Lecture 12
single-cell multi-omics integration

Outline

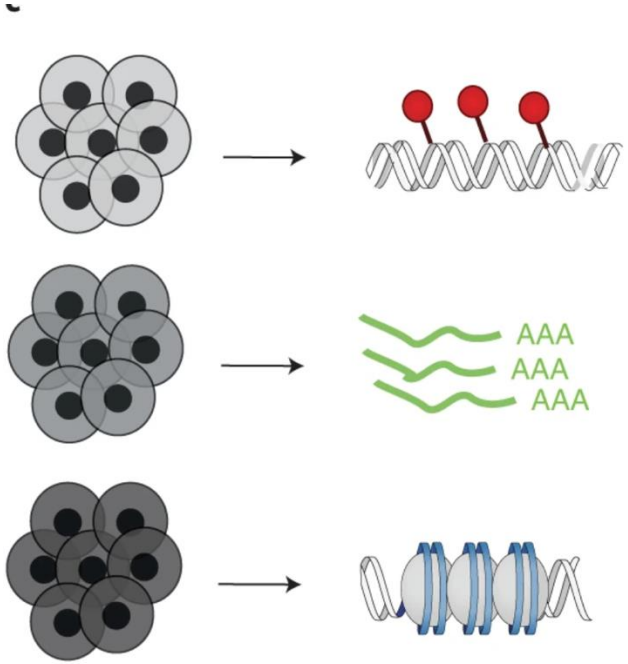
- Multi-omics data integration
  - Integrate unpaired multi-omics data
    - Integration of scATAC-seq and scRNA-seq
  - Integrate paired multi-omics data
  - Integrate unpaired multi-omics data using paired data as bridges

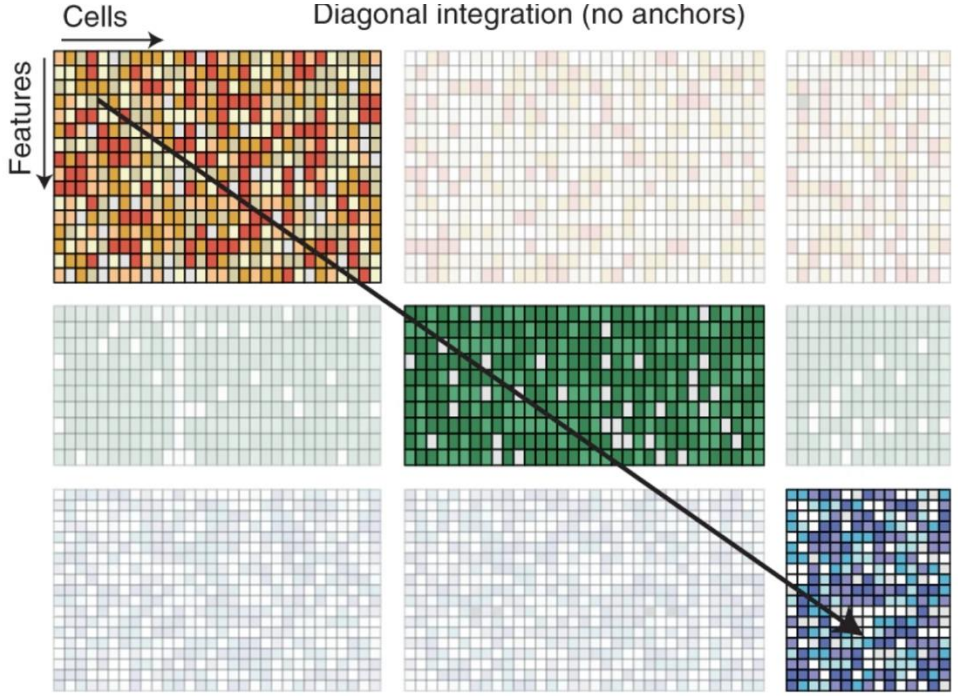
Integration between scRNA-seq and scATAC-seq

Why do we integrate?

- Identify cell-specific regulatory network
- scATAC-seq data is extremely sparse → borrow information from scRNA-seq for better cell type annotation

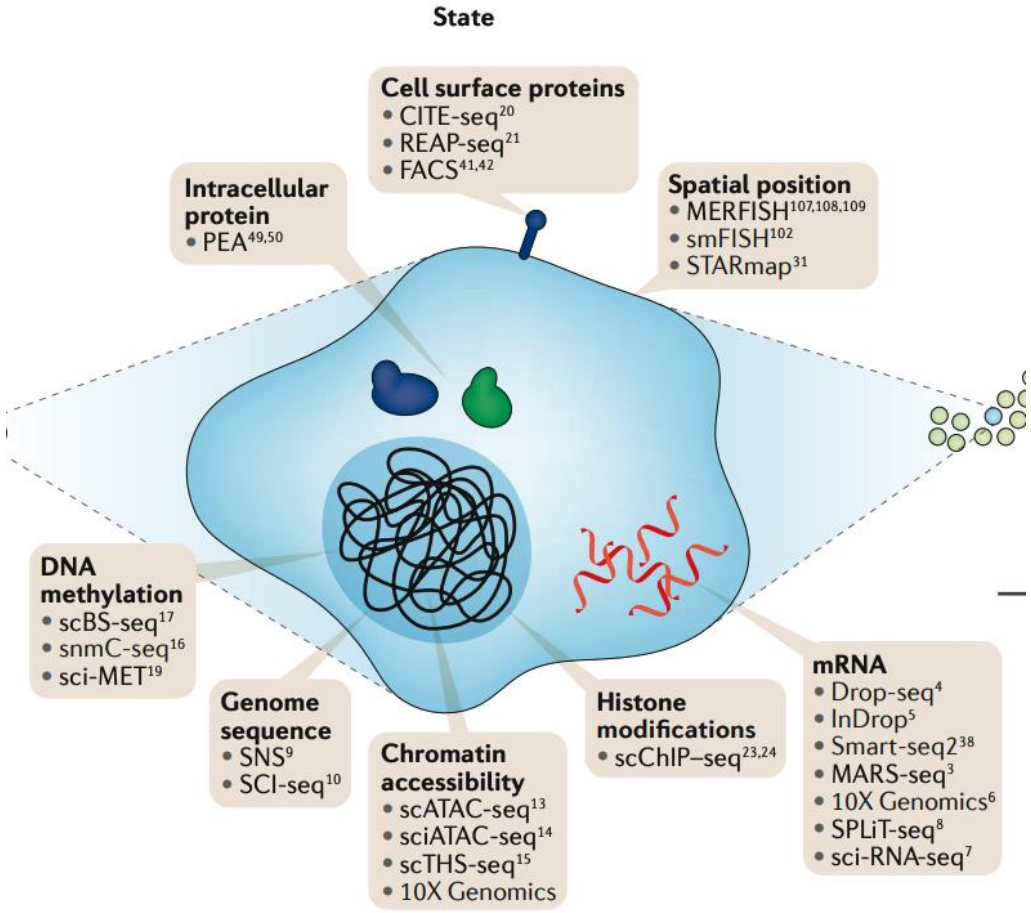
Challenge: require extra information about feature connections





Integrative single-cell analyses

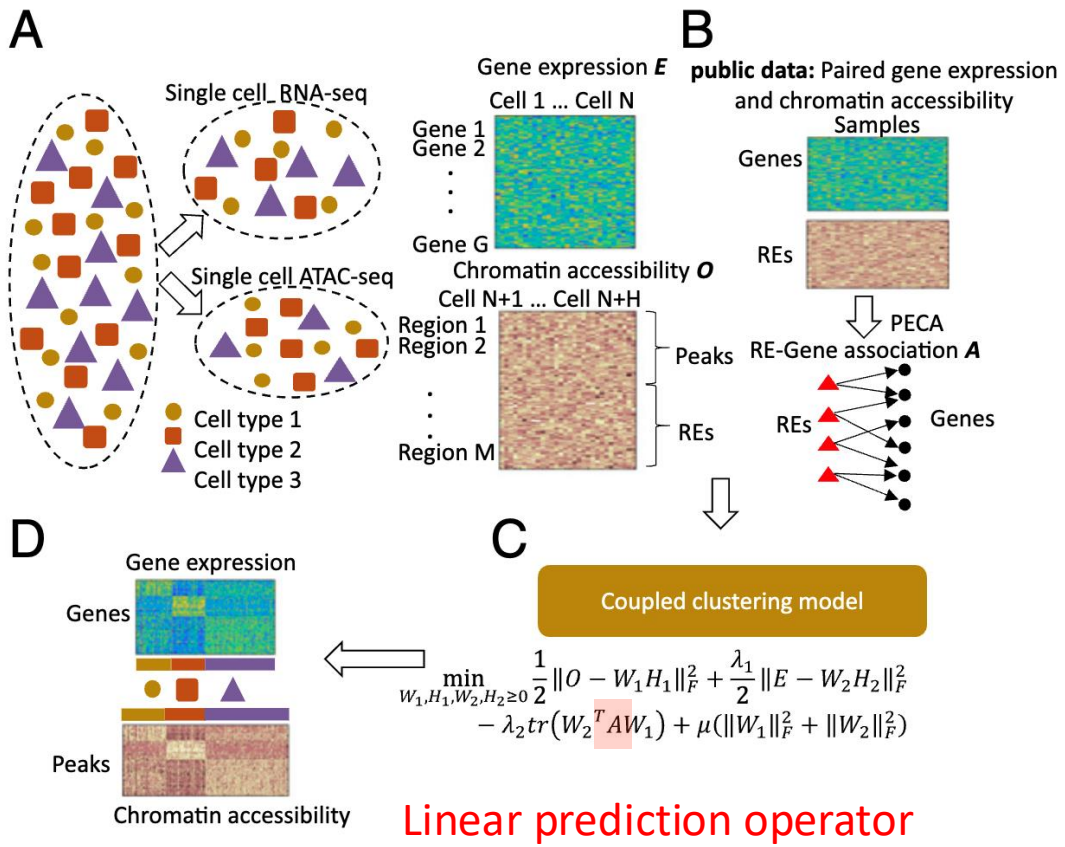
- Many technology only measure one modality of the single cells → unpaired multi-omics data
- Experimental methods have been developed to measure multiple modalities but can be more expensive



Integration of scRNA-seq and scATAC-seq

- Seurat v3 (Stuart et. al. Cell, 2019) :
  - Obtain gene activity matrix using Signac for scATAC-seq, treat as scRNA-seq data and integrate
  - Similar ideas used in scJoint (Lin et. al., Nature Biotech, 2022) and LIGER (Liu et. al., Nature Protocols, 2020)

- Coupled NMF (Daren et. al., PNAS, 2018)

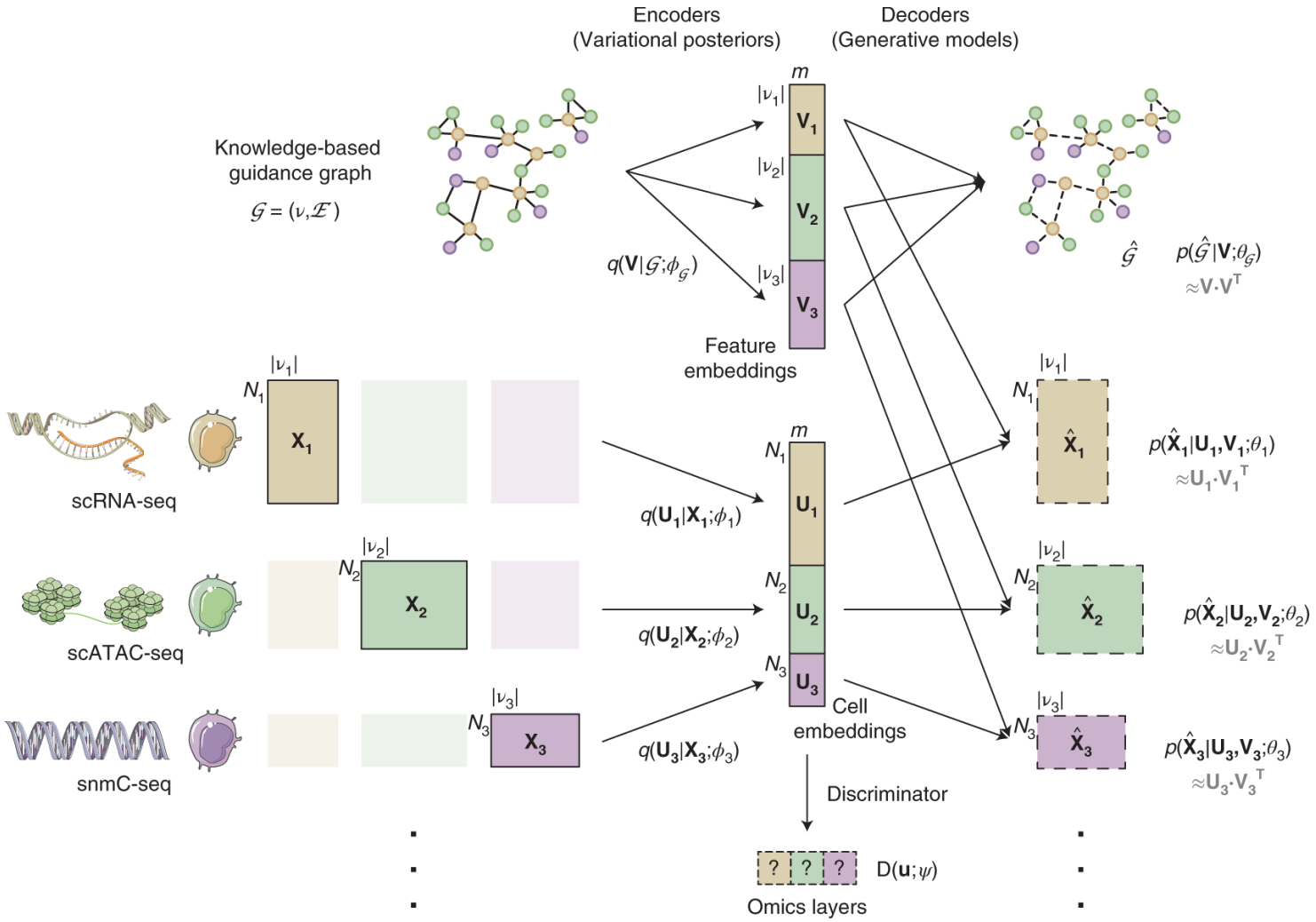


Linear prediction operator

- Core idea: perform coupled clustering, making sure that feature loadings are similar after transformations
- A: coupling matrix, gene-peak prediction matrix where each peak is predicted by sets of genes learnt from paired mRNA-ATACseq bulk data

- Challenges:
  - Single-cell and bulk level data can have platform specific biases
  - Can not guarantee that H_1 and H_2 can be properly merged

GLUE (Cao and Gao, Nature Biotech, 2022)



- General integration of unpaired multi-omics data
- Build a separate VAE for each modality data for cell embeddings
- Build feature embeddings using the variational graph auto-encoders (VGAE, Kipf and Welling, Arxiv, 2016)
- Build a guidance graph (signed and weighted, possibly multi edges between two nodes) based on prior knowledge on regulatory interactions across features from different modalities
  - Peak and gene are linked if they overlap with the gene body or proximal promoter regions

GLUE (Cao and Gao, Nature Biotech, 2022)

- Idea of VGAE

Data (input and output)

Definitions We are given an undirected, unweighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $N = |\mathcal{V}|$ nodes. We introduce an adjacency matrix $\mathbf{A}$ of $\mathcal{G}$ (we assume diagonal elements set to 1, i.e. every node is connected to itself) and its degree matrix $\mathbf{D}$. We further introduce stochastic latent variables $\mathbf{z}_i$, summarized in an $N \times F$ matrix $\mathbf{Z}$. Node features are summarized in an $N \times D$ matrix $\mathbf{X}$.

Target

Confounding covariates?

Inference model We take a simple inference model parameterized by a two-layer GCN:

$$q(\mathbf{Z} \,|\, \mathbf{X}, \mathbf{A}) = \prod_{i=1}^{N} q(\mathbf{z}_i \,|\, \mathbf{X}, \mathbf{A}) \,, \quad \text{with} \quad q(\mathbf{z}_i \,|\, \mathbf{X}, \mathbf{A}) = \mathcal{N}(\mathbf{z}_i \,|\, \boldsymbol{\mu}_i, \mathrm{diag}(\boldsymbol{\sigma}_i^2)) \,. \tag{1}$$

Here, $\boldsymbol{\mu} = \mathrm{GCN}_{\boldsymbol{\mu}}(\mathbf{X}, \mathbf{A})$ is the matrix of mean vectors $\boldsymbol{\mu}_i$; similarly $\log \boldsymbol{\sigma} = \mathrm{GCN}_{\boldsymbol{\sigma}}(\mathbf{X}, \mathbf{A})$. The two-layer GCN is defined as $\mathrm{GCN}(\mathbf{X}, \mathbf{A}) = \tilde{\mathbf{A}}\,\mathrm{ReLU}\big(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}_0\big)\mathbf{W}_1$, with weight matrices $\mathbf{W}_i$. $\mathrm{GCN}_{\boldsymbol{\mu}}(\mathbf{X}, \mathbf{A})$ and $\mathrm{GCN}_{\boldsymbol{\sigma}}(\mathbf{X}, \mathbf{A})$ share first-layer parameters $\mathbf{W}_0$. $\mathrm{ReLU}(\cdot) = \max(0, \cdot)$ and $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ is the symmetrically normalized adjacency matrix.

Generative model Our generative model is given by an inner product between latent variables:

$$p\left(\mathbf{A} \,|\, \mathbf{Z}\right) = \prod_{i=1}^{N} \prod_{j=1}^{N} p\left(A_{ij} \,|\, \mathbf{z}_i, \mathbf{z}_j\right) \,, \quad \text{with} \quad p\left(A_{ij} = 1 \,|\, \mathbf{z}_i, \mathbf{z}_j\right) = \sigma(\mathbf{z}_i^{\top} \mathbf{z}_j) \,, \tag{2}$$

where A_{ij} are the elements of $\mathbf{A}$ and $\sigma(\cdot)$ is the logistic sigmoid function.

Learning We optimize the variational lower bound $\mathcal{L}$ w.r.t. the variational parameters $\mathbf{W}_i$:

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{Z}|\mathbf{X},\mathbf{A})}\big[\log p\left(\mathbf{A} \,|\, \mathbf{Z}\right)\big] - \mathrm{KL}\big[q(\mathbf{Z} \,|\, \mathbf{X}, \mathbf{A}) \,||\, p(\mathbf{Z})\big] \,, \tag{3}$$

GLUE (Cao and Gao, Nature Biotech, 2022)

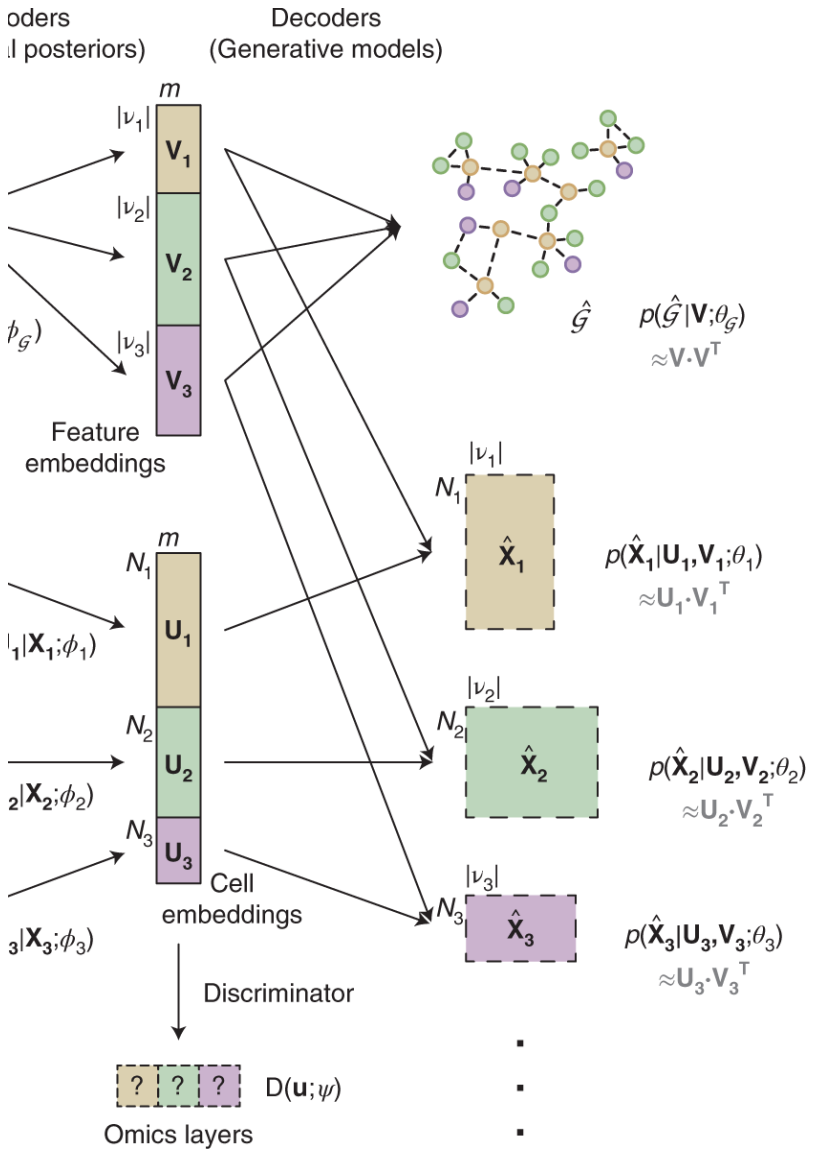
Some further details:

- GLUE is robust to corruption of the graph even 90% of the edges are random

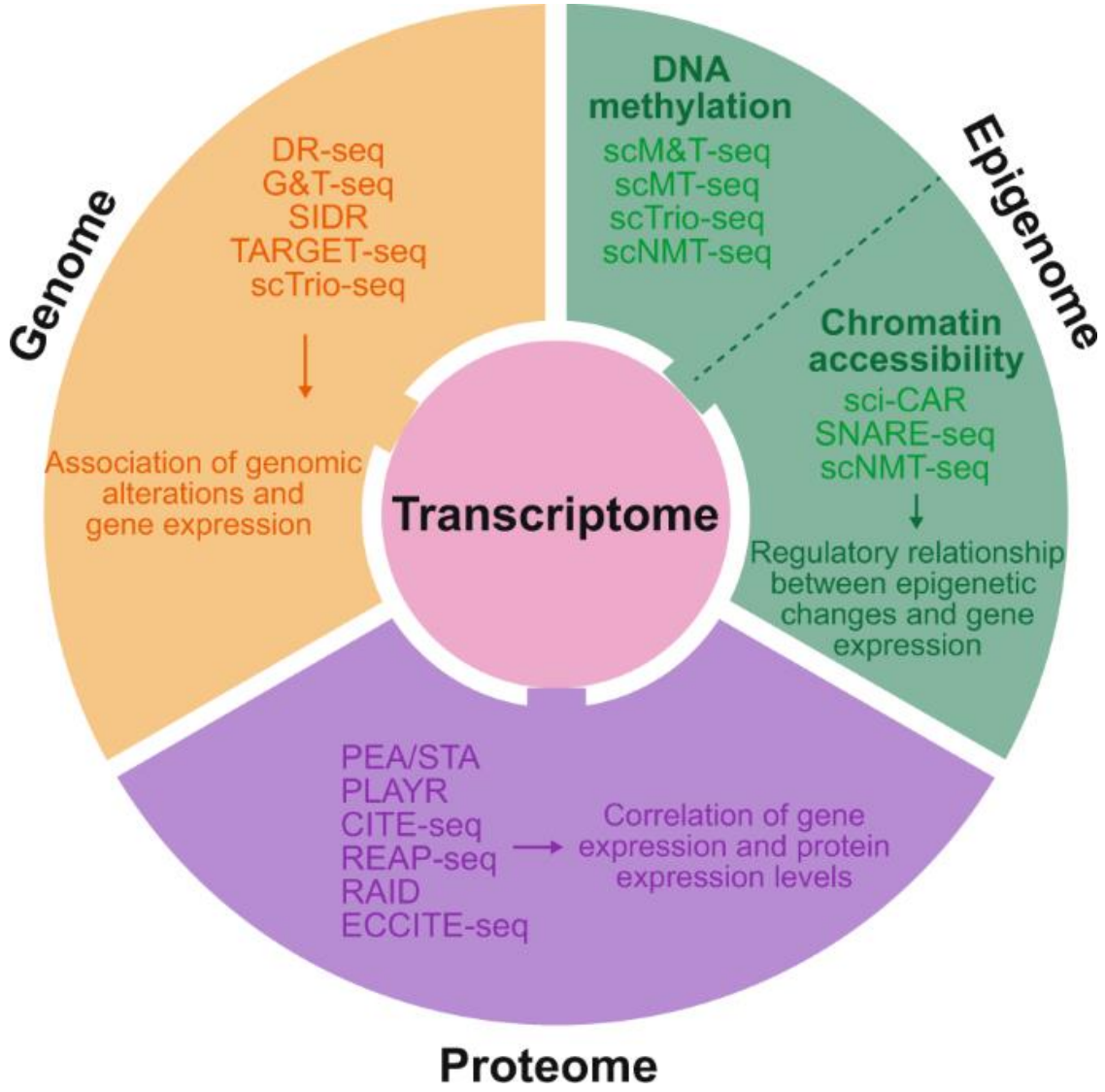
- How to combine the VGAE for feature embeddings and VAE for cell embeddings?
  - Cell embeddings are transformed based on feature embeddings
  - Linear decoder like SVD: for a cell i in dataset k, the predicted data has the form

$$\hat{\mu}_i^{(k)} = U_i \left(V^{(k)}\right)^T$$

- Need extra penalty to assure that cell embeddings are aligned across modalities (correct for batch effects)
  - Train a classifier (discriminator) to separate different datasets based on the cell embeddings
  - Penalize the loss if the discriminator has small classification error



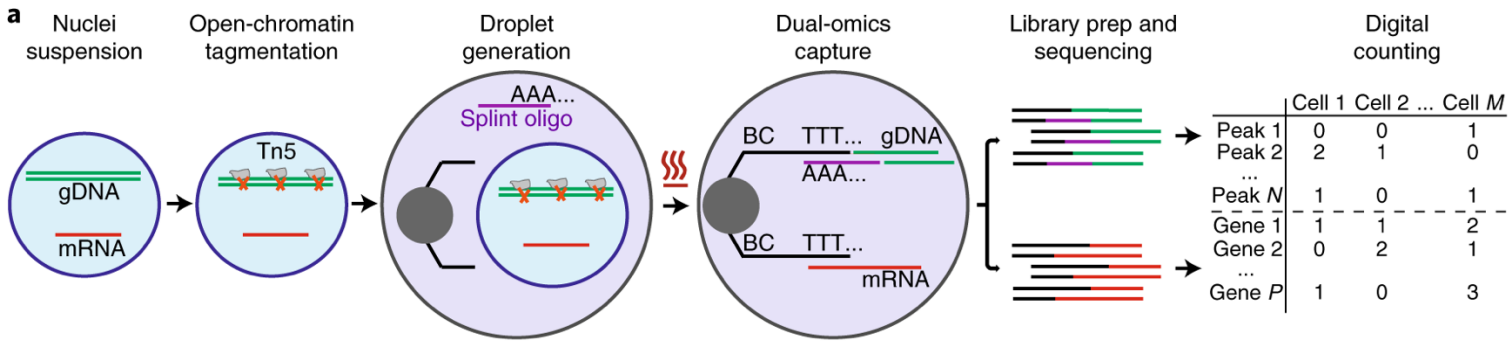
Single-cell multi-omics



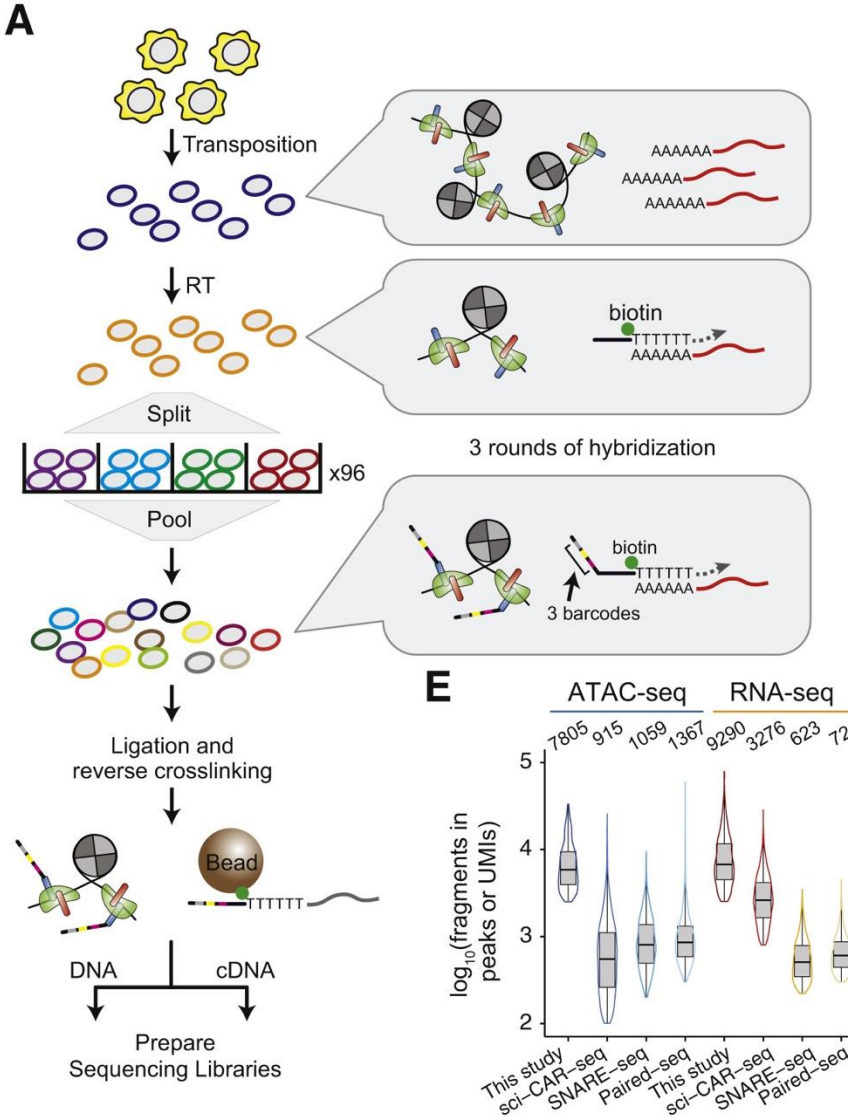
- Paired single-cell multi-omics can be used as bridges to learn feature relationships across modalities

Simultaneous measure of mRNA and chromatin accessibility

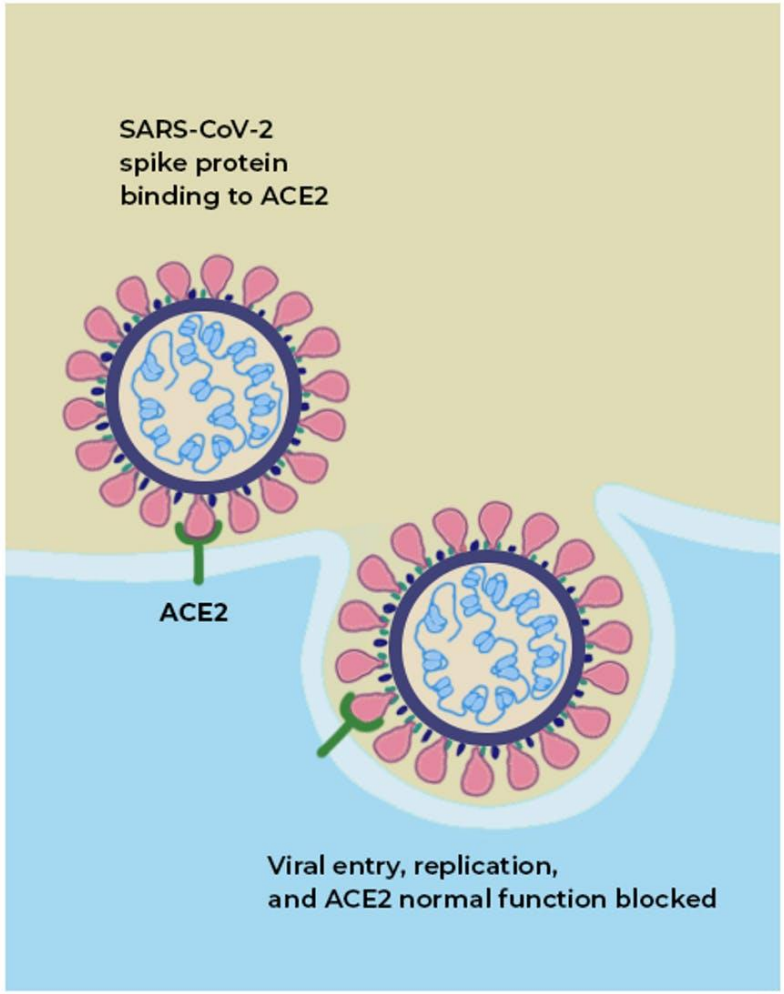
SNARE-seq (Chen et. al., Nature Biotech 2019)



Share-seq (Ma et. al., Cell 2020)

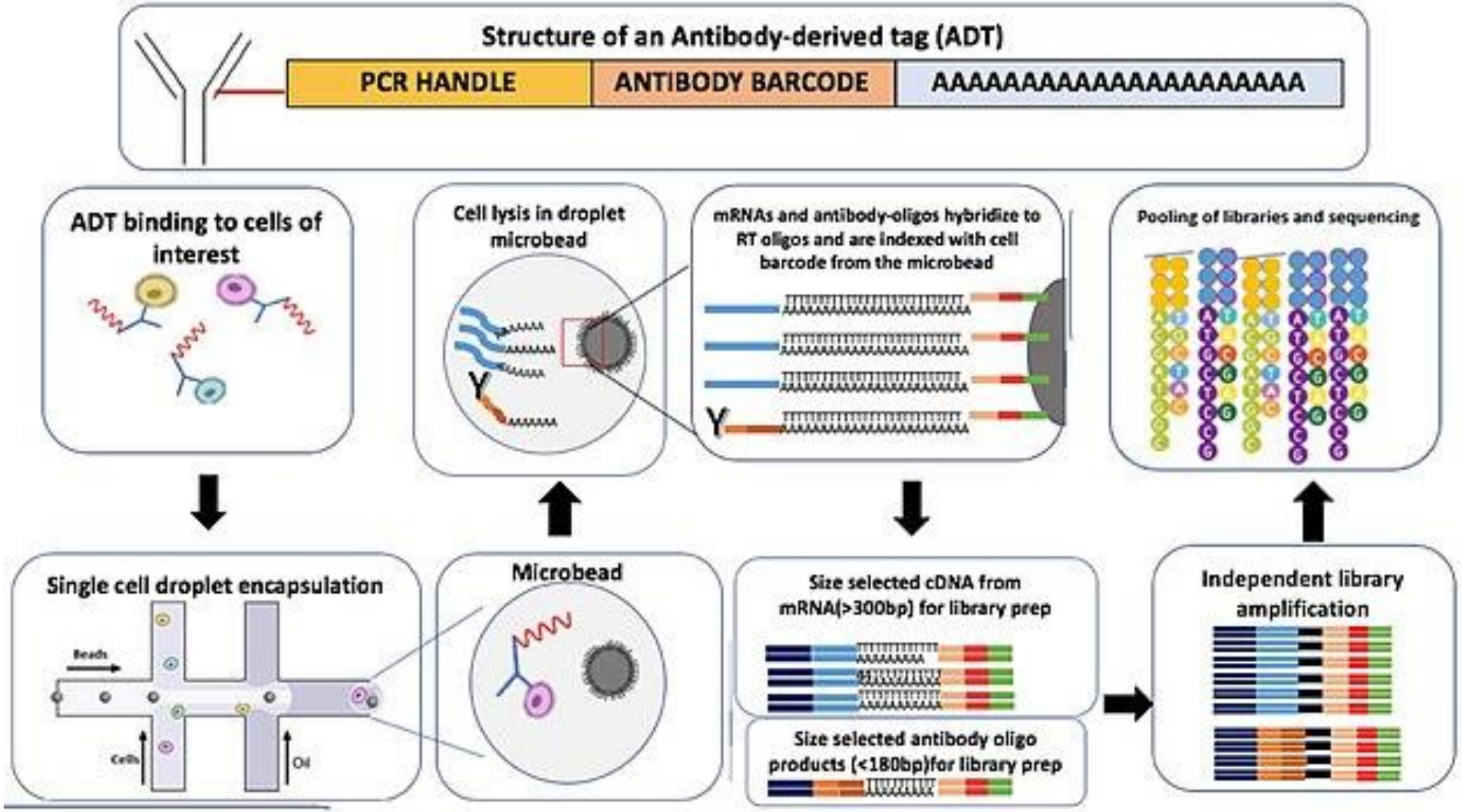


Simultaneous measure of mRNA and surface protein



- Proteins can more reliably indicate cellular activity and function
- Cell surface proteins: play crucial role in effective communication between the cell and its environment
- About 25% to 30% of human genes encode for membrane proteins
- Common technologies: REAP-seq (Peterson et. al., Nature Biotech 2017), CITE-seq (Stoeckius et. al., Nature Methods 2017)

CITE-seq workflow



Integrate paired single cell multi-omics data

- Seurat v4 (Hao et. al. Cell, 2021)

- Core challenge: need to consider multiple sets of features when calculating cell-cell similarity

- Core idea: calculate a weighted NN graph with cell-specific weights
  - Generate KNN graph within each modality
  - Within-modality and cross-modality prediction based on KNN (4 prediction values)
    - Calculate similarity between predicted values and observed values
      - For example:

$$\theta_{rna}\left(r_i, \hat{r}_{i,knn_r}\right) = \exp\left(\frac{-\max\left(d\left(r_i, \hat{r}_{i,knn_r}\right) - d\left(r_i, r_{knn_{r,i,1}}\right), 0\right)}{\sigma_{r,i} - d\left(r_i, r_{knn_{r,i,1}}\right)}\right)$$

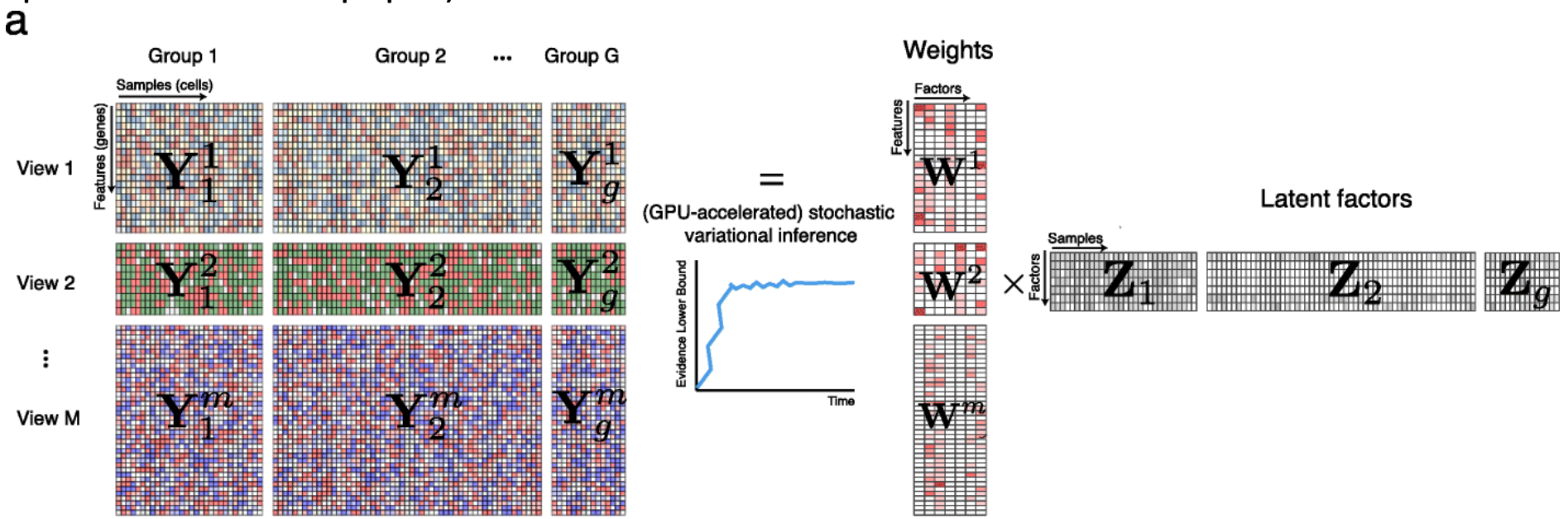
  - Calculated cell-specific modality weights: higher weights on protein if protein neighbors predict better than mRNA neighbors → the neighbors better reflect the molecular state of the cell

$$s_{rna}(i) = \frac{\theta_{rna}\left(r_i, \hat{r}_{i,knn_r}\right)}{\theta_{rna}\left(r_i, \hat{r}_{i,knn_p}\right) + \varepsilon}, \quad s_{protein}(i) = \frac{\theta_{protein}\left(p_i, \hat{p}_{i,knn_p}\right)}{\theta_{protein}\left(p_i, \hat{p}_{i,knn_r}\right) + \varepsilon}$$

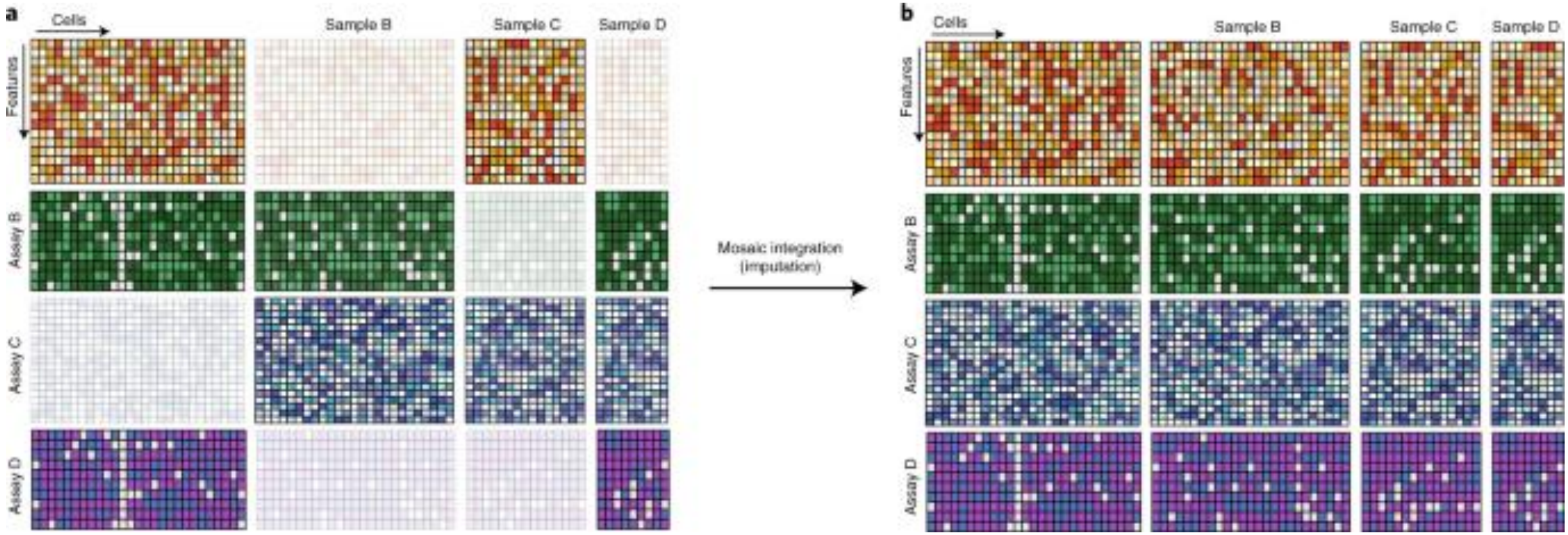
$$w_{rna}(i) = \frac{e^{s_{rna}(i)}}{e^{s_{rna}(i)} + e^{s_{protein}(i)}}, \quad w_{protein}(i) = \frac{e^{s_{protein}(i)}}{e^{s_{rna}(i)} + e^{s_{protein}(i)}}$$

MOFA+ (Argelaguet et. al., Genome Biology 2020)

- Apply Linear factor model on the data
- Apply spike-and-slab prior on both the feature factors and cell factors
  - Result in sparse feature factors and cell factors
  - Very challenging to solve, the authors used stochastic variational inference
  - Can deal with non-Gaussian likelihood, but very slow
- Should be (easy) to allow missing blocks (mosaic data) when performing the factor analysis (not implemented in the paper)

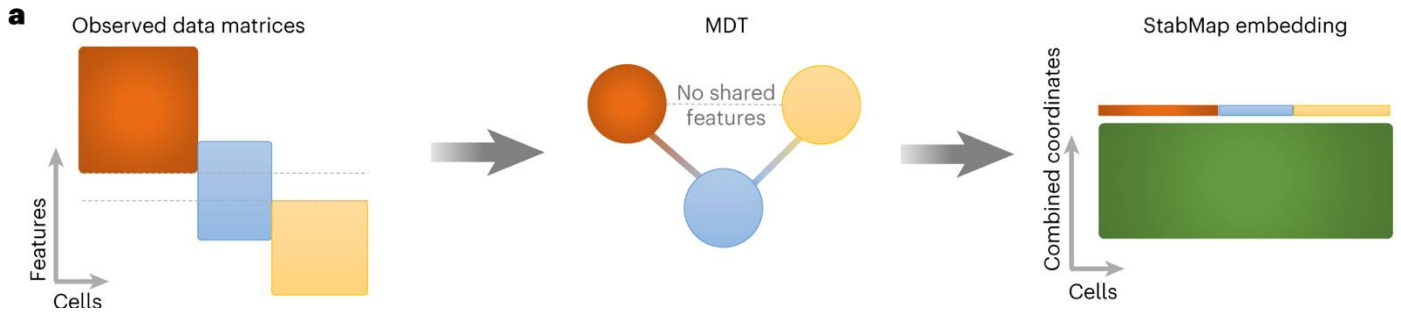


Multi-omics cells as bridges to integrate unpaired data



StabMap (Ghazanfar et. al., Nature Biotech, 2024)

- Essential idea: imputing the missing entries using linear factor analyses
  - Simpler example integrating three datasets, scRNA-seq, scATAC-seq, SNARE-seq



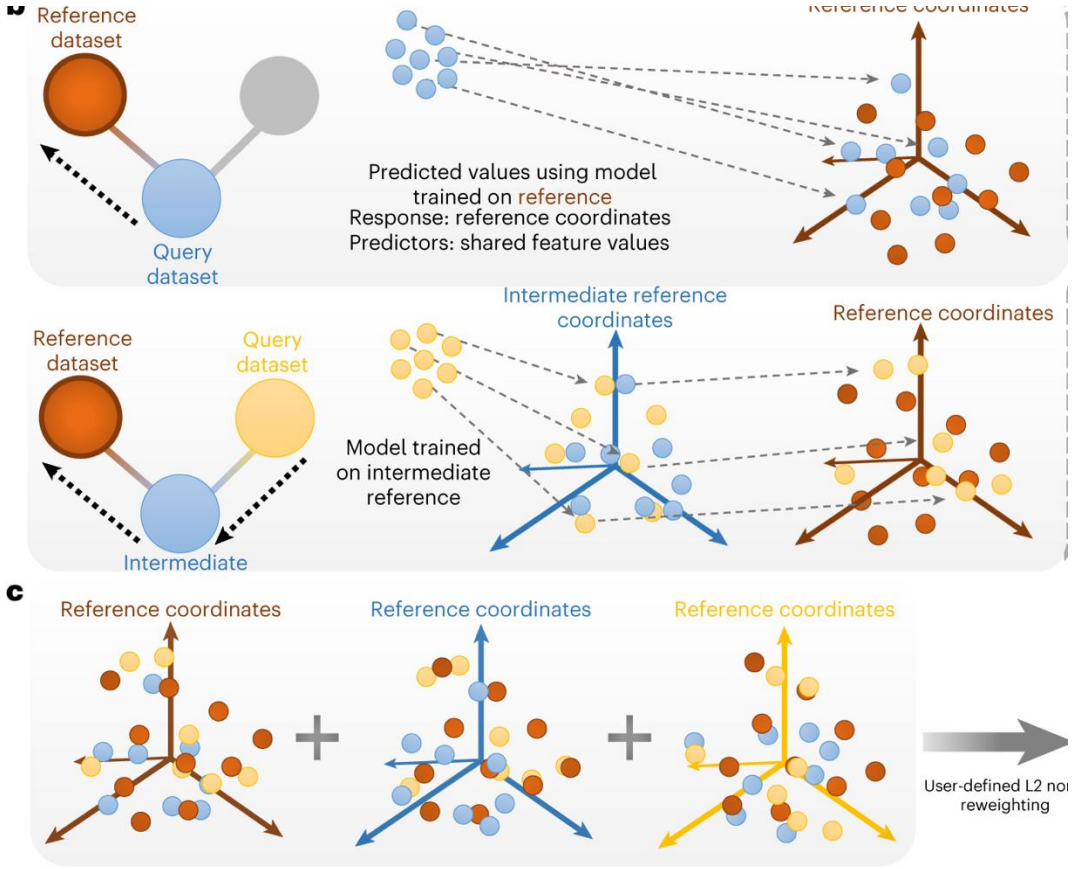
- Core steps:
  - For each reference data r (a reference data can have only one modality), obtain a linear embedding of the cells (use PCA [no cell labels] or LDA if cell labels are given)

$$S_r = D_r^T \times A_r$$

    - Dataset D_r (cell by gene), feature loading (embedding) A_r
  - For dataset i that only overlap part of the features with r,
    - Predict the cell embeddings S_i^r using the linear regression $S_i^r = X_i^T \times A_r$

StabMap (Ghazanfar et. al., Nature Biotech, 2024)

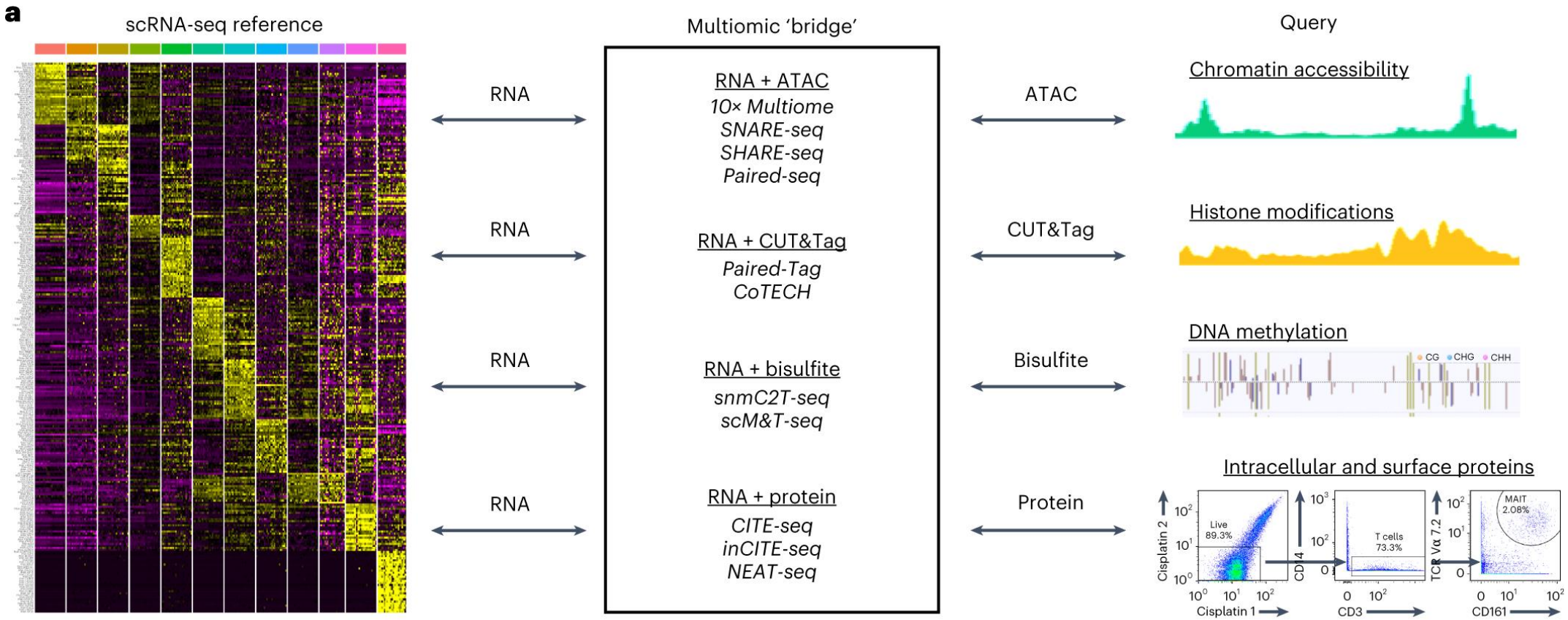
- Core steps:
  - For each reference data r obtain cell embeddings
  - For dataset i that overlap part of the features with r
    - predict the cell embeddings S_i^r using linear regression
  - If dataset i doesn't have overlapping features with r
    - estimate S_i^r iteratively through a sequence of datasets that have overlapping features with each other
  - For each dataset, concatenate all embeddings as the final embedding
  - Can choose various reference datasets and concatenate



- Still need to perform batch correction on the final embedding
- Regression may not be the best way to do factor analysis with missing entries
  - For example, one can directly perform missing value SVD

Seurat v5 (Hao et. al., Nature Biotech, 2024)

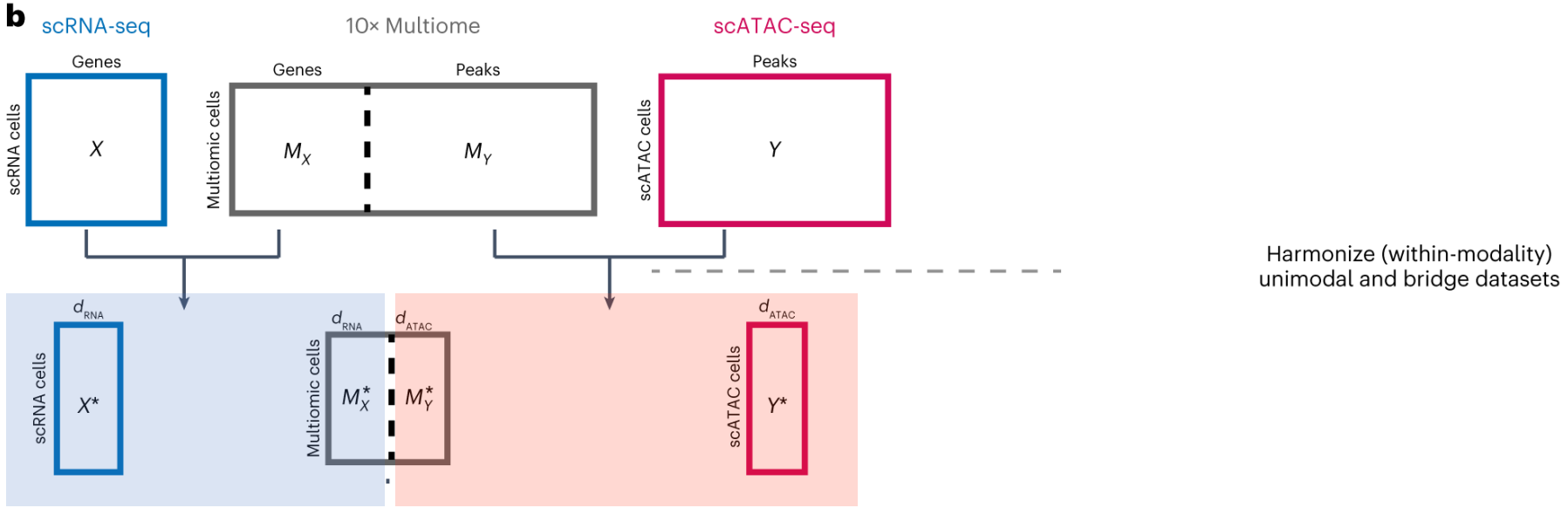
- Build reference using scRNA-seq and map cells of any modality onto a shared latent space



Seurat v5 (Hao et. al., Nature Biotech, 2024)

Core steps:

- Data integration within modality across all datasets (can use various methods for batch correction)



- Only need to integrate low-dimensional space.
- When merging between multiome and unimodal data, can use other modality as supervision in dimension reduction
  - Supervised PCA (sPCA): Construct a cell-cell similarity matrix L using both modalities
    - Find U that maximized the Hilbert-Schmidt Independence Criterion (HSIC):

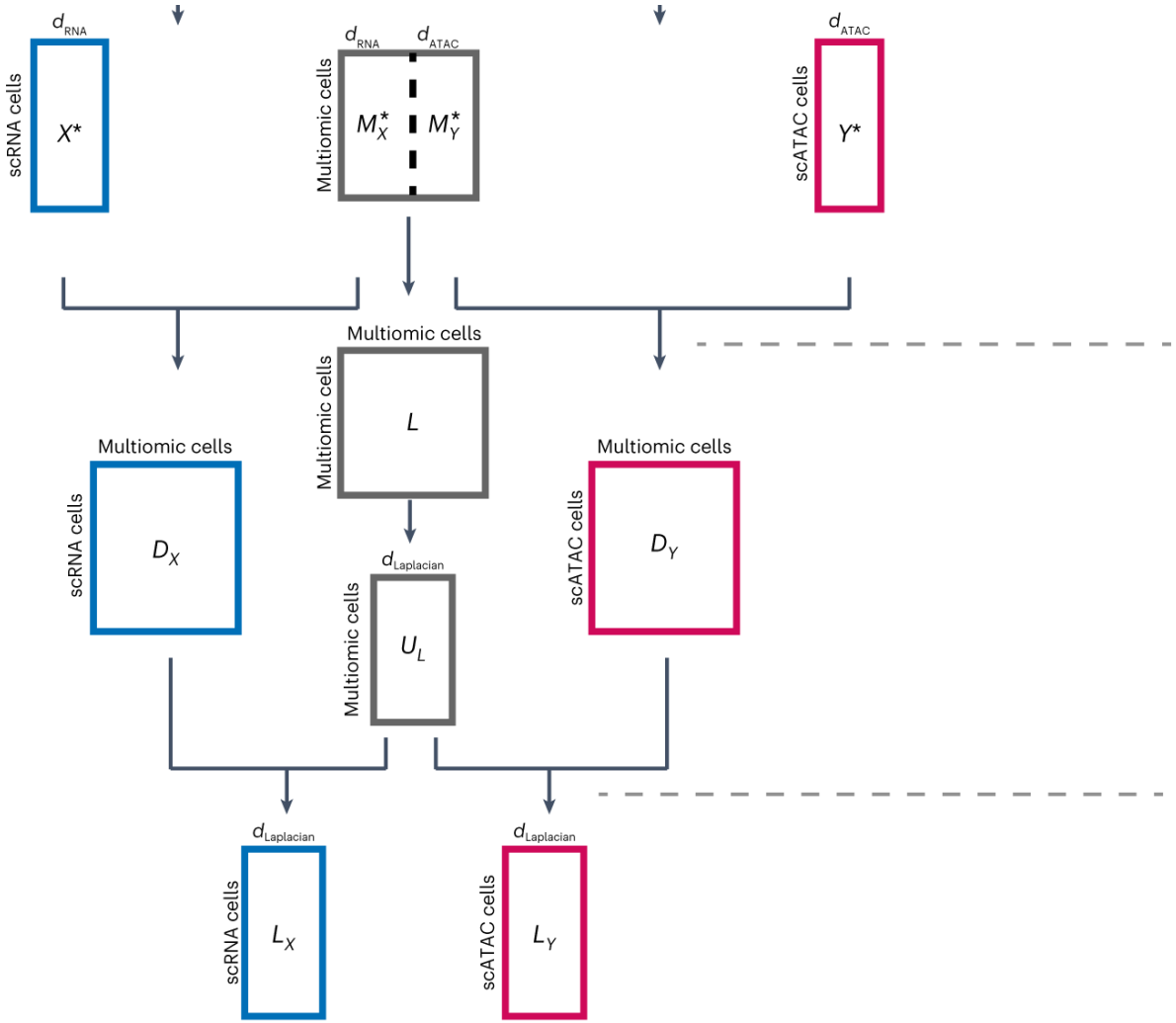
$$HSIC\left(\left(U^T X\right)^T U^T X, L\right)$$

$$= \frac{1}{(n-1)^2} tr\left(X^T U U^T X H L H\right)$$

Seurat v5 (Hao et. al., Nature Biotech, 2024)

Core steps:

- Construct dictionaries for each unimodal dataset



$$\arg\min_{D_X}(||D_X(M_X^*) - X^*||_F^2 + ||D_X||_F^2)$$

$$D_X = X^*(M_X^*)^\dagger$$

- Dimension reduction based on the multiomics data (G as KNN similarity defined based on M^*):

$$L = I - D^{-\frac{1}{2}} G D^{-\frac{1}{2}}$$

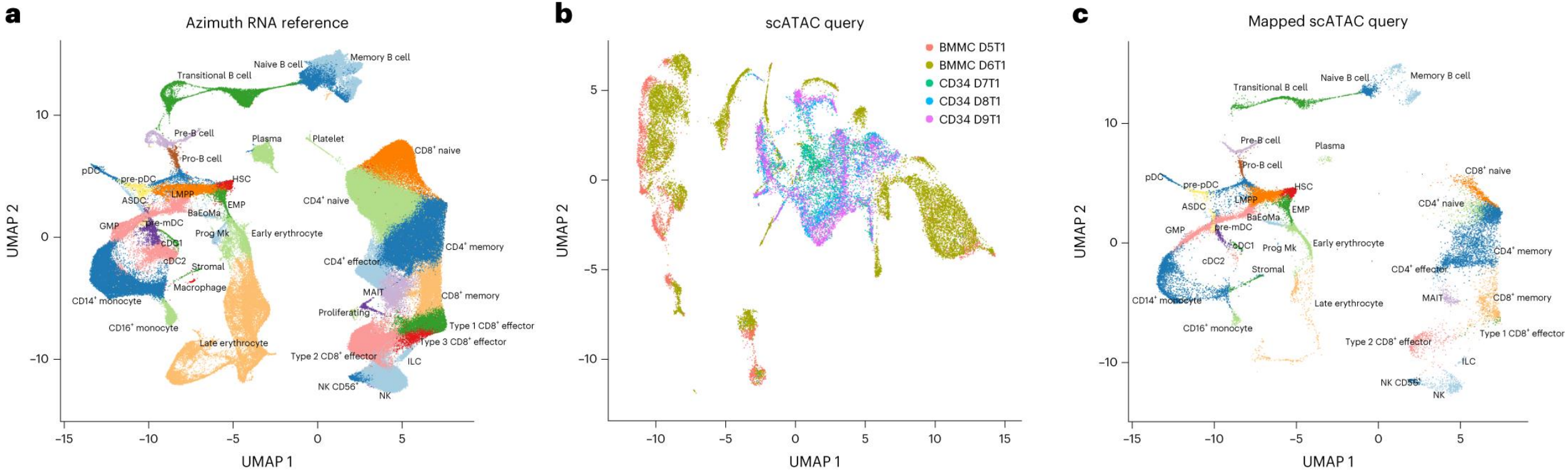
  - Find U_L as the eigenvectors of the k smallest eigenvalues (except 0) of L
  - Map the unimodal data as the weighted average of the multi-omics cells

$$L_X = D_X U_L = X^* \left((M_X^*)^\dagger U_L\right)$$

$$L_Y = D_Y U_L = Y^* \left((M_Y^*)^\dagger U_L\right)$$

- Align the two datasets

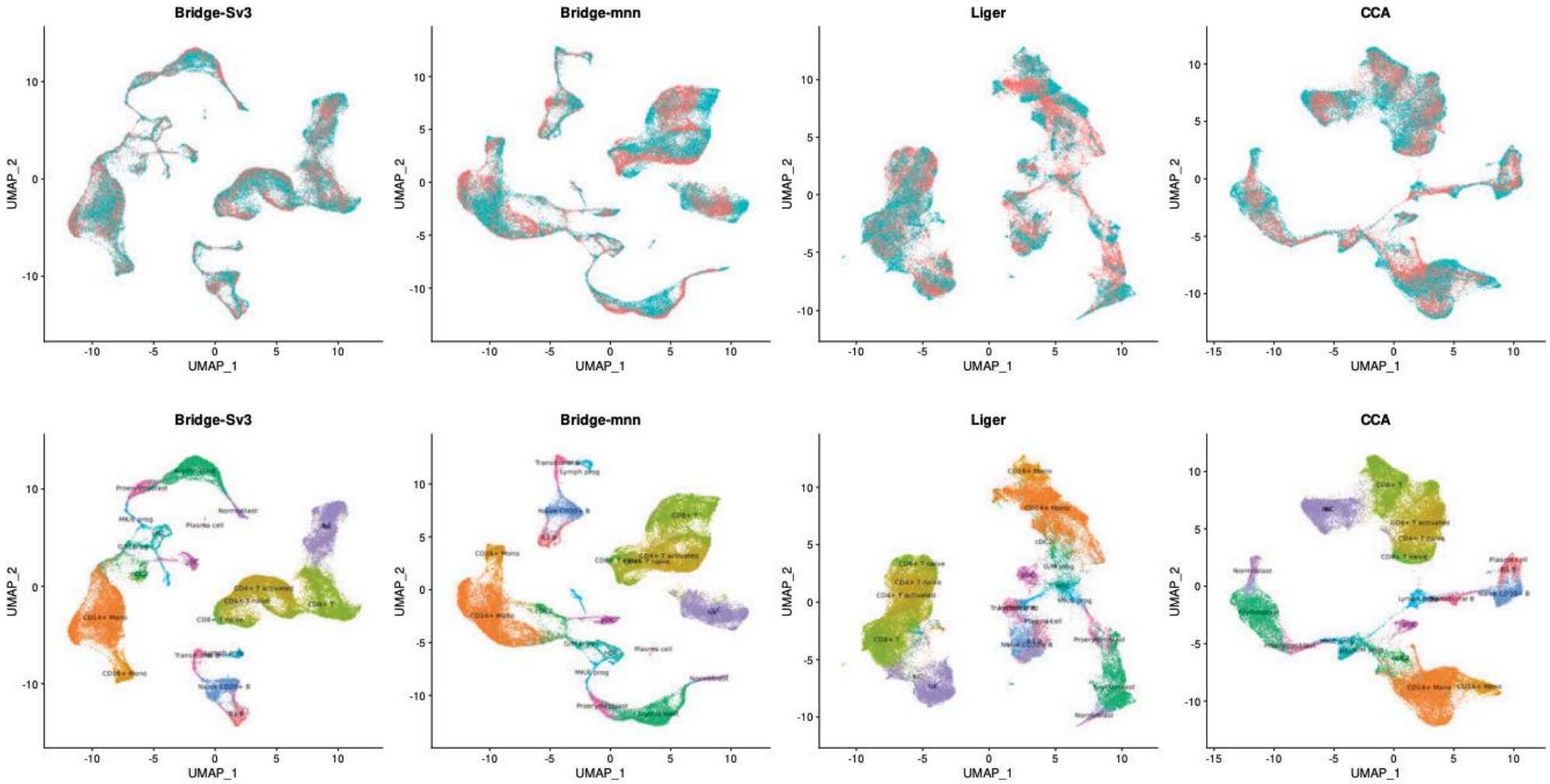
Seurat v5 (Hao et. al., Nature Biotech, 2024)



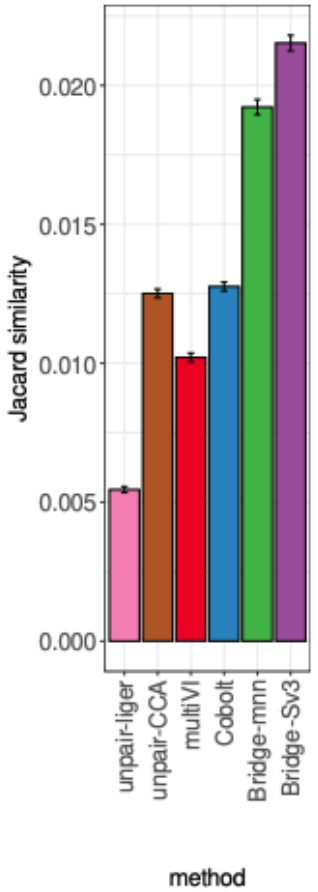
- Comparison with Seurat v3?

Seurat v5 (Hao et. al., Nature Biotech, 2024)

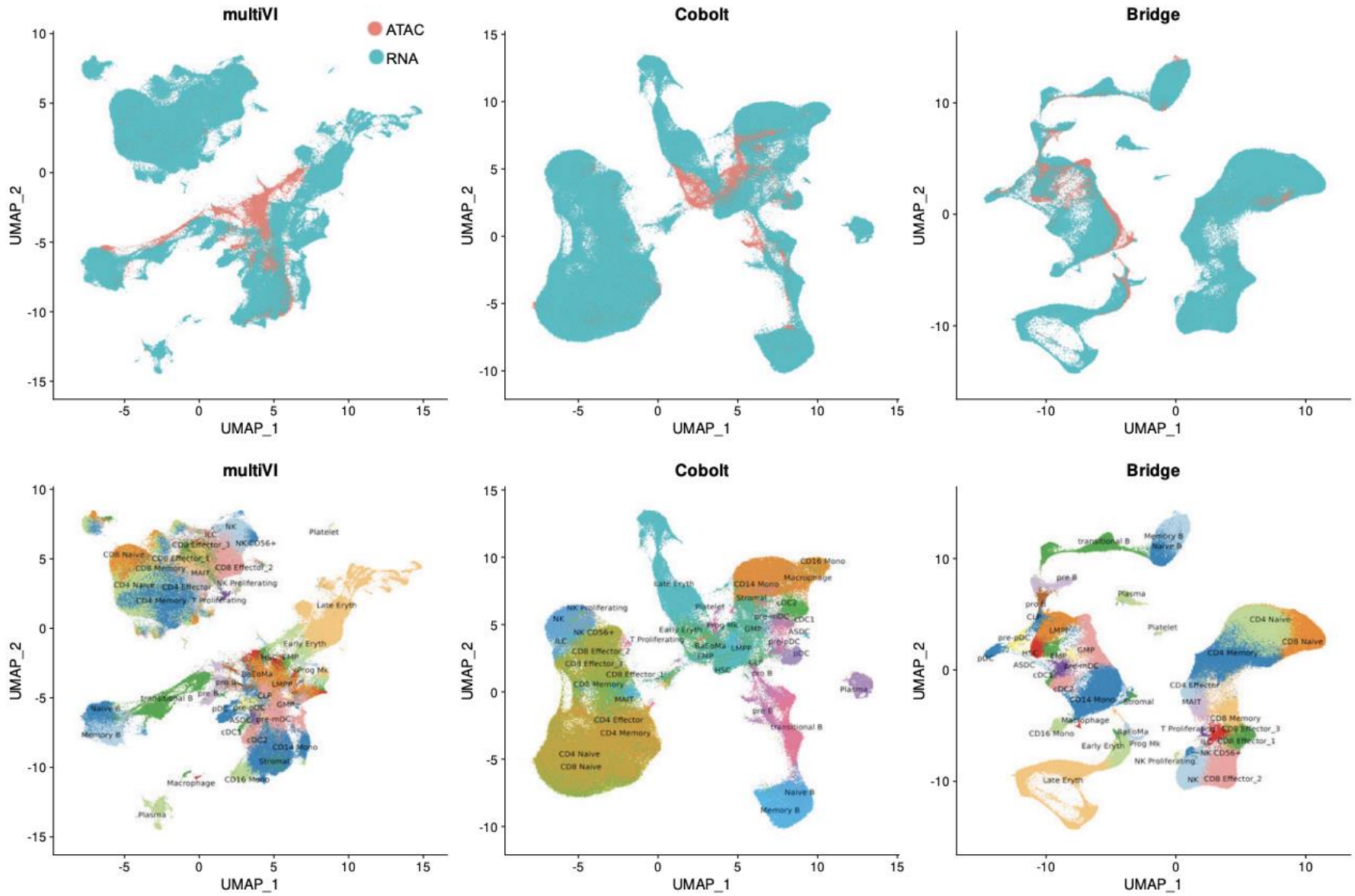
Seurat V3



Matched cell label similarity



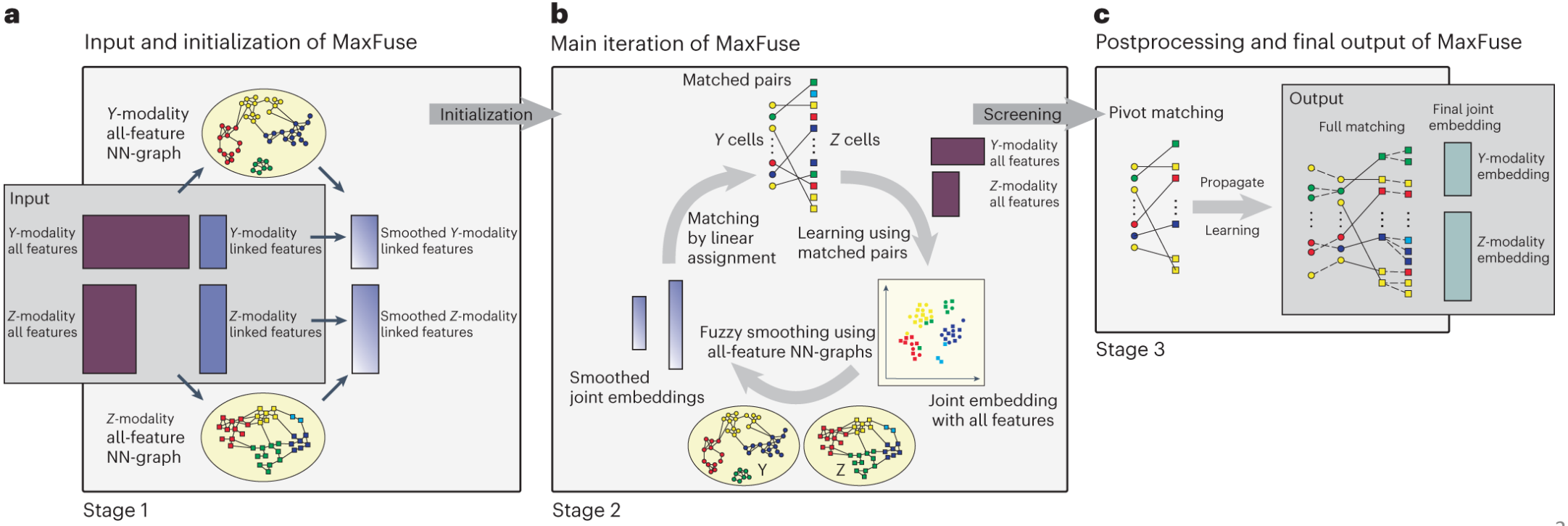
Seurat v5 (Hao et. al., Nature Biotech, 2024)



MaxFuse (Chen et. al., Nature Biotech, 2023)

- Core idea: smooth over similar cells and features to help find cell-cell pairs across modalities

- Inputs:
  - two unpaired single modality datasets
  - A pre-trained feature prediction model projecting both datasets on the same space
  - Noisy projection because the pre-trained model may not be reliable



MaxFuse (Chen et. al., Nature Biotech, 2023)

- Initial smoothing of the projected data
  - Create meta cells within modality by Louvain clustering if data is too sparse
  - Find KNN for within each dataset based on the original feature space
  - (fuzzy) smooth the projected data by similar cells within each modality
    - A weighted average between itself and the smoothed representations

$$\begin{aligned}\mathcal{S}_Y(A;w) &= wA + (1-w)\mathcal{A}_Y(A),\\ \mathcal{S}_Z(B;w) &= wB + (1-w)\mathcal{A}_Z(B).\end{aligned}$$

In this way, we define $\widetilde{Y}^{\circ}_{\mathrm{m}} = \mathcal{S}_Y(Y^{\circ}_{\mathrm{m}}; w_0)$ and $\widetilde{Z}^{\circ}_{\mathrm{m}} = \mathcal{S}_Z(Z^{\circ}_{\mathrm{m}}; w_0)$ with $w_0 \in [0,1]$.

$$\mathcal{A}_Y(Y_{\mathrm{m}}) = K_Y^{-1} G_Y Y_{\mathrm{m}} \text{ and } \mathcal{A}_Y(Y^{\circ}_{\mathrm{m}}) = K_Y^{-1} G_Y Y^{\circ}_{\mathrm{m}}$$

- G_Y: sparse Similarity matrix (KNN connectivity), $K_Y = \mathrm{diag}(k_1^Y, \ldots, k_n^Y)$: number of nearest neighbors

MaxFuse (Chen et. al., Nature Biotech, 2023)

- Initial smoothing of the projected data

- Find initial matched pairs by optimal matching
  - D^0: Euclidean distance between two cells cross modalities based on projected data

$$\begin{aligned} \text{minimize} \quad & \langle \Pi, D^{\circ} \rangle \\ \text{subject to} \quad & \Pi \in \{0,1\}^{n_y \times n_z} \\ & \sum_i \Pi_{ij} \le 1, \forall j, \quad \sum_j \Pi_{ij} \le 1, \forall i, \\ & \sum_{i,j} \Pi_{ij} = n_{\min}. \end{aligned}$$

- Given the matched pairs of cells, perform CCA of two datasets in the original feature space
  - CCA for the features instead of cells in Seurat
  - Perform PCA first within each dataset to reduce dimension
  - Obtain a new joint embedding of all cells from CCA

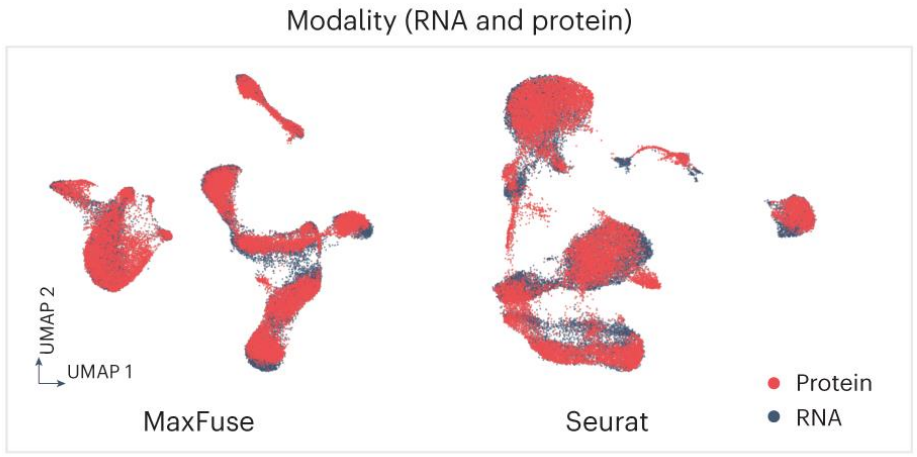
$$Y_{\mathtt{m}}^{\mathtt{cc}} = Y_{\mathtt{m}}^{\mathtt{r}} \widehat{C}_y \in \mathbb{R}^{n_y \times r_{\mathtt{cc}}} \text{ and } Z_{\mathtt{m}}^{\mathtt{cc}} = Z_{\mathtt{m}}^{\mathtt{r}} \widehat{C}_z \in \mathbb{R}^{n_z \times r_{\mathtt{cc}}}$$

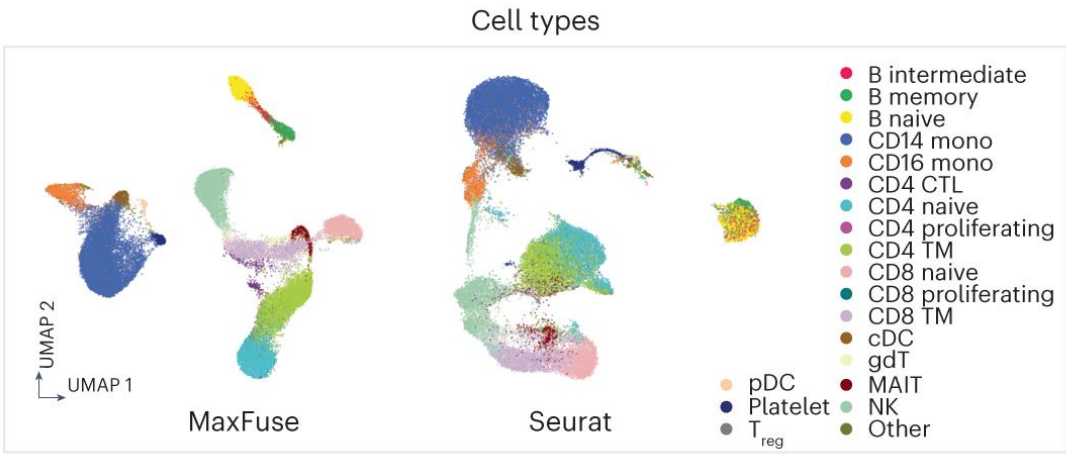
- Iterative refinement
  - Compute joint mapping via CCA using matched pairs of cells
  - (fuzzy) smoothing over similar cells
  - Apply optimal matching to find matched pairs of cells

- Similar to Seurat, only using a subset of pairs of cells as the anchor (pivot) pairs

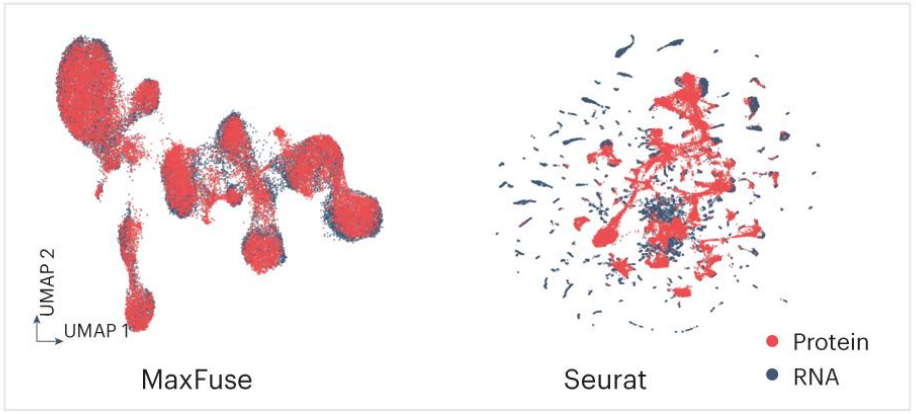
MaxFuse (Chen et. al., Nature Biotech, 2023)

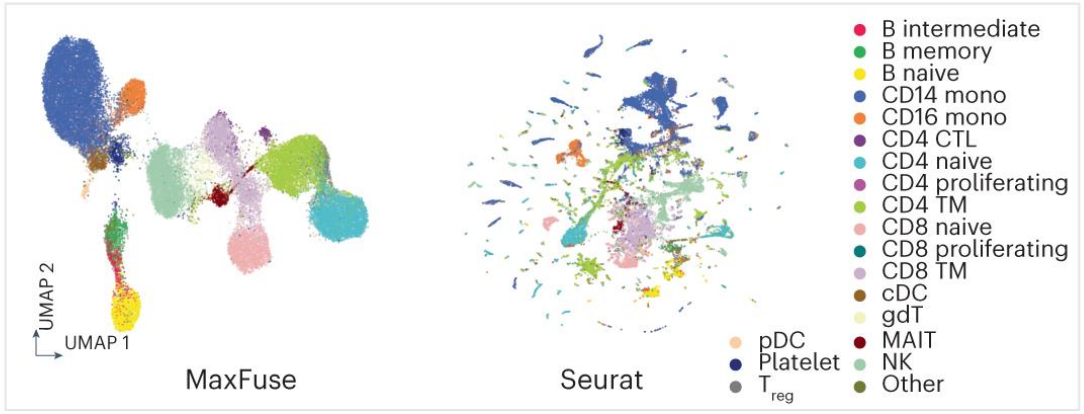
Using all 228 antibodies





Using 30 antibodies





Related papers

- Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., ... & Satija, R. (2019). Comprehensive integration of single-cell data. *cell*, *177*(7), 1888-1902.
- Lin, Y., Wu, T. Y., Wan, S., Yang, J. Y., Wong, W. H., & Wang, Y. R. (2022). scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. *Nature biotechnology*, *40*(5), 703-710.
- Liu, J., Gao, C., Sodicoff, J., Kozareva, V., Macosko, E. Z., & Welch, J. D. (2020). Jointly defining cell types from multiple single-cell datasets using LIGER. *Nature protocols*, *15*(11), 3632-3662.
- Duren, Z., Chen, X., Zamanighomi, M., Zeng, W., Satpathy, A. T., Chang, H. Y., ... & Wong, W. H. (2018). Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proceedings of the National Academy of Sciences*, *115*(30), 7723-7728.
- Cao, Z. J., & Gao, G. (2022). Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, *40*(10), 1458-1466.
- Chen, S., Lake, B. B., & Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, *37*(12), 1452-1457.
- Ma, S., Zhang, B., LaFave, L. M., Earl, A. S., Chiang, Z., Hu, Y., ... & Buenrostro, J. D. (2020). Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell*, *183*(4), 1103-1116.
- Peterson, V. M., Zhang, K. X., Kumar, N., Wong, J., Li, L., Wilson, D. C., ... & Klappenbach, J. A. (2017). Multiplexed quantification of proteins and transcripts in single cells. *Nature biotechnology*, *35*(10), 936-939.
- Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., ... & Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, *14*(9), 865-868.
- Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., ... & Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, *184*(13), 3573-3587.
- Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., & Stegle, O. (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*, *21*, 1-17.
- Ghazanfar, S., Guibentif, C., & Marioni, J. C. (2024). Stabilized mosaic single-cell data integration using unshared features. *Nature Biotechnology*, *42*(2), 284-292.
- Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., ... & Satija, R. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature biotechnology*, *42*(2), 293-304.
- Chen, S., Zhu, B., Huang, S., Hickey, J. W., Lin, K. Z., Snyder, M., ... & Ma, Z. (2023). Integration of spatial and single-cell data across modalities with weakly linked features. *Nature Biotechnology*, 1-11.