Genomics, Proteomics & Bioinformatics, 2025, **23(4)**, qzaf062
https://doi.org/10.1093/gpbjnl/qzaf062
Advance access publication: 4 August 2025
Method

ACE: A Versatile Contrastive Learning Framework for Single-cell Mosaic Integration

Xuhua Yan [1], Jinmiao Chen [2,3,4], Ruiqing Zheng [1], Min Li [1,*]

[1]School of Computer Science and Engineering, Central South University, Changsha 410083, China
[2]Center for Computational Biology and Program in Cancer and Stem Cell Biology, Duke-NUS Medical School, Singapore 169857, Singapore
[3]Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), Singapore 138671, Singapore
[4]Immunology Translational Research Program, Department of Microbiology and Immunology, Yong Loo Lin School of Medicine, National University of Singapore (NUS), Singapore 117545, Singapore

*Corresponding author: limin@mail.csu.edu.cn (Li M).

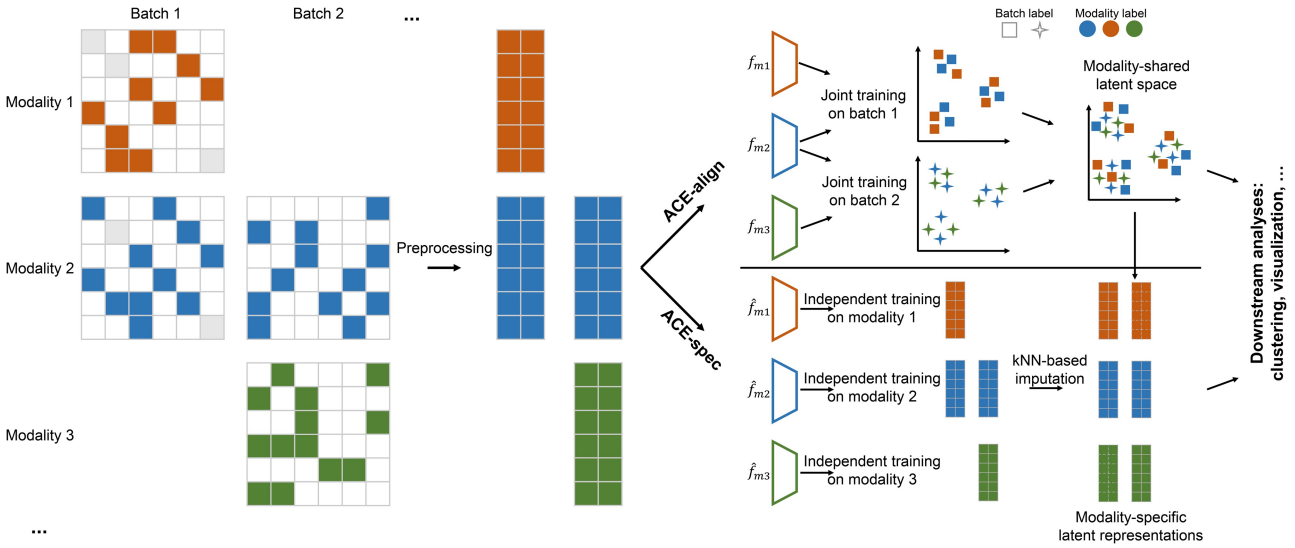
Handling Editor: Wei Lin

Abstract

The integration of single-cell multi-omics datasets is critical for deciphering cellular heterogeneities. Mosaic integration, the most general integration task, poses a greater challenge regarding disparity in modality abundance across datasets. Here, we present Align and CompletE (ACE), a mosaic integration framework that assembles two types of strategies to handle this problem: modality alignment-based strategy (ACE-align) and regression-based strategy (ACE-spec). ACE-align utilizes a novel contrastive learning objective for explicit modality alignment to uncover the shared latent representations behind modalities. ACE-spec combines the modality alignment results and modality-specific representations to construct complete multi-omics representations for all datasets. Extensive experiments across various mosaic integration scenarios demonstrate the superiority of ACE's two strategies over existing methods. Application of ACE-spec to bi-modal and tri-modal integration scenarios showcases that ACE-spec is able to enhance the representation of cellular heterogeneities for datasets with incomplete modalities. The source code of ACE can be accessed at https://github.com/CSUBioGroup/ACE-main.

Key words: Single cell; Multi-omics; Mosaic integration; Imputation; Contrastive learning.

Graphical abstract



Introduction

Single-cell RNA sequencing (scRNA-seq) allows for quantifying RNA molecular traits at the single-cell level, enabling applications from cellular heterogeneity characterization [1] to regulatory network inference [2]. However, biological processes in cells involve multiple molecular types, such as DNA, RNA, and protein, resulting in an intrinsic need for multimodal understanding. Advances in single-cell sequencing have enabled technologies that can capture multiple molecule types. For example, cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) captures RNA expression and cell surface protein abundance via antibody-derived tag (ADT) [3], 10X Genomics Multiome captures RNA expression alongside transposase-accessible DNA fragments (ATAC) [4], and DNA,

Received: 21 May 2024; **Revised:** 5 February 2025; **Accepted:** 19 February 2025.

Open chromatin, Gene expression, and surface protein Multiomics Assay sequencing (DOGMA-seq) simultaneously measures chromatin accessibility, gene expression, and protein abundance [5]. To gain biological insights from such rich resources, data integration has emerged as a key challenge.

Ricard et al. [6] categorize data integration tasks on single-cell data into four scenarios: horizontal, vertical, diagonal, and mosaic. Horizontal integration, also termed batch correction, refers to the scenario in which all data batches share the same modality. Vertical integration refers to the scenario where data batch is measured with multiple modalities. Diagonal integration refers to the scenario where data batches do not share any modality. Mosaic integration refers to the most general scenario where different batches are profiled with various modalities, like a grid that's concatenated using various horizontal, vertical, and diagonal scenarios. Among the four scenarios, mosaic integration is the most challenging one.

Similar to horizontal integration, mosaic integration also has the goal of preserving cellular heterogeneities and aligning data batches [7]. The distinction is that in mosaic integration, the inter-batch differences not only exist within the same modality but also come from difference in modality abundance. Different modalities may carry different information content and different batches can be measured with different modalities. Existing mosaic integration methods mainly adopt regression-based strategy to solve this problem. In particular, they attempt to recover the unmeasured modality for each batch using their measured modalities. For example, StabMap [8] sets the batch measured with multiple modalities as reference (bridge batch) and calculates its low-dimensional embeddings which contain information from multiple modalities. Then, on the bridge batch, it trains regression models to predict these embeddings based on single-modality profile and applies these regression models to other batches with single modality only. The core idea of Cobolt [9] is similar, which uses multimodal variational autoencoder (MVAE) [10] to embed all batches first and then trains Extreme Gradient Boosting (XGBoost) [11] to predict the reference embeddings from single modality. CLUE [12] also adopts the MVAE framework, but it doesn't predict the reference embeddings directly. Instead, CLUE predicts the unmeasured modality based on the measured information and aggregates all modalities for each batch. Some of mosaic integration methods adopt non-negative matrix factorization to decompose each batch into cell factors and feature-specific factors, such as UINMF [13] and scMoMaT [7]. Those cell factors are expected to represent shared heterogeneities across all batches.

Here, we develop a novel mosaic integration framework, Align and CompletE (ACE), which builds on a new integration strategy and can be flexibly adjusted to encompass the output of regression-based methods. Specifically, we propose to build a modality-aligned latent space in which cells with similar biological signals are concentrated rather than batch labels or modality labels. In other words, we aim to extract the shared latent representations behind modalities. To achieve this target, we utilize contrastive learning (CL) to align modalities and propose a new learning objective to address the modality gap phenomena caused by the commonly used Information Noise Contrastive Estimation (InfoNCE) loss [14,15]. Then, we can construct consensus latent representations across batches based on modality-aligned outputs. However, we also realize that modality alignment can erase the unique biological variations in each modality, especially when the disparity in the information content between modalities is significant. In this case, we think the regression-based strategy is more suitable. Accordingly, our method can be flexibly adapted to encompass regression-based outcomes: we construct modality-specific embeddings by learning independent modality encoders and then recover the embeddings of missing modalities through a cross-modality matching approach based on prior alignment results. Together, we propose a mosaic integration framework which can provide modality alignment-based outputs and regression-based outputs, facilitating its use in various integration scenarios. For convenience, we refer to the alignment-based outputs as ACE-align, and regression-based outputs as ACE-spec.

We evaluated the two types of outputs on four integration scenarios which cover different modality compositions (*e.g.*, CITE-seq and Multiome sequencing), different numbers of modalities (2 or 3 modalities), existence of horizontal batch effects, different cell type compositions across batches. Experimental results show that both outputs of ACE's framework can achieve superior mosaic integration performance. Moreover, we showcased that ACE-spec could enhance representation of cellular heterogeneities, help refine cell type annotations, and reveal differences between cell subpopulations. Finally, we validated the robustness of our framework across various aspects, including the loss function, sensitivity to hyperparameters, choice of batch correction methods and clustering methods.

Method

Overview of ACE

Let $X_{b_i}^{m_j} = \left[x_{b_i,1}^{m_j}, x_{b_i,2}^{m_j}, \ldots, x_{b_i,N_{b_i}}^{m_j}\right] \in R^{N_{b_i} \times D_{m_j}}$ denote inputs of one data batch b_i measured with modality m_j, where N_{b_i} denotes the number of cells and D_{m_j} denotes the input dimensions of modality m_j. $X_{b_i} = \left[X_{b_i}^{m_1}, \ldots, X_{b_i}^{m_{M_b}}\right] \in R^{N_{b_i} \times \sum_j D_{m_j}}$ denotes one data batch b_i measured with $\{m_1, \ldots, m_{M_b}\}$ modalities and $\{m_1, \ldots, m_{M_b}\} \in \{\text{RNA, ATAC, Histone mark, Protein}\}$. The general target of the mosaic integration task is to project each data batch b_i into a consensus latent space $Z_{b_i} = [z_{b_i,1}, z_{b_i,2}, \ldots, z_{b_i,N_{b_i}}]$, in which cells with similar biological signatures are clustered while cells with dissimilar biological signatures are separated.

ACE requires the existence of "bridge" batches in a mosaic dataset. More specifically, if each modality is seen as a node, one batch that's measured with two modalities, RNA and protein, connects nodes of RNA and protein. ACE requires all modalities in a dataset to be connected through several batches. These bridge batches lay the foundation for modality alignment. To remove within-modality batch effects across batches (if they exist), ACE applies Harmony [16] to perform horizontal integration within each modality. Those low-dimensional representations are used as inputs of different modalities for each batch.

As mentioned in previous section, our framework generates two types of outputs: ACE-align and ACE-spec. ACE-align is derived from a modality alignment strategy which creates a shared latent space to align representations across modalities while preserving biological variations within each modality. Building on ACE-align, ACE-spec extends this

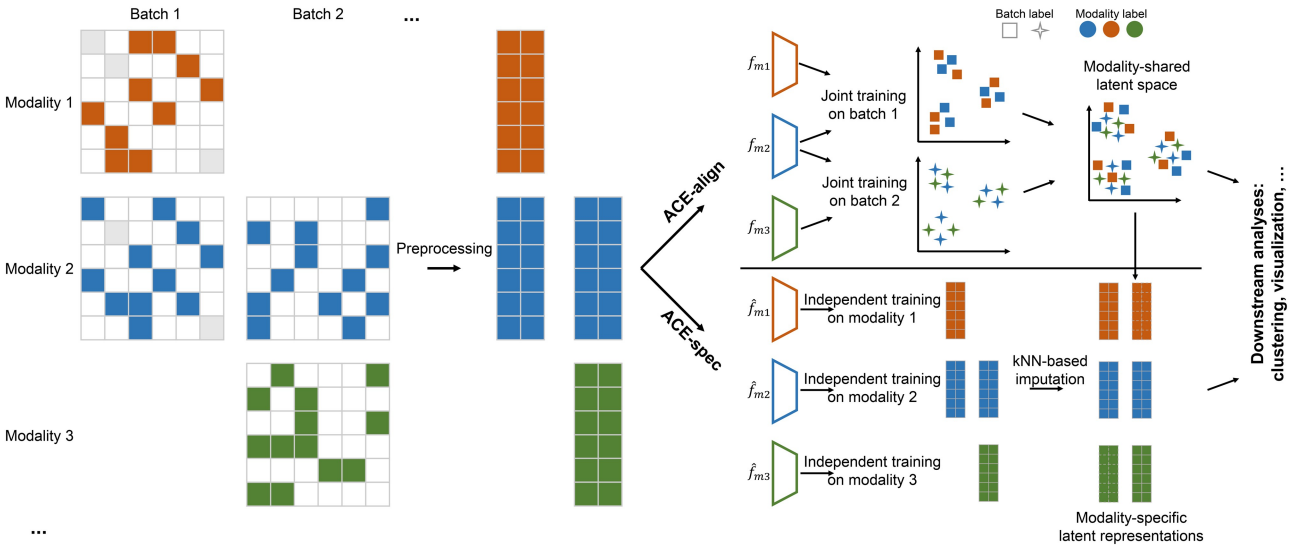


Figure 1 Architecture of ACE

ACE assembles two types of outputs: ACE-align and ACE-spec. Before model training, each modality profile is preprocessed through dimension reduction and batch correction (Harmony), respectively. ACE-align takes the low-dimensional representations as input and jointly trains the modality encoders on bridge batches to achieve a modality-shared latent space. ACE-spec independently trains modality encoders on corresponding modality inputs, and then utilizes the modality-aligned latent space to impute the missing modality-specific representations. Finally, both outputs eliminate the inter-batch differences in horizontal batch effects and modality abundance. ACE, Align and CompletE; kNN, k-nearest neighbors.

framework by constructing modality-specific representations and imputing missing modalities through a cross-modality matching-based strategy. An overview of ACE is shown in Figure 1.

ACE-align

To build a modality-shared latent space, we utilize CL to perform modality alignment. CL is a popular self-supervised learning framework which learns representations by concentrating predefined positive pairs while separating negative pairs [15]. CL loss has been validated as an effective objective for representation alignment because it can well align batches while preserving variations between cells [17], which motivates us to apply CL to perform modality alignment. Before us, NOVEL [18] and MatchCLOT [19] also used the common CL loss, InfoNCE [15], to learn aligned representations between two modalities of the same cells. However, NOVEL and MatchCLOT can only integrate datasets with two modalities and cannot be directly extended to datasets with three modalities. Additionally, the InfoNCE loss they adopted encounters the modality gap problem, which has been observed in the field of computer vision [14]. In other words, representations from different modalities are still clearly separable after alignment, indicating a poor quality of the modality-shared latent space.

For ACE-align, embeddings of different modalities from the same cell form positive pairs. For instance, if a cell i is measured with three modalities m_1, m_2, and m_3 (with embeddings $z_i^{m_1}, z_i^{m_2}$, and $z_i^{m_3}$), we define three positive pairs $(z_i^{m_1}, z_i^{m_2})$, $(z_i^{m_1}, z_i^{m_3})$, and $(z_i^{m_2}, z_i^{m_3})$. Negative pairs are constructed by randomly sampling embeddings from different cells, irrespective of modality (Figure S1A). For convenience, we first discuss the two-modality situation. During training, one mini-batch of cells with size n is sampled from a bridge batch: $B = \{x_{b,1}^{m_1}, x_{b,2}^{m_1}, \ldots, x_{b,n}^{m_1}\} \cup \{x_{b,1}^{m_2}, x_{b,2}^{m_2}, \ldots, x_{b,n}^{m_2}\}$. Each modality is projected by a separate encoder f_{m_j} into a low-dimensional latent space $z^{m_j} \in R^d$. CL loss is used to train all encoders. InfoNCE is a commonly used CL loss function in cross-modality alignment research [20]. It is defined as [15]:

$$l(i) = \log \frac{\exp\left(\frac{sim\left(z_{b,i}^{m_1}, z_{b,i}^{m_2}\right)}{\tau}\right)}{\sum_{l=1}^{n} \exp\left(\frac{sim\left(z_{b,i}^{m_1}, z_{b,l}^{m_2}\right)}{\tau}\right)} + \log \frac{\exp\left(\frac{sim\left(z_{b,i}^{m_2}, z_{b,i}^{m_1}\right)}{\tau}\right)}{\sum_{l=1}^{n} \exp\left(\frac{sim\left(z_{b,i}^{m_2}, z_{b,l}^{m_1}\right)}{\tau}\right)} \quad (1)$$

$$L_{nce} = -\frac{1}{2n} \sum_{i=1}^{n} l(i) \quad (2)$$

where $sim(z_{b,i}^{m_1}, z_{b,i}^{m_2})$ denotes the cosine similarity between two embeddings and τ can be a positive constant or a learnable parameter. By optimizing the aforementioned objective, embeddings of different modalities from the same cell are pushed closer than embeddings of different modalities from different cells. However, one critical limitation of InfoNCE loss is that it ignores the relationships within intra-modality embeddings. Specifically, Equation 1 only optimizes inter-modality embeddings of the same cell to be closer than inter-modality embeddings between different cells, indicating that the intra-modality embeddings between different cells can still be closer than inter-modality embeddings of the same cell. As a result, the modality gap remains. Our solution is to add intra-modality embeddings between different cells as

negative pairs in Equation 1. For convenience, we define $s_{i,l}^{1,2}$ as a shorthand for $sim(z_{b,i}^{m_1}, z_{b,l}^{m_2})$. Then, our proposed CL loss is defined as:

$$l(i) = \log \frac{\exp\left(\frac{s_{i,i}^{1,2}}{\tau}\right)}{\sum_{l=1}^{n} \exp\left(\frac{s_{i,l}^{1,2}}{\tau}\right) + \sum_{l \neq i}^{n} \exp\left(\frac{s_{i,l}^{1,1}}{\tau}\right)} + \log \frac{\exp\left(\frac{s_{i,i}^{2,1}}{\tau}\right)}{\sum_{l=1}^{n} \exp\left(\frac{s_{i,l}^{2,1}}{\tau}\right) + \sum_{l \neq i}^{n} \exp\left(\frac{s_{i,l}^{2,2}}{\tau}\right)} \tag{3}$$

$$L_{nce} = -\frac{1}{2n} \sum_{i=1}^{n} l(i) \tag{4}$$

When there are M $(M > 2)$ modalities to be aligned, we generalize the aforementioned objective to:

$$l(i,j) = \sum_{p \neq j}^{M} \log \frac{\exp\left(\frac{s_{i,i}^{j,p}}{\tau}\right)}{\sum_{v \neq j}^{M} \sum_{l=1}^{n} \exp\left(\frac{s_{i,l}^{j,v}}{\tau}\right) + \sum_{l \neq i}^{n} \exp\left(\frac{s_{i,l}^{j,j}}{\tau}\right)} \tag{5}$$

$$L_{nce} = -\frac{1}{M \cdot n} \sum_{i=1}^{n} \sum_{j=1}^{M} l(i,j) \tag{6}$$

By optimizing this objective, all modalities' embeddings of one cell are concentrated without a modality gap, and embeddings from different cells are separated. After modality alignment, we can build the consensus representation across batches by combining embeddings from measured modalities. One common approach is weighted average:

$$z_{b,i} = \sum_{m_j \in T_b} w_{mj} \cdot z_{b,i}^{m_j} \tag{7}$$

where T_b denotes the set of modalities that batch b is measured with; $w_{m_j} \in [0,1]$ denotes the weight of modality m_j and $\sum_{m_j \in T_b} w_{m_j} = 1$.

ACE-spec

We are aware that modality alignment can bring loss of unique information inherent in each modality. Especially when the differences in biological variations are large between modalities, the modality with richer information will be biased to modality with poorer information [21]. To address this problem, we extend ACE-align to ACE-spec. Our approach is straightforward: learning modality-specific representations effectively preserves their original variations but fails to address inter-batch differences in modality abundance; thus, we design a strategy to impute the missing modalities for each batch based on the results of ACE-align.

Within each modality m_j, we employ CL to train an encoder network, denoted as $\widehat{f}_{m_j}$ to project inputs into a modality-specific latent space. The training is conducted independently for each modality. Positive pairs are defined as each cell paired with itself, while negative pairs are constructed by randomly sampling embeddings from different cells within the same modality (Figure S1B). For a mini batch of n cells sampled from modality m_j, denoted as $\{x_1^{m_j}, x_2^{m_j}, \ldots, x_n^{m_j}\}$, the loss function is defined as:

$$L = -\frac{1}{n} \sum_{i=1}^{n} \log \frac{\exp\left(\frac{sim\left(\widehat{z}_i^{m_j}, \widehat{z}_i^{m_j}\right)}{\tau}\right)}{\sum_{l=1}^{n} \exp\left(\frac{sim\left(\widehat{z}_i^{m_j}, \widehat{z}_l^{m_j}\right)}{\tau}\right)} \tag{8}$$

where $\widehat{z}_i^{m_j} = \widehat{f}_{m_j}(x_i^{m_j})$ and $sim\left(\widehat{z}_i^{m_j}, \widehat{z}_i^{m_j}\right) = 1$, indicating that the objective is to maximize the separation between negative pairs. In other words, optimizing this function enables the model to learn discriminative features between different cells, which we believe helps better preserve biological variations within each modality.

To impute embeddings for missing modalities, we combine the modality-shared embeddings z and modality-specific embeddings $\widehat{z}$ to infer the embeddings for missing modalities. In detail, let $Z^{m_j} = \{z_1^{m_j}, \ldots, z_{N_{m_j}}^{m_j}\}$ denote the aligned embeddings of modality m_j from all batches, where N_{m_j} denotes the number of cells that are measured with modality m_j. $\widehat{Z}^{m_j} = \{\widehat{z}_1^{m_j}, \ldots, \widehat{z}_{N_{m_j}}^{m_j}\}$ denotes the modality-specific embeddings of m_j from all batches. Let T_{b_i} denote the set of modalities that batch b_i is measured with. $\bar{T}_{b_i} = \{m_1, \ldots, m_M\} - T_{b_i}$ denotes the set of missing modalities in batch b_i. For each modality $m_j \in \bar{T}_{b_i}$, we first perform cross-modality matching between Z^{m_j} and $\left\{z_{b_i,l}^{m_t} | m_t \in T_{b_i}\right\}$. Specifically, for each $z_{b_i,l}^{m_t}$, we find its k nearest neighbors $Q = \{q_1, q_2, \ldots, q_k\}$ in Z^{m_j}. Then, modality-specific embedding of m_j imputed from modality m_t is defined as:

$$\widehat{z}_{b_i,l}^{\,m_t \to m_j} = \frac{1}{k} \sum_{q \in Q} \widehat{z}_q^{m_j} \tag{9}$$

where $m_t \to m_j$ denotes that the embedding is imputed based on modality m_t. For each modality in T_{b_i}, we repeat the aforementioned imputation process, and the final imputed embedding for missing modality m_j of batch b_i is:

$$\widehat{z}_{b_i,l}^{m_j} = \frac{1}{|T_{b_i}|} \sum_{m_t \in T_{b_i}} \widehat{z}_{b_i,l}^{\,m_t \to m_j} \tag{10}$$

where $|T_{b_i}|$ denotes the number of elements in T_{b_i}. After the imputation, we average the representations from all modalities to get the consensus representation for each batch, which is similar to Equation 7.

The workflow of ACE-spec is illustrated in Figure S2. Essentially, the computation process of ACE-spec is similar to the regression-based methods, but our method completely decouples the modality-specific learning process and cross-modality learning process, ensuring better preservation of unique information in each modality. Notably, our proposed imputation strategy can also be applied to impute raw omics features. The modification is to change $\widehat{Z}^{m_j}$ to the expression profiles of modality m_j.

Implementation detail

ACE-align employs separate encoder networks to project inputs from each modality into a shared embedding space.

Each encoder network consists of three fully connected layers. The RNA-specific and ATAC-specific encoders have output dimensions of 1024, 512, and 256 for the three layers, while the protein-specific encoder has output dimensions of 512, 2048, and 256 for the three layers. In each encoder, the first two fully connected layers are followed by an exponential linear unit (ELU) activation [22] and dropout regularization ($P = 0.2$) [23]. The CL loss is directly computed on the embeddings. The temperature parameter for CL is fixed at 0.1. ACE-align is trained using the Adam optimizer [24] with a learning rate of 2E−4, a batch size of 512, and 100 training epochs. ACE-spec shares the same network architecture and the temperature parameter with ACE-align. ACE-spec is trained using the Adam optimizer with a learning rate of 1.75E−4, a batch size of 512, and 10 training epochs. The models are implemented using PyTorch [25]. For embedding imputation, the number of nearest neighbors (k) is set to 2. When combining modality-specific embeddings to generate the final embeddings for ACE-spec, equal weights are assigned to all modalities. For ACE-align, the weights of RNA embeddings (or protein embeddings) are set to 1 for multimodal batches. In single-modal batches, ACE-align assigns a weight of 1 to the respective modality.

Datasets

We collected six publicly available datasets which covered different sequencing technologies, different numbers of modalities, and different scenarios of missing modalities. CITE and Multiome datasets are NeurIPS 2021 multimodality competition [26] datasets sequenced by two types of technologies, CITE-seq and 10X Genomics Multiome. CITE dataset contains two modalities: gene expression (RNA) and protein abundance (ADT), while Multiome dataset contains two modalities: gene expression (RNA) and chromatin accessibility (ATAC). Following the data splitting scheme in the competition, we first divided all 12 batches in the CITE dataset into two sets: training and testing. Training and testing sets contain 9 and 3 batches, respectively. Then, the RNA and ADT profiles of the testing set were split into two parts: RNA-modal part and ADT-modal part, and they were treated as originating from different experiments. The task on the CITE dataset is to integrate multimodal (RNA + ADT) part (training set), RNA-modal part, and ADT-modal part. Similar to the splitting scheme on CITE, Multiome dataset was split into three parts: multimodal part (10 batches), RNA-modal part (3 batches), and ATAC part (3 batches). Bone marrow dataset [27] was sequenced using CITE-seq (referred to as BM-CITE dataset), consisting of 30,672 cells measured alongside a panel of 25 antibodies from bone marrow. BM-CITE consists of two batches. Following the split scheme of CITE and Multiome datasets, the two batches were split into three parts: multimodal part (batch 1), RNA-modal part (from batch 2), and ADT-modal part (from batch 2). PBMC-Mult is a publicly available dataset on the 10X website (https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k), where paired transcriptomes and ATAC profiles were measured in 10,412 peripheral blood mononuclear cells (PBMCs). PBMC-Mult consists of one batch, and we randomly split it into three parts: multimodal part, RNA-modal part, and ATAC-modal part. DOGMA dataset [5] contains two batches profiled by DOGMA-seq, which measures RNA, ATAC, and ADT data simultaneously. Following the split scheme, these two batches were split into four parts: multimodal part (from "control" batch), RNA-modal batch (from "stim" batch), ATAC-modal part (from "stim" batch), and ADT-modal part (from "stim" batch). CITE-ASAP dataset [5] contains four batches, in which two of them were profiled with CITE-seq and the other two were profiled with Assay for Single-cell multi-omics chromatin Accessibility and Protein sequencing (ASAP-seq). ASAP-seq measures RNA and ATAC. CITE-ASAP dataset was not split. Cell type labels of all datasets were collected from original studies.

Evaluation metrics

Following existing studies [7,8], we evaluated mosaic integration methods from three aspects: biological preservation, batch correction, and modality alignment. In brief, biological preservation metrics compromise Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI). NMI and ARI both evaluate the overlap of two clusterings. Score 0 corresponds to random clustering and score 1 corresponds to the perfect match for both NMI and ARI.

We used graph inverse Local Inverse Simpson's Index (iLISI) [28] as the batch correction metric. iLISI score is a diversity score to assess batch mixing degree. Following StabMap [8], we not only used batch labels (iLISI_{batch}) to compute iLISI but also used modality labels (iLISI_{mod}). For example, if one batch is measured with RNA and ADT, its modality label is RNA + ADT, whereas if one batch is measured with RNA, then its modality label is RNA. It can be regarded as evaluation at two levels of resolution, in which the modality label is at coarse-grained resolution. Score 0 corresponds to separation of batches and score 1 corresponds to the perfect mixing of batches.

Modality alignment evaluation metrics compromise Fraction of Samples Closer Than the True Match (FOSCTTM) [29] and matching score (MS) [26]. Both metrics evaluate the alignment between embeddings of different modalities for the same cells. Score 1 indicates perfect modality alignment for both metrics. The details of calculating each metric can be found in Note A1 in File S1.

Following the single-cell Integration Benchmarking (scIB) [28], we aggregated all metrics into three scores: bio-conservation score, batch correction score, and modality alignment score. For each method q_i on a dataset, its three scores were calculated via:

$$\begin{aligned} S_{bio} &= \frac{\text{NMI}_{q_i} + \text{ARI}_{q_i}}{2} \\ S_{batch} &= \frac{\text{iLISI}_{batch,q_i} + \text{iLISI}_{mod,q_i}}{2} \\ S_{modAlign} &= \frac{\text{FOSCTTM}_{q_i} + \text{MS}_{q_i}}{2} \end{aligned} \tag{11}$$

Following scIB, each metric was min-max scaled among compared methods before metric aggregation so that all metrics have equal weights [28]. Finally, three scores were aggregated into the overall score as follows:

$$S_{overall} = w_{bio} \cdot S_{bio} + w_{batch} \cdot S_{batch} + w_{modAlign} \cdot S_{modAlign} \tag{12}$$

If the dataset doesn't support evaluation of modality alignment metrics, we used $w_{bio} = 0.6$ and $w_{batch} = 0.4$, following

the scIB benchmarking approach [28]. This weighting emphasizes the importance of preserving biological information while accounting for batch effects. When the dataset supports evaluation of modality alignment metrics, we used $w_{bio} = 0.4$, $w_{batch} = 0.3$, and $w_{modAlign} = 0.3$, following the single-cell Multimodal Integration Benchmarking (scMIB) benchmarking proposed by MIDAS [30]. This setup similarly prioritizes biological conservation by giving it a higher weight.

Results

ACE shows superior performance in bi-modal mosaic integration

We organized four cases to evaluate mosaic integration methods across various batch effects and cell type composition scenarios. Case 1 includes the BM-CITE and PBMC-Mult datasets, which lack intra-modality batch effects and have similar cell type compositions across batches. Case 2 includes the CITE and Multiome datasets, which exhibit intra-modality batch effects but maintain similar cell type compositions. In case 3, we created multiple datasets by selecting different cell types from each part of the BM-CITE dataset and adjusted the proportion of shared cell types among batches (proportion = 0.1, 0.2, 0.4, and 0.8) to simulate variations in cell type composition without batch effects. For case 4, we applied the same selection strategy to the CITE dataset to generate multiple datasets with both intra-modality batch effects and varying cell type compositions. We compared ACE with five state-of-the-art bi-modal mosaic integration methods: Cobolt, CLUE, MatchCLOT, scMoMaT, and StabMap. We used Harmony as a post-processing step for Cobolt, CLUE, scMoMaT, and StabMap to enable a more thorough benchmarking, which led to four additional methods: Cobolt-Harmony, CLUE-Harmony, scMoMaT-Harmony, and StabMap-Harmony. Detailed method settings can be found in Note A2 in File S1.

On both datasets of case 1, ACE-spec achieves the highest overall scores, followed by ACE-align (**Figure 2A**, Figure S3A and B). ACE-spec has the highest bio-conservation scores and the highest modality alignment scores. Its batch correction scores are also comparable to the top performers. ACE-spec outperforms ACE-align by 6% with respect to ARI on average, indicating that using modality-specific representations can help preserve cellular heterogeneity. We visualized the latent embeddings of all methods on both datasets using Uniform Manifold Approximation and Projection (UMAP) [31] (Figure 2B, Figure S4A), and the visualization shows that in ACE-spec outputs, batches are well mixed and cell types are clearly separated. For example, CD4 naïve and CD8 naïve are well separated in ACE-spec's outputs while for other methods, those two types are basically connected. Notably, MatchCLOT has the third highest modality alignment scores on BM-CITE dataset, but its embeddings show clear separation between batches that are measured with different modalities, which is the modality gap phenomenon and explains its 0 batch correction scores on this dataset. However, such phenomena are not observed in ACE-align's embeddings, demonstrating that our proposed CL loss can better align embeddings between batches and is more suitable for mosaic integration tasks (refer to Note B in File S1 for further details). For the reason why MatchCLOT's embeddings on PBMC-Mult dataset do not show the modality gap (Figure S4B), we think it's probably because of greater information content disparity between protein and RNA, compared to that between ATAC and RNA.

In case 2 where each dataset contains a larger number of cells and more cell types than case 1, ACE-spec still achieves the highest overall scores, followed by ACE-align (Figure 2A, Figures S3C, S3D, and S5). The bio-conservation scores of ACE-spec are also the highest and its modality alignment scores rank the first and the second on two datasets, respectively. Its batch correction performance is the best on the CITE dataset and competitive on the Multiome dataset. Despite ACE-align not achieving the highest scores on these datasets, its performance remains comparable to the leading method. UMAP visualizations also show that ACE-spec produces well mixed batches and better separation of cell types (Figure 2C, Figure S6). Similar to case 1, we observed modality gap phenomena in MatchCLOT's results for the CITE dataset while it does not appear in ACE-align's results.

In case 3 and case 4, we did not evaluate the modality alignment score because the number of matched test cell pairs can be scarce. With different proportions of shared types, ACE-align and ACE-spec consistently attain the top overall scores (Figure 2D, Figure S7A and B). When the proportion above certain thresholds, ACE-spec shows superior performance over other methods. When the proportions are small, ACE-align and ACE-spec both rank within the top 5 among all methods, indicating their good generalization capabilities. Notably, when the proportions are small (< 0.1), ACE-align shows higher bio-conservation scores (and overall scores) than ACE-specin in both cases. The reason is that when the number of shared cell types among batches is small, it's very likely to match cells with different cell types, eroding the original signal of cellular heterogeneity.

Finally, we evaluated the robustness of all methods against the number of cells in bridge batches. Specifically, we randomly removed certain proportions of cells from the bridge batches in the CITE dataset (proportion = 0.1, 0.2, 0.4, and 0.8, remaining 60,000, 53,000, 40,000, and 13,000 cells, respectively) and evaluated all the methods on these new datasets. Generally, with the proportion of removed cells increasing, all methods' performance decreased with respect to specific metrics (Figure S7C). However, ACE-align and ACE-spec still attain the top 3 overall scores among all methods (Figure 2E).

ACE achieves competitive performance in tri-modal mosaic integration

We further applied ACE to tri-modal integration. DOGMA dataset and CITE-ASAP dataset were used in this experiment. They both exhibit intra-modality batch effect, and following bi-modal settings, we organized tri-modal case 2 and tri-modal case 4 on them. Specifically, case 2 includes the DOGMA and CITE-ASAP datasets. In case 4, following bi-modal settings, we randomly sampled different cell types from each part in the DOGMA dataset, respectively, and constructed multiple datasets with various proportions of shared types among batches. We used Cobolt, CLUE, scMoMaT, and StabMap for comparisons. MatchCLOT was not included because it cannot handle tri-modal integration situations. We also added Harmony as the post-processing step for CLUE, Cobolt, scMoMaT, and StabMap. Note that the CITE-ASAP dataset does not contain test cell pairs, so we didn't evaluate modality alignment on it.

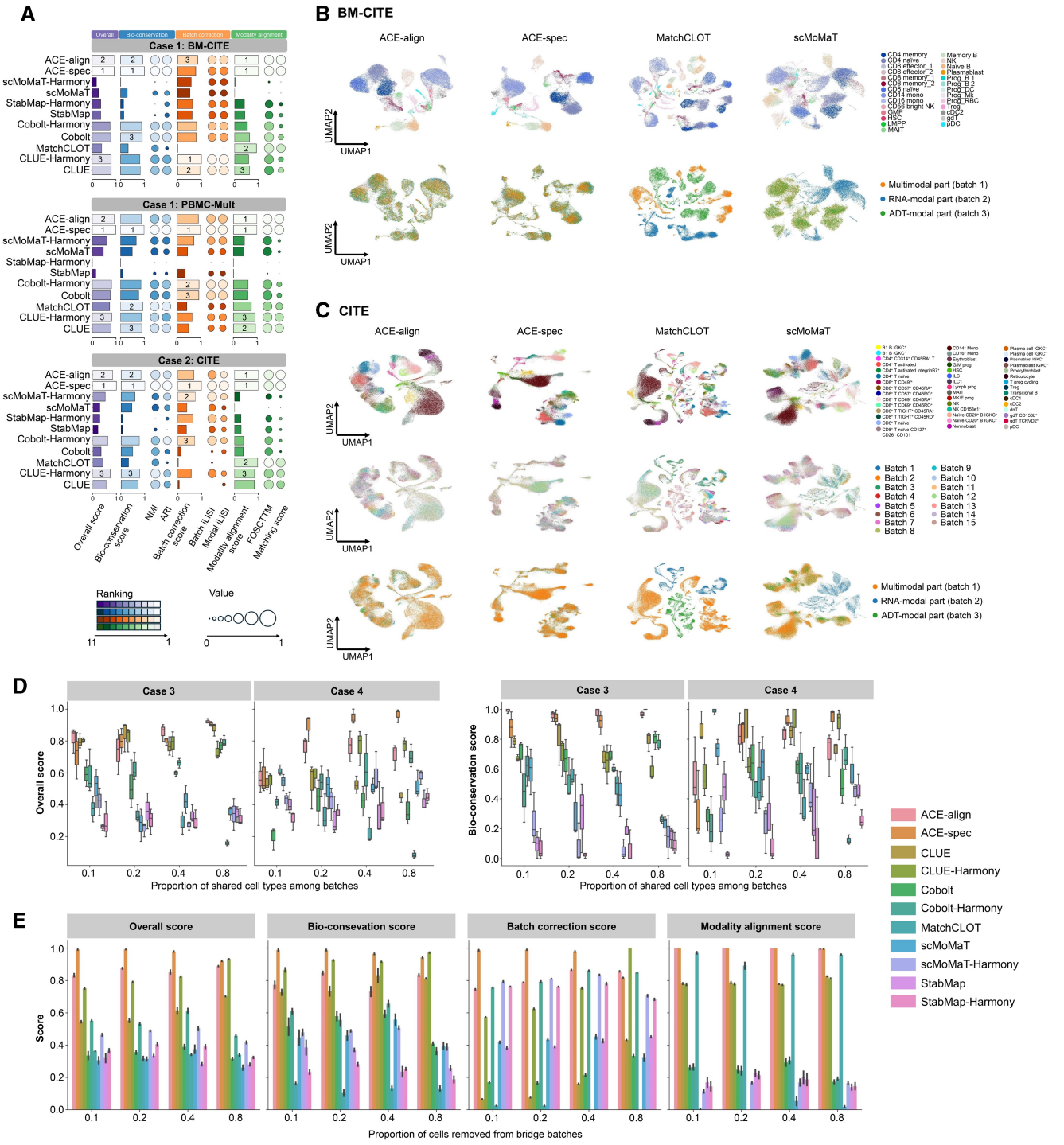


Figure 2 Bi-modal mosaic integration benchmark

A. Overall scores for all compared methods in case 1 and case 2. The top 3 methods for each score are labeled. **B**. UMAP plots for embeddings of ACE-align, ACE-spec, MatchCLOT, and scMoMaT on the BM-CITE dataset. Cells in the first row are colored by cell types and colored by modality labels (and batch labels) in the second row. **C**. UMAP plots for embeddings of ACE-align, ACE-spec, MatchCLOT, and scMoMaT on the CITE dataset. Cells are colored by cell types in the first row, colored by batch labels in the second row, and colored by modality labels in the third row. **D**. Overall scores and bio-conservation scores of all methods in case 3 and case 4. **E**. Benchmarking of all methods' robustness against the number of cells in bridge batches. CITE, PBMC-Mult, and BM-CITE are experimental datasets. UMAP, Uniform Manifold Approximation and Projection; ARI, Adjusted Rand Index; NMI, Normalized Mutual Information; iLISI, inverse Local Inverse Simpson's Index; FOSCTTM, Fraction of Samples Closer Than the True Match.

In tri-modal case 2, ACE-spec achieves the highest overall score on the DOGMA dataset, followed by ACE-align (Figure 3A, Figure S8A). Their bio-conservation scores rank in the top 2 and their batch correction scores rank in the top 3 among all methods. UMAP plots also show that within ACE's outputs, batches are well mixed and cell types are clearly separated (Figure 3B, Figure S9A). Surprisingly, on the CITE-ASAP dataset, StabMap reaches much higher bio-

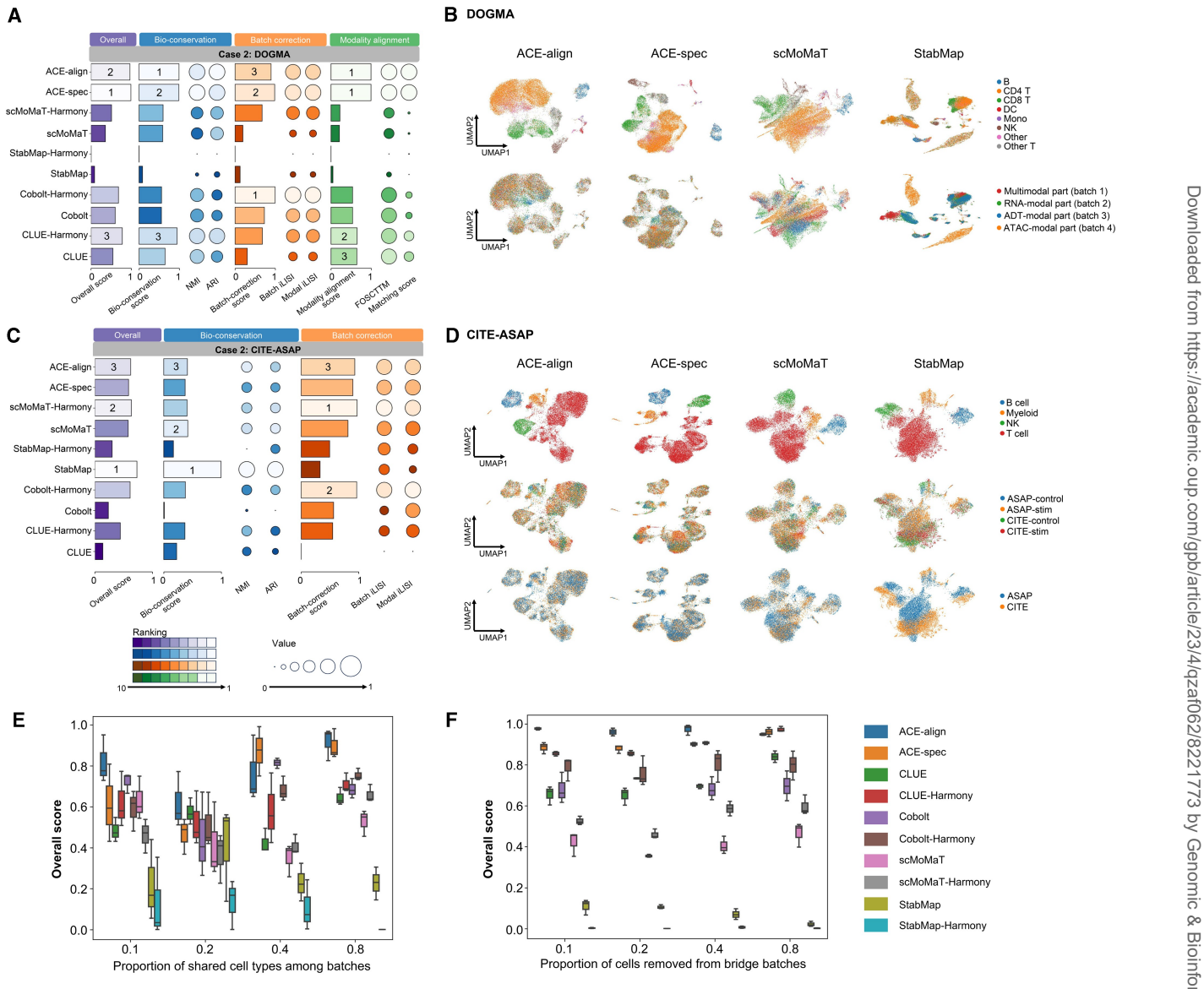


Figure 3 Tri-modal mosaic integration benchmark

A. Overall scores of all compared methods on the DOGMA dataset in case 2. The top 3 methods for each score are labeled. **B**. UMAP plots for embeddings of ACE-align, ACE-spec, scMoMaT, and StabMap on the DOGMA dataset. Cells in the first row are colored by cell types and colored by modality labels (batch labels) in the second row. **C**. Overall scores of all methods on the CITE-ASAP dataset in case 2. **D**. UMAP plots for embeddings of ACE-align, ACE-spec, scMoMaT, and StabMap on the CITE-ASAP dataset. Cells are colored by cell types in the first row, colored by batch labels in the second row, and colored by modality labels in the third row. **E**. Overall scores of all methods in tri-modal case 4. **F**. Benchmarking of all methods' robustness against the number of cells in bridge batches. DOGMA and CITE-ASAP are experimental datasets.

conservation scores than other methods, helping it achieve the highest overall score (Figure 3C, Figure S8B). UMAP plots show that StabMap indeed clearly separates four cell types in the CITE-ASAP dataset whereas ACE-align and ACE-spec separate T cells and myeloid cells into several subpopulations (Figure 3D, Figure S9B). The superior performance of StabMap on the CITE-ASAP dataset can be attributed to the low resolution of annotations, which simplifies the task of preserving cellular heterogeneities.

In tri-modal case 4, ACE-align consistently ranks within the top 2 in terms of overall scores across various proportions of shared types (Figure 3E, Figure S10A and B). With the proportion increasing, ACE-spec's overall scores gradually improve and reach similar or even exceed ACE-align, which is similar to the case 4 in bi-modal integration. The main reason is that low proportions of shared types among batches can easily result in incorrect matching of cells. We also evaluated the robustness of all methods against the number of cells in bridge batches. Specifically, we randomly removed certain proportions of cells from the bridge batches in the DOGMA dataset (proportion = 0.1, 0.2, 0.4, and 0.8, remaining 6900, 6100, 4600, and 1500 cells, respectively) and evaluated all the methods on these new datasets. Despite all methods' performance decreasing with the removed cells increasing (Figure S10C), ACE-align and ACE-spec consistently achieve the top overall scores (Figure 3F, Figure S10D).

ACE-spec helps refine cell type annotations of CITE-ASAP dataset

Note that in the integration task of the CITE-ASAP dataset, StabMap achieves top bio-conservation scores whereas on the rest of datasets, its scores are at a low level. This is

because the annotation resolution of the CITE-ASAP dataset is so low that preservation of biological variations is easily achieved. However, the embeddings of ACE-spec show stronger cellular heterogeneity than the original annotations, which inspires us to use ACE to refine the cell type annotations. We applied Louvain clustering algorithm on the embeddings of four batches using SCANPY [32] (resolution = 1.0) and obtained 17 clusters (**Figure 4A**). For a comparison, we also performed clustering on other methods' embeddings with the same resolution.

As shown in Figure 4A, all methods' embeddings successfully separate the four annotated cell types. StabMap-Harmony produces the largest number of clusters, but many appear to be outliers because there are significant discrepancies between the outputs of StabMap and StabMap-Harmony. When comparing the clustering results of ACE-spec, scMoMaT, scMoMaT-Harmony, and Cobolt-Harmony, we can observe that ACE-spec achieves the most compact clusters and clearest separation boundaries between clusters. The mapping of ACE-spec's clustering labels onto the UMAP of other methods is largely consistent with their own clustering results, but ACE-spec provides a finer granularity in clustering resolution. We also applied weighted nearest neighbors (WNN) analysis [33] on the batches measured with CITE-seq and batches measured with ASAP-seq, respectively. When projecting ACE's cluster labels to the WNN UMAP plots, we observed that the majority of cluster labels are consistent with the grouping in WNN analysis results (Figure S11A and B). The differences are as follows: in the batches measured with CITE-seq, WNN further separates clusters 1 and 2 into two groups, respectively, but mixes clusters 7 and 13 as well as clusters 2 and 12. In the batches measured with ASAP-seq, WNN mixes clusters 1 and 9 as well as clusters 6 and 14.

Next, we tried to annotate these clusters. Taking the two batches measured with CITE-seq as an example, cluster 4 corresponds to the original natural killer (NK) cell and cluster 5 corresponds to the original B cell. The upregulated genes *PRF1*, *GNLY*, and *KLRD1* [33] ($P < 1 \times 10^{-10}$ based on Wilcoxon test) and *MS4A1*, *CD79A*, and *RALGPS2* [33] ($P < 1 \times 10^{-10}$ based on Wilcoxon test) validate the original annotations for clusters 4 and 5, respectively (Figure 4B). The annotation refinement mainly happens within T cell and myeloid cell populations. Clusters 0, 1, 2, 3, 7, 8, 9, 10, 12, 13, and 16 have upregulated expression of T cell markers [33] (*CD3E*, *CD3D*, and *CD3G*, $P < 0.05$), making them T cells (Figure 4B). Clusters 0, 1, 3, 8, and 9 have high expression of *CD4*, indicating they are $CD4^+$ T cells (Figure 4B). Clusters 2, 7, 12, and 16 with high expression of *CD8A* and *CD8B* are $CD8^+$ T cells (Figure 4B). Clusters 10 and 13 belong to other T cells. Notably, cluster 11 consists of T cells in the original annotations, but any of *CD3D*, *CD3E*, and *CD3G* is not highly expressed within them ($P > 0.05$). Instead, cluster 11 has upregulated expression of *CD11C* ($P < 1 \times 10^{-10}$) and *CD1C* [33,34] ($P < 0.05$), and thus we annotated it as conventional dendritic cell (cDC).

Within $CD4^+$ T cells, clusters 0, 3, and 8 have high expression of *TCF7* and *LEF1* [33], making them naïve $CD4^+$ T cells (Figure 4C); cluster 1 has high expression of *IL7R* and *IL32* [33], making it $CD4^+$ memory T cells (Figure 4C); cluster 9 has high expression of *RTKN2*, *FOXP3*, *IL2RA*, and *TIGIT* [33], making it T regulatory cells (Tregs) (Figure 4B). Within $CD8^+$ T cells, clusters 2 and 12 with high expression of *CCR7*, *LINC02446*, and *LEF1* [33] are naïve $CD8^+$ T cells (Figure 4C); clusters 7 and 16 with high expression of *IL7R*, *CCL5*, and *CST7* are $CD8^+$ memory T cells [33] (Figure 4C). Within other T cells, cluster 13 has high expression of *KLRB1* [33] and the surface proteins T-cell receptor (TCR) Vα7.2 and CD161 [33,34] (Figure 4B and C), making it mucosal-associated invariant T (MAIT) cells. Cluster 10 shows high expression of *TRDC* and *TRGC1* [33] as well as the surface protein TCR-γδ [35] (Figure 4B and C), making it gamma delta T (gdT) cells.

Within the original myeloid cells, cluster 6 has high expression of *CD14*, *S100A9*, and *LYZ* [33] (Figure 4B and C), making it CD14 monocytes. Cluster 14 is identified as plasmacytoid dendritic cell (pDC) due to its high expression of *ITM2C*, *SERPINF1*, and *IL3RA* [33] (Figure 4B). Cluster 15 is identified as hematopoietic stem and progenitor cell (HSPC) due to its high expression of *PRSS57*, *EGFL7*, and *GATA2* [33] as well as the surface protein CD34 [33] (Figure 4B and C).

For each cluster, we performed Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis on the top 200 differentially expressed genes (DEGs; based on RNA data from CITE-seq) using clusterProfiler [36]. The results demonstrate a consistency between the enriched pathways and our cell type annotations (Figure S12). For instance, NK cell-mediated cytotoxicity is the most significantly enriched pathway in cluster 4, supporting our annotation of NK cells. In cluster 5, pathways such as B cell receptor signaling and intestinal immune network for IgA production are enriched, further validating the cluster's identification as B cells. Those clusters annotated as T cells display enrichment in pathways relevant to T cell function, including TCR signaling, primary immunodeficiency, and cell adhesion molecules. Additionally, cluster 6, annotated as $CD14^+$ monocytes, shows enrichment in pathways like phagosome and lysosome, which are critical for the phagocytic activity of monocytes [37]. For cluster 15, annotated as HSPCs, we observed enrichment in the hematopoietic cell lineage pathway.

Moreover, analysis on the other two batches measured with ASAP-seq can also validate our annotations. For example, the surface proteins CD1C and CD11C are highly expressed within cluster 11 ($P < 0.01$) (Figure 4D), consistent with the results mentioned above. Cluster 13 has high expression of TCR Vα7.2 and CD161, and cluster 15 has high expression of CD34 (Figure 4D, Figure S11C), confirming the annotations of MAIT cells and HSPCs. We further used chromVar [38] to infer accessibility scores for known motifs. We found that peaks in cluster 11 are highly enriched for motifs for the transcription factor JUNB [39] ($P < 1 \times 10^{-10}$) (Figure 4D) that is essential for cDC identity [39], confirming our annotation for cluster 11. Peaks in our annotated MAIT cells are highly enriched for motifs for the pro-inflammatory transcription factor RORgammat ($P < 1 \times 10^{-10}$) (Figure 4D), consistent with existing studies [33]. Other motifs' activity scores also support our annotations (Figure S11D).

ACE-spec helps discriminate cellular heterogeneity in viral pneumonia datasets

To further demonstrate ACE-spec's ability in enhancing the representation of cellular heterogeneity, we considered a more complex task of mosaic integration for atlas-scale datasets from viral pneumonia patients (referred to as VP

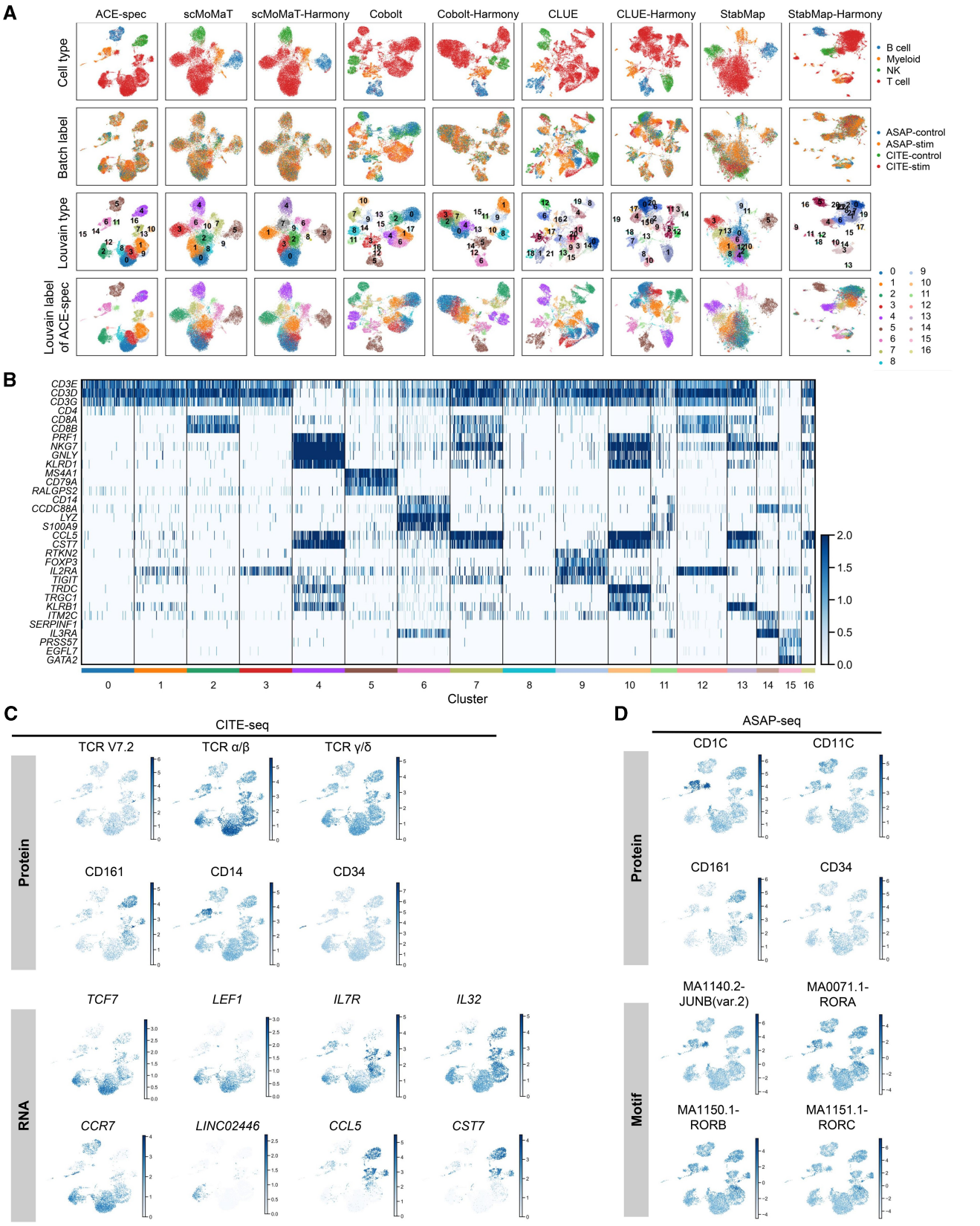


Figure 4 Analysis of ACE-spec's results on the CITE-ASAP dataset

A. UMAP plots of embeddings from different integration methods on this dataset. Cells are colored by original annotations (the first row), batch labels (the second row), Louvain clustering labels (the third row), and clustering labels from ACE-spec (the last row). **B**. Expression heatmap of marker genes of known cell types in all clusters. **C**. Expression heatmap of known marker genes and marker proteins in the batches measured with CITE-seq. **D**. Expression heatmap of known marker genes and marker motifs in the batches measured with ASAP-seq. CITE-seq and ASAP-seq are experimental datasets.

dataset). We collected a public scRNA-seq dataset [40] (referred as VP-RNA dataset) which analyzed the transcriptome of PBMCs from 17 viral pneumonia patients with moderate disease ($n = 5$), acute respiratory distress syndrome (ARDS) (severe, $n = 6$), or recovering from ARDS (recovering, $n = 6$). This dataset contains 69,983 cells and 11 manually annotated cell types. A single-cell resolution mass cytometry (CYTOF) dataset [41] spanning 160 patients and a total of 7.11 million cells was collected, which was generated from granulocyte depleted whole blood of viral pneumonia patients, sepsis patients, and healthy volunteers. We used the CITE-seq dataset (referred as CITE2 dataset) of 161,764 PBMCs from healthy donors [33] as the bridge dataset. Following the study by Hao et al. [42], we removed cells from individuals with sepsis in the CYTOF dataset, resulting in 116 samples and 5.17 million cells. The CYTOF dataset was then downsampled to 1000 cells per sample, and finally it contains 116,000 cells in total. The CITE2 dataset shares 31 surface protein features with the CYTOF dataset and shares 19,668 gene features with the VP-RNA dataset. UMAP visualizations for each dataset and their embeddings from ACE-spec are shown in **Figure 5**A and Figure S13.

ACE-spec enhances the representation of cellular heterogeneity for the VP-RNA dataset (Figure 5B) (clustering is performed on the embeddings using SCANPY with 1.5 resolution). For example, within originally annotated NK cells, ACE-spec separates them into two clusters, 4 and 17. DEG analysis between these two clusters shows that *GZMK* is the most significantly upregulated gene in cluster 17 ($P < 1 \times 10^{-10}$) and *GZMB* is one of the most significantly upregulated genes in cluster 4 ($P < 1 \times 10^{-10}$) (Figure 5C). The expression difference between two clusters may be correlated with their proportion of cells from different patients. Specifically, within cluster 17, cells from moderate patients occupy 37% and cells from severe (severe and recovering) patients occupy 63%, while within cluster 4, they occupy 44% and 56%, respectively. Increasing number of cells from moderate patients within cluster 4 may lead to higher

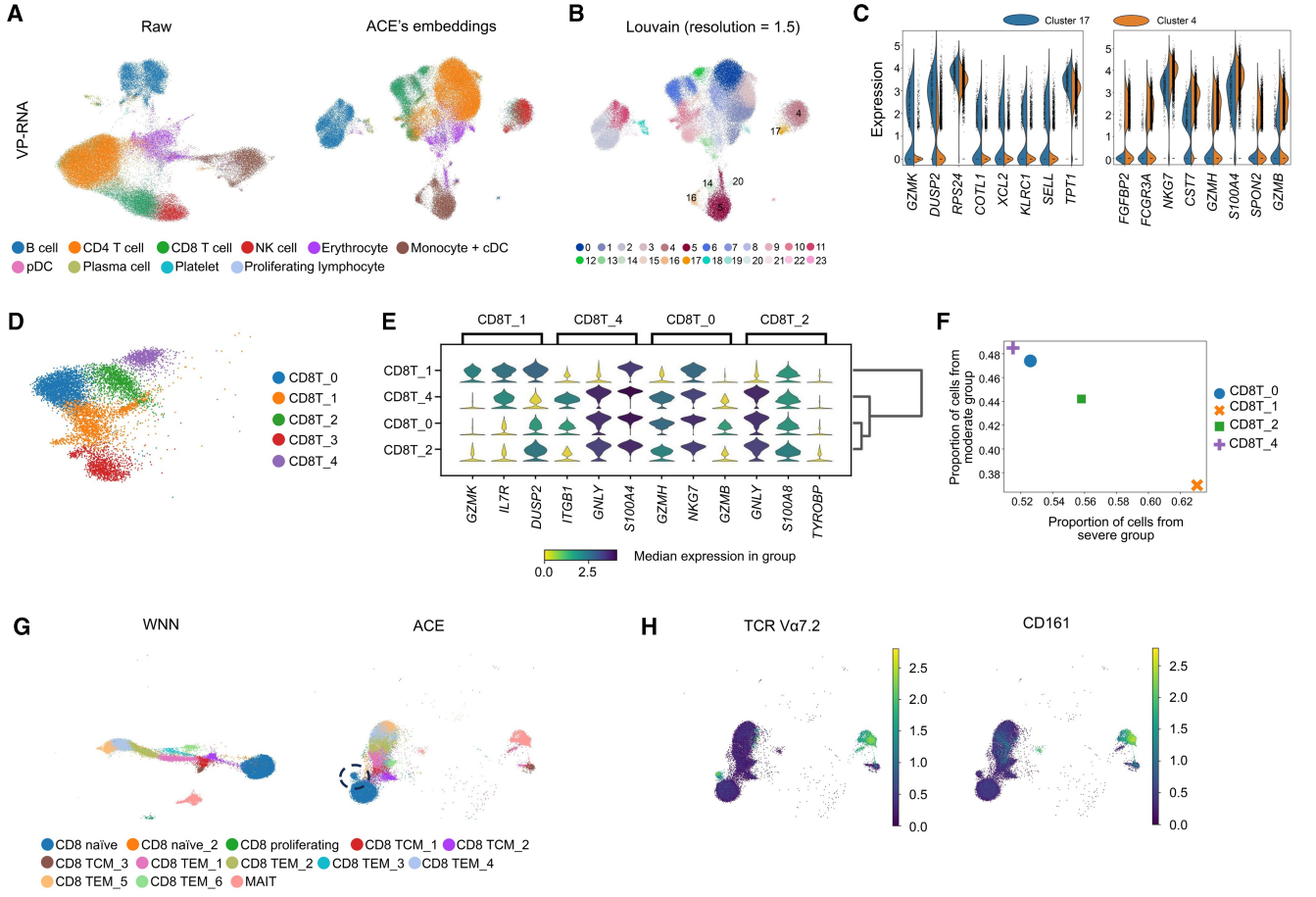


Figure 5 Analysis of ACE-spec's results on the viral pneumonia-related datasets
A. UMAP plots for raw expression profiles of the VP-RNA dataset and its embeddings from ACE-spec. Cells are colored by original annotations. **B**. UMAP plot of ACE's embeddings for the VP-RNA dataset. Cells are colored by clustering labels through performing Louvain on the embeddings. **C**. The most significant DEGs between clusters 17 and 4. **D**. Clustering labels within original $CD8^+$ T populations by performing Louvain on the embeddings of $CD8^+$ T populations. **E**. Stacked violin plot displaying the top DEGs among four clusters. **F**. Scatter plot showing the proportion of cells from different severity groups within four clusters. **G**. UMAP plots of $CD8^+$ T populations and MAIT cells in the bridge dataset using embeddings from WNN analysis and ACE. The black dotted circle highlights a separated cluster from the CD8 naïve cluster. Cells are colored by the level-3 annotations in the bridge dataset. **H**. Heatmap of expression of TCR Vα7.2 and CD161 for cells in (G). VP, viral pneumonia; DEG, differentially expressed gene; MAIT, mucosal-associated invariant T cell; TCR, T-cell receptor.

expression of *GZMB* and lower expression of *GZMK*, which is consistent with the finding that *GZMB* has increased expression in NK cells of mild patients and *GZMK* has predominantly increased expression in NK cells of severe patients [43].

Within the originally annotated monocytes/cDCs, ACE-spec separates them into clusters 5, 14, 16, and 20. It's noteworthy that ACE-spec groups a part of original erythrocytes into cluster 5. We observed that this small part of cells has high expression of *S100A8*, *S100A9*, *LYZ*, and *CD14* [33], indicating CD14 monocytes (Figure S14). Cluster 14 also highly expresses these genes, making it CD14 monocytes. Cluster 16 highly expresses *CDKN1C*, *FCGR3A*, *MS4A7*, and *HES4* [33], making it CD16 monocytes (Figure S14). Cluster 20 has high expression of *FCER1A*, *HLA-DQA1*, *CLEC10A*, and *CD1C* [33], making it cDCs (Figure S14).

Within the original $CD8^+$ T cells, we observed that ACE-spec separates them into several groups, and we performed Louvain algorithm among them using SCANPY (resolution = 0.3), resulting in 5 groups (Figure 5D). Cluster CD8T_3 highly expresses *CCR7*, *LINC02446*, *LEF1*, and *OXNAD1* [33], making it naïve $CD8^+$ T cells (Figure S14). Clusters CD8T_0, CD8T_1, CD8T_2, and CD8T_4 highly express *CD8A*, *CCL5*, *GZMH*, and *KLRD1* [33], making them $CD8^+$ T effector memory (CD8 TEM) cells (Figure S14). Within these four clusters, CD8T_1 has notable expression differences compared to the other three (Figure 5E), which may be correlated with its much higher proportion of cells from severe patients (Figure 5F).

Moreover, for the bridge dataset, it also obtains enhanced representations. For example, within its level-3 annotations, ACE-spec isolates a group of cells ($n = 742$) from the CD8 naïve cluster (Figure 5G). We found that this group of cells has high expression of TCR Vα7.2 but low expression of CD161, marking them TCR $V\alpha7.2^+$ $CD161^-$ T cells (a subtype differing from MAIT cells and TCR $V\alpha7.2^-$ conventional T cells) (Figure 5H) [44]. Together, ACE-spec enhances the representation of cellular heterogeneity, not only helping define fine-grained cell type annotations but also revealing changes in phenotype.

Parameter sensitivity and scalability

To test how ACE's performance is affected by different parameter settings, we evaluated ACE on the CITE dataset as in bi-modal case 2 and on multiple downsampled CITE datasets as in bi-modal case 4. The hyperparameters in ACE mainly include the temperature parameter τ, the latent dimension d, and the number of nearest neighbors k for embedding imputation in ACE-spec. The learning rate and training epochs were fixed across experiments. Figure S15A and B show that ACE-align is generally insensitive to the choice of latent dimension, and it performs better when τ is not greater than 0.1. When τ is greater than 0.1, modality alignment scores of ACE-align will decrease. Also, ACE-spec gets higher modality alignment scores when τ is not greater than 0.1. The latent dimension has a notable impact on ACE-spec's bio-conservation scores and batch correction scores. Specifically, batch iLISI scores get higher when d decreasing whereas bio-conservation scores get higher when d increasing. However, the influence of d on the batch correction scores is comparatively minor compared to its impact on the bio-conservation scores. So, we recommend using higher d (*e.g.*, 256) for ACE-spec. As for the reason why bio-conservation scores favor large latent dimension, we believe that it's because better discrimination of cellular heterogeneity requires more feature dimensions. Varying k has little impact on the NMI and ARI scores of ACE-spec, while smaller values of k lead to better iLISI scores of ACE-spec (Figure S15C). This is likely because the imputed embeddings that are averaged over more neighbors' embeddings will exhibit reduced fidelity to the real ones, introducing batch effects within intra-modality embeddings. We recommend setting k to 2, as this value consistently attains superior batch correction and robust bio-conservation performance for ACE-spec.

We compared the scalability of ACE (align + spec) against other methods on the CITE2 dataset which simultaneously measures RNA and protein expression in 161,764 PBMCs from healthy donors [33]. We randomly selected cells from them with various proportions (1%, 10%, 20%, 40%, 80%, 100%) and then randomly partitioned them into three groups: multimodal, RNA-modal, and protein-modal parts. We evaluated the running time of all methods on a Linux server equipped with Intel^R Core^{TM} i9-10980XE Central Processing Unit (CPU), 128 GB memory, and a GeForce RTX 3090 Graphics Processing Unit (GPU). Figure S15D demonstrates that ACE is the most time-efficient method and completes its run on 160,000 cells within 3 min, suggesting its scalability for larger-scale datasets.

Discussion

In this study, we present ACE, a mosaic integration framework for single-cell multi-omics data analysis. ACE assembles two strategies, ACE-align and ACE-spec, to handle the disparity in modality abundance across datasets. ACE-align applies CL for explicit modality alignment to construct a shared latent space across modalities, thereby bridging the gap between modalities. We propose a novel CL loss, which addresses the modality gap problem and achieves better modality alignment. ACE-spec utilizes the results of ACE-align to impute the missing modality-specific representations, which helps preserve and better represent cellular heterogeneities.

We evaluated ACE-align and ACE-spec under various data integration scenarios using comprehensive metrics, and experimental results showed that both strategies achieved superior performance compared to other state-of-the-art methods. Each strategy has its own suitable applications. Generally, ACE-align fits scenarios with a small proportion of shared cell types across batches, whereas ACE-spec is more appropriate for scenarios with a large proportion of shared cell types across batches. Particularly, ACE-spec demonstrates outstanding performance in capturing cell-to-cell variations, rendering it an advantageous integration method for enhancing cellular representations of existing datasets.

We conducted extensive ablation studies to investigate the robustness of our proposed framework. First, we validated that the observed cellular heterogeneity was robust regardless of the choice of UMAP parameters (Note C in File S1). Second, we validated that our proposed loss function significantly outperformed the InfoNCE loss in terms of batch correction and bio-conservation performance (Note D in File S1). Third, we found that ACE was generally robust to the choice of batch correction methods applied prior to model training (Note E in File S1). Fourth, we observed that ACE achieved higher NMI and ARI scores when using shared

nearest neighbor (SNN)-based clustering algorithms [45,46] compared to *K*-means, while ACE-spec showed lower unsupervised metric scores under the same conditions (Note F in File S1).

In most scenarios, the goal of data integration is to extract biological insights from the datasets, which is why the bio-conservation score is given a larger weight [28,30]. However, in some cases, the objective may shift toward aligning batches or modalities. For these scenarios, a larger weight should be assigned to batch correction or modality alignment scores to better reflect the integration target. Based on our experiment results, ACE-align and ACE-spec demonstrate superior performance in biological conservation and modality alignment. Therefore, assigning larger weights to bio-conservation or modality alignment scores maximizes the overall performance of ACE's framework, whereas placing a higher weight on batch correction reduces ACE's comprehensive performance.

ACE can be used not only for data representation, but also for reconstruction of raw omics features of missing modalities. We performed detailed comparison between our method and other reconstruction methods (Note G in File S1), including scVAEIT [47], totalVI [48], and MultiVI [49]. Experimental results showed that our simple reconstruction strategy even achieved comparable performance to the state-of-the-art method, scVAEIT, and outperformed totalVI and MultiVI with respect to overall scores and robustness. Moreover, UMAP plots confirmed that our reconstructed features showed better separation of cell types compared to scVAEIT.

Overall, ACE is a valuable tool for understanding single-cell multimodal data. We envisage that ACE will serve as a promising tool for the community of single-cell multi-omics data analysis.

Code availability

The source code used in this study is available on GitHub (https://github.com/CSUBioGroup/ACE-main). The code has also been submitted to BioCode at the National Genomics Data Center (NGDC), China National Center for Bioinformation (CNCB) (BioCode: BT007620), which is publicly accessible at https://ngdc.cncb.ac.cn/biocode/tools/BT007620.

Data availability

The processed datasets have been deposited to Zenodo and can be accessed at https://zenodo.org/records/10851161.

CRediT author statement

Xuhua Yan: Methodology, Software, Formal analysis, Writing – original draft. **Jinmiao Chen:** Formal analysis, Visualization, Writing – review & editing. **Ruiqing Zheng:** Methodology, Formal analysis, Writing – review & editing. **Min Li:** Conceptualization, Supervision, Validation, Writing – review & editing. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Supplementary material

Supplementary material is available at *Genomics, Proteomics & Bioinformatics* online (https://doi.org/10.1093/gpbjnl/qzaf062).

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62225209) and the Hunan Provincial Science and Technology Program (Grant Nos. 2019CB1007 and 2021RC4008).

ORCID

0000-0002-3183-3342 (Xuhua Yan)
0000-0001-7547-6423 (Jinmiao Chen)
0000-0001-6372-6798 (Ruiqing Zheng)
0000-0002-0188-1394 (Min Li)

References

[1] Shekhar K, Lapan SW, Whitney IE, Tran NM, Macosko EZ, Kowalczyk M, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell 2016;166:1308–23.e30.
[2] Aibar S, Gonzalez-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods 2017;14:1083–6.
[3] Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods 2017;14:865–8.
[4] Ma S, Zhang B, LaFave LM, Earl AS, Chiang Z, Hu Y, et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. Cell 2020;183:1103–16.e20.
[5] Mimitou EP, Lareau CA, Chen KY, Zorzetto-Fernandes AL, Hao Y, Takeshima Y, et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nat Biotechnol 2021;39:1246–58.
[6] Argelaguet R, Cuomo ASE, Stegle O, Marioni JC. Computational principles and challenges in single-cell data integration. Nat Biotechnol 2021;39:1202–15.
[7] Zhang Z, Sun H, Mariappan R, Chen X, Chen X, Jain MS, et al. scMoMaT jointly performs single cell mosaic integration and multi-modal bio-marker detection. Nat Commun 2023;14:384.
[8] Ghazanfar S, Guibentif C, Marioni JC. Stabilized mosaic single-cell data integration using unshared features. Nat Biotechnol 2024;42:284–92.
[9] Gong B, Zhou Y, Purdom E. Cobolt: integrative analysis of multimodal single-cell sequencing data. Genome Biol 2021;22:351.
[10] Wu M, Goodman N. Multimodal generative models for scalable weakly-supervised learning. Proceedings of the 32nd International Conference on Neural Information Processing Systems 2018:5580–90.
[11] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016:785–94.
[12] Tu X, Cao ZJ, Xia C, Mostafavi S, Gao G. Cross-linked unified embedding for cross-modality representation learning. Proceedings of the 36th International Conference on Neural Information Processing Systems 2022:15942–55.
[13] Kriebel AR, Welch JD. UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. Nat Commun 2022;13:780.
[14] Liang VW, Zhang Y, Kwon Y, Yeung S, Zou JY. Mind the gap: understanding the modality gap in multi-modal contrastive

representation learning. Proceedings of the 36th International Conference on Neural Information Processing Systems 2022:17612–25.
[15] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. Proc 37th Int Conf Mach Learn 2020:1597–607.
[16] Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods 2019;16:1289–96.
[17] Yan X, Zheng R, Wu F, Li M. CLAIRE: contrastive learning-based batch correction framework for better balance between batch mixing and preservation of cellular heterogeneity. Bioinformatics 2023;39:btad099.
[18] Lance C, Luecken MD, Burkhardt DB, Cannoodt R, Rautenstrauch P, Laddach A, et al. Multimodal single cell data integration challenge: results and lessons learned. Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track 2021 Compet Demonstr Track 2022:162–76.
[19] Gossi F, Pati P, Chouvardas P, Martinelli AL, Kruithof-de Julio M, Rapsomaniki MA. Matching single cells across modalities with contrastive learning and optimal transport. Brief Bioinform 2023;24:bbad130.
[20] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. Proc 38th Int Conf Mach Learn 2021:8748–63.
[21] Trosten DJ, Lokse S, Jenssen R, Kampffmeyer M. Reconsidering representation alignment for multi-view clustering. Proc IEEE/CVF Conf CVPR 2021:1255–65.
[22] Clevert DA, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (ELUs). arXiv 2015;1511.07289.
[23] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58.
[24] Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv 2014;1412.6980.
[25] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. Proceedings of the 33th International Conference on Neural Information Processing Systems 2019:8026–37.
[26] Luecken MD, Burkhardt DB, Cannoodt R, Lance C, Agrawal A, Aliee H, et al. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. 35th Conference NeuralIPS Datasets and Benchmarks Track (Round 2) 2021.
[27] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al. Comprehensive integration of single-cell data. Cell 2019;177:1888–902.e21.
[28] Luecken MD, Buttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods 2022;19:41–50.
[29] Singh R, Demetci P, Bonora G, Ramani V, Lee C, Fang H, et al. Unsupervised manifold alignment for single-cell multi-omics data. ACM BCB 2020;2020:1–10.
[30] He Z, Hu S, Chen Y, An S, Zhou J, Liu R, et al. Mosaic integration and knowledge transfer of single-cell multimodal data with MIDAS. Nat Biotechnol 2024;42:1594–605.
[31] McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv 2018;1802.03426.
[32] Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 2018;19:15.
[33] Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell 2021;184:3573–87.e29.
[34] Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, et al. CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res 2019;47:D721–8.
[35] Zhao Y, Niu C, Cui J. Gamma-delta (γδ) T cells: friend or foe in cancer development? J Transl Med 2018;16:3.
[36] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284–7.
[37] Lim JJ, Grinstein S, Roth Z. Diversity and versatility of phagocytosis: roles in innate immunity, tissue remodeling, and homeostasis. Front Cell Infect Microbiol 2017;7:191.
[38] Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat Methods 2017;14:975–8.
[39] Novoszel P, Drobits B, Holcmann M, Fernandes CDS, Tschismarov R, Derdak S, et al. The AP-1 transcription factors c-Jun and JunB are essential for CD8α conventional dendritic cell identity. Cell Death Differ 2021;28:2404–20.
[40] Yao C, Bora SA, Parimon T, Zaman T, Friedman OA, Palatinus JA, et al. Cell-type-specific immune dysregulation in severely ill COVID-19 patients. Cell Rep 2021;34:108590.
[41] COMBAT Consortium. A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. Cell 2022;185:916–38.e58.
[42] Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol 2024;42:293–304.
[43] Malengier-Devlies B, Filtjens J, Ahmadzadeh K, Boeckx B, Vandenhaute J, De Visscher A, et al. Severe COVID-19 patients display hyper-activated NK cells and NK cell-platelet aggregates. Front Immunol 2022;13:861251.
[44] Park D, Kim HG, Kim M, Park T, Ha HH, Lee DH, et al. Differences in the molecular signatures of mucosal-associated invariant T cells and conventional T cells. Sci Rep 2019;9:7094.
[45] Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 2019;9:5233.
[46] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Meth 2008;2008:P10008.
[47] Du JH, Cai Z, Roeder K. Robust probabilistic modeling for single-cell multimodal mosaic integration and imputation via scVAEIT. Proc Natl Acad Sci U S A 2022;119:e2214414119.
[48] Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat Methods 2021;18:272–82.
[49] Ashuach T, Gabitto MI, Koodli RV, Saldi GA, Jordan MI, Yosef N. MultiVI: deep generative model for the integration of multimodal data. Nat Methods 2023;20:1222–31.

Genomics, Proteomics & Bioinformatics, 2025, 23, 1–14
https://doi.org/10.1093/gpbjnl/qzaf062
Method