales robustly to large datasets, and readily accommodates additional omics layers. These results collectively underscore scMRDR as a flexible and scalable framework for large-scale unpaired single-cell data integration and the discovery of complex biological mechanisms.

We summarize the contributions as follows: (1) Through in-depth analysis of existing works in the field of multi-omics integration, we identify their limitations in flexibility and scalability, and propose a generative framework with a unified single encoder-decoder β-VAE to disentangle latent representations; (2) We propose a joint optimization goal, incorporating isometric regularization, adversarial

training, and masked loss to facilitate modality fusion while preserving biological signals; (3) We validate scMRDR through extensive experiments on multiple real-world datasets, demonstrating its strong flexibility and scalability on large-scale datasets and more complex multi-omics integration tasks, as well as its practical significance in downstream biological analyses (such as spatial position imputation).

2 Related work

Integrating multi-omics data has been extensively studied, with methods typically falling into two broad categories. Joint dimension reduction methods, including statistical models such as factor analysis based method like MOFA [1], canonical correlation analysis (CCA)-based methods like Seurat v3 [30], as well as deep generative models such as scVI-based [24] adaptations, assume access to matched (paired) measurements across omics layers, enabling supervised or semi-supervised learning of joint representations. For example, MultiVI [2] integrates paired multi-omic data by directly averaging latent embeddings inferred by encoders of respective modalities. However, the need for explicit pairing limits their applicability in cases where cross-modality correspondences are incomplete or noisy. JAMIE [11] incorporates the manifold alignment approach into the VAE framework, and UniPort [6] takes advantage of partially paired features and coupled VAE, but they still confront computational intensity in dealing with large-scale clinical or experimental datasets.

Another prominent line of work aligns modalities through unsupervised manifold matching, optimizing for geometric consistency between latent spaces. UnionCom [5] aligns modalities by constructing a kNN graph and applying unsupervised linear manifold alignment, while CMOT [39] adopts nonlinear manifold alignment with partial supervision. SCOT [14] leverages Gromov-Wasserstein optimal transport (GWOT) on similarity matrices and uses barycentric projection for integration, with following revised versions such as Pamona [7] and SCOTv2 [13] by imposing regularizations on GWOT, while SCOOTR [15] aligns both samples and features using co-optimal transport (COOT). These methods typically require constructing pairwise similarities and a global coupling matrix, restricting scalability to large datasets due to computational bottlenecks, and in practice, they are often validated only on limited sample sizes or toy examples. Moreover, they often focus on the integration of two modalities, leaving the integration of more omics layers an underexplored challenge.

Recent works, such as scTFBridge [34], scMaui [20], and InClust+ [35], have discussed integration based on latent decomposition. However, they still rely on designs like partial pairing supervision, stacked encoders, and cross-modal contrastive learning, limiting their scalability in completely unpaired and multi-omics contexts. In contrast, we aim to achieve the decoupling via a unified β-VAE composed of a single encoder-decoder, treating observations in different omics equally as a single sample, thereby ensuring flexibility and scalability in completely unpaired data across multiple omics. Theoratically, such disentangled subspaces are unidentifiable (i.e., not unique) without additional constraints [21]. We leverage this unidentifiability and, by imposing isometric and adversarial regularization, constrain the modality-shared subspace to be the one that preserves the maximum sample structure information from the entire space while aligning different modalities.

3 Methods

3.1 Preliminary: Disentangled VAE

Variational Autoencoders (VAEs) [22] are a class of generative models that learn a probabilistic mapping between observed data x and latent variable space z via variational inference by introducing an encoder network to parametrize the variational posterior $q_\phi(z \mid x)$ and a decoder network to reconstruct the generative process $p_\theta(x \mid z)$. The model is trained by maximizing the evidence lower bound (ELBO) on the marginal log-likelihood:

$$\mathrm{ELBO}_{\mathrm{VAE}}(x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - \mathrm{KL}(q_\phi(z \mid x) \parallel p(z)), \tag{1}$$

where $\mathrm{KL}(q \parallel p)$ denotes the Kullback–Leibler (KL) divergence between distributions q and p. The first term encourages faithful data reconstruction, while the second regularizes the latent space to align with the prior.

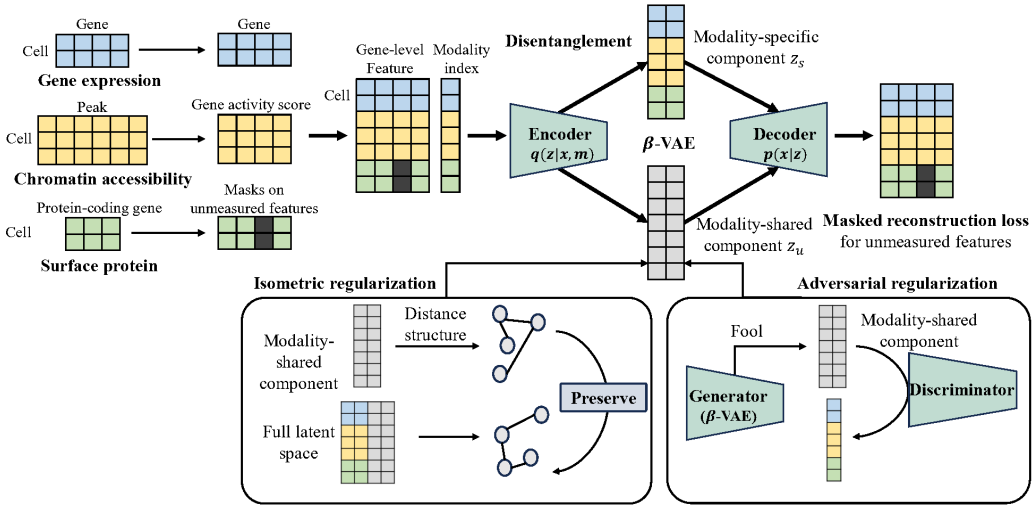


Figure 2: Overview of the proposed scMRDR. We employ β-VAE to disentangle omics-specific and omics-shared latent representations, and impose isometric loss and adversarial training as regularization to encourage modality integration and bio-conservation.

Classical VAE often conflates modality-shared and modality-specific signals in the latent space, impeding interpretability and downstream analyses. To improve disentanglement in the learned latent representations, β-VAE [19] introduces a hyperparameter $\beta > 1$ to upweight the KL divergence term in the VAE objective

$$\text{ELBO}_{\beta\text{-VAE}}(x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - \beta \cdot \text{KL}(q_\phi(z \mid x) \parallel p(z)). \tag{2}$$

A higher value of β enforces a stronger constraint on the latent space, promoting disentangled and interpretable representations at the cost of reduced reconstruction accuracy.

Simply disentangling the latent space does not guarantee that the shared latent components from different omics data are fully aligned in distribution, nor does it ensure that the shared latent components preserve the structural information present in the original data. Such structural information is captured by the VAE in the entire latent space and may not necessarily be retained within the shared subspace. We will impose additional regularization on the disentangled latent space in a generative model designed for single-cell multi-omics to address the issue.

3.2 Disentangled generative model for multi-omics data

To achieve flexible and scalable integration, we propose a generative model tailored for single-cell multi-omics data (Fig.3). We assume that observations $x^{(m)}$ in the omics m are generated from latent embeddings lying in two independent subspaces, i.e., common latent variables z_u shared across modalities and modality-specific latent variables $z_s^{(m)}$, and we have

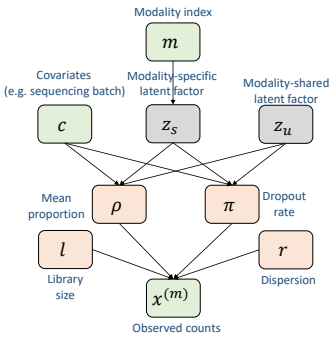


Figure 3: Graphical illumination of the single-cell multi-omics data generative model.

$$p(z|m) = p(z_u)p(z_s|m) \tag{3}$$

and

$$p(x) = p(x|z,c)p(z|m)p(m)p(c) = p(x|z,c)p(z_u)p(z_s|m)p(m)p(c) \tag{4}$$

where c represents other covariates, such as the experimental batch during sequencing. Batch effects are systematic variations introduced by non-biological factors, including differences in experimental runs, reagent lots, or operators. These effects can obscure true biological signals or introduce spurious

patterns in the data [32]. To mitigate such confounding influences, batch information is commonly included as a covariate in the modeling process.

Sequencing reads of scRNA, protein, and other omics data can all be mapped to the gene level (e.g., gene activity scores derived from peak aggregation in single-cell ATAC-seq). Since the read matrix are typically sparse (with both biological and technical dropouts) count data with over-dispersion (i.e., the variance exceeds the mean), we parametrize the generative process $p(x|z,c)$ using a zero-inflated negative binomial (ZINB) distribution [24, 29] (which can be changed to ordinary Gaussian model if normalized scores or more non-count data included), i.e.,

$$x_{ng} \sim \pi_g \delta_0 + (1-\pi_{ng})\mathrm{NB}(l_n \rho_{ng}, r_g) \overset{d}{=} \mathrm{ZINB}(l_n \rho_{ng}, r_g, \pi_g) \tag{5}$$

where δ_0 is point mass at zero, l_n represents the library size of cell n, ρ_{ng} is the mean proportion of the corresponding measurement (RNA expressions, activity score, protein level, etc.) of gene g in cell n, r_g is the dispersion factor of gene g, π_{ng} is the dropout rate of gene g in cell n. We parametrize the parameters by some non-linear neural networks as follows $h = f_h(z,c)$, $\rho_{ng} = f_\rho(h)$, $\pi_{ng} = f_\pi(h)$.

Prior distributions of latent factors z are assumed as isotropic multivariate Gaussian distributions, i.e., $p(z_s|m) \overset{d}{=} \mathcal{N}(\mu_m, \sigma_m^2 I)$ and $p(z_u|t) \overset{d}{=} \mathcal{N}(0, I)$. We employ variational posteriors $q(z_u|x) = \mu_u(x) + \sigma_u(x) \odot \mathcal{N}(0,I)$ and $q(z_s|x,m) = \mu_s(x,m) + \sigma_s(x,m) \odot \mathcal{N}(0,I)$ to approximate the prior, and the loss function (negative ELBO) of β-VAE is

$$\begin{aligned} \mathcal{L}_{\beta\text{-VAE}} &= -\mathrm{ELBO} = \mathcal{L}_{\text{recon}} + \beta \mathcal{L}_{\mathrm{KL}} \\ &= -\mathbb{E}_{z\sim q(z|x,m)} \log p(x|z,c) + \beta \left[\mathrm{KL}(q(z_u|x) \| p(z_u)) + \mathrm{KL}(q(z_s|x,m) \| p(z_s|m)) \right] \end{aligned} \tag{6}$$

where $\beta > 1$ to encourage the disentanglement of z_u and z_s. For the ZINB model, the reconstruction loss, i.e., the expected log likelihood under the variational posterior, is

$$\mathbb{E}_{z\sim q(z|x,m)} \log p(x|z,c) = \frac{1}{N} \sum_{n=1}^{N} \sum_{g=1}^{G} \log \left[P_{\mathrm{ZINB}}(x_n; l_n \rho_{ng}, r_g, \pi_{ng}) \right] \tag{7}$$

where $P_{\mathrm{ZINB}}(X = x; \mu, r, \pi) = \pi \mathbb{I}_{x=0} + (1-\pi) P_{\mathrm{NB}}(x; \mu, r)$, and $P_{\mathrm{NB}}(x; \mu, r)$ stands for the probability mass of negative binomial distribution $NB(\mu, r)$ at x, i.e., $P_{\mathrm{NB}}(x; \mu, r) = \frac{\Gamma(x+r)}{x!\,\Gamma(r)} \left(\frac{r}{r+\mu}\right)^r \left(\frac{\mu}{r+\mu}\right)^x$. In particular, the probability mass at zero $P_{\mathrm{NB}}(0; \mu, r) = (\frac{r}{r+\mu})^r$.

3.3 Adversarial regularization for omics integration

By disentangling the latent space, we obtain, in general, modality-invariant latent variable z_u lying in a shared subspace. To further encourage the alignment of distributions $z_u^{(m)}$ from different omics, we impose an additional adversarial regularization [17, 16] by introducing a m-class discriminator $D(z_u) : \mathbb{R}^{d_u} \to \{0, 1, \ldots, m\}$ to distinguish z_u of samples from different omics and try to optimize its capacity, i.e.,

$$\min_D \mathcal{L}_{\text{discriminator}} = \min_D \left[-\sum_m m \log(D(z_u^{(m)})) \right], \; z_u^{(m)} \sim q(z_u|x) \tag{8}$$

while training the VAE encoders $q(z_u|x)$ to fool the discriminator as much as possible by optimizing in the opposite direction

$$\max_{q(z_u|x)} \inf_D \mathcal{L}_{\text{discriminator}} \tag{9}$$

which is equivalent to

$$\min_{q(z_u|x)} \mathcal{L}_{\text{alignment}} = \min_{q(z_u|x)} \sup_D \left[\sum_m m \log \left(D(z_u^{(m)}) \right) \right] \tag{10}$$

achieving a proper alignment of embeddings from different omics ultimately.

3.4 Isometric loss for structure preservation

To ensure that z_u captures the biological differences between samples (e.g., cell types), we introduce an additional unsupervised structure-preserving regularization, since cell type labels are typically

unavailable. Though the original feature matrices are high-dimensional, the full latent representation $z = (z_u, z_s)$ learned by the generative model effectively preserves intra-modality structure. Hence, we reformulate the problem as encouraging z_u to retain the structural information of the full latent space. Specifically, since the latent embeddings already reside in a low-dimensional space, we apply an isometric loss [31] that minimizes the discrepancy between the pairwise Euclidean distance matrices computed from z and from z_u, for each modality, i.e.,

$$\mathcal{L}_{\text{preserve}} = \sum_{m} \sum_{i,j \in X^{(m)}} \left[||\mu_{z_u}(x_i) - \mu_{z_u}(x_j)||_2 - ||\mu_z(x_i) - \mu_z(x_j)||_2 \right]^2 \tag{11}$$

where $\mu_{z_u}(x)$ is the posterior mean of variational approximation $q(z_u|x)$ and $\mu_z(x)$ is the posterior mean of total latent embeddings $(q(z_u|x), q(z_s|x, m))$.

And the total optimization goal for VAE becomes

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \beta \mathcal{L}_{\text{KL}} + \lambda \mathcal{L}_{\text{alignment}} + \gamma \mathcal{L}_{\text{preserve}} \tag{12}$$

In the training process, we first update the discriminator by optimizing $\mathcal{L}_{\text{discriminator}}$, then update VAE with respect to the total loss $\mathcal{L}_{\text{total}}$ in turn in each mini-batch.

3.5 Masked reconstruction loss for missing features

In unpaired multi-omics datasets, different modalities are measured separately and often originate from distinct sources. Although it is possible to align features across modalities at the gene level, severe missing features still exist due to different sequencing coverages, especially for antibody-based protein profiling techniques such as CITE-seq, which typically covers only a few hundred proteins due to the limited availability of antibody markers [4], while tens of thousands of genes can be measured in other omics. Restricting the analysis to the overlapping features across all modalities would lead to substantial information loss, whereas naively imputing unmeasured features with zeros would severely distort the data distribution. To address this, we introduce a binary mask $\mathbf{b} \in \{0, 1\}^G$ indicating feature availability that prevents gradients from back-propagating through unmeasured features in the reconstruction loss for each modality, and then scale by the proportion of available features to ensure that the reconstruction loss for each sample is on a comparable scale, i.e.,

$$\mathcal{L}_{\text{recon}} = -\frac{1}{N} \sum_{n=1}^{N} \left\{ \frac{G}{\sum_{g=1}^{G} \mathbf{b}_{ng}} \sum_{g=1}^{G} \mathbf{b}_{ng} \log \left[P_{\text{ZINB}}(x_n; l_n \rho_{ng}, r_g, \pi_{ng}) \right] \right\} \tag{13}$$

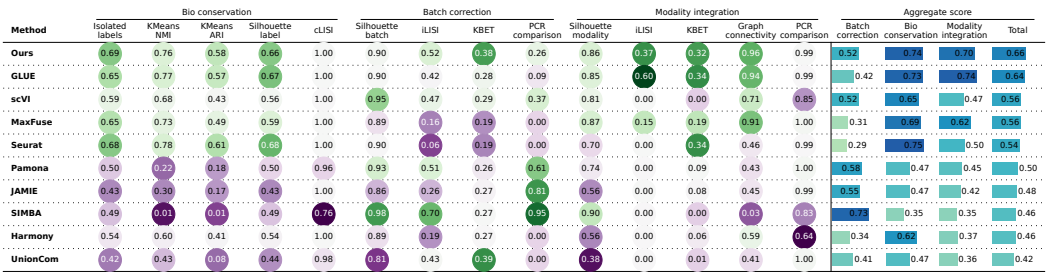
The masked loss strategy ensures that the model can fully utilize the available information while preserving the integrity of the original data distribution.

4 Results

4.1 Setup and evaluation metrics

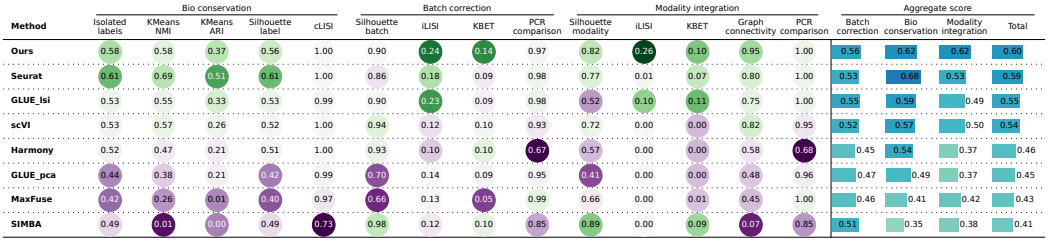
To verify the effectiveness of our proposed method, we comprehensively evaluate scMRDR through a series of experiments, beginning with standard benchmarks and then scaling up to more complex single-cell and multi-omics scenarios. We employ publicly available datasets from previous researches with curated cell-type annotations [28, 37, 25]. Detailed setups are shown in Appendix A.1 and Table 2. We compare to state-of-the-art baselines, including GLUE, scVI, Seurat v5, Harmony, JAMIE, and so on, and evaluate in terms of cell-type clustering, modality integration, and batch removal. Cell-type labels in different omics has been aligned in evaluation. It should be emphasized that we did not use cell-type labels during training. UMAP visualizations are presented for qualitative comparison. For quantitative evaluation, the following commonly used metrics (Appendix A.2) are included: F1 isolated label scores, k-means NMI, k-means ARI, cell-type Silhouette, and cell-type separation LISI (cLISI) to evaluate the performance in cell type conservation, modality Silhouette, modality integration LISI (iLISI), kBET, Principal Component Regression (PCR) R^2, and graph connectivity to evaluate the performance in modality integration, as well as batch Silhouette, batch integration LISI (iLISI), kBET, and PCR R^2 to evaluate the performance in batch effect correction [26, 36].

We evaluate and visualize all the above metrics based on the package scib-metrics [26]. The overall score is calculated as a weighted average of all metrics on bio-conservation (40% weights), modality integration (30% weights), and batch correction (30% weights).



Method	Bio conservation					Batch correction				Modality integration					Aggregate score			
	Isolated labels	KMeans NMI	KMeans ARI	Silhouette label	cLISI	Silhouette batch	iLISI	KBET	PCR comparison	Silhouette modality	iLISI	KBET	Graph connectivity	PCR comparison	Batch correction	Bio conservation	Modality integration	Total
Ours	0.69	0.76	0.58	0.66	1.00	0.90	0.52	0.38	0.26	0.86	0.37	0.32	0.96	0.99	0.52	0.74	0.70	0.66
GLUE	0.65	0.77	0.57	0.67	1.00	0.90	0.42	0.28	0.09	0.85	0.60	0.34	0.94	0.99	0.42	0.73	0.74	0.64
scVI	0.59	0.68	0.43	0.56	1.00	0.95	0.47	0.29	0.37	0.81	0.00	0.00	0.71	0.85	0.52	0.65	0.47	0.56
MaxFuse	0.65	0.73	0.49	0.59	1.00	0.89	0.16	0.19	0.00	0.87	0.15	0.19	0.91	1.00	0.31	0.69	0.62	0.56
Seurat	0.68	0.78	0.61	0.68	1.00	0.90	0.06	0.19	0.00	0.70	0.00	0.34	0.46	0.99	0.29	0.75	0.50	0.54
Pamona	0.50	0.22	0.18	0.50	0.96	0.93	0.51	0.26	0.61	0.74	0.00	0.09	0.43	1.00	0.58	0.47	0.45	0.50
JAMIE	0.43	0.30	0.17	0.43	1.00	0.86	0.26	0.27	0.81	0.56	0.00	0.08	0.45	0.99	0.55	0.47	0.42	0.48
SIMBA	0.49	0.01	0.01	0.49	0.76	0.98	0.70	0.27	0.95	0.90	0.00	0.00	0.03	0.83	0.73	0.35	0.35	0.46
Harmony	0.54	0.60	0.41	0.54	1.00	0.89	0.19	0.27	0.00	0.56	0.00	0.06	0.59	0.64	0.34	0.62	0.37	0.46
UnionCom	0.42	0.43	0.08	0.44	0.98	0.81	0.43	0.39	0.00	0.38	0.00	0.01	0.41	1.00	0.41	0.47	0.36	0.42

Figure 4: Performance comparisons on two-omics integration, where unscaled metrics calculated via scIB are reported.



Method	Bio conservation					Batch correction				Modality integration					Aggregate score			
	Isolated labels	KMeans NMI	KMeans ARI	Silhouette label	cLISI	Silhouette batch	iLISI	KBET	PCR comparison	Silhouette modality	iLISI	KBET	Graph connectivity	PCR comparison	Batch correction	Bio conservation	Modality integration	Total
Ours	0.58	0.58	0.37	0.56	1.00	0.90	0.24	0.14	0.97	0.82	0.26	0.10	0.95	1.00	0.56	0.62	0.62	0.60
Seurat	0.61	0.69	0.51	0.61	1.00	0.86	0.18	0.09	0.98	0.77	0.01	0.07	0.80	1.00	0.53	0.68	0.53	0.59
GLUE_lsi	0.53	0.55	0.33	0.53	0.99	0.90	0.23	0.09	0.98	0.52	0.10	0.11	0.75	1.00	0.55	0.59	0.49	0.55
scVI	0.53	0.57	0.26	0.52	1.00	0.94	0.12	0.10	0.93	0.72	0.00	0.00	0.82	0.95	0.52	0.57	0.50	0.54
Harmony	0.52	0.47	0.21	0.51	1.00	0.93	0.10	0.10	0.67	0.57	0.00	0.00	0.58	0.68	0.45	0.54	0.37	0.46
GLUE_pca	0.44	0.38	0.21	0.42	0.99	0.70	0.14	0.09	0.95	0.41	0.00	0.00	0.48	0.96	0.47	0.49	0.37	0.45
MaxFuse	0.42	0.26	0.01	0.40	0.97	0.66	0.13	0.05	0.99	0.66	0.00	0.01	0.45	1.00	0.46	0.41	0.42	0.43
SIMBA	0.49	0.01	0.00	0.49	0.73	0.98	0.12	0.10	0.85	0.89	0.00	0.09	0.07	0.85	0.51	0.35	0.38	0.41

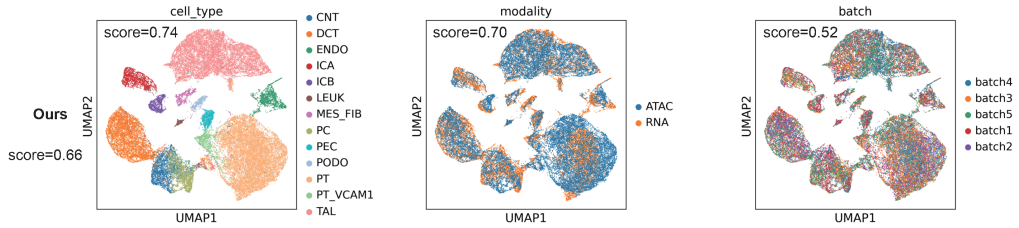
Figure 5: Performance comparisons on two-omics integration with large-scale dataset, where unscaled metrics calculated via scIB are reported. The default preprocessing method 'scglue.data.lsi' for GLUE fails to handle the large-scale data, and substituting it with PCA leads to severe performance degradation, although using 'TruncatedSVD' as an approximation of LSI can alleviate this issue.

4.2 Benchmarking performance on two-omics integration

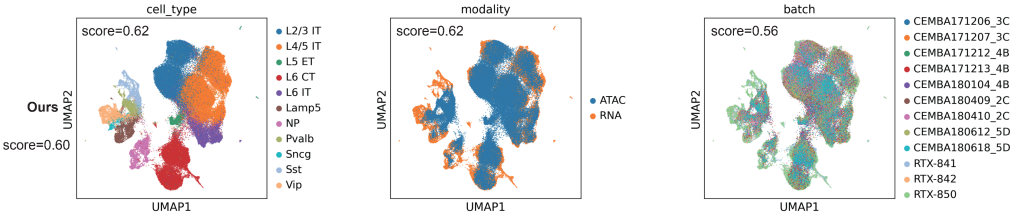
We first compare scMRDR with 9 existing methods, including Seurat v5 [18], Harmony [23], scVI [24], scGLUE [8], JAMIE [11], UnionCom [5], Pamona [7], MaxFuse [10] and SIMBA [9] on a unpaired scRNA and scATAC dataset of human kidney tissue [28]. Among these, scVI can be regarded as a baseline counterpart of our model, with no disentanglement or regularization applied to the latent space. The dataset contains scRNA-seq with 27,146 genes on 19,985 cells and scATAC-seq with 99,019 peaks on 24,205 cells. Peak signals in scATAC are aggregated to gene-level activity score in scMRDR by the package episcanpy [12]. Gene activity scores are also used in integration by Seurat, Harmony, and scVI, while others use raw peak signals directly. We choose the highly variable genes for each omics and take the union as input. Shown in Fig.4, scMRDR outperforms the existing methods, exhibiting an excellent performance in modality integration, bio-conservation, and batch correction. As shown in Fig.6a, without the incorporation of explicit cell type annotations in training, our method yields well-separated embedding clusters corresponding to distinct cell types in an unsupervised way, thereby preserving the underlying biological heterogeneity. Meanwhile, samples from different omics and batches fuse and align well in the latent space, demonstrating successful correction of modality-specific variations and technical noises like batch effects. In contrast, some other methods such as Harmony, scVI and JAMIE (Fig.9 in Appendix A.5) can preserve biological differences between distinct cell types, but fail to integrate the distributions of the two different omics modalities.

4.3 Scalability on integration with large-scale dataset

To validate the scalability of scMRDR on larger datasets, we evaluate the performance on an large-level dataset on mouse primary motor cortex with more cells available [37]. This large-scale data includes 69,727 cells with measurements on 27,123 genes by scRNA-seq and 54,844 cells with measurements on 148,814 chromatin peaks by scATAC-seq. Methods based on optimal transport or other unsupervised manifold alignment, such as JAMIE, UnionCom, and Pamona, fail to run on datasets with such scale due to errors in memory or optimization. We compare the performance of the rest methods. Shown in Fig.5, some methods that perform well on small-scale datasets suffer significant performance drops on larger ones (Fig.10 in Appendix A.5). For example, GLUE exhibits



(a) UMAP visualization of small-sized two-omics dataset



(b) UMAP visualization of large-scale two-omics dataset

Figure 6: UMAP visualization of the latent representations obtained by scMRDR in the two-omics integration task. Latent embeddings of other methods are shown in Appendix A.5. An effective method should yield well-separated clusters corresponding to distinct cell types, and fuse samples from different modalities and experimental batches in sequencing. Noted that the annotated cell type labels are not incorporated in the unsupervised learning but only used in evaluation.

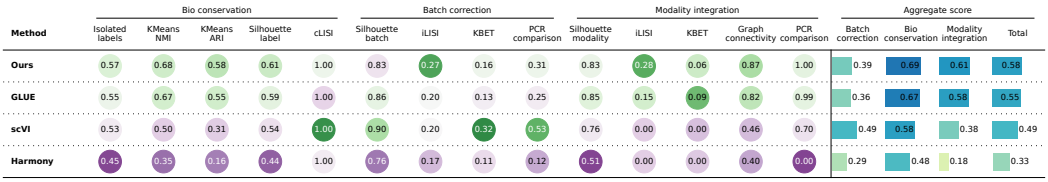


Figure 7: Performance comparisons on triple-omics integration, where unscaled metrics calculated via scIB are reported.

strong susceptibility to the choice of preprocessing strategy which itself markedly dependent on data scale, whereas our method maintains stable performance on large-scale datasets (Fig.6b). This makes it highly scalable for large-level single-cell data integration, enabling the informative linking among different omics and providing more valuable biological insights.

4.4 Scalability on triple-omics integration

Most existing methods do not support integrating datasets with more than two omics, while scMRDR can naturally extend to the integration of triple-omics or even more. To illuminate it, we conduct a case study on integrating scRNA, scATAC, and CITE-seq measured sc-protein levels, using a dataset on human bone marrow [25]. We conducted integration on 30,486 cells with scRNA-seq on 13,431 genes, 10,330 cells with scATAC-seq on 116,490 peaks, and 18,052 cells with 134 surface proteins measured by CITE-seq. Some methods that perform relatively well on two-omics datasets, such as Seurat v5, are not suitable for integrating more modalities, as they require dictionary learning and bridge integration between two omics. As a result, we compare the performance of the existing methods that support triple-omics integration, including GLUE, scVI, and Harmony. Shown in Fig.7, scMRDR shows a consistently excellent performance in the integration task with more modalities On the contrary, methods like GLUE fail to align latent distributions over three omics, especially when one of the omics modalities (proteomics here) has significantly fewer measured features than the others (Fig.8).

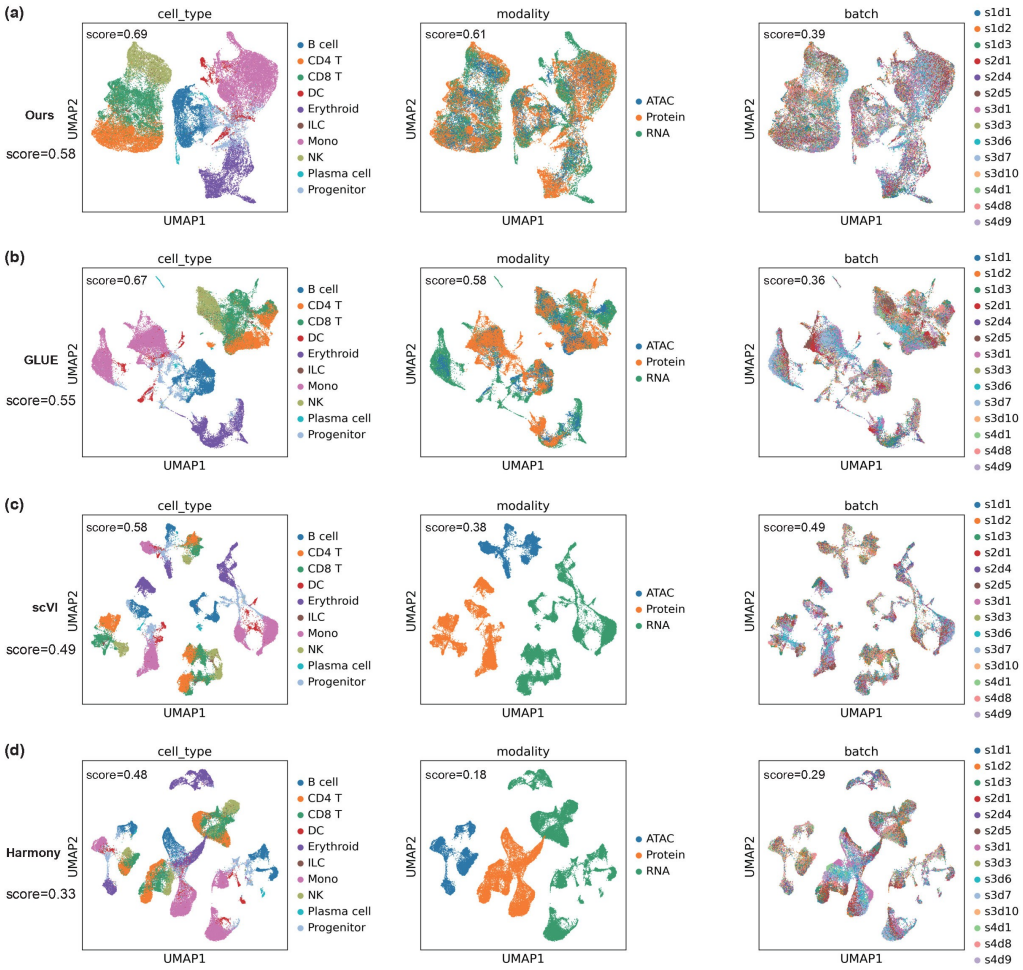


Figure 8: UMAP visualization of the unified embeddings in triple-omics data.

4.5 Ablation study and sensitivity analysis on the regularized beta-VAE

To demonstrate the effects of isometric and adversarial regularization within the β-VAE framework, we conducted ablation experiments and sensitivity analysis with a range of different hyperparameter combinations on the two-omics human kidney datasets. Shown in Table 1, with different hyperparameters, scMRDR performs consistently better than the ablation results when removing any individual component, and the worst model is the baseline without any regularization. The absence of isometric constraints or modality-adversarial regularization leads to a substantial drop in performance, where isometry is more essential to bio-conservation while adversary more vital to modality alignment. The influence of β is relatively minor as the structured prior imposed on z inherently promotes disentanglement. Nevertheless, employing a moderately large β can increase the conditional independence between z_u and z_s, thereby enhancing the effectiveness of latent disentanglement. While different hyperparameter combinations have a certain impact due to trade-offs among various losses and the influence of randomness, the performance are generally robust to the choices.

We also conduct ablation studies on a triple-omics dataset to evaluate the effectiveness of the masked loss. The results (Table 4 in Appendix A.4) show that, keeping all other hyperparameters unchanged, simply setting the missing features in the proteomics to zero without applying a loss mask leads to a significant performance drop. This highlights the importance of our masked loss strategy in effectively handling integration tasks where some omics layers (e.g., scProtein) contain substantially fewer features compared to other modalities.

Table 1: Performance in ablation studies and sensitivity analysis, where unscaled scIB aggregate scores are reported

β	γ	λ	Total Score	Batch-correct	Bio-conserve	Modal-integrate	Notes
2	5	5	0.66	0.52	0.74	0.70	
2	7	5	0.65	0.54	0.69	0.72	
2	5	3	0.65	0.55	0.69	0.71	
2	2	2	0.65	0.50	0.72	0.70	
2	5	7	0.65	0.49	0.70	0.73	
2	3	5	0.65	0.47	0.71	0.74	
2	1	5	0.64	0.49	0.68	0.75	
3	5	5	0.64	0.48	0.71	0.71	
4	5	5	0.64	0.47	0.71	0.72	
2	10	5	0.64	0.52	0.68	0.71	
1	5	5	0.62	0.50	0.67	0.68	$\beta = 1$
2	5	0	0.61	0.53	0.69	0.60	$\lambda = 0$
2	0	5	0.61	0.51	0.60	0.72	$\gamma = 0$
1	0	0	0.59	0.48	0.63	0.66	Baseline

4.6 Biological significance of integrating single-cell and spatial omics

To better evaluate the biological significance of scMRDR, we integrated scRNA [37], scATAC [37], and spatial transcriptomics (merFISH) [38] of mouse primary motor cortex using our method, and then used the aligned latent representation to interpolate the missing spatial locations in single-cell data by optimal transport. Visualization shows that this interpolation performs well, where inferred locations of cells align well with the provided cortex layers annotations (Fig. 12 in Appendix).

Due to the low coverage of merFISH (only 254 genes measured), only 103 genes are detected as spatial variable genes (SVGs) by SPARKX ($P_{\text{adj}} < 10^{-20}$). We leveraged the spatially interpolated scRNA data (26069 genes) and 4095 SVGs ($P_{\text{adj}} < 10^{-20}$) are detected. We replicated 83 out of 101 SVGs detectable by merFISH (like *Lamb5*, *Calb1*, *Cux2*), and also revealed new SVGs (like *Hs3st4*, *Cpa6*, *Zfhx4*). Similarly, using scATAC with imputed spatial locations, we identified 142 SVGs in gene activity scores ($P_{\text{adj}} < 10^{-20}$), including several key transcription factors like *Zfhx4*, *Cux1*, *Cux2*, *Gpc5*. This will further support the investigation of spatially specific regulatory mechanisms.

5 Discussion

Conclusion. In this paper, we propose a principled and feasible generative model named scMRDR to integrate unpaired multi-omics single-cell data into a unified latent space. We employ β-VAE to disentangle latent embedding into modal-shared and modal-specific subspace, incorporating isometric regularization to ensure the conservation of biological information within each omics, and adversarial loss to encourage the fusion of different modalities. Masked loss are adopted to address the feature-missing issue in different modalities. Via empirical experiments and comprehensive comparison with existing methods, scMRDR exhibits an excellent performance in modality integration, bio-conservation, and batch correction, and demonstrates strong adaptability for scaling to larger, large-level datasets with more omics modalities to integrate.

Limitations. There are still some limitations to our approach. The regularized β-VAE introduces the trade-off between different optimization goals, reflecting in the choices of hyperparameters β, λ, and γ. In particular, the introduction of the adversarial loss, i.e., the min-max optimization objective, increases the training difficulty. Besides, we map features of all modalities to the gene level, such as aggregating scATAC-seq peak signals into gene activity scores. Although selective masking in the loss allows for partially unmatched features, it may still introduce potential information loss.

Outlooks. In addition to measuring various omics features within cells, emerging spatial multi-omics technologies allow us to capture the spatial coordinates of cells, further enriching cellular information. Meanwhile, perturbation sequencing enables the observation of cellular responses across omics layers to various chemical treatments and genetic perturbations (e.g., CRISPR). Integrating spatial and dynamic information across modalities under different conditions will be an important direction for future exploration and extension of scMRDR.

References

[1] Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, and Oliver Stegle. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21:1–17, 2020.

[2] 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.

[3] Alev Baysoy, Zhiliang Bai, Rahul Satija, and Rong Fan. The technological landscape and applications of single-cell multi-omics. *Nature Reviews Molecular Cell Biology*, 24(10):695–713, 2023.

[4] Hayley M Bennett, William Stephenson, Christopher M Rose, and Spyros Darmanis. Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. *Nature Methods*, 20(3):363–374, 2023.

[5] 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.

[6] Kai Cao, Qiyu Gong, Yiguang Hong, and Lin Wan. A unified computational framework for single-cell data integration with optimal transport. *Nature Communications*, 13(1):7419, 2022.

[7] 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.

[8] 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.

[9] Huidong Chen, Jayoung Ryu, Michael E Vinyard, Adam Lerer, and Luca Pinello. Simba: single-cell embedding along with features. *Nature Methods*, 21(6):1003–1013, 2024.

[10] Shuxiao Chen, Bokai Zhu, Sijia Huang, John W Hickey, Kevin Z Lin, Michael Snyder, William J Greenleaf, Garry P Nolan, Nancy R Zhang, and Zongming Ma. Integration of spatial and single-cell data across modalities with weakly linked features. *Nature Biotechnology*, 42(7):1096–1106, 2024.

[11] Noah Cohen Kalafut, Xiang Huang, and Daifeng Wang. Joint variational autoencoders for multimodal imputation and embedding. *Nature Machine Intelligence*, 5(6):631–642, 2023.

[12] Anna Danese, Maria L Richter, Kridsadakorn Chaichoompu, David S Fischer, Fabian J Theis, and Maria Colomé-Tatché. EpiScanpy: integrated single-cell epigenomic analysis. *Nature Communications*, 12(1):5228, 2021.

[13] Pinar Demetci, Rebecca Santorella, Manav Chakravarthy, Bjorn Sandstede, and Ritambhara Singh. Scotv2: Single-cell multiomic alignment with disproportionate cell-type representation. *Journal of Computational Biology*, 29(11):1213–1228, 2022.

[14] Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. SCOT: single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology*, 29(1):3–18, 2022.

[15] Pinar Demetci, Quang Huy Tran, Ievgen Redko, and Ritambhara Singh. Jointly aligning cells and genomic features of single-cell multi-omics data with co-optimal transport. *bioRxiv*, pages 2022–11, 2022.

[16] Ayse B Dincer, Joseph D Janizek, and Su-In Lee. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics*, 36(Supplement_2):i573–i582, 2020.

[17] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

[18] Yuhan Hao, Tim Stuart, Madeline H Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, 42(2):293–304, 2024.

[19] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[20] Yunhee Jeong, Jonathan Ronen, Wolfgang Kopp, Pavlo Lutsik, and Altuna Akalin. scmaui: a widely applicable deep learning framework for single-cell multiomics integration in the presence of batch effects and missing data. *BMC Bioinformatics*, 25(1):257, 2024.

[21] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[23] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12):1289–1296, 2019.

[24] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

[25] Malte D Luecken, Daniel Bernard Burkhardt, Robrecht Cannoodt, Christopher Lance, Aditi Agrawal, Hananeh Aliee, Ann T Chen, Louise Deconinck, Angela M Detweiler, Alejandro A Granados, et al. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[26] Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, 2022.

[27] Zhen Miao, Benjamin D Humphreys, Andrew P McMahon, and Junhyong Kim. Multi-omics integration in the age of million single-cell data. *Nature Reviews Nephrology*, 17(11):710–724, 2021.

[28] Yoshiharu Muto, Parker C Wilson, Nicolas Ledru, Haojia Wu, Henrik Dimke, Sushrut S Waikar, and Benjamin D Humphreys. Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nature Communications*, 12(1):2190, 2021.

[29] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(1):284, 2018.

[30] 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.

[31] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[32] Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome Biology*, 21:1–32, 2020.

[33] Katy Vandereyken, Alejandro Sifrim, Bernard Thienpont, and Thierry Voet. Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics*, 24(8):494–515, 2023.

[34] Feng-ao Wang, Ruikun He, Junwei Liu, and Yixue Li. sctfbridge: A disentangled deep generative model informed by tf-motif binding for gene regulation inference in single-cell multi-omics. *bioRxiv*, pages 2025–01, 2025.

[35] Lifei Wang, Rui Nie, Xuexia Miao, Yankai Cai, Anqi Wang, Hanwen Zhang, Jiang Zhang, and Jun Cai. Inclust+: the deep generative framework with mask modules for multimodal data integration, imputation, and cross-modal generation. *BMC Bioinformatics*, 25(1):41, 2024.

[36] Chuxi Xiao, Yixin Chen, Qiuchen Meng, Lei Wei, and Xuegong Zhang. Benchmarking multi-omics integration algorithms across single-cell RNA and ATAC data. *Briefings in Bioinformatics*, 25(2):bbae095, 2024.

[37] Zizhen Yao, Hanqing Liu, Fangming Xie, Stephan Fischer, Ricky S Adkins, Andrew I Aldridge, Seth A Ament, Anna Bartlett, M Margarita Behrens, Koen Van den Berge, et al. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature*, 598(7879):103–110, 2021.

[38] Meng Zhang, Stephen W Eichhorn, Brian Zingg, Zizhen Yao, Kaelan Cotter, Hongkui Zeng, Hongwei Dong, and Xiaowei Zhuang. Spatially resolved cell atlas of the mouse primary motor cortex by merfish. *Nature*, 598(7879):137–143, 2021.

[39] Yan Zhou, Qingkai Fang, and Yang Feng. CMOT: Cross-modal mixup via optimal transport for speech translation. *arXiv preprint arXiv:2305.14635*, 2023.

A Technical Appendices and Supplementary Material

A.1 Details on experiments

The species, tissues, sample sizes, modalities, and references (the original sources) of the single-cell datasets used in the experiments are shown in Table 2. In our experiments, we set the batch size to 128, the epoch number to 200, and the dimensions of both the shared latent component (d_u) and modality-specific component (d_s) to 20. Parameters are optimized via Adam optimizer and the learning rate is started from 0.001 with a cosine annealing scheduler. Other hyperparameter settings are summarized in Table 2. 10% of the data was used as a validation set for early stopping during training based on the total loss on validation set. For scATAC-seq data, peak count signals are converted to gene-level gene activity scores using EpiScanpy. For datasets integrating two modalities, we select highly variable genes that are measured in both scRNA-seq and scATAC-seq (based on gene activity scores) as input features, and do not apply masked loss. For tri-modal integration, we use highly variable genes from scRNA-seq and scATAC-seq, as well as all genes with nonzero measurements in scProtein as input features, and apply masked loss according to the available gene list associated with each modality.

Competing methods are used with their respective default settings. Specifically, Seurat, Harmony, and scVI use gene activity scores for scATAC-seq, while all other methods operate on the original peak counts.

We run all empirical experiments on a single NVIDIA RTX 4080 GPU. The runtime of our method for 200 epochs is also reported in Table 2.

A.2 Evaluation metrics

F1 isolated label scores. The optimal F1 score by optimizing the cluster assignment of the isolated label using the F1 score across Louvain clustering (resolutions 0.1–2, step 0.1). The metric is averaged across all isolated cell-type labels.

Silhouette scores. The global silhouette width measures (ASW) between isolated and non-isolated labels on the PCA embedding, scaled to [0, 1]. ASW measures how similar a data point is to its own cluster compared to other clusters, with higher ASW indicating more compact and well-separated clusters. For the bio-conservation, ASW was computed and averaged across cell identity labels; while for modality or batch integration, ASW was computed and averaged across batch or modality labels, and then subtracted from 1 to ensure higher scores indicate better integration.

Kmeans NMI. Normalized Mutual Information (NMI) measures the similarity between clustering results and known labels, accounting for label permutations. We compute NMI between KMeans clusters and ground truth labels, with scores ranging from 0 (no overlap) to 1 (perfect match).

Kmeans ARI. Adjusted Rand Index (ARI) evaluates clustering accuracy by considering both agreements and disagreements between predicted and true labels, adjusted for chance. We compare KMeans clusters with ground truth labels, where 0 indicates random labeling and 1 indicates perfect agreement.

Graph LISI. Graph LISI is an extension of the original LISI metric that is computed from neighborhood lists per node from integrated kNN graphs. Instead of relying on a fixed number of nearest neighbors, Graph LISI computes shortest-path distances on the integrated graph to consistently define neighborhoods for each cell, providing a stable diversity score even when the underlying graph has variable connectivity. The resulting scores are rescaled to a 0–1 range, where higher values indicate better batch or modality integration (iLISI) or better cell-type separation (cLISI).

Principal component regression. Principal Component Regression (PCR) is used to quantify batch effects by assessing how much variance in the data can be attributed to batch variables or modality differences, which is computed by multiplying the variance explained by each principal component (PC) with the R^2 value from regressing the batch on that PC and then summing over all PCs.

kBET. k-nearest neighbor Batch Effect Test (kBET) assesses data mixing by checking whether the local batch label distribution in each cell's neighborhood matches the global distribution. It reports a rejection rate across tested neighborhoods, where a lower rate indicates better batch mixing.

Table 2: Details on the single-cell datasets and experimental settings.

Dataset	Species	Tissue	Modality	Sample size	Reference	Hyperparameters				Running time
						β	γ	λ	Masked loss	
small-size two-omics integration	Human	Kidney	scRNA-seq scATAC-seq	19,985 24,205	[28]	2	5	5	No	20min
large-scale two-omics integration	Mouse	Primary motor cortex	scRNA-seq scATAC-seq	69,727 54,844	[37]	2	5	5	No	120min
triple-omics integration	Human	Bone marrow	scRNA-seq scATAC-seq scProtein (CITE-seq)	30,486 10,330 18,052	[25]	5	10	10	Yes	50min

Graph connectivity. Graph connectivity, ranging from 0 to 1, measures how well cells of the same identity are connected within the integrated kNN graph. For each cell type, it computes the fraction of cells in the largest connected component of that type's subgraph.

A.3 Metric values in experiments

Unscaled metric values in all experiments are shown in Table 3. We directly used unscaled values to calculate aggregate scores.

A.4 Ablation study and sensitivity analysis

We conducted ablation experiments and sensitivity analysis on the two-omics human kidney datasets to demonstrate the effects of applying isometric loss and adversarial learning regularization within the β-VAE framework. The results have been shown and discussed in the main text (Table 1).

To evaluate the effectiveness of the Masked loss, we also conduct ablation studies on a triple-omics dataset. The results (Table4) indicate that simply setting the missing features in the proteomics to zero without applying a loss mask leads to a significant performance drop. This highlights the importance of our masked loss strategy in effectively handling integration tasks where some omics layers (e.g., scProtein) contain substantially fewer features compared to other modalities.

A.5 UMAP visualization of experimental results

We employ UMAP to visualize the results of multi-omics integration, where each point denotes the low-dimensional representation of an individual sample. An effective integration method should yield well-separated clusters corresponding to distinct cell types, thereby preserving the underlying biological heterogeneity. Concurrently, samples originating from different omics modalities and batches should be well-aligned in the embedding space, reflecting successful correction of modality-specific and batch-specific technical variations.

A.6 Spatial location imputation via integration of single-cell and spatial omics

We integrated scRNA, scATAC, and spatial transcriptomics (merFISH) data in mouse primary motor cortex using scMRDR, and then interpolated the missing spatial locations in single-cell data by conducting optimal transport between the aligned latent representation z_u of samples with spatial locations (merFISH) and samples without spatial information (scRNA and scATAC). We visualized the imputed spatial locations labeled by the cortex layer annotations provided by the original datasets (Fig. 12).

Table 3: Unscaled metric values in experiments

	aggregate score				bio-conservation					batch correction				modality integration				
	Overall score	Batch correct	Bio conserve	Modal integrate	IL	NMI	ARI	ASW	cLISI	ASW	iLISI	kBERT	PCR	ASW	iLISI	kBERT	GC	PCR
two-omicsintegration																		
Ours	0.66	0.52	0.74	0.70	0.69	0.76	0.58	0.66	1.00	0.90	0.52	0.38	0.26	0.86	0.37	0.32	0.96	0.99
GLUE	0.64	0.42	0.73	0.74	0.65	0.77	0.57	0.67	1.00	0.90	0.42	0.28	0.09	0.85	0.60	0.34	0.94	0.99
scVI	0.56	0.52	0.65	0.47	0.59	0.68	0.43	0.56	1.00	0.95	0.47	0.29	0.37	0.81	0.00	0.00	0.71	0.85
MaxFuse	0.56	0.31	0.69	0.62	0.65	0.73	0.49	0.59	1.00	0.89	0.16	0.19	0.00	0.87	0.15	0.19	0.91	1.00
Seurat	0.54	0.29	0.75	0.50	0.68	0.78	0.61	0.68	1.00	0.90	0.06	0.19	0.00	0.70	0.00	0.34	0.46	0.99
Pamona	0.50	0.58	0.47	0.45	0.50	0.22	0.18	0.50	0.96	0.93	0.51	0.26	0.61	0.74	0.00	0.09	0.43	1.00
JAMIE	0.48	0.55	0.47	0.42	0.43	0.30	0.17	0.43	1.00	0.86	0.26	0.27	0.81	0.56	0.00	0.08	0.45	0.99
SIMBA	0.46	0.73	0.35	0.35	0.49	0.01	0.01	0.49	0.76	0.98	0.70	0.27	0.95	0.90	0.00	0.00	0.03	0.83
Harmony	0.46	0.34	0.62	0.37	0.54	0.60	0.41	0.54	1.00	0.89	0.19	0.27	0.00	0.56	0.00	0.06	0.59	0.64
UnionCom	0.42	0.41	0.47	0.36	0.42	0.43	0.08	0.44	0.98	0.81	0.43	0.39	0.00	0.38	0.00	0.00	0.41	1.00
large-scaletwo-omicsintegration																		
Ours	0.60	0.56	0.62	0.62	0.58	0.58	0.37	0.56	1.00	0.90	0.24	0.14	0.97	0.82	0.26	0.10	0.95	1.00
Seurat	0.59	0.53	0.68	0.53	0.61	0.69	0.51	0.61	1.00	0.86	0.18	0.09	0.98	0.77	0.01	0.07	0.80	1.00
GLUE(LSI)	0.55	0.55	0.59	0.49	0.53	0.55	0.33	0.53	0.99	0.90	0.23	0.09	0.98	0.52	0.10	0.11	0.75	1.00
scVI	0.54	0.52	0.57	0.50	0.53	0.57	0.26	0.52	1.00	0.94	0.12	0.10	0.93	0.72	0.00	0.00	0.82	0.95
Harmony	0.46	0.45	0.54	0.37	0.52	0.47	0.21	0.51	1.00	0.93	0.10	0.11	0.67	0.57	0.00	0.00	0.58	0.68
GLUE(PCA)	0.45	0.47	0.49	0.37	0.44	0.38	0.21	0.42	0.99	0.70	0.14	0.09	0.95	0.41	0.00	0.00	0.48	0.96
MaxFuse	0.43	0.46	0.41	0.42	0.42	0.26	0.01	0.40	0.97	0.66	0.13	0.05	0.99	0.66	0.00	0.01	0.45	1.00
SIMBA	0.41	0.51	0.35	0.38	0.49	0.01	0.00	0.49	0.73	0.98	0.12	0.10	0.85	0.89	0.00	0.10	0.07	0.85
triple-omicsintegration																		
Ours	0.58	0.39	0.69	0.61	0.57	0.68	0.58	0.61	1.00	0.83	0.27	0.16	0.31	0.83	0.28	0.06	0.87	1.00
GLUE	0.55	0.36	0.67	0.58	0.55	0.67	0.55	0.59	1.00	0.86	0.20	0.13	0.25	0.85	0.15	0.09	0.82	0.99
scVI	0.49	0.49	0.58	0.38	0.53	0.50	0.31	0.54	1.00	0.90	0.20	0.32	0.53	0.76	0.00	0.00	0.46	0.70
Harmony	0.33	0.29	0.48	0.18	0.45	0.35	0.16	0.44	1.00	0.76	0.17	0.11	0.12	0.51	0.00	0.00	0.40	0.00

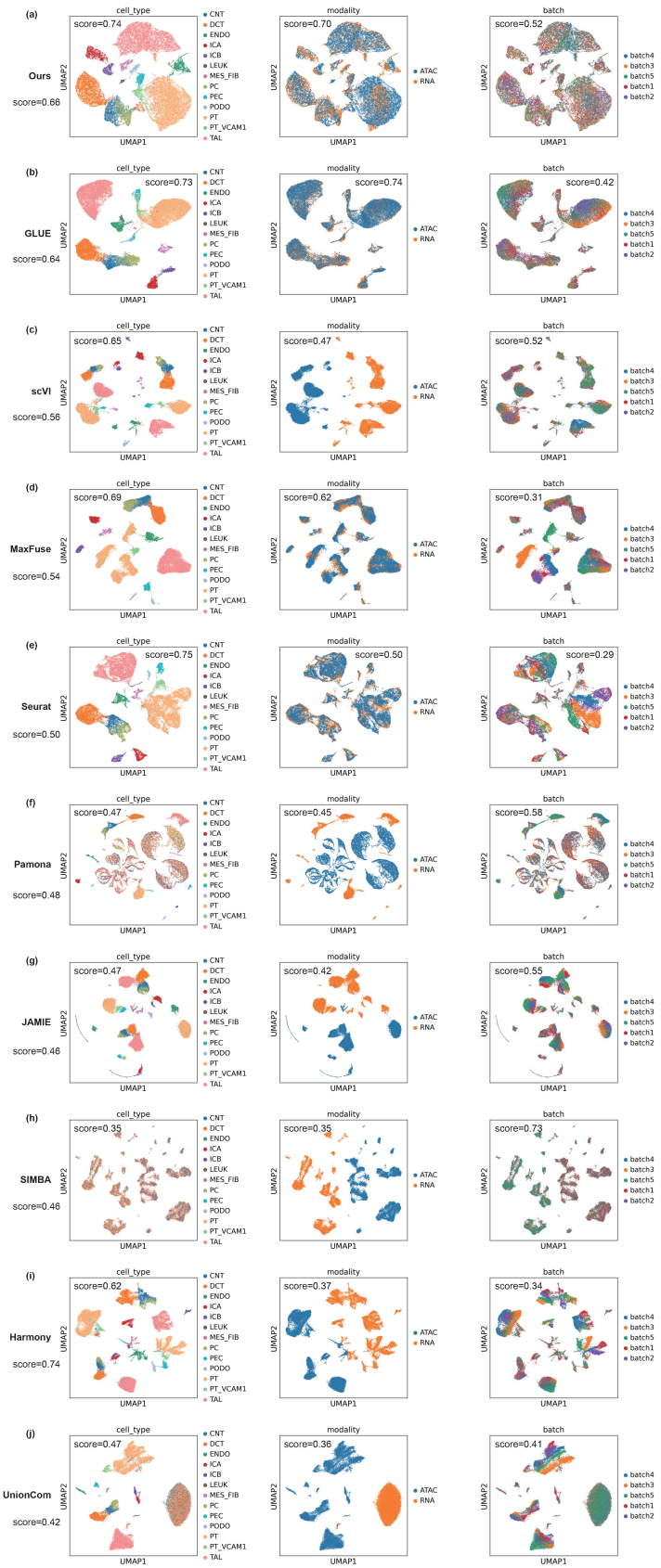


Figure 9: UMAP visualization of the unified embeddings in small-scale two-omics data integration.

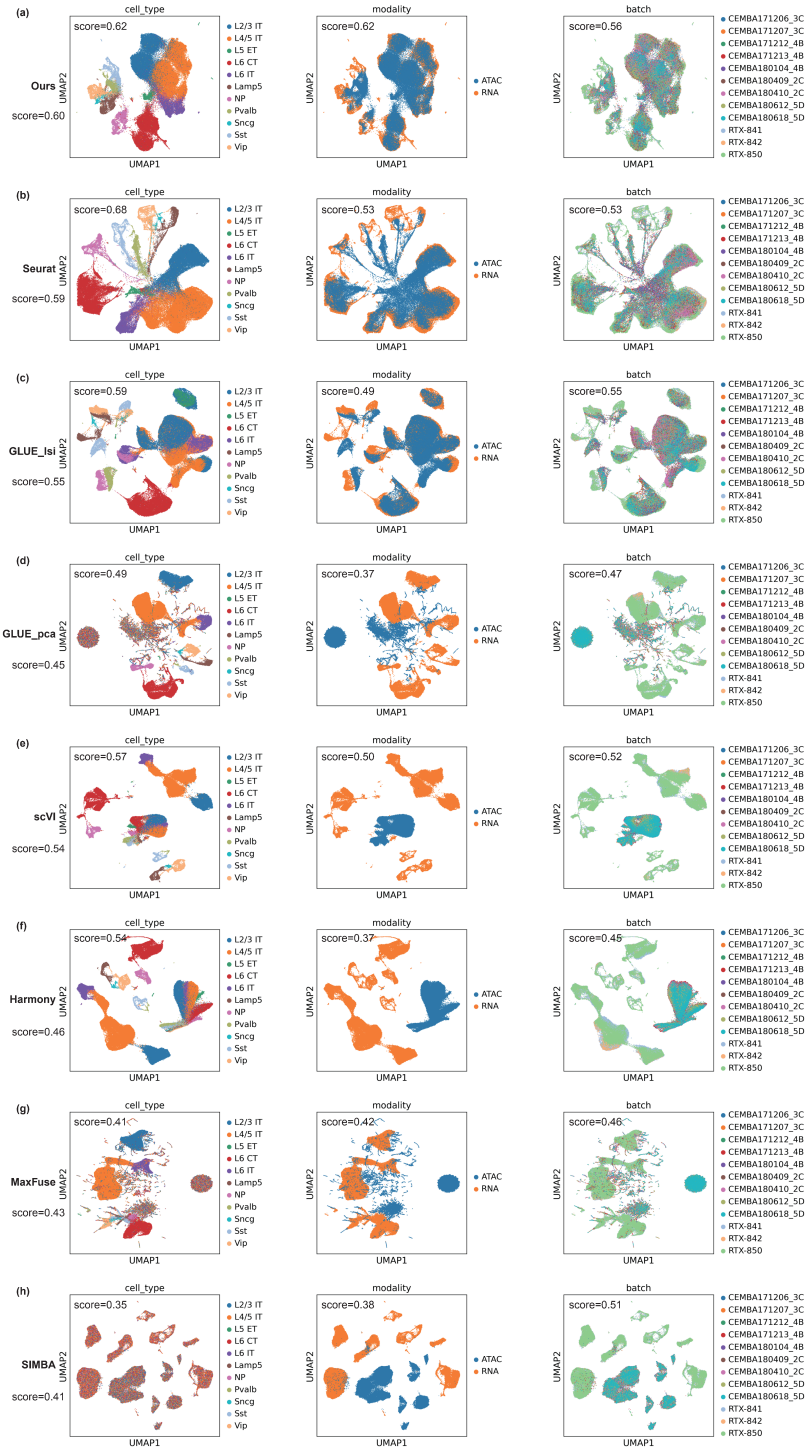


Figure 10: UMAP visualization of the unified embeddings in large-scale two-omics data integration.

Table 4: Ablation study in triple-omics data, where unscaled scIB aggregate scores are reported.

	Overall score	Batch correct	Bio conserve	Modal integrate
Ours	0.58	0.39	0.69	0.61
beta=1	0.51	0.39	0.58	0.52
lambda=0	0.48	0.32	0.60	0.49
gamma=0	0.53	0.40	0.58	0.58
w/o masked loss	0.44	0.33	0.58	0.37

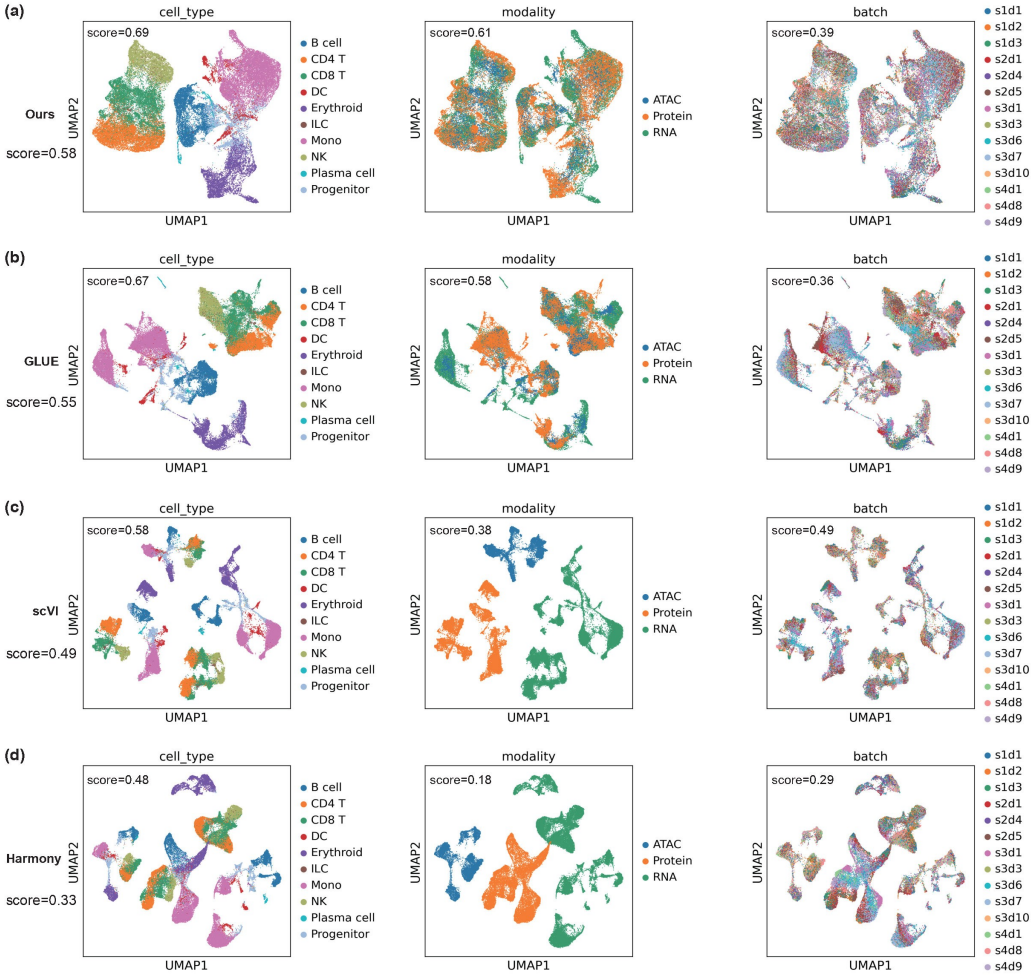
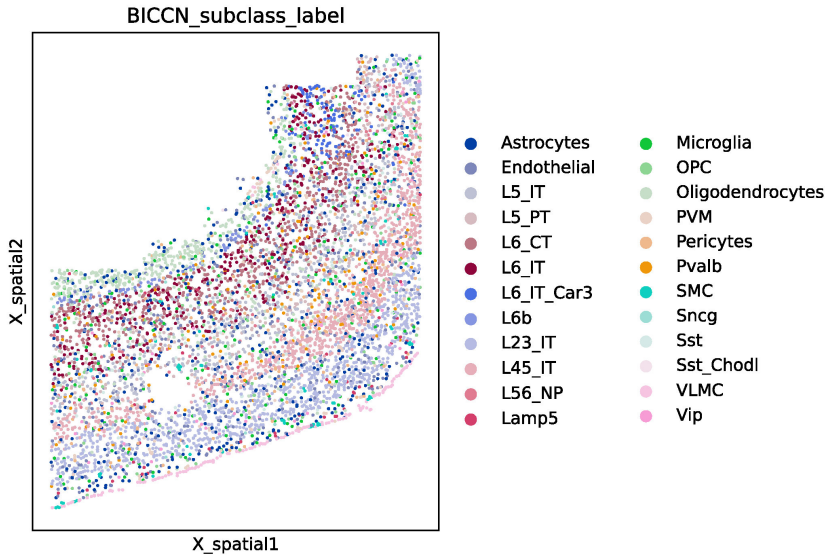
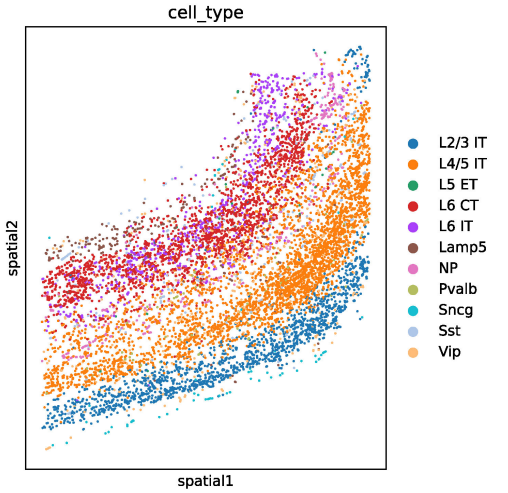


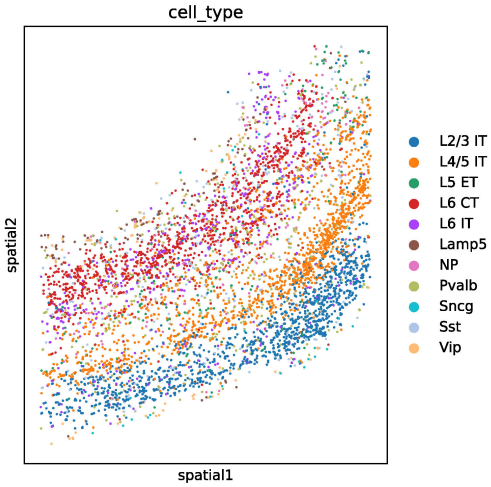
Figure 11: UMAP visualization of the unified embeddings in triple-omics data.



(a) Spatial plot of the merFISH data with originally measured spatial locations.



(b) Spatial plot of the scRNA data with imputed spatial locations.



(c) Spatial plot of the scATAC data with imputed spatial locations.

Figure 12: Spatial plots of merFISH and spatially imputed scRNA and scATAC data.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We summarize our contribution, establishing a scalable new method for multi-omics data integration, and highlight the main procedures in the abstract and introduction section.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitations in terms of unstable training, hyperparameters tuning, and feature alignment in the discussion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This work is not a theoretical work but aims at proposing a new method in computational biology.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide a detailed explanation of the underlying principles and implementation details of our method to ensure its reproducibility. Furthermore, we have packaged our approach as a module built upon Scanpy—a widely used Python library for single-cell data analysis—which will be released upon the completion of the anonymous review process.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Biological data used in our analysis are all publicly available, and we provided the original sources in the appendix table 1. Source codes are provided in the supplementary files. We will further integrate the codes into a Python package and will publish it after anonymous review.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental details are provided in Appendix A1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: This is an unsupervised learning task and no train-test split. 10% of the data was used as a validation set for early stopping during training based on the total loss on validation set. No statistical tests are used.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute workers and running times are reported in appendix A1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We make sure to preserve anonymity and follow other NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential impacts and limitations of our methods in biomedical research in the introduction and discussion section.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the codes for comparing methods are open sourced, public available and have been cited in the manuscript.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new asset is introduced

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: All the analysis are based on public available, second-handed, anonymous data, and no human participants involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: All the analysis are based on public available, second-handed, anonymous data, and no human participants involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM is only used for writing, and does not impact core methodology, scientific rigorousness, or originality of the research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.