multiHIVE: Hierarchical Multimodal Deep Generative Modeling for Single-cell Multiomics

Anirudh Nanduri[1,†], Musale Krushna Pavan[1,†], Kushagra Pandey[2], and Hamim Zafar[1,3,4,*]

[1]Department of Computer Science and Engineering, Indian Institute of Technology Kanpur, India
[2]Department of Computer Science, University of California, Irvine
[3]Department of Biological Sciences and Bioengineering, Indian Institute of Technology Kanpur, India
[4]Mehta Family Center for Engineering in Medicine, Indian Institute of Technology Kanpur, India
[†]*Authors contributed equally*
[*]*Corresponding Author:* Hamim Zafar (hamim@iitk.ac.in)

April 16, 2026

Contents

1 Supplementary Notes

Supplementary Note 1: Details of multiHIVE model

Overview of multiHIVE Model

Figure 1 provides an overview of the multiHIVE model, which is designed to integrate features from distinct modalities into a unified latent space. The model architecture consists of two primary components. The first component employs a hierarchical variational autoencoder (h-VAE) with two hierarchically stacked latent variables [1]. This h-VAE processes the combined expression data from all modalities, embedding this information into a lower-dimensional latent space that captures characteristics across all modalities in two distinct layers of representation $\boldsymbol{z^{s_1}}, \boldsymbol{z^{s_2}}$. The second component of multiHIVE features private encoders, each encoder dedicated to one modality, embedding the modality into their respective private latent spaces of RNA $\boldsymbol{z^r}$, chromatin $\boldsymbol{z^a}$, and protein $\boldsymbol{z^p}$ [2]. The original expression data for each modality is reconstructed using the respective decoders for each modality. Each decoder utilises a concatenated private representation of that modality along with the shared latent space as input. For example RNA specific decoder takes $[\boldsymbol{z^{s_1}} \| \boldsymbol{z^r}]$. These decoders output the parameters of the respective data distributions. Batch information S, such as experimental batch or donor from each dataset, is incorporated into all encoders and decoders, ensuring the model accounts for batch effects. The decoders are shared between both hierarchical and non-hierarchical VAEs.

multiHIVE models the RNA, protein and chromatin data with the appropriate distributions such that they take in contemplation of the technicalities of the data. RNA data is modelled by the Negative Binomial distribution, effectively considering the dropouts. protein data is modelled using a Negative Binomial Mixture consisting of background and foreground intensities of proteins. [3]. Chromatin data is modelled using a Bernoulli distribution as the counts are mostly binary, depicting whether a region is accessible or not. Along with cell and region-specific scaling factors. multiHIVE is trained using the Evidence Lower Bound (ELBO) loss function (ref methods), incorporating all the latent variables along with a guided prior from the hierarchical model. This hierarchical latent space structure encourages a close relationship between the variational posteriors and the generative segments of the decoders. This inductive bias aids in the latent representation, enabling it to capture information from different modalities. The inclusion of private latent variables ensures that the modality-specific information is preserved in the final latent representation $[\boldsymbol{z^{s_1}} \| \boldsymbol{z^r} \| \boldsymbol{z^p} \| \boldsymbol{z^a}]$, which is the combination of private and shared latent representations used for downstream tasks such as clustering and visualisation.

The multiHIVE architecture addresses several challenges in multimodal data integration tasks, as outlined in our analysis. First, multiHIVE facilitates the integration of multiple multimodal datasets. Second, the decoder outputs provide an imputed expression space, also predicting the missing modality. Third, the different levels of latent embeddings are useful for capturing and analyzing various biological processes.

Generative Process for RNA Modality

$$\boldsymbol{z_n^{s_2}} \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{I}) \quad \text{(shared latent at level 2)} \tag{1}$$

$$\boldsymbol{z_n^{s_1}} = f_{l1}(\boldsymbol{z_n^{s_2}}) \quad \text{(derived shared latent at level 1)} \tag{2}$$

$$\boldsymbol{z_n^{r}} \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{I}) \quad \text{(RNA-specific latent variable)} \tag{3}$$

$$\boldsymbol{\rho_n} = f_\rho(\boldsymbol{z_n^{s_1}}, \boldsymbol{z_n^{r}}, s_n) \quad \text{(normalized gene expression)} \tag{4}$$

$$\ell_n \sim \text{LogNormal}(\boldsymbol{\ell_\mu^T} s_n, \boldsymbol{\ell_{\sigma^2}^T} s_n) \quad \text{(library size per cell)} \tag{5}$$

$$w_{ng} \sim \text{Gamma}(\ell_n \rho_n^g, \theta_g) \quad \text{(latent count rate)} \tag{6}$$

$$x_{ng} \sim \text{Poisson}(w_{ng}) \quad \text{(observed RNA count)} \tag{7}$$

Integrating out auxiliary variable (w_{ng}) from the Gamma–Poisson mixture yields the following negative binomial distribution for RNA counts:

$$x_{ng} \sim \text{NegativeBinomial}(\ell_n \rho_n^g, \theta_g) \tag{8}$$

Here, $\boldsymbol{z_n^{s_1}}$ and $\boldsymbol{z_n^{r}}$ capture shared and RNA-specific latent representations, f_{l1} and f_ρ are neural networks, $\boldsymbol{\ell_\mu^T}$ and $\boldsymbol{\ell_{\sigma^2}^T}$ define the batch-wise library size distribution empirically estimated from the data, and θ_g is the inverse dispersion parameter learnt during inference. The gamma distribution in the generative process is parameterized by $\ell_n \rho_n^g$ (mean) and θ_g (shape parameter), where ℓ_n acts as a library

size scaling factor, and ρ_n^g denotes the normalized gene expression ($\sum_g \rho_n^g = 1$). The RNA decoder f_ρ maps the latent variables to the normalized expression space.

For $\boldsymbol{z_n^{s_2}}$ and $\boldsymbol{z_n^r}$, the priors are chosen as isotropic normal distributions. The shared latent representation at level 1, $\boldsymbol{z_n^{s_1}}$, is constructed from $\boldsymbol{z_n^{s_2}}$ via the neural network f_{l1}. The shared latent variables $\boldsymbol{z_n^{s_1}}$ and $\boldsymbol{z_n^{s_2}}$ are common across all the modalities for a cell n.

Generative Process for Protein Modality

$$\boldsymbol{z_n^p} \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{I}) \quad \text{(protein-specific latent variable)} \tag{9}$$

$$\boldsymbol{\pi_n} = f_\pi(\boldsymbol{z_n^p}, \boldsymbol{z_n^{s_1}}, s_n) \quad \text{(probability of background expression)} \tag{10}$$

$$\boldsymbol{\alpha_n} = f_\alpha(\boldsymbol{z_n^p}, \boldsymbol{z_n^{s_1}}, s_n) \quad \text{(foreground/background intensity ratio)} \tag{11}$$

$$\beta_{nh} \sim \text{LogNormal}(\boldsymbol{c_h^T} s_n, \boldsymbol{d_h^T} s_n) \quad \text{(background intensity per batch)} \tag{12}$$

$$v_{nh} \mid \boldsymbol{z_n^p}, \boldsymbol{z_n^{s_1}}, s_n \sim \text{Bernoulli}(\pi_{nh}) \quad \text{(indicator for background vs foreground)} \tag{13}$$

$$r_{nh} \mid \boldsymbol{z_n^p}, \boldsymbol{z_n^{s_1}}, v_{nh}, \beta_{nh}, s_n \sim \text{Gamma}(v_{nh}\beta_{nh} + (1 - v_{nh})\beta_{nh}\alpha_{nh}, \phi_h) \tag{14}$$

$$y_{nh} \mid r_{nh} \sim \text{Poisson}(r_{nh}) \quad \text{(observed protein count)} \tag{15}$$

Integrating out r_{nh} gives

$$y_{nh} \sim \text{NegativeBinomial}(v_{nh}\beta_{nh} + (1 - v_{nh})\beta_{nh}\alpha_{nh}, \phi_h) \tag{16}$$

$$y_{nh} \sim \text{NegativeBinomialMixture}(\beta_{nh}, \beta_{nh}\alpha_{nh}, \pi_{nh}, \phi_h) \tag{17}$$

Here, $\boldsymbol{z_n^{s_1}}$ is the shared latent representation at level 1 (ref RNA generative process) of h-VAE, and $\boldsymbol{z_n^p}$ captures protein-specific variation. f_π and f_α are neural networks producing the mixture weight parameter $\boldsymbol{\pi_n}$ and foreground/background intensity ratio $\boldsymbol{\alpha_n}$, respectively. β_{nh} models baseline background expression with parameters $\boldsymbol{c_h}$ and $\boldsymbol{d_h}$ learned per protein during inference. v_{nh} determines whether a count arises from background or foreground, and ϕ_h is the inverse dispersion parameter of the negative binomial distribution in the generative process learnt during inference. This formulation models protein counts as a mixture of background and foreground signals following [3].

The components of the protein mixture (background β_{nh}, foreground $\beta_{nh}\alpha_{nh}$) as well as the mixture weights v_{nh} are obtained based on the shared $\boldsymbol{z_n^{s_1}}$ and protein-specific latent representations $\boldsymbol{z_n^p}$ given the batch s_n.

Generative Process for Chromatin Accessibility Modality

$$\boldsymbol{z_n^a} \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{I}) \quad \text{(accessibility-specific latent variable)} \tag{18}$$

$$d_n = f_d(\boldsymbol{c}_n) \quad \text{(cell-specific scaling from chromatin profile)} \tag{19}$$

$$b_{ni} = f_c(\boldsymbol{z_n^{s_1}}, \boldsymbol{z_n^a}, s_n) \quad \text{(probability of accessibility)} \tag{20}$$

$$c_{ni} \sim \text{Bernoulli}(b_{ni}.d_n.r_i) \quad \text{(binary accessibility observation)} \tag{21}$$

Here, $\boldsymbol{z_n^a}$ is a latent variable for accessibility-specific modality sampled from a standard normal prior. d_n denotes the cell-specific scaling factor computed from the observed chromatin profile $\boldsymbol{c}_n$ using neural network f_d. The neural network f_c maps the shared $\boldsymbol{z_n^{s_1}}$, and accessibility specific latent representations $\boldsymbol{z_n^a}$ given the batch s_n to b_{ni} the estimated probability of accessibility. c_{ni}, the observed chromatin accessibility is sampled from the Bernoulli distribution parametrized as a product $b_{ni}.d_n.r_i$. $\boldsymbol{z_n^{s_1}}$ represents the shared latent embedding across modalities, r_i denotes the region-specific scaling factor, which is learnt during inference and is specific to each region. The Bernoulli variable $c_{ni} \in \{0, 1\}$ indicates whether region i is accessible in cell n. This formulation follows [4], where accessibility observations are modeled as probabilistic binary events governed by both shared and modality-specific latent variables.

Supplementary Note 2: Evaluation metrics

To evaluate the performance of multimodal dataset integration, we employed a combination of biological conservation and batch effect correction metrics [5].

Biological Conservation Metrics

Biological conservation was quantified by comparing clustering structure in the latent space $[\boldsymbol{z}^{\boldsymbol{s_1}} \| \boldsymbol{z}^{\boldsymbol{r}} \| \boldsymbol{z}^{\boldsymbol{p}} \| \boldsymbol{z}^{\boldsymbol{a}}]$ against ground-truth cell type annotations. Leiden clustering was applied across 20 resolutions from 0.1 to 2, and for each metric, the median value across resolutions was reported.

- **Normalized Mutual Information (NMI).** This measures agreement between predicted clusters K and true labels C:

$$\mathrm{NMI}(C, K) = \frac{2I(C;K)}{H(C) + H(K)}, \tag{22}$$

  where $I(C;K)$ denotes mutual information and $H(\cdot)$ denotes entropy. NMI ranges from 0 to 1 (best).

- **Adjusted Rand Index (ARI).** This evaluates clustering similarity while correcting for chance:

$$\mathrm{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}, \tag{23}$$

  where n_{ij} denotes the overlap between true cluster i and predicted cluster j, with $a_i = \sum_j n_{ij}$ and $b_j = \sum_i n_{ij}$. ARI ranges from -1 to 1 (best).

- **Fowlkes–Mallows Index (FMI).** FMI captures the geometric mean of precision and recall at the pairwise level:

$$\mathrm{FMI} = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}, \tag{24}$$

  where TP, FP, and FN denote true positives, false positives, and false negatives, respectively, for every pair of cells. FMI ranges from 0 to 1 (best).

Batch Effect Correction Metrics

Batch correction performance was assessed by quantifying batch mixing in the latent representation.

- **Average Silhouette Width–Batch (ASW-batch).** For each cell i, the silhouette coefficient is defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \qquad \text{ASW-batch} = \frac{1}{n}\sum_{i=1}^{n} s(i), \tag{25}$$

  where $a(i)$ is the mean distance to cells in the same batch and $b(i)$ the minimum mean distance to cells in other batches. Values range from -1 to 1 (best).

- **Graph Connectivity (GC).** GC evaluates whether cells of the same identity form connected components in the kNN graph. For each cell type c, a subgraph $G(N_c, E_c)$ was constructed and GC computed as

$$\mathrm{GC} = \frac{1}{|C|}\sum_{c \in C} \frac{|\mathrm{LCC}(G(N_c, E_c))|}{|N_c|}, \tag{26}$$

  where $\mathrm{LCC}(\cdot)$ denotes the largest connected component. GC ranges in $(0, 1]$.

Protein Imputation Evaluation Metrics

Protein imputation accuracy was evaluated using metrics capturing both correlation and absolute error.

- **Pearson Correlation Coefficient (PCC).** For each protein,

$$\mathrm{PCC} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{27}$$

  where x_i and y_i denote imputed and ground-truth expression, respectively. PCC ranges from -1 to 1.

- **Root Mean Square Error (RMSE).** RMSE measures the magnitude of imputation error:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}, \tag{28}$$

where n is the number of cells. Lower RMSE indicates better performance.

Supplementary Note 3: Preprocessing of Multi-Omics Data

The quality control of RNA-seq datasets is performed by filtering low-quality cells and genes based on the unique molecular identifier (UMI) counts and mitochondrial gene expressions. Subsequently, UMI counts were normalized using the normalize_total method with a scaling factor of 10,000. From the normalized data, the top 4,000 highly variable genes (HVGs) were selected using the Seurat v4 HVG selection method. The Normalized UMI counts, along with the top 4,000 genes, are used as input to the model. For the protein modality, antibody-derived tag (ADT) counts were processed by filtering out low-quality proteins with minimal expression. The filtered ADT counts were then used as protein input to the model. For chromatin accessibility data (ATAC), we retained regions expressed in at least 1% of cells and binarized the resulting count matrix to represent accessibility status. The final input to the multiHIVE model consisted of the normalized RNA UMI matrix (restricted to 4,000 HVGs), filtered ADT protein counts, and binarized ATAC matrix

Supplementary Note 4: Downstream tasks

For downstream tasks including clustering, visualization, and trajectory inference, we use a joint embedding, which is obtained by concatenating private and shared latent representations $[\boldsymbol{z^{s_1}} \| \boldsymbol{z^r} \| \boldsymbol{z^p} \| \boldsymbol{z^a}]$.

Trajectory inference

We used MARGARET [6] for trajectory inference based on the joint embedding of RNA and protein inferred by multiHIVE. We ran MARGARET with default parameters except for the "obsm_data_key" parameter which was set to the combined embedding $[\boldsymbol{z^{s_1}} \| \boldsymbol{z^r} \| \boldsymbol{z^p}]$ inferred by multiHIVE and "metric_clusters" parameter was set to the cell type annotation. The "compute_pseudo_time()" function was used to infer the pseudotime of cells. To infer the lineage trend of marker genes along the pseudotime, we used the $\boldsymbol{z^{s_1}}$ embeddings for reconstructing RNA expression which were then fed to the "plot_lineage_trends()" function of MARGARET.

Inference of gene expression programs

To investigate the expression programs captured by each of the lower-dimensional embeddings inferred by multiHIVE, we first trained the multiHIVE model to obtain the lower-dimensional embeddings $\boldsymbol{z_n^{s_1}}$ and $\boldsymbol{z_n^{s_2}}$. These embeddings are then mapped to the RNA expression space using the trained multiHIVE's RNA decoder network, producing the reconstructed expression matrices $\hat{X}_1$ and $\hat{X}_2$ corresponding to $\boldsymbol{z_n^{s_1}}$ and $\boldsymbol{z_n^{s_2}}$, respectively. Consensus non-negative matrix factorization (cNMF) [7] was then used to infer the gene expression programs (GEPs) from the RNA expression matrix reconstructed by multiHIVE. cNMF models the $N \times G$ gene expression matrix as a product of two lower-rank matrices, program usage per cell matrix ($N \times K$) and a gene expression program matrix ($K \times G$), where K is the number of GEPs. The first matrix gives information about the usage of each program in the cell. The second matrix models the contribution of each gene to the program. The value of K was determined based on stability and error criteria as suggested by the authors of cNMF [7].

Pathway analysis

We used fast gene set enrichment analysis (fGSEA) [8] for evaluating the pathways associated with a GEP. The rows of the gene expression program matrix inferred by cNMF gives us a list of genes corresponding to a GEP where the genes are ranked according to their contributions (weights) to the GEP. The ranked list of genes is used as input to fGSEA. In addition, we used the Molecular Signatures Database (MSigDB)[9] to extract the GOBP, REACTOME, and HALLMARK (`https://www.gsea-msigdb.org/gsea/msigdb/human/collections.jsp`) related pathways (gene sets) and used them as input to fGSEA.

2 Supplementary Tables

Supplementary Table 1: Encoder architecture summary. BN: BatchNorm; LN: LayerNorm. G: Number of Genes, H: Number of Proteins, I: Number of accessible regions, S: number of batches

Module	**Layers**	**Details**
Shared encoder (r_1)	[G+H+I+S] $\rightarrow$ 128 $\rightarrow$ 128	BN, ReLU, Dropout(0.2)
Latent Δ_1	128 $\rightarrow$ (20)*2	Mean & log-variance heads
Shared encoder (r_2)	128 $\rightarrow$ 128 $\rightarrow$ 128	BN, ReLU, Dropout(0.2)
Latent z_n^{s2}	128 $\rightarrow$ (20)*2	Mean & log-variance heads
Latent z_n^{s1}	20 $\rightarrow$ 128 $\rightarrow$ (20)*2	BN + ReLU + Dropout(0.2)
Gene encoder	[G+S] $\rightarrow$ 128 $\rightarrow$ 128	BN, ReLU, Dropout(0.2)
Gene latent z_n^r	128 $\rightarrow$ (20)*2	Mean & log-variance heads
Protein encoder	[H+S] $\rightarrow$ 128 $\rightarrow$ 128	BN, ReLU, Dropout(0.2)
Protein latent z_n^p	128 $\rightarrow$ (20)*2	Mean & log-variance heads
Accessibility encoder	[I+S] $\rightarrow$ 128 $\rightarrow$ 128	BN, LN, LeakyReLU, Dropout(0.1)
Accessibility latent z_a	128 $\rightarrow$ (20)*2	Mean & variance heads
Accessibility library-size	[I+S] $\rightarrow$ 128 $\rightarrow$ 128 $\rightarrow$ 1	BN, LN, LeakyReLU; Sigmoid output

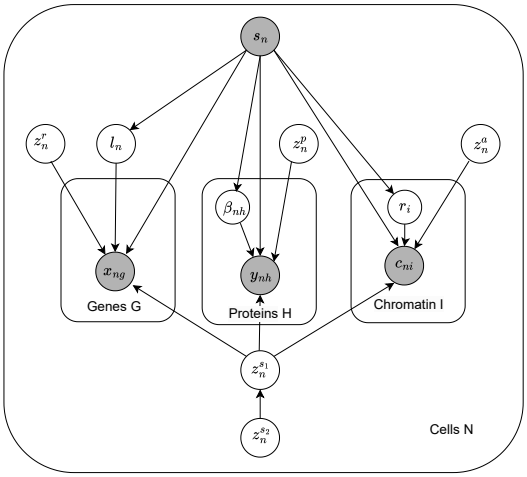
Supplementary Table 2: Decoder architecture summary.

Module	**Layers**	**Details**
Gene decoder MLP	$[z_n^{s1}, z_n^r, S] \rightarrow 128$ [+S] $\rightarrow 128$	BN, ReLU, Dropout(0.1)
Gene expected frequency	$[z_n^{s1}, z_n^r, S, 128] \rightarrow$ G	Linear output, Softmax
Gene dropout decoder	$[z_n^{s1}, z_n^r, S, 128] \rightarrow$ G	Linear output
Protein background decoder	$[z_n^{s1}, z_n^p, S] \rightarrow 256$ [+S] $\rightarrow 256$	BN, ReLU, Dropout(0.1)
Protein background mean log alpha	$[256, z_n^{s1}, z_n^p, S] \rightarrow$ H	Mean log-α
Protein background mean log beta	$[256, z_n^{s1}, z_n^p, S] \rightarrow$ H	Mean log-β
Protein foreground	$[z_n^{s1}, z_n^p, S] \rightarrow 256$ [+S] $\rightarrow 256$	BN, ReLU, Dropout(0.1)
Protein foreground scale decoder	$[256, z_n^{s1}, z_n^p, S] \rightarrow$ H	ReLU
Protein background mixing	$[z_n^{s1}, z_n^p, S] \rightarrow$ H	π_n,
Accessibility decoder	$[z_n^{s1}, z_n^a, S] \rightarrow 128 \rightarrow 128$	BN, LeakyReLU
Accessibility output	128 $\rightarrow$ I	Sigmoid activation

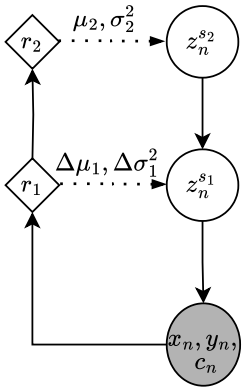
Supplementary Table 3: Pathways associated with cell type-specific gene expression programs corresponding to multiHIVE embedding $\boldsymbol{z^{s_1}}$ for Thymocyte development dataset and their corresponding references.

Cell type	Pathway	Reference
DP (Q1), DP (Q2)	TRANSLOCATION_OF_ZAP_70_TO_IMMUNOLOGICAL SYNAPSE	[10]
	PHOSPHORYLATION_OF_CD3_AND_TCR_ZETA_CHAINS	[11, 12]
DP (Sig.)	TNFA_SIGNALING_VIA_NFKB	[13, 14]
Mature CD4 T cells, Mature CD8 T cells	CD8_POSITIVE_ALPHA_BETA_T_CELL_ACTIVATION	[15]
	T_CELL_MEDIATED_CYTOTOXICITY	[16]
	IL2_STAT5_SIGNALING	[17, 18]
$\gamma\delta$ T cells	TRANSMEMBRANE_RECEPTOR_PROTEIN_TYROSINE KINASE_SIGNALING_PATHWAY	[19, 20]

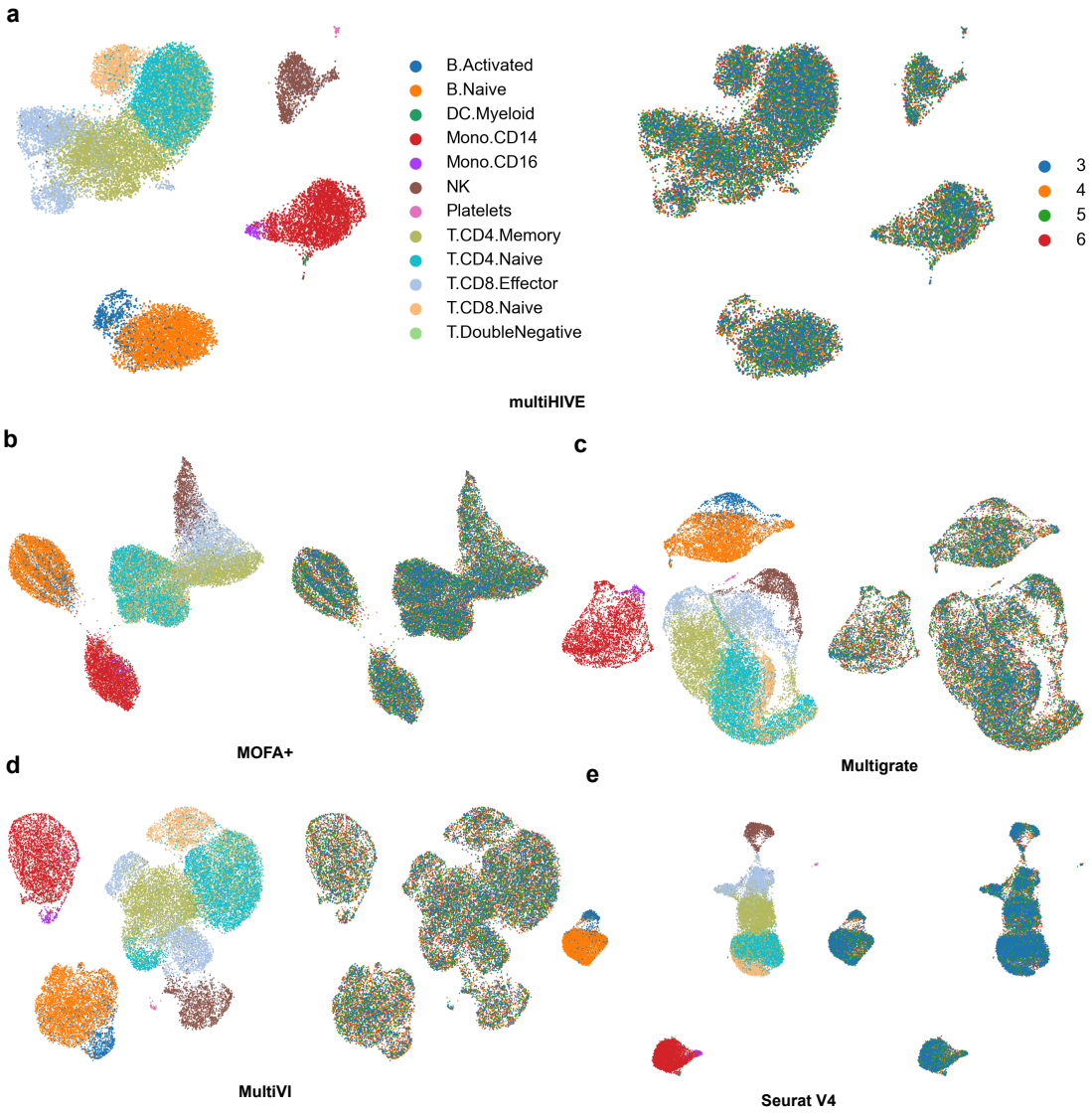
3 Supplementary Figures



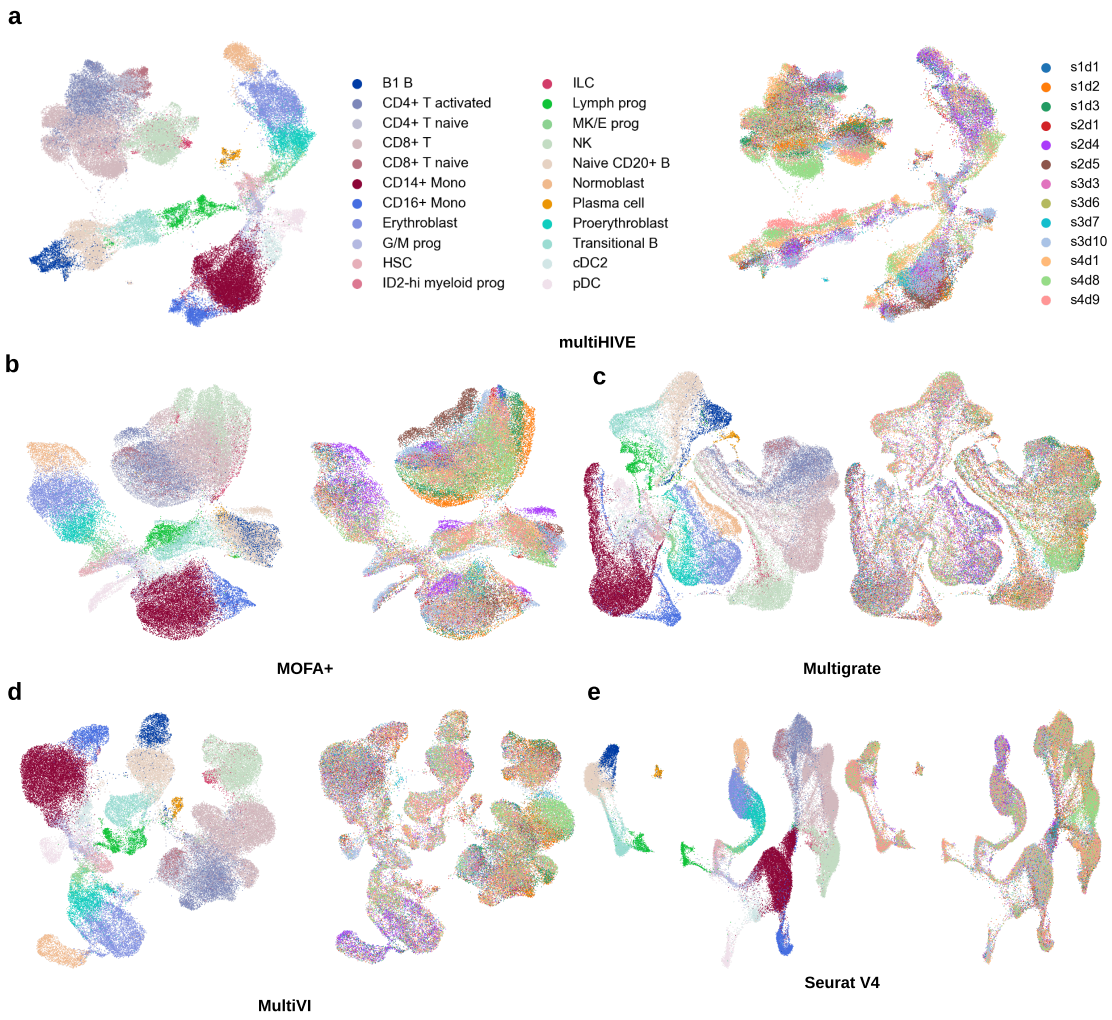
Supplementary Figure 1: multiHIVE's probabilistic hierarchical graphical model. Shaded nodes represent observed random variables. Unshaded nodes represent latent variables. Edges denote conditional independence in the direction shown. Rectangles ("plates") represent independent replication of variables inside.



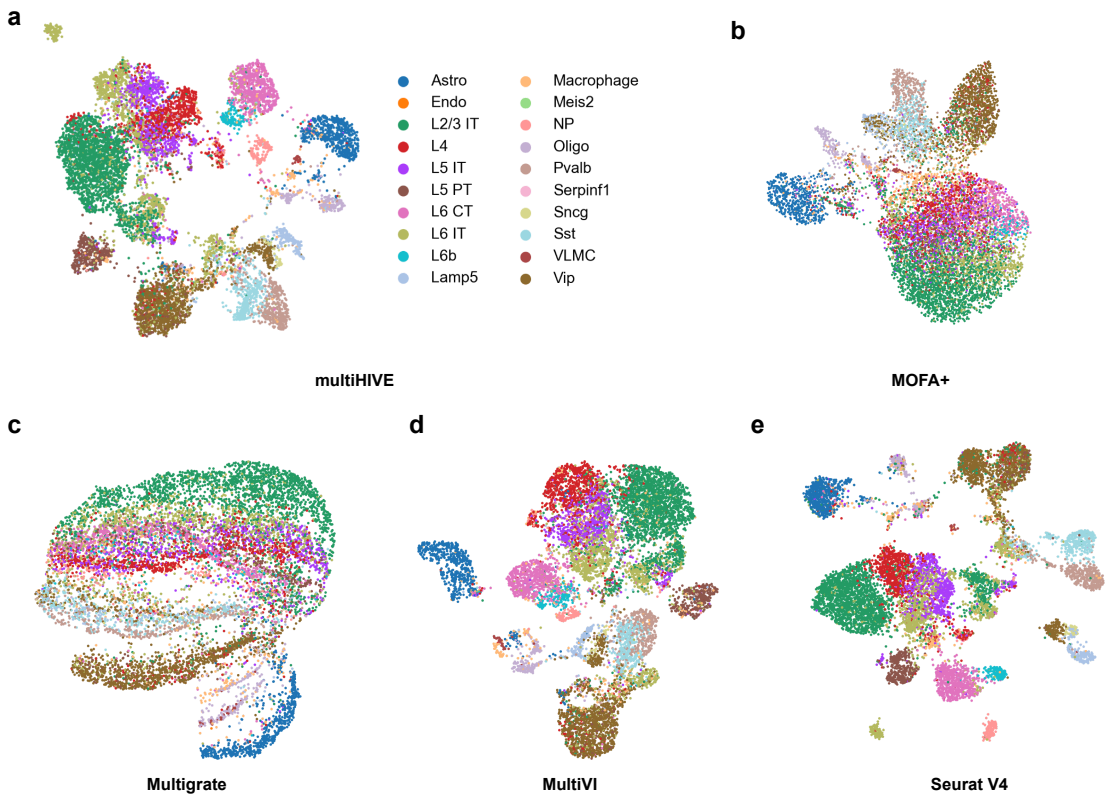
Supplementary Figure 2: multiHIVE's architecture for joint latent space representation. The left side of the figure consists of a bottom-up deterministic path $[(\boldsymbol{x_n}, \boldsymbol{y_n}, \boldsymbol{c_n}) \rightarrow \boldsymbol{r_1} \rightarrow \boldsymbol{r_2}]$, and to the right, there is the top-down stochastic path $[\boldsymbol{z_n^{s_2}} \rightarrow \boldsymbol{z_n^{s_1}} \rightarrow (\boldsymbol{x_n}, \boldsymbol{y_n}, \boldsymbol{c_n})]$. The diamond shape indicates neural network transformations. The circle indicates a random variable. The shaded circle indicates observed random variable.



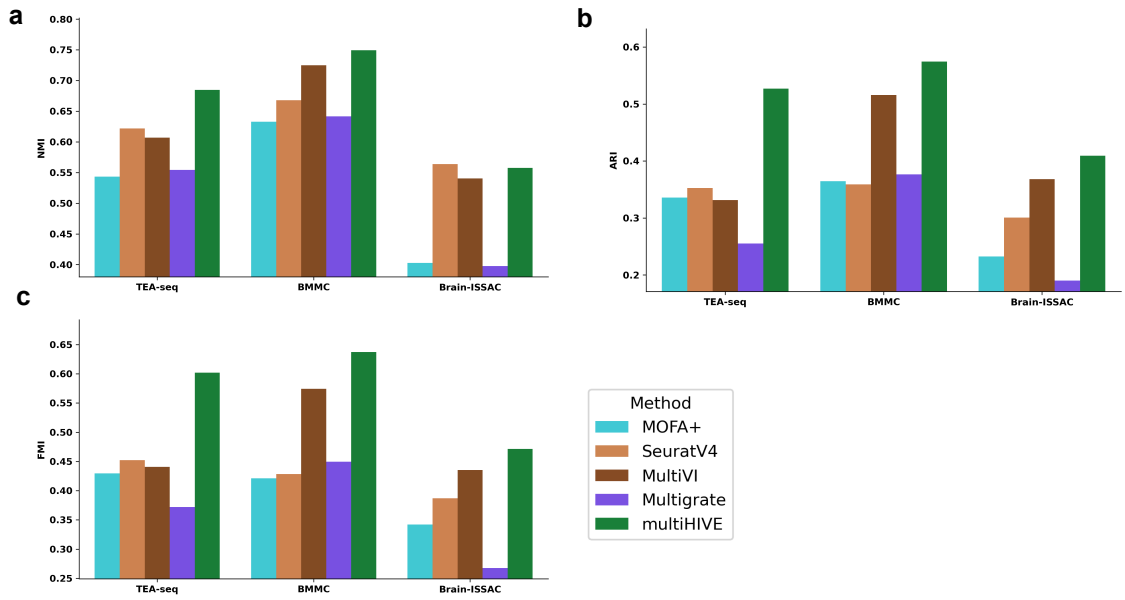
Supplementary Figure 3: Visualization of latent space embeddings for the TEA-seq data of RNA and ATAC post-integration. Cells are colored by different cell types (left) and batch information (right), for different integration algorithms: (a) multiHIVE, (b) MOFA+, (c) Multigrate (d) MultiVI and (e) SeuratV4



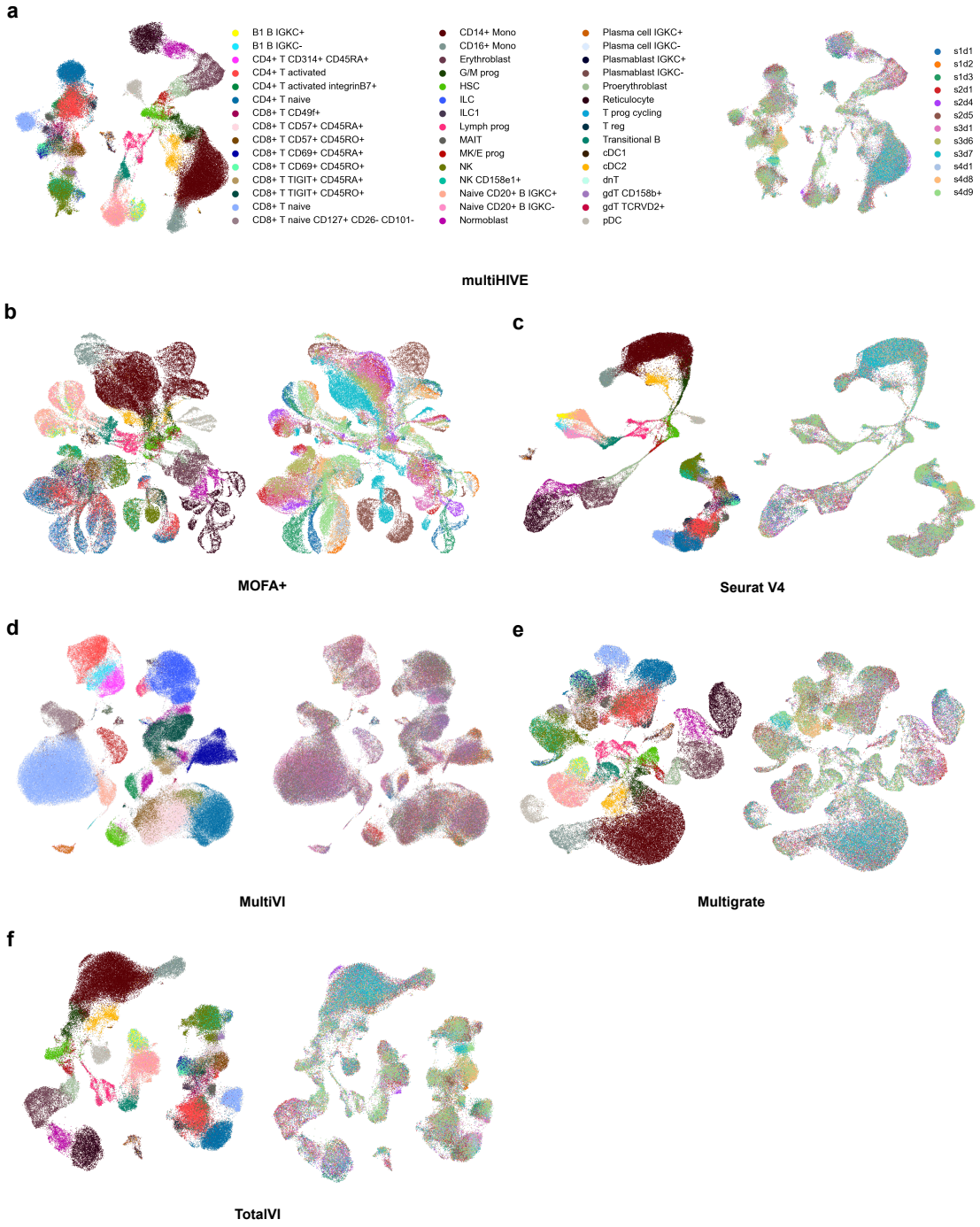
Supplementary Figure 4: Visualization of latent space embeddings for the BMMC dataset post-integration. Cells are colored by different cell types (left) and batch information (right), for different integration algorithms: (a) multiHIVE, (b) MOFA+, (c) Multigrate (d) MultiVI and (e) SeuratV4



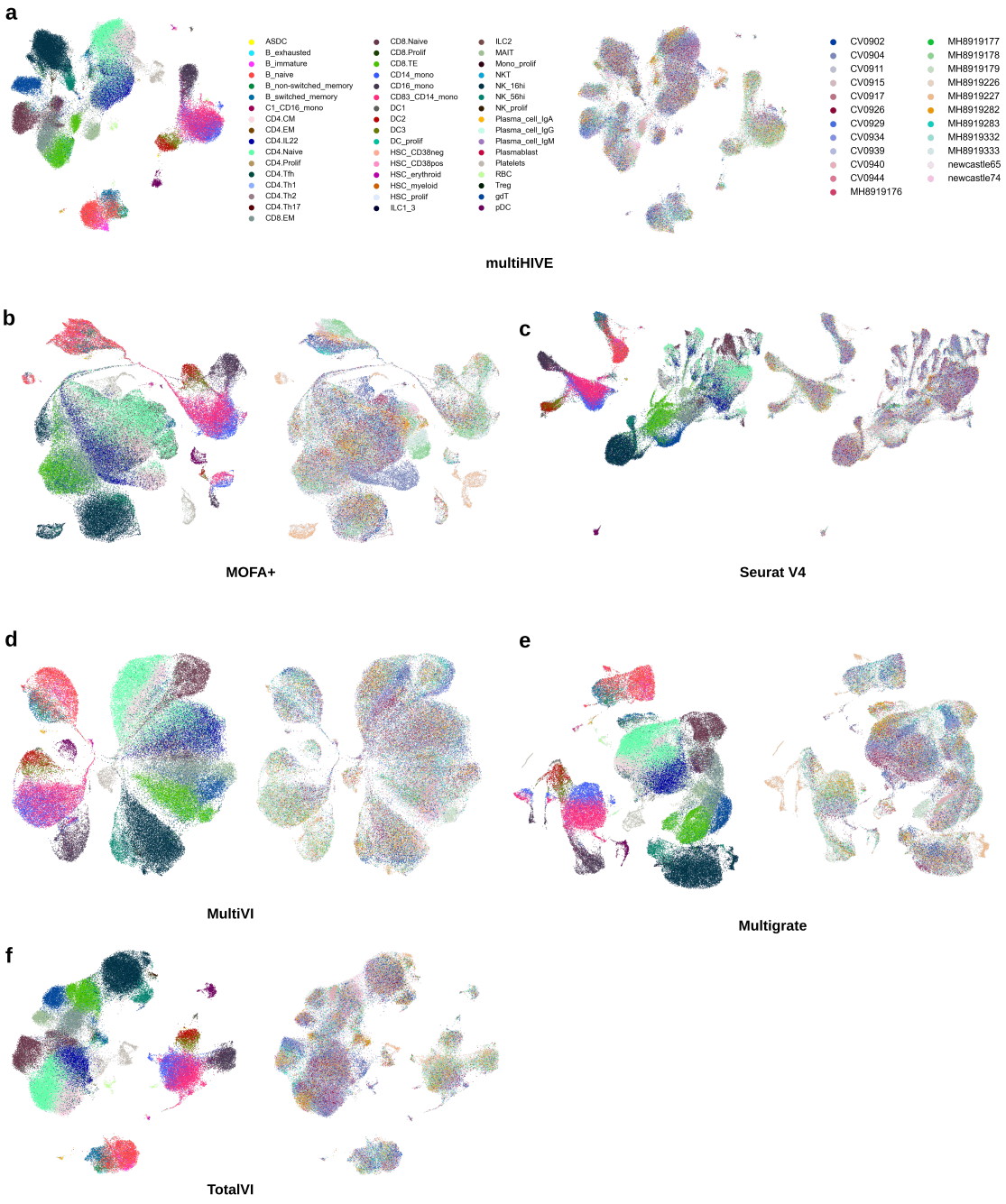
Supplementary Figure 5: Visualization of latent space embeddings for the Brain-ISSAAC dataset post-integration. Cells are colored by different cell types, for different integration algorithms: (a) multiHIVE, (b) MOFA+, (c) Multigrate (d) MultiVI and (e) SeuratV4



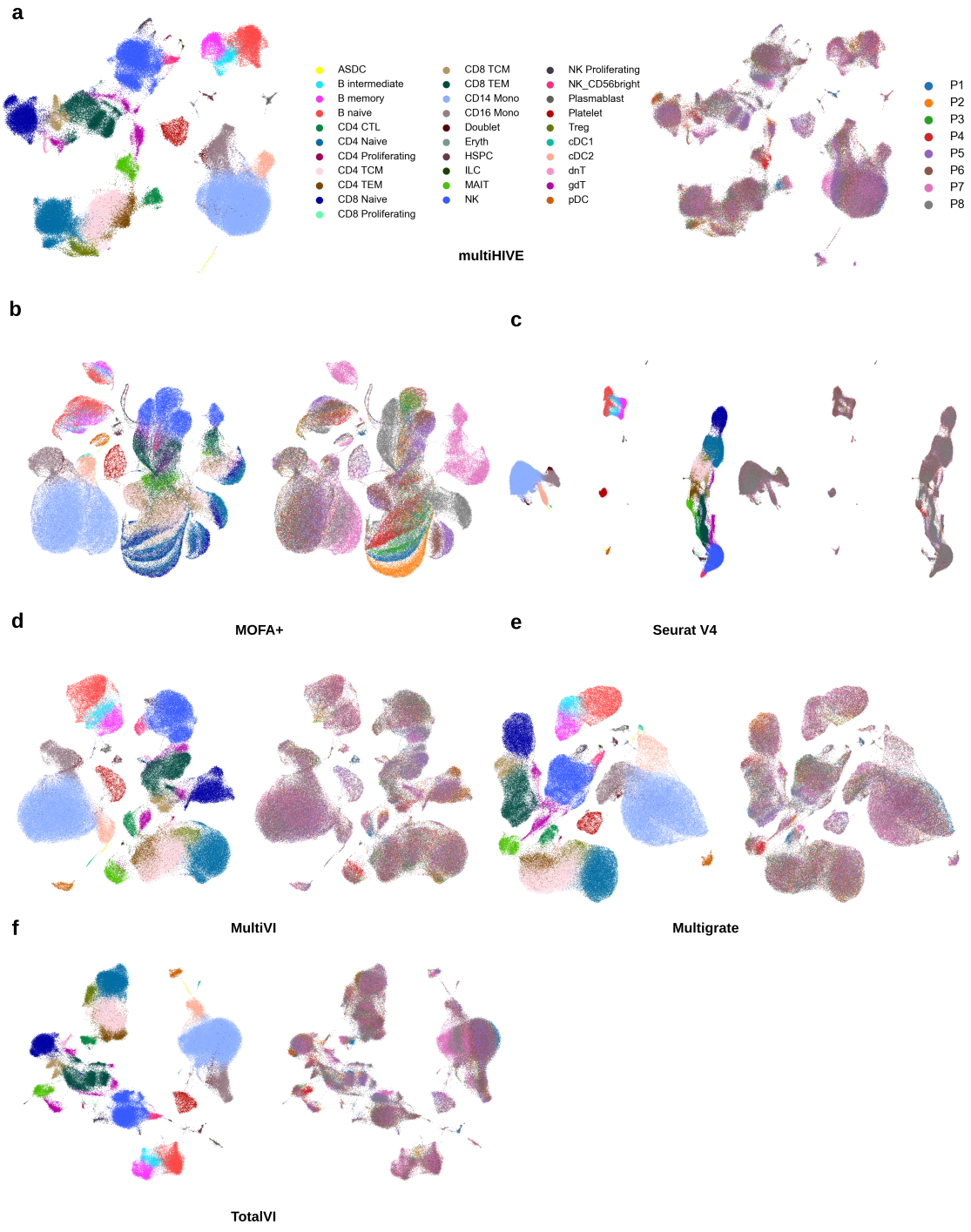
Supplementary Figure 6: **Quantitative assessment of biological conservation metrics for RNA + ATAC modality.** Comparison of raw (a) NMI, (b) ARI, and (c) FMI values across the methods MOFA+, SeuratV4, Multigrate, MultiVI, and multiHIVE for the TEA-seq, BMMC, and Brain-ISSAAC datasets.



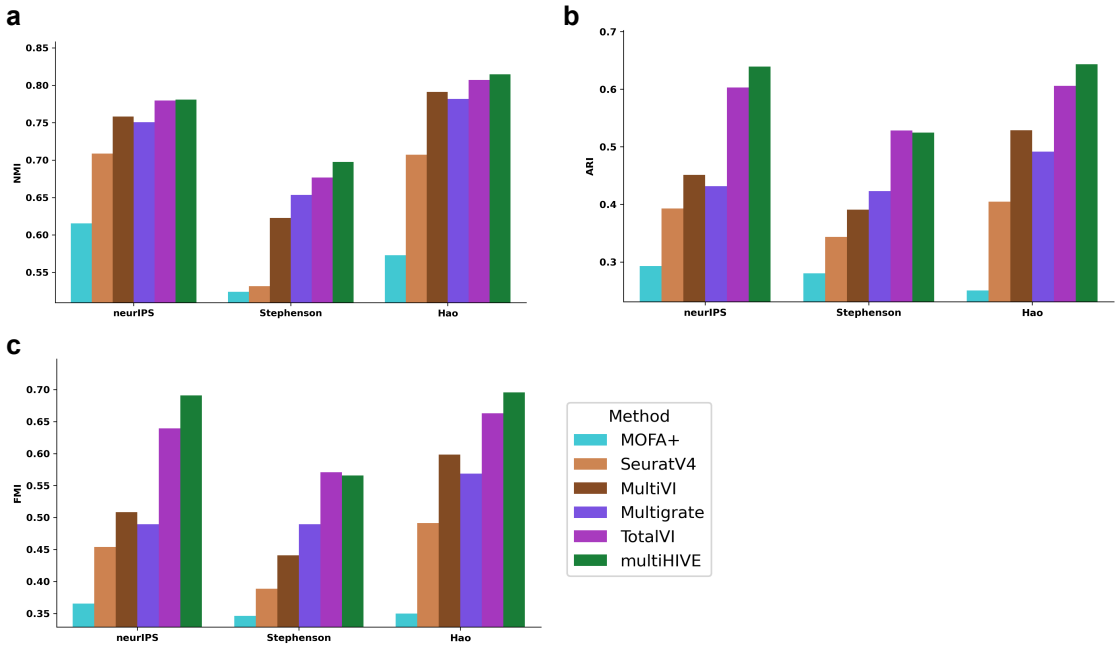
Supplementary Figure 7: Visualization of latent space embeddings for the neurIPS dataset post-integration. Cells are colored by different cell types (left) and batch information (right), for different integration algorithms: (a) multiHIVE, (b) MOFA+, (c) SeuratV4, (d) MultiVI, (e) Multigrate, and (f) TotalVI.



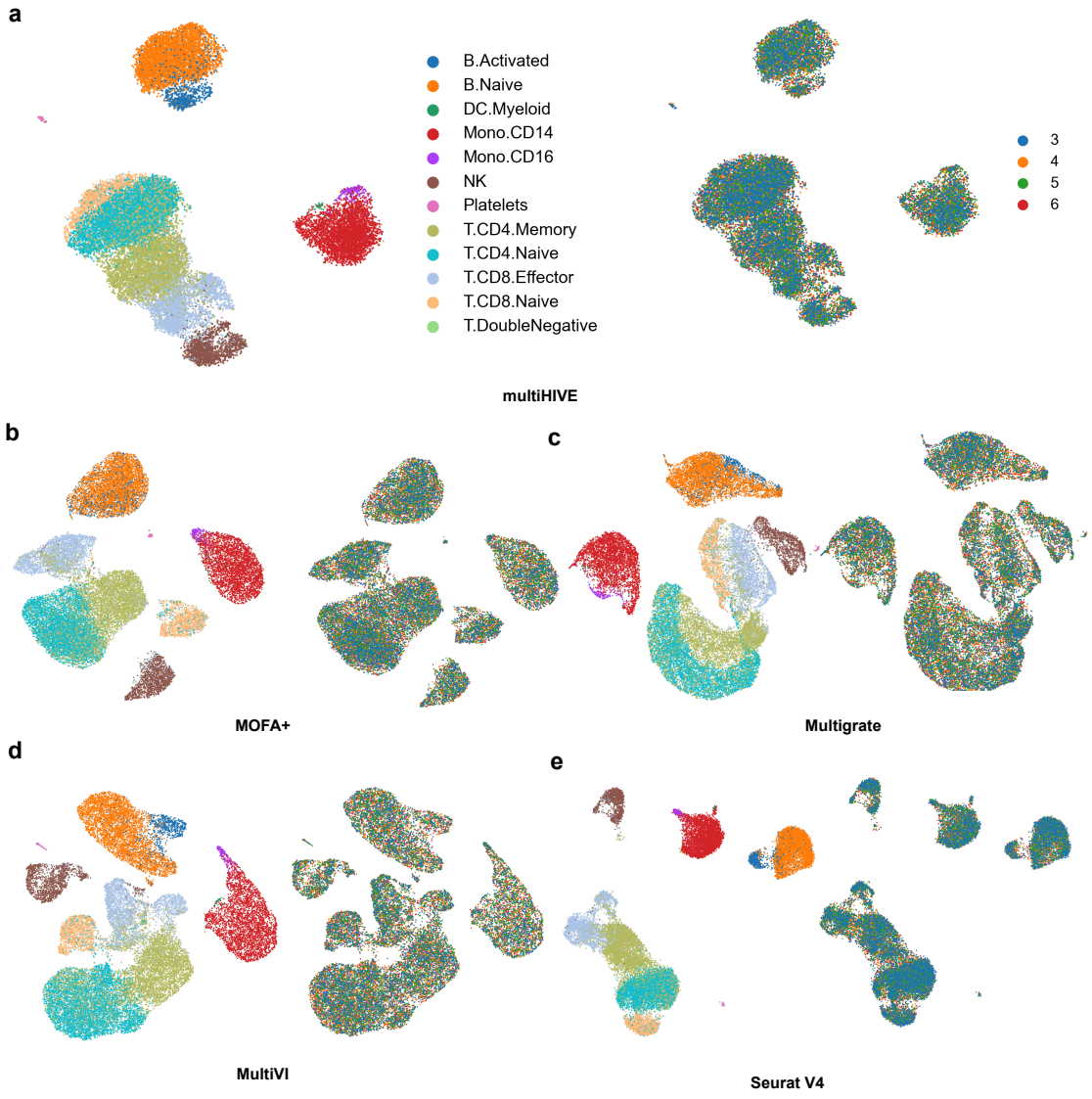
Supplementary Figure 8: Visualization of latent space embeddings for the Stephenson dataset post-integration. Cells are colored by different cell types (left) and batch information (right), for different integration algorithms: (a) multiHIVE, (b) MOFA+, (c) SeuratV4, (d) MultiVI, (e) Multigrate, and (f) TotalVI.



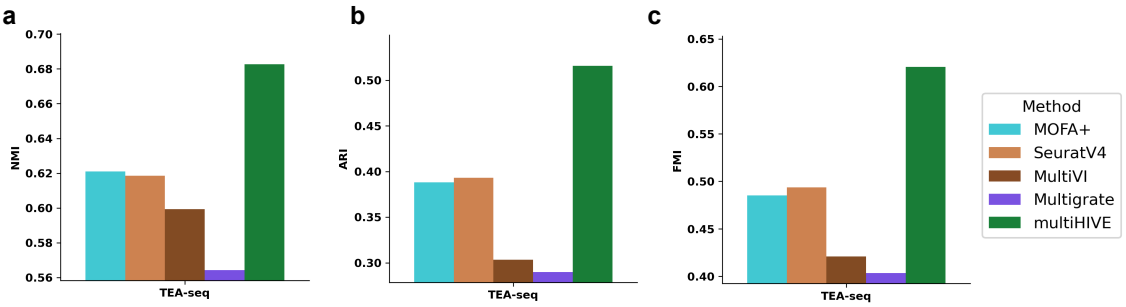
Supplementary Figure 9: Visualization of latent space embeddings for the Hao dataset post-integration. Cells are colored by different cell types (left) and batch information (right), for different integration algorithms: (a) multiHIVE, (b) MOFA+, (c) SeuratV4, (d) MultiVI, (e) Multigrate, and (f) TotalVI.



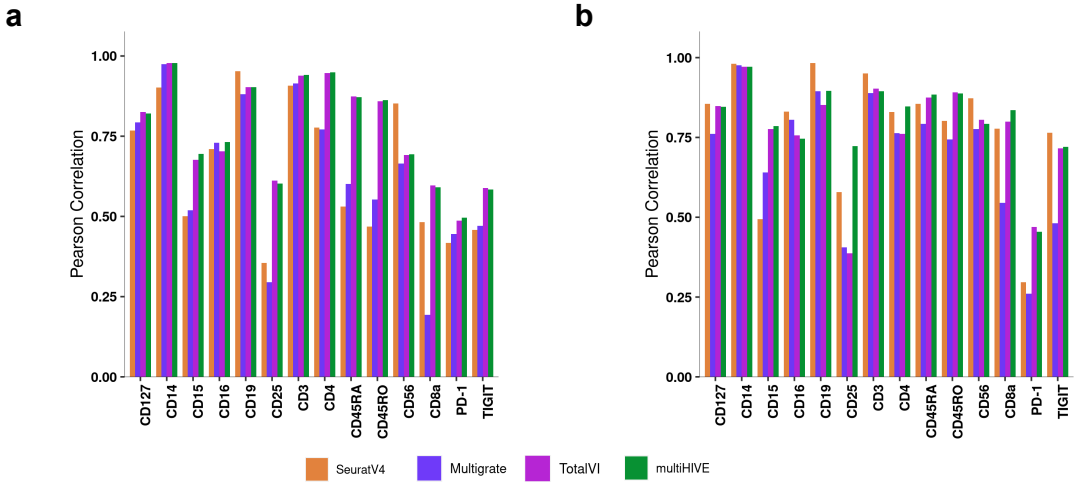
Supplementary Figure 10: **Quantitative assessment of biological conservation metrics for RNA + Protein modality.** Comparison of raw (a) NMI, (b) ARI, and (c) FMI values across the methods MOFA+, SeuratV4, Multigrate, MultiVI, TotalVI, and multiHIVE for the neurIPS, Stephenson, and Hao datasets.



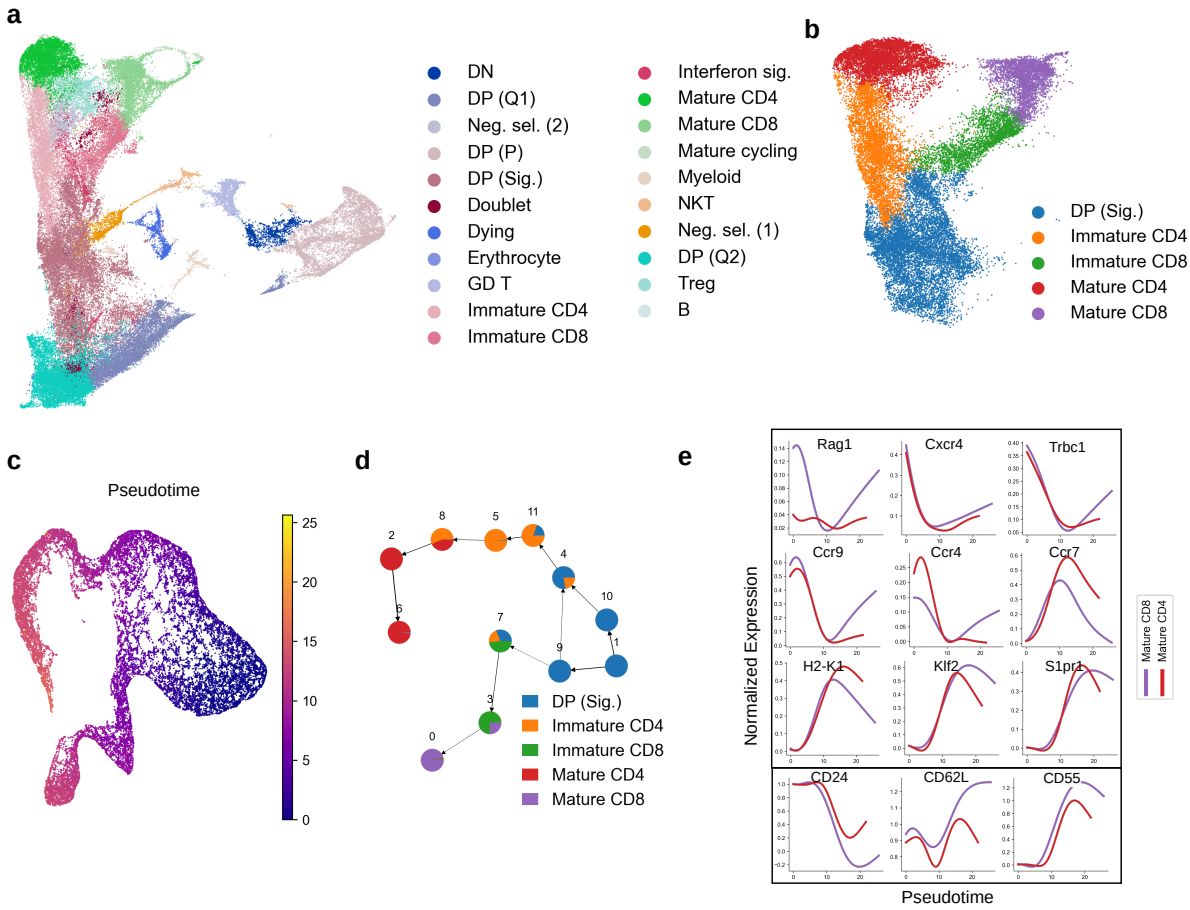
Supplementary Figure 11: Visualization of latent space embeddings for the TEA-seq dataset comprising three modalities—RNA, ATAC, and protein—post-integration. Cells are colored by different cell types (left) and batch information (right), for different integration algorithms: (a) multiHIVE, (b) MOFA+, (c) Multigrate, (d) MultiVI, and (e) SeuratV4.



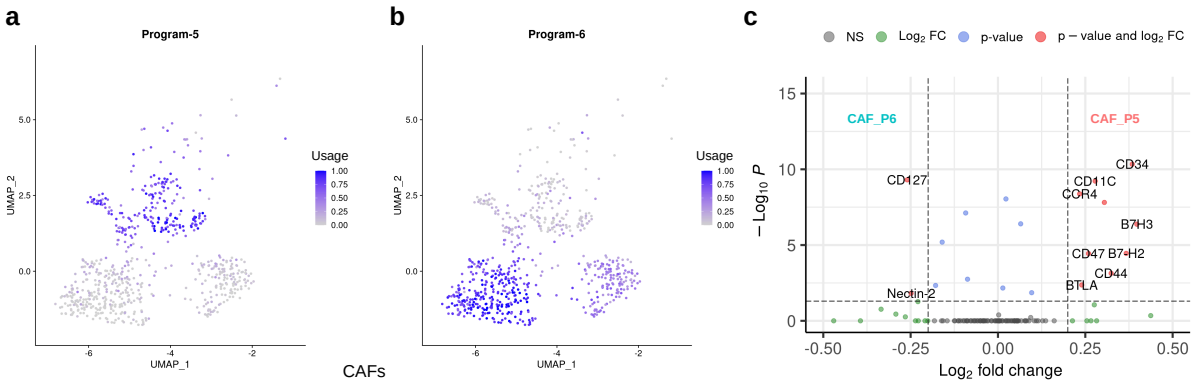
Supplementary Figure 12: **Quantitative assessment of biological conservation metrics for RNA + ATAC + Protein modality.** Comparison of raw (a) NMI, (b) ARI, and (c) FMI values across the methods MOFA+, SeuratV4, Multigrate, MultiVI, and multiHIVE for the TEA-seq dataset.



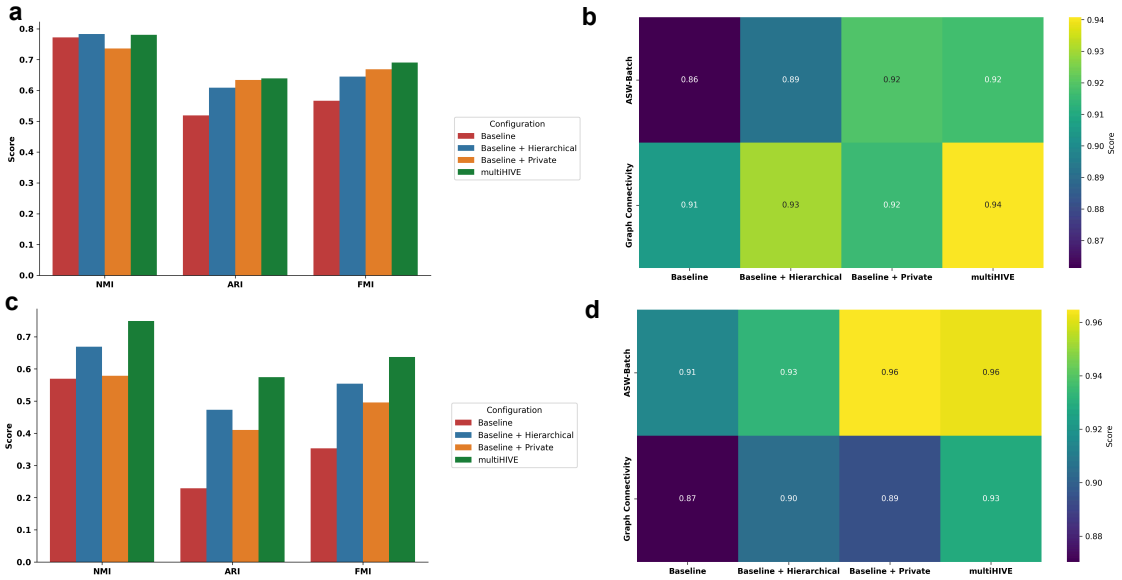
Supplementary Figure 13: Barplot comparison of Pearson's correlation coefficient of imputed protein expression for (a) 5k batch and (b) 10k batch of the PBMC dataset. multiHIVE's performance is compared against that of SeuratV4, Multigrate and TotalVI.



Supplementary Figure 14: Application of TotalVI on a thymocyte development dataset. (a) UMAP visualization of TotalVI embeddings for the Thymocyte development dataset, with cells colored according to cell type. (b) Subset of positive selection cells from (a) for trajectory inference. (c) Trajectory of positively selected CD4+ and CD8+ T cells inferred by MARGARET based on TotalVI-inferred embeddings. (d) UMAP plot of trajectory delineating the progression of pseudotime. (e) Gene expression trends for known marker genes for the CD4+ and CD8+ T cell lineages using the denoised expression values from TotalVI, scaled per gene.



Supplementary Figure 15: UMAP plot of cancer associated fibroblasts (CAFs) showing the usage of (a) Program-5, and (b) Program-6 encoded in the shared latent embeddings ($\boldsymbol{z}^{s_1}$) inferred by multiHIVE for the breast cancer dataset. (c) Volcano plot of differentially expressed proteins in the CAF subpopulations (CAF_P5 and CAF_P6).



Supplementary Figure 16: **Ablation study across configurations**. (a, b) Ablation analysis of the RNA + Protein modality on the neurIPS dataset. (c, d) Ablation analysis of the RNA + ATAC modality on the TEA-seq dataset. (a, c) report biological conservation metrics, while (b, d) report batch correction metrics across different model configurations: (1) without private representations and hierarchical structure, (2) without private representations, and (3) without hierarchical structure, in comparison to multiHIVE.

Supplementary References

[1] Vahdat, A. & Kautz, J. NVAE: A deep hierarchical variational autoencoder. *Advances in neural information processing systems* **33**, 19667–19679 (2020).

[2] Kingma, D. P. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[3] Gayoso, A. *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature methods* **18**, 272–282 (2021).

[4] Ashuach, T., Reidenbach, D. A., Gayoso, A. & Yosef, N. PeakVI: A deep generative model for single-cell chromatin accessibility analysis. *Cell Reports Methods* **2**, 100182 (2022). URL `https://www.sciencedirect.com/science/article/pii/S2667237522000376`.

[5] Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nature methods* **19**, 41–50 (2022).

[6] Pandey, K. & Zafar, H. Inference of cell state transitions and cell fate plasticity from single-cell with MARGARET. *Nucleic Acids Research* **50**, e86–e86 (2022).

[7] Kotliar, D. *et al.* Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife* **8**, e43803 (2019).

[8] Korotkevich, G. *et al.* Fast gene set enrichment analysis. *biorxiv* 060012 (2016).

[9] Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
[10] Palacios, E. H. & Weiss, A. Distinct roles for Syk and ZAP-70 during early thymocyte development. *The Journal of experimental medicine* **204**, 1703–1715 (2007).
[11] Levelt, C., Carsetti, R. & Eichmann, K. Regulation of thymocyte development through CD3. II. Expression of T cell receptor beta CD3 epsilon and maturation to the CD4+ 8+ stage are highly correlated in individual thymocytes. *The Journal of experimental medicine* **178**, 1867–1875 (1993).
[12] Brodeur, J.-F., Li, S., Damlaj, O. & Dave, V. P. Expression of fully assembled TCR–CD3 complex on double positive thymocytes: synergistic role for the PRS and ER retention motifs in the intra-cytoplasmic tail of CD3ε. *International immunology* **21**, 1317–1327 (2009).
[13] Webb, L. V., Ley, S. C. & Seddon, B. TNF activation of NF-κB is essential for development of single-positive thymocytes. *Journal of Experimental Medicine* **213**, 1399–1407 (2016).
[14] Liu, T., Zhang, L., Joo, D. & Sun, S.-C. NF-κB signaling in inflammation. *Signal transduction and targeted therapy* **2**, 1–9 (2017).
[15] Lambolez, F., Kronenberg, M. & Cheroutre, H. Thymic differentiation of TCR$\alpha\beta$+ CD8$\alpha\alpha$+ IELs. *Immunological reviews* **215**, 178–188 (2007).
[16] Overgaard, N. H., Jung, J.-W., Steptoe, R. J. & Wells, J. W. CD4+/CD8+ double-positive T cells: more than just a developmental stage? *Journal of Leucocyte Biology* **97**, 31–38 (2015).
[17] Moriggl, R. *et al.* Stat5 is required for IL-2-induced cell cycle progression of peripheral T cells. *Immunity* **10**, 249–259 (1999).
[18] Mahmud, S. A., Manlove, L. S. & Farrar, M. A. Interleukin-2 and STAT5 in regulatory T cell development and function. *Jak-Stat* **2**, e23154 (2013).
[19] Parker, M. E. & Ciofani, M. Regulation of $\gamma\delta$ T cell effector diversification in the thymus. *Frontiers in immunology* **11**, 42 (2020).
[20] Muro, R., Takayanagi, H. & Nitta, T. T cell receptor signaling for $\gamma\delta$T cell development. *Inflammation and Regeneration* **39**, 1–11 (2019).