Integrating unpaired scRNA-seq and scATAC-seq data with equal or unequal cell type compositions

Ziqi Zhang [1] **Haoran Sun** [2] **Ragunathan Mariappan** [3] **Xi Chen** [4] **Mika S Jain** [5] **Mirjana Efremova** [6] **Sarah A Teichmann** [5] **Vaibhav Rajan** [3] **Xiuwei Zhang** [1]

Abstract

Single cell multi-omics technology is able to measure multiple data modalities at single cell resolution, such as gene expression level (using single cell RNA-sequencing) and chromatin accessibility (using single cell ATAC-sequencing). Integrating scRNA-seq and scATAC-seq data profiled from different cells is a challenging problem. Existing methods often require that the scRNA-seq and scATAC-seq data cover the same cell types, that is, the same clusters. However, this is often not true for many existing datasets. Here we propose a joint matrix tri-factorization algorithm `scJMT` that is capable of integrating and clustering cells from both modalities of data in the case where the two data modalities do not share exactly the same cell types. The tri-factorization framework also allows us to obtain clusters of genes and chromatin regions, and the association matrices between cell clusters and gene or region clusters. We show that `scJMT` is superior to two state-of-the-art methods under both scenarios where the two modalities have the same or different cluster compositions.

1. Introduction

The availability of single cell multimodal omics data provides a comprehensive view of each single cell. Single cell RNA-sequencing (scRNA-seq) and single cell ATAC-sequencing (scATAC-seq) respectively measures the gene-expression and chromatin accessibility profiles of cells, each being considered as an important aspect of a cell. Recently, techniques which can measure both gene-expression and chromatin accessibility in the same cells have been proposed (Chen et al., 2019; Cao et al., 2018; Ma et al., 2020), but these technologies are still not widely used, and they can suffer from low sensitivity of one of the data modalities. To make use of the enormous amount of existing data, computational methods have been proposed to integrate scRNA-seq and scATAC-seq data obtained separately in different batches (Stuart et al., 2019; Welch et al., 2019; Duren et al., 2018; Korsunsky et al., 2019; Cao et al., 2020), by removing the batch effect and grouping the cells with similar biological identity together across data modalities.

Existing methods which can integrate scRNA-seq and scATAC-seq data from different cells use various optimization objectives. Some methods aim to integrate datasets and learn their latent embedding such that the latent embedding can reconstruct the original datasets (Welch et al., 2019; Duren et al., 2018; Argelaguet et al., 2020). Some use manifold alignment (Cui et al., 2014) and aim to learn the latent embedding by enforcing the latent embedding to preserve the pairwise distances of cells in the original high-dimensional space (Singh et al., 2020; Cao et al., 2020). `Seurat` (Stuart et al., 2019) maps a query dataset to a reference dataset, and obtain a new data matrix for the query dataset based on the reference dataset. However, most of these methods were designed with the assumption that the cell type composition is the same for both data modalities and it is not clear how they perform when the two data modalities do not share exactly the same cell types.

Here we design a joint matrix tri-factorization framework (`scJMT`), which decomposes a single cell data matrix into a cell factor, a feature factor, and an association matrix representing the correspondence between the cell and feature factors. By confining the factor representation of each entity to be simplex, the latent factor can be interpreted as a joint soft clustering result and each latent dimension corresponds to one cluster identity. In order to connect the features in different data modalities (in our case, the genes in scRNA-seq dataset and regions in scATAC-seq dataset), we use a pre-defined gene activity matrix, which has been used in

[1]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA [2]School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia, USA [3]Department of Information Systems and Analytics, National University of Singapore, Singapore [4]Department of Biology, Southern University of Science and Technology, China [5]Wellcome Sanger Institute, United Kingdom [6]Cancer Research UK Barts Center. Correspondence to: Xiuwei Zhang <xiuwei.zhang@gatech.edu>.

The 2021 ICML Workshop on Computational Biology.

existing work (Welch et al., 2019; Duren et al., 2018; Stuart et al., 2019) (Fig. 1).

scJMT outperforms both Liger (Welch et al., 2019) and CoupleNMF (Duren et al., 2018) in terms of integrating the modalities and correctly clustering cells into cell types. scJMT also has unique features compared to existing method which is that it can learn feature (gene or region) clusters and association matrices while learning the cell clusters, which will be benchmarked in our future work.

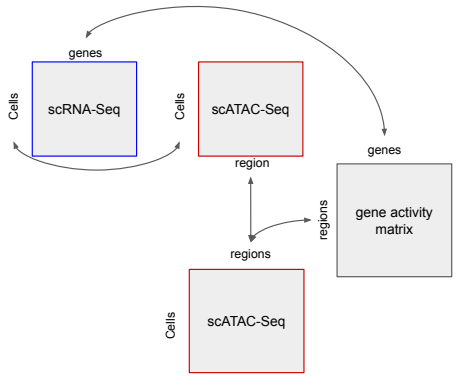

Figure 1. scJMT considers all the data matrices as relationship matrices between entities, and the latent factor of the same entity is shared between relationship matrices.

2. Methods

2.1. Matrix tri-factorization

We consider a data matrix where the two entities both have cluster structure. A matrix tri-factorization approach can cluster the two entities while learning the association relationship between the clusters of the two entities. For example, in a scRNA-seq count matrix, the two entities are cells and genes, each forming certain number of clusters. Furthermore, there can be correspondence between the cell and gene clusters, where the clusters of genes corresponding to a cluster of cells are considered as marker genes for that cluster. In a matrix tri-factorization model, a data matrix $\mathbf{X}$ is assumed to be the product of three matrices $\mathbf{C}_i\boldsymbol{\Sigma}\mathbf{C}_j$, where $\mathbf{C}_i$ and $\mathbf{C}_j$ are the cluster identification matrices of the two entities, and $\boldsymbol{\Sigma}$ is the association matrix between the clusters of the two entities. We find estimates of $\mathbf{C}_i$, $\boldsymbol{\Sigma}$ and $\mathbf{C}_j$ that minimize the reconstruction error:

$$\hat{\mathbf{C}}_i, \hat{\boldsymbol{\Sigma}}, \hat{\mathbf{C}}_j = \arg\min_{\mathbf{C}_i;\mathbf{C}_j;\boldsymbol{\Sigma}} \|\mathbf{X} - \mathbf{C}_i\boldsymbol{\Sigma}\mathbf{C}_j\|_F^2 \quad (1)$$

We apply tri-factorization to both the scRNA-seq and the scATAC-seq data matrices. We assume that one cell or feature only belongs to one cluster when performing clustering, where one cell only has one cell type identity, and the same is true for a feature. This assumption makes each row of $\mathbf{C}_i$ or $\mathbf{C}_j$ to be a one-hot cluster indicator vector for each cell. Finding such binary $\mathbf{C}_i$ and $\mathbf{C}_j$ matrices is computationally intractable (Anagnostopoulos et al., 2008). So, we relax the problem into a soft clustering problem by considering the cluster membership of each cell as a probability instead of a binary value. As a result, each row of $\mathbf{C}_i$ or $\mathbf{C}_j$ is constrained to be a simplex. Using soft clustering relaxation provides another advantage: soft clustering can better preserve the cell heterogeneity within each clusters, even cells in the same clusters have slightly different cluster indicator vector, and this vector can be used as a latent cell embedding.

2.2. Single Cell Joint matrix tri-factorization (scJMT)

We develop a single cell joint matrix tri-factorization (scJMT) approach to integrate unpaired scRNA-seq and scATAC-seq data, where cells from both data matrices, genes, and regions are jointly clustered, and associations between cell and gene clusters (from the scRNA-seq data), and between cell and region clusters (from the scATAC-seq data) are learned.

As shown in Fig. 1, when integrating scRNA-Seq and scATAC-Seq data with cells from different batches, the binary relationships between chromatin regions and genes which represent which region affect which gene are needed. This is often in the form of a *gene activity matrix*, and is typically determined by the relative distance between genes and regions on the genome. Denoting the scRNA-Seq data matrix by $\mathbf{G}$ where rows correspond to cells and columns correspond to genes, the scATAC-Seq matrix by $\mathbf{R}$ where rows correspond to cells and columns correspond to regions, and the gene activity matrix by $\mathbf{A}$ where rows and columns correspond to respectively genes and regions. The objective function of scJMT is:

$$\begin{aligned} &\arg\min_{\mathbf{C}_1;\mathbf{C}_2;\mathbf{C}_g;\mathbf{C}_r;\boldsymbol{\Sigma}_g;\boldsymbol{\Sigma}_r;\mathbf{b}_g;\mathbf{b}_r} L \\ &s.t. \\ &\quad \mathbf{C}_i\mathbf{1} = \mathbf{1},\ \mathbf{C}_i \geq 0,\ i = 1,2,g,r \end{aligned} \quad (2)$$

where

$$\begin{aligned} L = {} & \alpha_1 k\mathbf{R} - \mathbf{C}_1\boldsymbol{\Sigma}_r\mathbf{C}_r - \mathbf{b}_1 - \mathbf{b}_r^T\|_F^2 \\ & + \alpha_2 k\mathbf{G} - \mathbf{C}_2\boldsymbol{\Sigma}_g\mathbf{C}_g - \mathbf{b}_2 - \mathbf{b}_g^T\|_F^2 \\ & \alpha_3 \frac{tr(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g^T)}{k\boldsymbol{\Sigma}_r k_F k\boldsymbol{\Sigma}_g k_F} \quad \alpha_4 \frac{tr(\mathbf{C}_g^T\mathbf{A}\mathbf{C}_r)}{k\mathbf{C}_g k_F k\mathbf{A}\mathbf{C}_r k_F} \end{aligned} \quad (3)$$

In the equation above, $\mathbf{C}_1$ and $\mathbf{C}_2$ are the cell factors for scATAC-Seq data and scRNA-Seq data. $\mathbf{C}_g$ is the gene factor and $\mathbf{C}_r$ is the region factor. $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_g$ are the association matrices of scATAC-Seq and scRNA-Seq data. $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_g, \mathbf{b}_r$ are vectors that correspond to cell- and feature- (gene and chromatin region) specific bias. In addition to the tri-factorization term, we include two cosine similarity term to enforce the matching of cell and feature clusters between modalities. The gene activity matrix $\mathbf{A}$ is used in the cosine similarity term $\frac{tr(\mathbf{C}_g^T\mathbf{A}\mathbf{C}_r)}{k\mathbf{C}_g k_F k\mathbf{A}\mathbf{C}_r k_F}$ in order

to match the latent factors of genes and chromatin regions. The term $\frac{tr(\mathbf{\Sigma}_r \mathbf{\Sigma}_g^T)}{k\mathbf{\Sigma}_r k_F k\mathbf{\Sigma}_g k_F}$ forces the cell clusters from the two modalities to have similar association property with features between data modalities.

When dealing with the scenario where the two modalities have different cluster compositions, we include two additional regularization terms to guide the framework to learn the correct clusters. Assume the cells from the two data modalities each has N_1 and N_2 clusters, but only k (k $\min\{N_1, N_2\}$) of them are shared across modalities. We assign latent dimension of $N_1 + N_2$ k to all cluster indicator matrices (C_1, C_2, C_g and C_r). The first k dimensions correspond to the shared clusters, dimensions from $k+1$ to N_1 correspond to the batch specific clusters for the first batch, and the remaining dimensions correspond to that specific to the second batch. We then would like to include two additional regularization terms, $k\mathbf{C}_1(:,(N_1+1):(N_1+N_2 \quad k))k_F^2 + k\mathbf{C}_2(:,(k+1):N_1)k_F^2$, into the loss function (Eq. 3). By minimizing the first term, we enforce that no cell in the scRNA-seq data falls into clusters N_1+1 to N_1+N_2 k. And by minimizing the second term, we enforce that no cell in the scATAC-seq data falls into clusters $k+1$ to N_1.

We minimize the loss function using mini-batch stochastic gradient descent. Within each step, we pick one parameter matrix from $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_g$, $\mathbf{C}_r$, $\mathbf{\Sigma}_g$, $\mathbf{\Sigma}_r$ and fix all the other parameters. Then, we update a mini-batch of the selected parameter matrix using gradient descent. We loop through all the parameters and update them one by one. We update the bias terms ($\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_g, \mathbf{b}_r$) using closed form solution by setting the gradient of it to zero, and we take latent factors into a softmax function in order to enforce the simplex constraint. Since the optimization problem is non-convex, gradient descent can converge to a local minimum point. So we run the algorithm independently multiple times with different random seeds, and select the one with the smallest loss value. The pseudo-code of the algorithm is as follows:

Algorithm 1 single cell joint matrix tri-factorization

function scJMT($\mathbf{G}, \mathbf{R}, \mathbf{A}$, batchsize)
  Initialize $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_g$, $\mathbf{C}_r$, $\mathbf{\Sigma}_g$ and $\mathbf{\Sigma}_r$ using $N(\mathbf{0}, \mathbf{I})$
  for t **in** $1, 2, \quad, T$ **do**
    //Sample mini-batch
    $\mathbf{R}_m, \mathbf{G}_m, \mathbf{A}_m$ = Sample($\mathbf{R}$, $\mathbf{G}$, $\mathbf{A}$, batchsize)
    for $\mathbf{X}$ **in** $\{\mathbf{C}_i, \mathbf{b}_i, \mathbf{\Sigma}_g, \mathbf{\Sigma}_r\ (i = 1, 2, g, r)\}$ **do**
      if $\mathbf{X} \in \mathbf{b}_i$ **then**
        $\mathbf{X} = \mathbf{X} \quad r_{\mathbf{X}} L(\mathbf{R}_m, \mathbf{G}_m, \mathbf{A}_m)$
      else
        $\mathbf{X} = \arg\min_{\mathbf{X}} L(\mathbf{R}_m, \mathbf{G}_m, \mathbf{A}_m)$
  return $\mathbf{C}_1$, $\mathbf{C}_2$, $\mathbf{C}_g$, $\mathbf{C}_r$, $\mathbf{\Sigma}_g$ and $\mathbf{\Sigma}_r$

3. Results

We generate simulated datasets using an extended version of SymSim (Zhang et al., 2019). SymSim is able to generate scRNA-Seq data that resembles real scRNA-Seq data. We extend SymSim such that it is also able to generate scATAC-Seq data, given a gene activity matrix. The data simulation process is similar to that described in (Zhang et al., 2021). The process in (Zhang et al., 2021) was used to generate continuous populations and in this paper we generate discrete populations. We simulate two scenarios. In the first scenario, cells from different batches have the same cell type composition. And in the second scenario, one cell type is missing in one of the data batches.

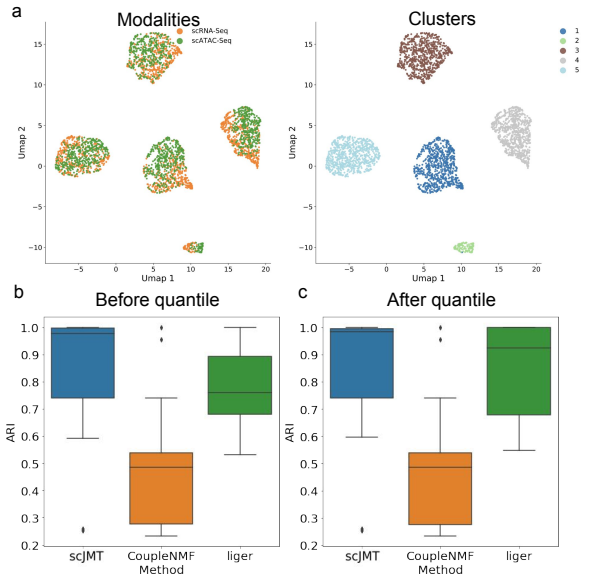

Figure 2. (a) The latent factor of cells in one sample dataset learned by `scJMT` visualized using UMAP. In the plot on the left, cells are colored according to modalities, and on the right, cells are colored according to ground truth cluster labels. (b-c) The ARI scores of `scJMT`, `Liger` and `CoupleNMF` on simulated datasets. (b) Scores are calculated on the cell factors before quantile normalization. (c) Scores are calculated on the cell factors after quantile normalization.

In the first scenario, we generate 6 simulated datasets, each with 5 clusters. `scJMT` is able to group cells of the same cell type into the same cluster regardless of different modalities (the UMAP (McInnes et al., 2018) visualization of learned cell factor on one of the simulated datasets is shown in Fig 2a). We then measure the accuracy of the inferred cluster ID using Adjusted Rank Index (ARI), and compare the ARI score of `scJMT` with that of `Liger` and `CoupleNMF`. `Liger` uses an additional post-processing step after matrix factorization to further improve the clustering accuracy, termed *quantile normalization*. In order to make an unbiased comparison, we measure the ARI score of all three methods

in both cases before and after applying the quantile normalization. The result (Fig. 2b, c) shows that our model consistently achieves a better score compared to `CoupleNMF` and `Liger`.

In the second scenario, we generate 12 simulated datasets where one data modality (scRNA-seq) has three cell clusters ($N_1 = 3$) and the other (scATAC-seq) has two cell clusters ($N_2 = 2$). Only two clusters between modalities are matched ($k = 2$). We compare the performance of our model with `CoupleNMF` and `Liger`. The cell latent factor learned by `scJMT` in one simulated dataset is shown in Fig. 3a,b. `scJMT` successfully integrates cluster 2 and 3 between two modalities and leave unpaired cluster 1, the unique cluster identity in scRNA-Seq. The ARI scores of `scJMT`, `Liger` and `CoupleNMF` on all 12 datasets are shown in Fig. 3c, `scJMT` has a higher score compared to `Liger` and `CoupleNMF`.

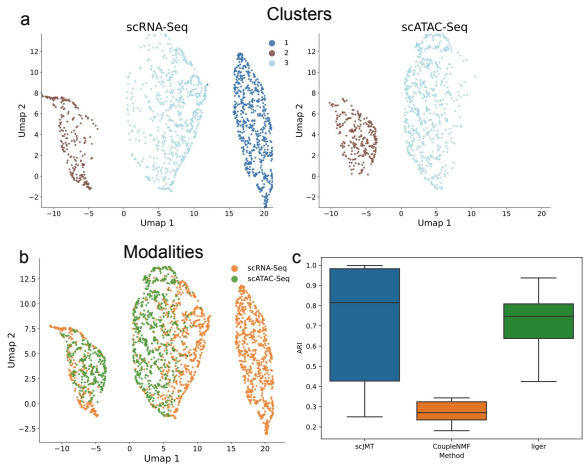

Figure 3. (a) The latent factors of cells in one sample dataset learned by `scJMT`. In the sample data one cluster is missing in scATAC-Seq. Cells are colored according to ground truth cell labels. UMAP was performed on the latent factors $\mathbf{C}_1$ and $\mathbf{C}_2$ jointly, and the embeddings of $\mathbf{C}_1$ and $\mathbf{C}_2$ are separately plotted in order to show the missing cluster clearly. (b) The latent factor of cells in one sample dataset learned by `scJMT`. The cells are colored according to different modalities. (c) The ARI scores of `scJMT`, `CoupleNMF` and `Liger` (after quantile normalization) on 12 simulated datasets.

Finally, we test our model on a real dataset, where scRNA-seq and scATAC-seq experiments were performed on mouse spleen (Chen et al., 2018; Jain et al., 2021). Some single cell data integration methods based on matrix factorization frameworks conduct post-processing steps on the learned factorized matrices to obtain reasonably good results (Welch et al., 2019). In the results shown on simulated data we do not perform any post-processing but directly take the cluster membership indicated by the factor matrices of cells. For our results on real data, we conduct a simple post-processing step similar to (Zhang et al., 2021) to match the latent space better. Basically, we construct a mutual nearest neighbor graph between cell factors $\mathbf{C}_1$ and $\mathbf{C}_2$. For every pair of cells p and q, the weight between them in the graph is:

$$\mathbf{W}_{pq} = \exp\left(\quad k\mathbf{C}_1(p) \quad \mathbf{C}_2(q)k_2^2 \right) \tag{4}$$

We update $\mathbf{C}_1$ for every cell p as follows:

$$\mathbf{C}_1(p)^0 = \frac{\sum_{q2\text{neigh}(p)} \mathbf{W}_{pq}\mathbf{C}_2(q)}{\sum_{q2\text{neigh}(p)} \mathbf{W}_{pq}} \tag{5}$$

We visualize the cell latent factor (Fig. 4) using UMAP, since there is no ground truth label on real dataset, we use the reference label provided by the original data paper for visualization. From Fig. 4, cells with the same reference labels are well integrated into the same clusters.

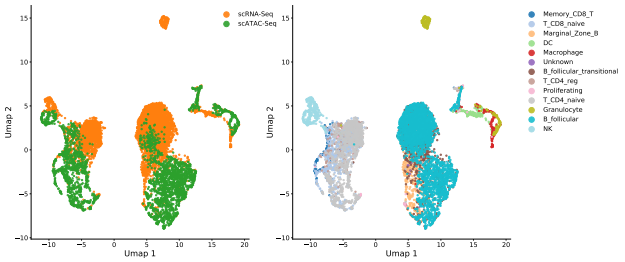

Figure 4. The latent factor of cells in Mouse spleen dataset learned by `scJMT` visualized with UMAP. In the left plot, cells are colored according to the modalities, and in the right plot, cells are colored according to the reference cluster labels.

4. Conclusion

Integrating single-cell multi-omics data is a challenging problem, especially when cells from different data modalities are not jointly profiled. Some existing methods integrate such data using the topological similarity between data manifolds, which requires different data batches to have similar manifold structure. However, it is common in real datasets for different data modalities to not cover exactly the same cell states. In this case the topologies are not matched between modalities. Our method, `scJMT`, explores the possibility of integrating data modalities where part of the biological process is not matched between data modalities. The tri-factorization framework also allows us to learn the clusters of the features (genes and regions) as well as the association between cell and feature clusters. Future work can evaluate other unique abilities of our model such as the inferred gene clusters and their association matrices. This framework can also be generalized to include more matrices, including paired scRNA-seq and scATAC-seq data.

References

Anagnostopoulos, A., Dasgupta, A., and Kumar, R. Approximation algorithms for co-clustering. In *Proceedings of the 27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 201–210, 2008.

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., and Stegle, O. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.*, 21(1):111, May 2020.

Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., McFaline-Figueroa, J. L., Packer, J. S., Christiansen, L., Steemers, F. J., Adey, A. C., Trapnell, C., and Shendure, J. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385, September 2018.

Cao, K., Bai, X., Hong, Y., and Wan, L. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics*, 36(Supplement_1):i48–i56, July 2020.

Chen, S., Lake, B. B., and Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.*, 37(12):1452–1457, December 2019.

Chen, X., Miragaia, R. J., Natarajan, K. N., and Teichmann, S. A. A rapid and robust method for single cell chromatin accessibility profiling. *Nat. Commun.*, 9(1):5345, December 2018.

Cui, Z., Chang, H., Shan, S., and Chen, X. Generalized unsupervised manifold alignment. *Adv. Neural Inf. Process. Syst.*, 27:2429–2437, 2014.

Duren, Z., Chen, X., Zamanighomi, M., Zeng, W., Satpathy, A. T., Chang, H. Y., Wang, Y., and Wong, W. H. Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc. Natl. Acad. Sci. U. S. A.*, 115(30):7723–7728, July 2018.

Jain, M. S., Conde, C. D., Polanski, K., Chen, X., Park, J., Botting, R. A., Stephenson, E., Haniffa, M., Lamacraft, A., Efremova, M., and Teichmann, S. A. MultiMAP: Dimensionality reduction and integration of multimodal data. February 2021.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-R., and Raychaudhuri, S. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods*, 16(12): 1289–1296, December 2019.

Ma, S., Zhang, B., LaFave, L. M., Earl, A. S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V. K., Tay, T., Law, T., Lareau, C., Hsu, Y.-C., Regev, A., and Buenrostro, J. D. Chromatin potential identified by shared Single-Cell profiling of RNA and chromatin. *Cell*, 183(4):1103–1116.e20, November 2020.

McInnes, L., Healy, J., and Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. February 2018.

Singh, R., Demetci, P., Bonora, G., Ramani, V., Lee, C., Fang, H., Duan, Z., Deng, X., Shendure, J., Disteche, C., and Noble, W. S. Unsupervised manifold alignment for single-cell multi-omics data. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 1–10. Association for Computing Machinery, New York, NY, USA, September 2020.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, 3rd, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. Comprehensive integration of Single-Cell data. *Cell*, 177(7):1888–1902.e21, June 2019.

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E. Z. Single-Cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887.e17, June 2019.

Zhang, X., Xu, C., and Yosef, N. Simulating multiple faceted variability in single cell RNA sequencing. *Nat. Commun.*, 10(1):2611, June 2019.

Zhang, Z., Yang, C., and Zhang, X. Learning latent embedding of multi-modal single cell data and cross-modality relationship simultaneously. April 2021.