

# MOJITOO: a fast and universal method for integration of multimodal single-cell data

Mingbo Cheng, Zhijian Li and Ivan G. Costa  \*

Institute for Computational Genomics, Joint Research Center for Computational Biomedicine, RWTH Aachen University Medical School, 52074 Aachen, Germany

\*To whom correspondence should be addressed.

## Abstract

**Motivation:** The advent of multi-modal single-cell sequencing techniques have shed new light on molecular mechanisms by simultaneously inspecting transcriptomes, epigenomes and proteomes of the same cell. However, to date, the existing computational approaches for integration of multimodal single-cell data are either computationally expensive, require the delineation of parameters or can only be applied to particular modalities.

**Results:** Here we present a single-cell multi-modal integration method, named **Multi-mOdal Joint IntegraTion of cOmpOnents (MOJITOO)**. MOJITOO uses canonical correlation analysis for a fast and parameter free detection of a shared representation of cells from multimodal single-cell data. Moreover, estimated canonical components can be used for interpretation, i.e. association of modality-specific molecular features with the latent space. We evaluate MOJITOO using bi- and tri-modal single-cell datasets and show that MOJITOO outperforms existing methods regarding computational requirements, preservation of original latent spaces and clustering.

**Availability and implementation:** The software, code and data for benchmarking are available at <https://github.com/CostaLab/MOJITOO> and <https://doi.org/10.5281/zenodo.6348128>.

**Contact:** [ivan.costa@rwth-aachen.de](mailto:ivan.costa@rwth-aachen.de)

**Supplementary information:** [Supplementary data](#) are available at *Bioinformatics* online.

## 1 Introduction

The technological advances of high-throughput single-cell sequencing enable us to characterize cellular heterogeneity of complex tissues for distinct molecular players of cells such as transcripts, proteins and chromatin (Efremova and Teichmann, 2020). The advent of multimodal technologies allow us to simultaneously measure two or more modalities at the same cells, i.e. RNA and open chromatin (Hu *et al.*, 2020; Cao *et al.*, 2018; Ma *et al.*, 2020); RNA and protein (Stoeckius *et al.*, 2017); and RNA, open chromatin and protein (Mimitou *et al.*, 2021; Swanson *et al.*, 2021). These methods allow us to access how genetic information is associated at distinct molecular levels, i.e. the effect of DNA accessibility changes on gene expression or the expression of genes to proteins. However, data produced by each modality has quite distinct characteristics regarding their numerical values (e.g. low counts for open chromatin and variable count values for RNA and proteins levels), dimensionality (dozens for proteins, tens of thousands for genes, hundreds of thousands for open chromatin) and levels of data sparsity (Argelaguet *et al.*, 2021; Li *et al.*, 2021). These make the integrative analysis of multi-modal data a challenging task.

Here, we are interested in the problem of estimating a shared latent space from parallel multiomic approaches, where two or more modalities are measured in the same cells. A few methods have been proposed for this problem. These follow two main frameworks: metric learning and latent variable learning. Weighted nearest neighbors (WNN) (Hao *et al.*, 2021) and Schema (Singh *et al.*, 2021)

explore, respectively, nearest neighbors and quadratic programming to estimate a single distance matrix representing the integrated multimodal data. Both approaches explore efficient algorithms, but do not explicitly provide models associating molecular features to the 'latent space'. MOFA (Argelaguet *et al.*, 2020), scAI (Jin *et al.*, 2020), totalVI (Gayoso *et al.*, 2021) and LIGER (Kriebel and Welch, 2022) explore distinct methods for matrix factorization and estimation of shared latent spaces between modalities. Moreover, estimated matrices can be used for model interpretation, i.e. decomposed matrices can be used to associate molecular features with the latent space. Overall, these methods have a large number of free parameters including the size of the latent space (or rank of the low dimensional matrices). These methods require the optimization of the size of the latent space, which in turn increases computational costs. Of note, the main application of LIGER is the integration of experiments within a modality (Welch *et al.*, 2019). The data integration evaluated here refers to a latter version of LIGER (Kriebel and Welch, 2022), which can only be used for two modalities and a subset of the molecular features need to be common in both modalities. Also, the implementation of some methods (totalVI (Gayoso *et al.*, 2021) and scAI (Jin *et al.*, 2020)) only allow integration of particular modalities (i.e. scRNA-seq and protein abundance for totalVI; scRNA-seq and scATAC-seq for scAI).

Two recent works have explored canonical correlation analysis (CCA) for problems close to the one addressed here. Symphony (Kang *et al.*, 2021), which has as main objective to build reference single-cell atlas, used CCA to obtain a joint space in a case study

with CITE-seq data. However, this approach does not support more than two modalities and was not evaluated concerning its interpretability or computational requirements. DIABLO (Singh *et al.*, 2019) extends generalized canonical analysis for integration of bulk multimodal datasets with respect to different phenotypes. It requires class labels for the samples, which are used for feature selection. However, such class labels are usually not provided in single-cell multi-modal datasets, as explored here.

## 2 Approach

Here, we propose Multi-mOdal Joint IntegraTion of cOMPonents (MOJITOO), an efficient method that is based on CCA to learn a shared latent space for any single-cell multimodal data protocol. The canonical components can be interpreted as factors and be used to characterize feature relevance by relating features across modalities (Fig. 1). In contrast to matrix factorization methods, MOJITOO does not require the definition of parameters such as the rank of the matrix. Furthermore, it provides an approach to estimate the size of the latent space after a single execution of CCA. MOJITOO is provided as an R package and is compatible with common single-cell pipelines for RNA, proteins (Seurat; Hao *et al.*, 2021) and ATAC modalities (Signac; Stuart *et al.*, 2021). MOJITOO builds upon a simple CCA procedure presented in a case study in Symphony (Kang *et al.*, 2021) and differs from DIABLO (Singh *et al.*, 2019), as it does not require labels for data integration.

We evaluate MOJITOO and competing methods (WNN, MOFA, scAL, LIGER, Schema, DIABLO, Symphony) in two bi-modal datasets with RNA and protein measurements (Buus *et al.*, 2021; Stuart *et al.*, 2021), two bi-modal datasets with RNA and ATAC-seq measurements (Ma *et al.*, 2020) and two tri-modal datasets with RNA, proteins and ATAC-seq measurements (Mimitou *et al.*, 2021; Swanson *et al.*, 2021) in regards to their ability to recover a shared latent space. The latent spaces are then evaluated with measures regarding the accuracy of clustering (adjusted Rand index), distance (silhouette score) and structure preservation, i.e. relation between shared space and original space of individual modalities (Jain *et al.*, 2021). Altogether, results show a superior performance of MOJITOO in both computational requirements and accuracy of estimated latent spaces. Moreover, we illustrate how estimated canonical components can be used to interpret the underlying single-cell data.

## 3 Materials and methods

### 3.1 MOJITOO

MOJITOO takes as input a set of matrices from  $m$  modalities:

$$\mathcal{X} = \{X^{(1)}, \dots, X^{(m)}\} \quad (1)$$

where  $X^{(i)} \in \mathbb{R}^{n \times s^{(i)}}$  represents the data of a particular single-cell modality,  $n$  represents the number of cells and  $s^{(i)}$  represents the number of features in modality  $i$ . Here, we focus on multimodal data, where the cells are the same across matrices and there is no direct relation between the features of the distinct modalities.

#### 3.1.1 Reducing the dimension for each modality

We first obtain a dimension reduced matrix for each modality independently using a modality-specific approach:

$$Y^{(i)} = f^{(i)}(X^{(i)}) \quad (2)$$

where  $Y^{(i)} \in \mathbb{R}^{n \times p^{(i)}}$  represents the low-dimensional matrix for modality  $i$ ,  $p^{(i)}$  represents the number of dimensions and  $f^{(i)}$  represents the specific dimension reduction method for this modality. MOJITOO uses latent semantic indexing (LSI) for scATAC-seq and principal component analysis (PCA) for other modalities, as is usual in the literature (Granja *et al.*, 2021; Hao *et al.*, 2021; Stuart *et al.*,

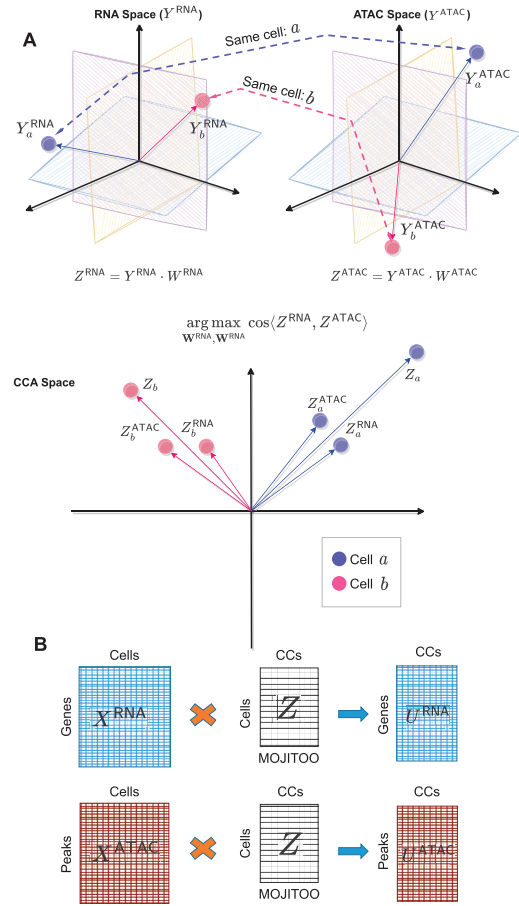


Fig. 1. Schematic MOJITOO. (A) MOJITOO receives as input two (or more) dimensional reduced matrices, where each matrix represents a particular molecular modality describing the same cells. In this example, we assume RNA and open chromatin (as measured by ATAC-seq) modalities are given. The main idea of MOJITOO is to use Canonical Correlation Analysis to find a set of canonical vectors  $W^{ATAC}$  and  $W^{RNA}$ . Exploring a geometrical interpretation of CCA, MOJITOO finds canonical vectors such that the cosine similarity between latent dimensions in  $Z^{RNA}$  and  $Z^{ATAC}$  is maximized. A final representation  $Z$  can be obtained by adding the modality-specific latent spaces. In the example, we show vectorial representations of two cells ( $a$  and  $b$ ) in both original and latent spaces. (B) An association between original features for each modality ( $U^{RNA}$  and  $U^{ATAC}$ ) can be obtained by multiplying original data representation per modality ( $X^{RNA}$  and  $X^{ATAC}$ ) with the shared latent space  $Z$ .

2021). The rationale behind the use of dimension reduction is twofold. First, low-dimensional matrices reduce the computing time of the CCA analysis without impacting accuracy for a minimum number of dimensions. Moreover, it allows to work directly on batch-corrected data, which is usually represented in a low-dimensional space (Hao *et al.*, 2021; Korsunsky *et al.*, 2019). Of note, batch correction is recommended previous to MOJITOO, whenever the data are affected by batch effects.

#### 3.1.2 Learning a shared space with canonical correlation analysis with two modalities

MOJITOO aims to learn a shared latent space  $Z$  from the set of low dimensional matrices  $\mathcal{Y} = \{Y^{(1)}, \dots, Y^{(m)}\}$

$$Z = \text{MOJITOO}(Y^{(1)}, \dots, Y^{(m)}), \quad (3)$$

where  $Z \in \mathbb{R}^{n \times k}$  represents the cells,  $n$  is the number of cells and  $k$  is the dimension of this latent space. When  $\mathcal{Y}$  has two modalities,

we first use CCA (This notation is based on a geometrical interpretation of CCA.) to project the matrices  $Y^{(1)}$  and  $Y^{(2)}$  to vectors  $\mathbf{z}_1^{(1)}$  and  $\mathbf{z}_1^{(2)}$ :

$$\begin{aligned} \mathbf{z}_1^{(1)} &= Y^{(1)} \mathbf{w}_1^{(1)}, \\ \mathbf{z}_1^{(2)} &= Y^{(2)} \mathbf{w}_1^{(2)}, \end{aligned} \quad (4)$$

where  $\mathbf{z}_1^{(1)}$  and  $\mathbf{z}_1^{(2)}$  represent canonical components (CC). The vectors  $\mathbf{w}_1^{(1)}$  and  $\mathbf{w}_1^{(2)}$  can be obtained by solving the following optimization problem:

$$\mathbf{w}_1^{(1)}, \mathbf{w}_1^{(2)} = \operatorname{argmax} \cos(\mathbf{z}_1^{(1)}, \mathbf{z}_1^{(2)}), \quad (5)$$

where  $\mathbf{w}_1^{(1)} \in \mathbb{R}^{p^{(1)}}$ ,  $\mathbf{w}_1^{(2)} \in \mathbb{R}^{p^{(2)}}$  represent the first canonical weight vectors, and  $\cos(\cdot)$  is the cosine similarity between two vectors  $a$  and  $b$  defined by:

$$\cos(a, b) = \frac{a \cdot b}{|a| \cdot |b|}. \quad (6)$$

This is repeated  $\hat{k} = \min(p^{(1)}, p^{(2)})$  times, such that new canonical vectors are orthogonal to previously estimated vectors. These provide the matrices:

$$\begin{aligned} W^{(1)} &= [\mathbf{w}_1^{(1)}, \dots, \mathbf{w}_k^{(1)}], \\ W^{(2)} &= [\mathbf{w}_1^{(2)}, \dots, \mathbf{w}_k^{(2)}]. \end{aligned} \quad (7)$$

These can be used to estimate the modality transformed space as

$$\begin{aligned} Z^{(1)} &= Y^{(1)} \cdot W^{(1)}, \\ Z^{(2)} &= Y^{(2)} \cdot W^{(2)}. \end{aligned} \quad (8)$$

A unique latent space is obtained as

$$Z = Z^{(1)} + Z^{(2)}, \quad (9)$$

where  $Z \in \mathbb{R}^{n \times k}$  and  $k$  is the number of canonical variables retained.

To further remove the noise from the latent space  $Z$ , we only keep highly correlated canonical components  $z_i^{(1)}$  and  $z_i^{(2)}$  by measuring the Pearson correlation and using a Student's  $t$ -test for significance. The  $P$ -values are then corrected using Benjamini Hochberg (BH) (Benjamini and Hochberg, 1995) and only canonical components with adjusted  $P$ -values  $< 0.05$  are retained.

MOJITOO uses an algorithm based on generalized eigenvector decomposition (Ramsay and Silverman, 1997) to estimate the canonical components. MOJITOO has a time complexity of  $\mathcal{O}(\max\{p^{(1)}, p^{(2)}\}^2 \times n)$  for computing covariance matrices and  $\mathcal{O}(\min\{p^{(1)}, p^{(2)}\} \times p^{(1)} \times p^{(2)})$  for the eigenvector decomposition. As  $n$  (number of cells) is usually 100 times larger than  $p^{(i)}$  (number of reduced dimensions in  $Y^{(i)}$ ) the first term dominates the complexity.

Of note, CCA is one of the several steps in the integration algorithm of an earlier version of Seurat (Butler et al., 2018). This had the objective to integrate distinct scRNA-seq experiments and CCA was performed in the common gene space, i.e. on transposed  $Y^{(i)}$  matrices and the objective was to find matching cells.

### 3.1.3 Learning a shared space for multiple modalities

For the case that  $\mathcal{Y}$  has more than two modalities, we perform the pairwise integration of modalities starting with the pair with highest dimensionality. The result of this CCA is then used for integration with the next modality. See Algorithm 1 for a brief description, which receives a set of matrices  $\{Y^{(1)}, \dots, Y^{(m)}\}$  with increasing dimensions  $p^{(i)} \geq p^{(i+1)}$  as input. This heuristic algorithm was adopted to avoid the high computational costs of

multiple CCA, which grows exponentially with the number of modalities.

---

#### Algorithm 1 Multimodal MOJITOO Algorithm

---

**procedure** MOJITOO( $Y^{(1)}, \dots, Y^{(m)}$ )

$i \leftarrow 2$

$Z^{(1)} \leftarrow Y^{(1)}$

**while**  $i \leq m$  **do**

$W^{(1)}, W^{(2)} \leftarrow \text{CCA}(Z^{(1)}, Y^{(i)})$

$Z^{(1)} \leftarrow Z^{(1)} \times W^{(1)}$

$Z^{(2)} \leftarrow Y^{(i)} \times W^{(2)}$

$Z \leftarrow Z^{(1)} + Z^{(2)}$

$Z \leftarrow Z[1:k] \triangleright$  only consider significantly correlated dimension

$Z^{(1)} \leftarrow Z$

$i \leftarrow i + 1$

**end while**

**return**  $Z$

**end procedure**

**end**

---

### 3.1.4 Association of molecular features with latent space

We can use the estimated latent spaces to associate molecular features to the latent space  $Z$ . For example, let  $X^{\text{RNA}} \in \mathbb{R}^{n \times s}$  be the gene expression matrix and  $X^{\text{ATAC}} \in \mathbb{R}^{n \times t}$  be the peak matrix, where  $n$  is the number of cells,  $s$  is the number of genes and  $t$  is the number of peaks. We can obtain a feature associating molecular features to the latent space by

$$\begin{aligned} U^{\text{RNA}} &= (X^{\text{RNA}})^T \cdot Z \\ U^{\text{ATAC}} &= (X^{\text{ATAC}})^T \cdot Z \end{aligned} \quad (10)$$

where  $U^{\text{RNA}} \in \mathbb{R}^{s \times k}$  and  $U^{\text{ATAC}} \in \mathbb{R}^{t \times k}$ . The  $i$ th column of matrix  $U^{\text{RNA}}$  represent the scores of features in the  $i$ th canonical component.

## 3.2 Datasets

We make use of public multimodal datasets with two or tri-modalities in our evaluation. The first dataset is single-cell CITE-seq data which measures single-cell RNA and surface proteins simultaneously. The human bone marrow mononuclear cells (BM-CITE) dataset contains full transcriptomes and 25 surface proteins for over 30 672 cells annotated in 27 cell types (Stuart et al., 2021). This data were obtained with the 'LoadData("bmcite")' command from package SeuratData. Next, we applied the pre-processing pipeline. Another CITE-seq data used were the human peripheral blood mononuclear cells from lung (LUNG-CITE) (Buus et al., 2021) with 52 surface proteins. It contains 10 470 cells annotated in 22 cell types. This data were obtained from [here](#).

The next dataset contains human peripheral blood mononuclear cells (PBMC-multiome) generated by the  $10 \times$  multiome technology to measure gene expression (scRNA-seq) and chromatin accessibility (scATAC-seq) on the same cells. This data contains 11 787 cells with 13 cell types annotated by  $10 \times$  Genomics. We use the scRNA-seq and scATAC-seq count matrices as provided by  $10 \times$  genomics

after processing with the cellranger pipeline obtained from the [here](#). We also use a dataset based on the SHARE-seq protocol measuring gene expression and chromatin accessibility of mouse skin cells (SKIN-SHARE) ([Ma et al., 2020](#)). This data contains 34 774 cells, which are annotated as 23 cell types. We obtain the skin scRNA-seq and scATAC-seq counts and fragments files from the Gene Expression Omnibus under accession number ([GSE140203](#)).

A tri-modal dataset of human PBMCs is measured with the DOGMA-seq protocol ([Mimitou et al., 2021](#)). This provides RNA, ATAC and epitope sequencing of the same cells (PBMC-DOGMA). We use data under low-loss lysis condition, which contains 13 763 cells in 27 cell types. We download count matrices as provided by the authors [here](#). A second tri-modal dataset is based on human PBMCs measured with the TEA-seq protocol ([Swanson et al., 2021](#)). It contains transcripts, epitopes and chromatin accessibility of 25 517 PBMCs grouped into 12 cell types (PBMC-TEA). For this dataset, we obtain original matrices and combine data from distinct wells from GEO ([GSE158013](#)). For scATAC-seq, we obtain an integrated matrix by combing peaks (allowing an extension of 250 bps). We finally intersect all barcodes from scRNA-seq, protein and scATAC-seq to obtain matrices in the same cell space. Characteristics of each of the six datasets are described in [Table 1](#).

### 3.2.1 Processing of single-cell sequencing data

We perform a uniform pre-processing of all previously datasets starting from their count matrices. For scRNA-seq matrices, we adopt the standard Seurat 4 [pipeline](#). First, we log normalize the data by calling the function `NormalizeData` with default parameters. Next we use `FindVariableFeatures` to find top 3000 variable features and run `ScaleData`. Finally, we use `RunPCA` to perform dimension reduction ([Hao et al., 2021](#)) by keeping the first 50 PCs. For scATAC-seq, we adopt the standard pipeline from [Signac](#) ([Stuart et al., 2021](#)). We first run TF-IDF (term frequency—inverse document frequency) on the peaks. Next, we use `RunSVD` on the top features calculated by function `FindTopFeatures` with parameter `min.cutoff = 'q0'`, which provides an LSI dimension reduced matrix. We keep the first 50 dimensions, but we discard the first dimension as this is highly correlated to the number of fragments. For protein/epitopes, we adopt the standard Seurat 4 [pipeline](#) ([Hao et al., 2021](#)). In short, we call `NormalizeData` with parameters `normalization.method = 'CLR'` and `margin = 2` followed by `ScaleData` and `RunPCA` with 30 PCs using default parameters.

PBMC-DOGMA is the only dataset evaluated here with two biological conditions (stimulated and unstimulated). To address the presence of batch, we apply the harmony integration ([Korsunsky et al., 2019](#)) for RNA-seq and epitope data independently to integrate control and stimulated samples. For scATAC-seq, integration is performed by ignoring the first LSI dimension, which has a high correlation with the stimulation. MOJITOO results with and without batch correction are shown in [Supplementary Figure S1](#). We provide these input matrices to MOJITOO, WNN, MOFA, DIABLO and Symphony. Other competing methods provide their own functionalities for normalization and dimension reduction, which are used accordingly (see below). Time and memory requirements of pre-processing data are considered for the benchmarking of the respective method.

## 3.3 Benchmarking of integration methods

We use three distinct metrics to measure the accuracy of the methods. The structure score measures the similarity between two latent space structures ([Jain et al., 2021](#)). It is based on the Pearson correlation of the pairwise Euclidean distance estimated on the shared ( $Z$ ) and latent spaces ( $Y^{(i)}$ ) for each individual modality. This score indicates how well the shared space is related to the modality and the average values indicate how well integration worked. This metric is also employed by Schema ([Singh et al., 2021](#)). We also evaluate the metrics concerning their distance representation using the silhouette score ([Rousseeuw, 1987](#)). For this, we use the labels as provided by the cluster of the respective dataset. We evaluate the use of Euclidean distance as ‘distance’ for the silhouette score. Finally, we evaluate the performance of methods regarding clustering. We perform Louvain clustering with varying resolution (parameter from 0.1 to 2.0) and estimate the adjusted Rand index (ARI) using cell labels ([Hubert and Arabie, 1985](#)).

## 3.4 Execution of competing methods

### 3.4.1 MOFA

MOFA+ ([Argelaguet et al., 2020](#)) uses Bayesian group factor analysis and variational inference to decompose individual modalities simultaneously by estimating a common latent factor matrix  $Z$ , as well as the weights for the transformation of the modalities to the latent space. MOFA+ includes a procedure to determine the optimal number of factors (dimension of the latent space) and has several hyper parameters for model regularization, detection of number of factors and learning rates. We execute MOFA with default parameters and followed their recommendations [tutorial](#) for the analysis of all data. However, we provide now PCA/LSI reductions as input for MOFA, as this improved its computational time ([Supplementary Table S1](#)), as well the clustering performance on MOFA’s latent space ([Supplementary Fig. S2](#)).

### 3.4.2 Schema

Schema ([Singh et al., 2021](#)) explores metrics learning to re-weight modality features through maximizing the agreement with other modalities. Specifically, it utilizes quadratic programming (QP) to learn a scaling transformation  $u$  for the primary matrix  $X$  such that pairwise distances of the transformation  $u * x_i$  (where  $*$  is coordinate-wise multiplication, for each  $x_i \in X$ ) are highly correlated in other modalities. Schema has two main parameters: minimum desired correlation and number of random pairs. We run Schema using default parameters as in [schema tutorial](#).

### 3.4.3 Seurat4 WNN

Weighted nearest neighbor (WNN) ([Hao et al., 2021](#)) constructs single unified representation across multiple modalities. It first creates  $k$ -nearest neighbor (KNN) graphs for each modality based on the latent representation of each feature matrix. Next, it calculates affinities using the exponential kernel between a cell and the average NN for each modality. The latter is used to weigh cells. WNN has two major free parameters: the number of neighbors and scaling factor of the neighborhood kernel. We execute WNN, which is part of Seurat4, using default parameters. WNN does not provide a shared

**Table 1.** Major characteristics of multiomics datasets

Dataset	Protocol	Species	Organ	Modalities	No. of cells	No. of cell types	No. of features (gene/peak/protein)
BM-CITE	CITE-seq	Human	Bone Marrow	RNA/protein	30 672	27	17 009/—/25
LUNG-CITE	CITE-seq	Human	PBMC&Lung	RNA/protein	10 470	22	33 514/—/52
PBMC-Multiome	Multiome	Human	PBMC	RNA/ATAC	11 787	13	36 610/108 377/—
Skin-SHARE	SHARE-seq	Mouse	Skin	RNA/ATAC	34 774	23	23 296/344 592/—
PBMC-TEA	TEA-seq	Human	PBMC	RNA/ATAC/epitope	25 517	12	36 601/128 853/47
PBMC-DOGMA	DOGMA-seq	Human	PBMC	RNA/ATAC/protein	13 763	27	36 495/68 963/210

latent space, but we can use the weighted nearest neighbors graph to build a distance matrix that can be used in all benchmarking evaluations.

#### 3.4.4 scAI

scAI simultaneously decomposes transcriptomic and epigenomic data into multiple biologically relevant factors (Clark et al., 2018). Its framework is similar to MOFA, but it can only cope with two modalities at a time. scAI uses a stability method to define the rank (size of the latent space) and has three main free parameters used for model regularization. We execute scAI in only bi-modal with RNA and ATAC-seq datasets with default parameters.

#### 3.4.5 LIGER

LIGER (Welch et al., 2019), which is based on non-negative matrix factorization, was originally proposed for data integration whenever modalities are in the same feature space. A newer variant of LIGER (Kriebel and Welch, 2022) is able to perform integration, whenever there is some overlap between the features across modalities (shared features), i.e. protein and RNA expression of the same gene or gene accessibility scores for ATAC-seq. LIGER estimates a gene accessibility (ATAC-seq) matrix by counting the total number of ATAC-seq reads within the gene body and promoter regions (3 kb upstream) for each gene per cell. An additional unshared feature matrix is further produced by binning the genome into bins of 100 000 bps and counting the overlap of these bins with peaks from the respective dataset. LIGER has two major parameters: a regularization term and the number of factors (dimensions of the latent space). Regions associated to ENCODE Blacklist regions (Amemiya et al., 2019) are removed. Moreover, LIGER can be only executed for bi-modal datasets.

#### 3.4.6 Symphony integration

Symphony (Kang et al., 2021) is a method to create single-cell reference atlas for subsequent annotation of new single-cell datasets. For a single case study with a multi-ome CITE-seq data, Symphony used canonical correlation analysis to find shared reference. This simple procedure differs from MOJITOO in several ways: it does not use dimension reduction as input; it is not able to cope with more than two modalities and it uses the latent space of only one modality (RNA) as shared space. Moreover, Symphony included the execution of batch correction with Harmony after the execution of CCA. The CCA analysis is not part of Symphony source code, but we implemented it based on a script provided by authors ([https://github.com/immunogenomics/TB\\_Tcell\\_CITEseq/blob/main/R/cca\\_analysis.R](https://github.com/immunogenomics/TB_Tcell_CITEseq/blob/main/R/cca_analysis.R)). Also, due to high computational requirements, we were only able to run this on the dimension reduced space (PCA/LSI) described above. We call this method Symp-Int to reflect the fact that this is not Symphony, but an integration method used in one of the analysis of Symphony manuscript.

#### 3.4.7 DIABLO

DIABLO (Singh et al., 2019) is based on a generalized canonical analysis to integrate multiple datasets in a supervised way. It was originally designed for bulk multi-omics data, where sample labels are available and explored for feature selection. Since labels are not available for single-cell multi-modal data evaluated here, we provide a distinct label for each cell. DIABLO is executed as indicated in their tutorial. Due to a prohibitive computational costs if raw matrices are used, we provide dimension reduced matrices as input. Another major parameter is the number of CC components, which is usually set to the number of classes—1. Here, we used 30 components. Also, DIABLO does not provide a common latent space. We therefore combined the CCs to obtain a final latent space.

## 4 Results

### 4.1 Benchmarking of multimodal integration methods

We evaluate MOJITOO and competing methods using six publicly available multimodal datasets with two or three modalities. These datasets have between 10 000 and 35 000 cells, 12 and 27 cell types and 25 to 344 492 features per modality (Table 1). We compare MOJITOO with MOFA (Argelaguet et al., 2020), WNN (Hao et al., 2021), Schema (Singh et al., 2021), scAI (Jin et al., 2020), LIGER (Kriebel and Welch, 2022), DIABLO (Singh et al., 2019) and Symp-Int (Kang et al., 2021). Of note, some methods (scAI, LIGER and Symp-Int) failed to be executed in some conditions, due to their inability to cope with more than 2 modalities or the lack of raw sequences for some of the evaluated datasets.

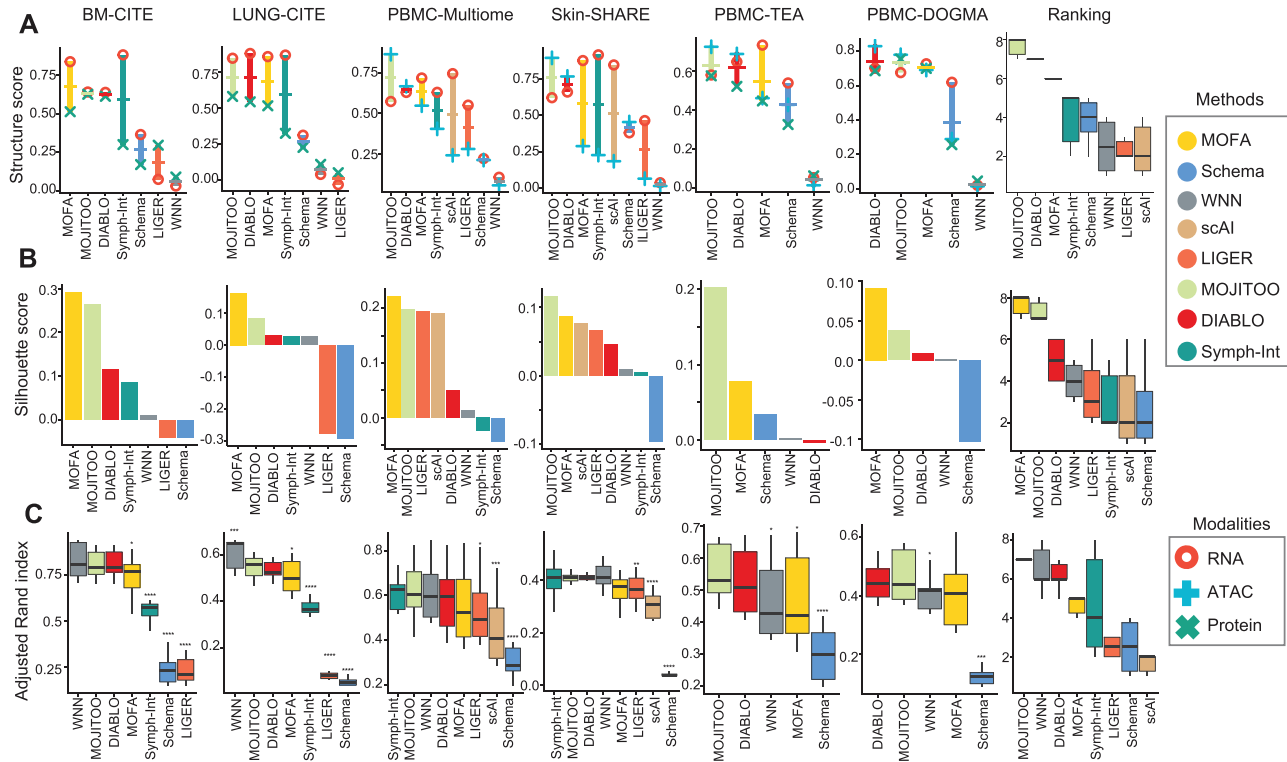
First, we evaluate algorithms regarding their structure preservation, i.e. the average similarity between the euclidean distances in the shared space and distances in the space of each modality (Jain et al., 2021). Results indicate highest structure scores for MOJITOO (4 out of 6) followed by DIABLO (1 out of 6) and MOFA (1 out of 6). A ranking of the structure scores indicates MOJITOO as the best algorithm followed by DIABLO, MOFA and Symp-Int (Fig. 2A). Interestingly, we observe that MOFA and Symp-Int tend to obtain higher structure scores for RNA and that MOJITOO and DIABLO have lower variance of structure scores across modalities. This suggests that the MOJITOO and DIABLO shared space captures information of all individual modalities more uniformly than competing methods. The lower performance of Symphony integration compared to other CCA methods (MOJITOO and DIABLO) is explained by the fact it only considers CCs from the RNA space as integrated embedding.

Next, we make use of the cell types reported in the original manuscripts introducing the single-cell datasets as true labels for benchmarking. First, we use these labels to evaluate the silhouette scores by contrasting class labels with Euclidean distance matrices estimates on the shared space. Regarding silhouette, MOFA is best in four out of six dataset, while MOJITOO is best in the other two datasets. MOJITOO obtains second rank in four out of six datasets and is ranked second in the overall ranking (Fig. 2B). Finally, we perform Louvain clustering at distinct resolutions (0.1–2.0) on the shared latent space. We then measure the agreement of clustering results with labels using the Adjusted Rand Index (ARI). WNN and Symphony have the highest ARI in two datasets each, while DIABLO and MOJITOO have highest values in one dataset each (Fig. 2C). MOJITOO has a higher rank than two in all dataset and has the highest overall rank followed by WNN and DIABLO. Moreover, we also perform a sensitivity analysis on MOJITOO to inspect if the dimension of the original PCA/LSI space has impact on its performance (Supplementary Fig. S3). We observe that if 50 or more components are used, MOJITOO obtain similar clustering and structure preservation scores. Examples of low dimensional embeddings obtained by distinct integration methods with the PBMC-Multiome dataset are provided in Supplementary Figure S4.

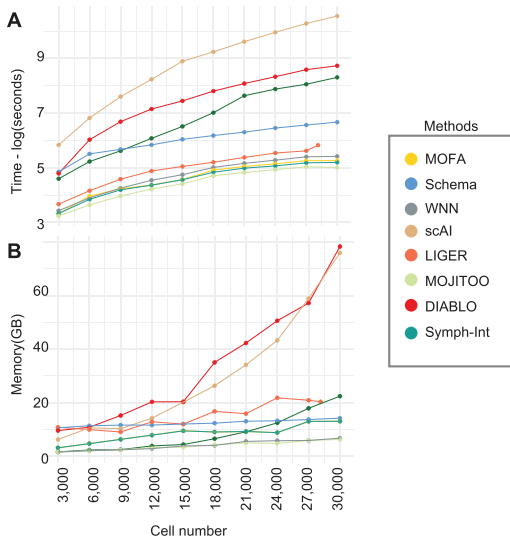
A crucial aspect of single-cell analysis is the computational resources needed for computation on an increasing number of cells. For this, we inspect the time and memory used in the largest datasets in our benchmark (SKIN-SHARE). To obtain curves, we down-sample the number of cells from 30 000 to 3000 (Fig. 3A and B and Supplementary Table S1–S2). We observe that MOJITOO has the overall lowest computational requirement (2.4 min and 6.3 GBs) followed closely by MOFA (3.21 min and 13.09 GBs) and WNN (3.74 min and 6.8 GBs). DIABLO, on the other hand, required up to 103 min and 78.3 GBs for 30 000 cells, while scAI required 637 min and 75 GB of memory. Altogether, results indicate MOJITOO has the best recovery of data structure and clustering results, while being the fastest and having the lowest memory footprint among all competing methods.

### 4.2 Canonical vectors support the interpretation of multiome data

Additionally, we explore the use of the dimensions of the latent space ( $Z$ ) as factors for interpreting the PBMC multiome data. We



**Fig. 2.** Benchmarking on data integration methods. (A) We show the average (trace) and modality-specific structure scores (dots) (y-axis) versus methods (x-axis) for the six datasets. The last graph shows the combined ranking of the method over all datasets, where the highest rank indicates the best performer. (B) Barplots showing silhouette score (y-axis) versus methods (x-axis) for six benchmark datasets. The last plot shows the combined ranking for competing methods. (C) Boxplots showing ARI scores (y-axis) versus methods (x-axis) for all six data-sets. Asterisks indicate  $P$ -values of  $<0.05$  (\*),  $<0.01$  (\*\*),  $<0.001$  (\*\*\*) obtained via  $t$ -test comparing the ARI values of MOJITOO versus other methods. The last boxplot shows the combined ranking for competing methods



**Fig. 3.** Time and memory consumption on the Skin-SHARE. (A) Line plots showing elapsed time (log of seconds) for each method (y-axis). (B) Line plots showing peak memory (Gigabytes) required by each method (y-axis). In both A–B, the x-axis shows the number of cells used (randomly sampled) from the Skin-SHARE data

denote the latent features as canonical components (CC). As shown in Figure 4, positive or negative values for the top CCs discern well all major cell types (Fig. 4). High values of CC1 are associated to myeloid cells (CD14+ and CD16+ monocytes and dendritic cells), while negative values are associated to T and NK cells (Fig. 4A). CC2 values discern B cell and plasmacytoid dendritic cells (pDC) from other cells, while CC3 differentiates B cells from pDCs (Fig. 4B and C). Further CCs capture subtle changes between major cell sub-

types (Fig. 4D and E). CC4 and CC5 capture changes between naive T cells and active T CD8 and active T CD4 cells respectively, while CC5 captures differences between naive monocytes (CD14+) and activated monocytes (CD16+). Other smaller cell types (dendritic cells, platelets, double negative T cells and pre-B and progenitor B cells) can be characterized with further CCs (Supplementary Fig. S5). We also evaluated CCs with low correlation, which were considered as noise by MOJITOO (Supplementary Fig. S6). We observe that these CC have high scores for a few cells and low concordance across modalities. This supports their low biological relevance.

Next, we explore the  $U$  matrices, which provide values associating molecular features with the latent dimensions (CCs). Indeed, the expression of genes with high CC1 values include monocyte genes as LYN and FCN1, while negative CC1 values are associated to T cell genes BCL11B and IL7R (Fig. 5A). Similarly, we observe that top ranked peaks with high or low CC1 scores have monocyte or T-cell-specific open chromatin. These include regions close to the T cell gene BCL11B (Fig. 5B). High CC2 value are associated with B cell genes IGHM and BCL11A, while low CC1 genes (BCBL11B and IL32) are associated with T cells (Fig. 5C). As before, we observe cell-specific open chromatin patterns on top ranked ATAC-seq peaks associated with high and low CC2 values. Altogether, these results indicates that MOJITOO CCs can be used to capture major cell types of peripheral blood cells as well as to detect modality-specific molecular features associated to these.

### 5 Conclusion

We present here MOJITOO, which is a fast and parameter free method based on canonical correlation analysis for integration of multimodal single-cell data of any protocol. A comprehensive analysis with six bi-modal and tri-modal multimodal datasets indicates that MOJITOO has the best performance regarding the preservation of the structures across modalities and the recovery of clusters, while

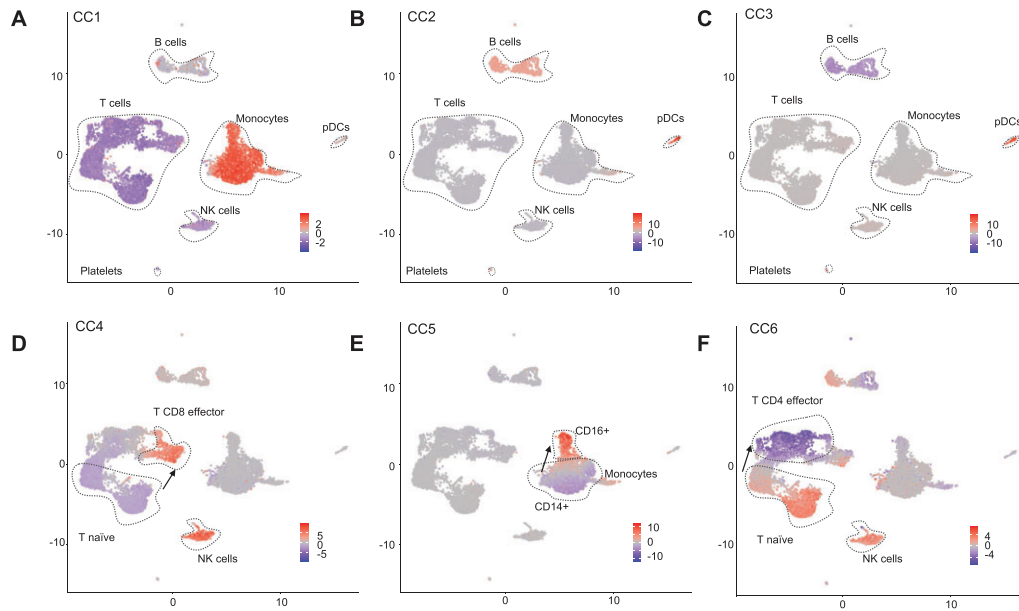


Fig. 4. (A-F) UMAP with the scores of CC1 to CC6. We highlight major cell types (or sub-types) associated to positive or negative CC scores and arrows indicate directions associated to the activation of particular immune cells

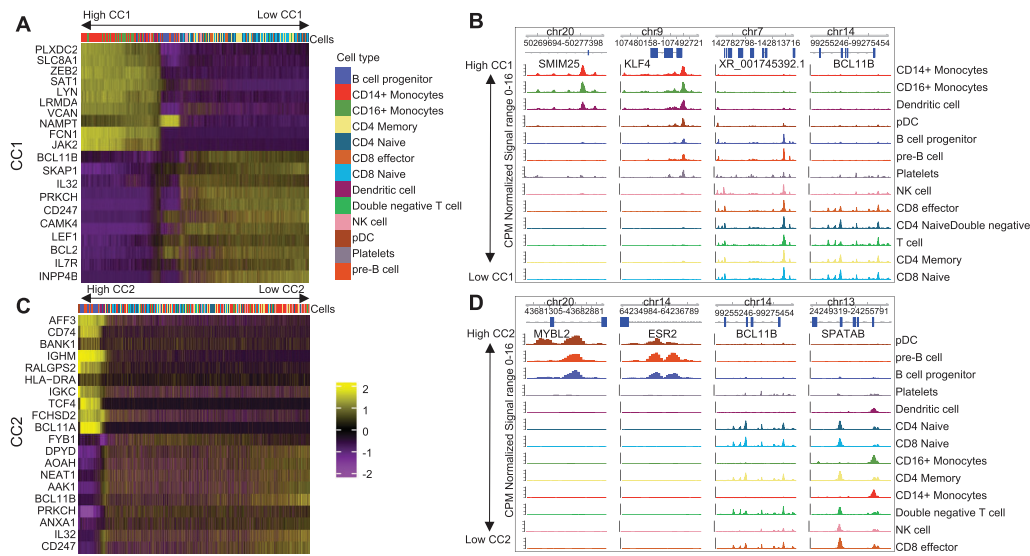


Fig. 5. (A) Heatmap with scores for the top 10 positive and negative genes for CC1 (y-axis) versus cells (x-axis). Cells are ordered by CC1 scores (high to low). (B) Genome browser tracks with top 2 positive and negative peaks for CC1. Tracks correspond to normalized cell-specific pseudo bulk ATAC-seq profiles generated by deepools (Ramirez et al., 2016). Cell-specific tracks are ordered by CC1 score (high to low). (C and D) show respectively the heatmap of top genes and the genome browser of top peaks for CC2

it is ranked second regarding distance representation. Moreover, MOJITOO has the lowest time and memory requirements requiring 2.5 min and 6.4 GB in the largest dataset with 30,000 cells. WNN, which is the standard method for integration in Seurat, performed well on the clustering problem (2nd after MOJITOO) and had a low computational time, but had a poor performance in the structure preservation and silhouette scores. Moreover, WNN, which outputs a distance matrix on the shared space, does not provide latent features as MOJITOO, DIABLO or MOFA. MOFA performed well on distance representation, but was ranked fourth in the clustering problem and third at structure preservation score. DIABLO, which is also based in CCA, had an good predictive performance (2nd for structure preservation and 3rd in clustering and distance preservation), but had an a poor computational performance being the 2nd slowest method.

An interesting result is the fact the structure preservation scores are more uniform across modalities for CCA-based methods MOJITOO and DIABLO, while runner-up methods (MOFA) obtained highest scores for the RNA modality. This is possibly rooted on the analytical frameworks of these methods. CCA analysis explicitly finds canonical vectors with high correlation across modalities, while matrix factorization methods (MOFA) do not explicitly guarantee factors are uniformly well represented across modalities. Of note, the CCA approach explored in Symphony was also biased toward structure preservation in RNA, due to the fact it only use RNA canonical vectors as a final latent representation. Another finding of this study is that matrix factorization methods (MOJITOO, MOFA, DIABLO), benefit from working with low dimensional matrices as input by having lower computational requirements. For MOFA, a comparison of the results with or without

dimension reduction even indicates positive effects of low dimensional inputs in the clustering problem.

It is important to point out that measures used in our benchmarking have their own limitations. The distance and clustering evaluation requires labels, which might be biased toward the methods used to derive the labels. Indeed, WNN does well on the clustering problem of the two CITE-seq data, which are originally evaluated with WNN. The Silhouette score has underlying distribution assumptions, which might not be met by the latent space generated by integration methods. Finally, the structure score requires a low-dimensional representation of the spaces. In this study, however, we consider methods performing well in all metrics, which mitigates biases of individual metrics.

Finally, we highlight how a simple inspection of CCA derived latent spaces supports the biological interpretation and detection of relevant molecular features, as exemplified in the multiome PBMC dataset. Future work includes further exploring the interpretability of MOJITOO, for example, by finding associations between molecular features across modalities as gene to peak links (Granja *et al.*, 2021). Here, it would be interesting to explore potential non-linearities of features, as commonly explored in scRNA-seq data (Van Dijk *et al.*, 2018). Finally, an interesting topic is to investigate if differences in the modality-specific canonical vectors detected by MOJITOO can indicate biological properties of those. For example, in the Skin SHARE-seq data (Ma *et al.*, 2020), authors show that cells with changes in chromatin preceding changes in gene expression indicates cell differentiation.

## Funding

This project was funded by the German Research Foundation (DFG) (project GE 2811/3-2) and the E: MED Consortia Fibromap funded by the German Ministry of Education and Science (BMBF).

*Conflict of Interest:* none declared.

## Data availability

Pre-processed publically available data used for benchmarking is available at ZENODO: <https://doi.org/10.5281/zenodo.6348128>.

## References

- Amemiya, H.M. *et al.* (2019) The encode blacklist: identification of problematic regions of the genome. *Sci. Rep.*, **9**, 1–5.
- Argelaguet, R. *et al.* (2020) MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.*, **21**, 1–17.
- Argelaguet, R. *et al.* (2021) Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.*, **39**, 1202–1215.
- Benjamini, Y., Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological)*, **57**, 289–300.
- Butler, A. *et al.* (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.
- Buus, T.B. *et al.* (2021) Improving oligo-conjugated antibody signal in multimodal single-cell analysis. *Elife*, **10**, e61973.
- Cao, J. *et al.* (2018) Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, **361**, 1380–1385.
- Clark, S.J. *et al.* (2018) scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.*, **9**, 1–9.
- Efremova, M., Teichmann, S.A. (2020) Computational methods for single-cell omics across modalities. *Nat. Methods*, **17**, 14–17.
- Gayoso, A. *et al.* (2021) Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods*, **18**, 272–282.
- Granja, J.M. *et al.* (2021) ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.*, **53**, 403–411.
- Hao, Y. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.e29.
- Hu, Y. *et al.* (2020) Single-cell RNA cap and tail sequencing (scRCAT-seq) reveals subtype-specific isoforms differing in transcript demarcation. *Nat. Commun.*, **11**, 1–11.
- Hubert, L., Arabie, P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.
- Jain, M.S. *et al.* (2021) MultiMAP: dimensionality reduction and integration of multimodal data. *Genome Biol.*, **22**, 1–26.
- Jin, S. *et al.* (2020) scAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles. *Genome Biol.*, **21**, 1–19.
- Kang, J.B. *et al.* (2021) Efficient and precise single-cell reference atlas mapping with symphony. *Nat. Commun.*, **12**, 1–21.
- Korsunsky, I. *et al.* (2019) Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods*, **16**, 1289–1296.
- Kriebel, A.R. and Welch, J.D. (2022) UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization. *Nat. Commun.*, **13**, 1–17.
- Li, Z. *et al.* (2021) Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen. *Nat. Commun.*, **12**, 1–14.
- Ma, S. *et al.* (2020) Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell*, **183**, 1103–1116.
- Mimitou, E.P. *et al.* (2021) Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.*, **39**, 1246.
- Ramírez, F. *et al.* (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**, W160–W165.
- Ramsay, J. and Silverman, B. (1997) *Functional Data Analysis*. Springer, New York, NY, pp. 199–216.
- Rousseeuw, P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.
- Singh, A. *et al.* (2019) DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, **35**, 3055–3062.
- Singh, R. *et al.* (2021) Schema: metric learning enables interpretable synthesis of heterogeneous single-cell modalities. *Genome Biol.*, **22**, 1–24.
- Stoeckius, M. *et al.* (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods*, **14**, 865–868.
- Stuart, T. *et al.* (2021) Single-cell chromatin state analysis with signac. *Nat. Methods*, **9**, 1–9.
- Swanson, E. *et al.* (2021) Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *Elife*, **10**, e63632.
- Van Dijk, D. *et al.* (2018) Recovering gene interactions from single-cell data using data diffusion. *Cell*, **174**, 716–729.
- Welch, J.D. *et al.* (2019) Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, **177**, 1873–1887.