Supplementary Materials for Integrated Analysis of Multimodal Single-Cell Data with Structural Similarity

Yingxin Cao[1,5,6⋆], Laiyi Fu[1,2 ⋆], Jie Wu[3], Qinke Peng[2], Qing Nie[4,5,6], Jing Zhang[1 ⋆⋆], and Xiaohui Xie[1 ⋆⋆]

[1] Department of Computer Science, University of California, Irvine, CA, 92697, USA
[2] Systems Engineering Institute, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shannxi, 710049, China
[3] Department of Biological Chemistry, University of California, Irvine, CA, 92697, USA
[4] Department of Mathematics, University of California, Irvine, CA, 92697, USA
[5] Center for Complex Biological Systems, University of California, Irvine, CA, 92697, USA
[6] NSF-Simons Center for Multiscale Cell Fate Research, University of California, Irvine, CA, 92697, USA

⋆ Equal contributions
⋆⋆ To whom correspondence should be addressed

Table of Contents

List of Tables

List of Figures

1 Results on Hyperparameter Stability

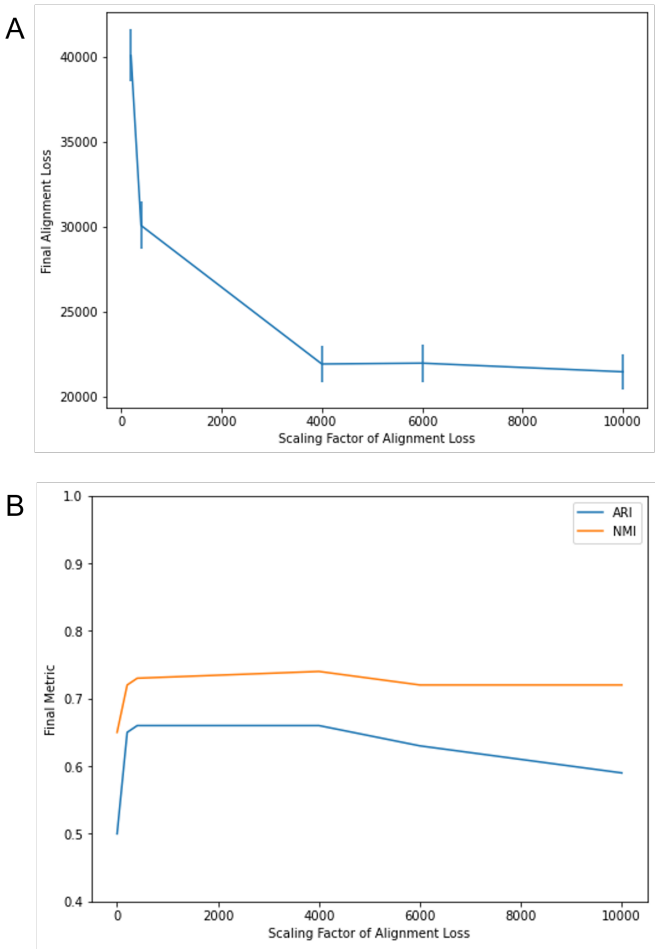

Fig. S1: Results on hyperparameter stability. (A) The alignment loss L_{Local} decreases as the scaling factor increases. (B) Clustering metrics ARI and NMI as a function of scaling factor.

2 Cell Type Specific Marker Genes

Table S1: List of cell-type specific marker genes used to visualize expressions.

Cell Type	Marker Gene Names
Pvalb	Erbb4, Cemip, Lrrc4c, Slit2, Cntnap4, Btbd11, Zfp536, Esrrg, Kcnc1,Cntnap5c
L4	Car10, Unc5d, Rorb, Pcdh15, Dcc, Gria4, Prkg1, Fstl4, Kcnh5, Cpne9
CD4 Naive	Bach2, Fhit, Igf1r, Ccr7, Ak5, Apba2, Lef1, Maml2, Sell,Satb1-as1
B Naive	Ighm, Ighd, Tcl1a, Bach2, Col19a1, Il4r, Skap1, Camk2D, Foxp1, Khdrbs2

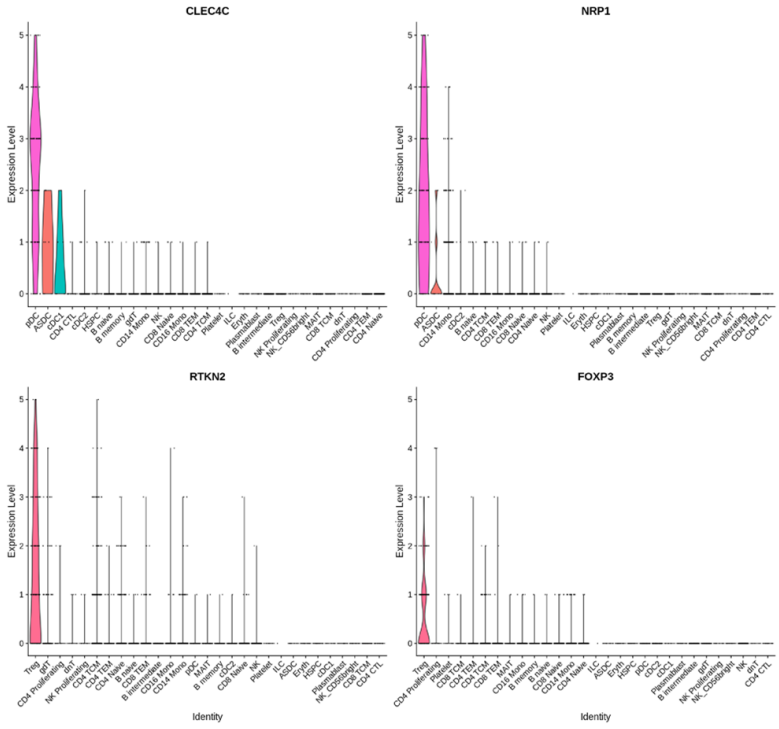

Fig. S2: Visualizations of marker gene expressions by inferred ground truth cell types in the PBMC 10k Dataset. CLEC4C and NRP1 are marker genes for pDC cells; RTKN2 and FOXP3 are marker genes for Treg cells.

In figure S3, each panel shows the mean normalized expression of marker genes of a cell type (Table S1) in the cells of the cluster labeled with the corresponding cell type. Boxplots show the mean as well 25% and 75% quantile expres-

sion of marker genes of different methods. Pairwise t-tests between SALIERX and other methods indicate whether the marker genes from SALIERX show significantly higher expression than those from other methods. T-test p-values are indicated by ns (p-value > 0.05, i.e., not significant), * (p-value < 0.05), ** (p-value < 0.01) and *** (p-value < 0.001). The plots show that the marker genes show overall higher expression in corresponding cell clusters discovered by SALIERX than those by other methods.

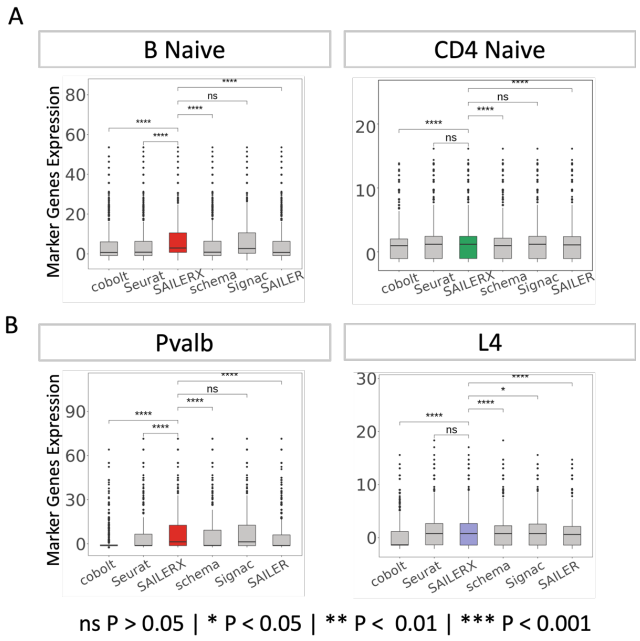

Fig. S3: Comparing the expression of marker genes in clusters derived by different methods. (A) Mean expression of marker genes of B Naive cells and CD4 Naive cells from PBMC 10k dataset. (B) Mean expression of marker genes of Pvalb cells and L4 cells from SNARE-seq dataset.

Table S2: Mean expressions of markers on cells clustered by different methods.

Cell Type	SAILERX	Seurat	Signac	Cobolt	Schema	SAILER
Pvalb	**7.10**	4.25	7.07	0.72	5.01	3.71
L4	1.05	**1.11**	0.97	-0.01	0.92	0.77
B naive	**6.29**	3.60	6.09	3.55	3.68	3.60
CD4	**1.36**	1.34	1.34	1.05	1.07	1.23

3 Extra Results on Clustering

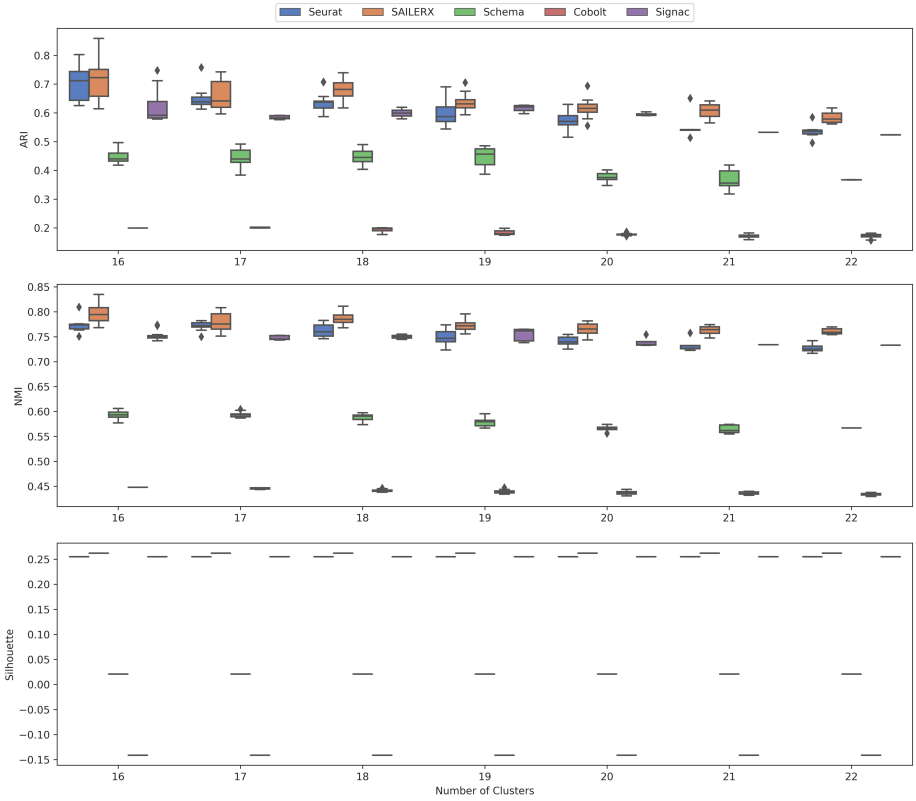

Fig. S4: Clustering scores of PBMC 10k dataset by different number of identified clusters.

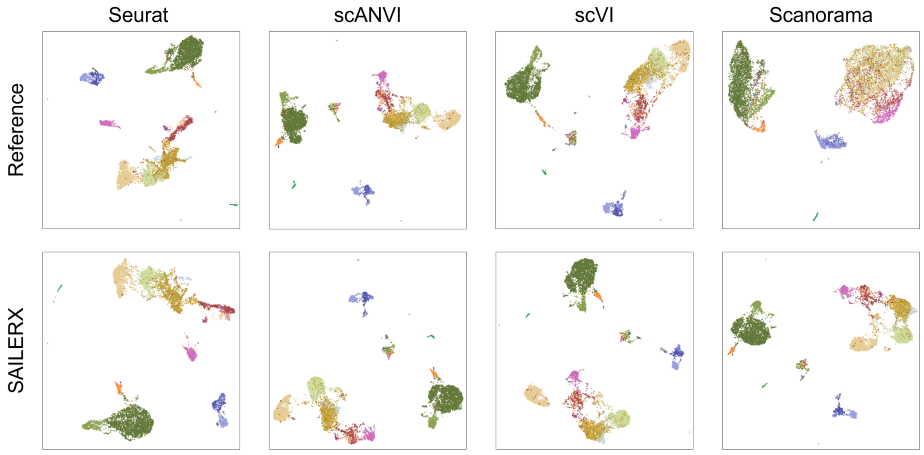

Fig. S5: UMAP Visualizations of reference embeddings vs SAILERX embeddings. Top row: UMAP visualizations of reference gene expression embeddings generated by different methods. Bottom row: joint embeddings generated by SAILERX after training.

4 Results on Share-seq Dataset

Fig. S6: Results on Share-seq dataset. Cells colored by ground truth label. (A) UMAP visualizations of embeddings on mouse skin Share-seq dataset generated by different methods. (B) Quantitative metrics of ARI, NMI and Silhouette Score on clustering.

5 Results on Batch Correction

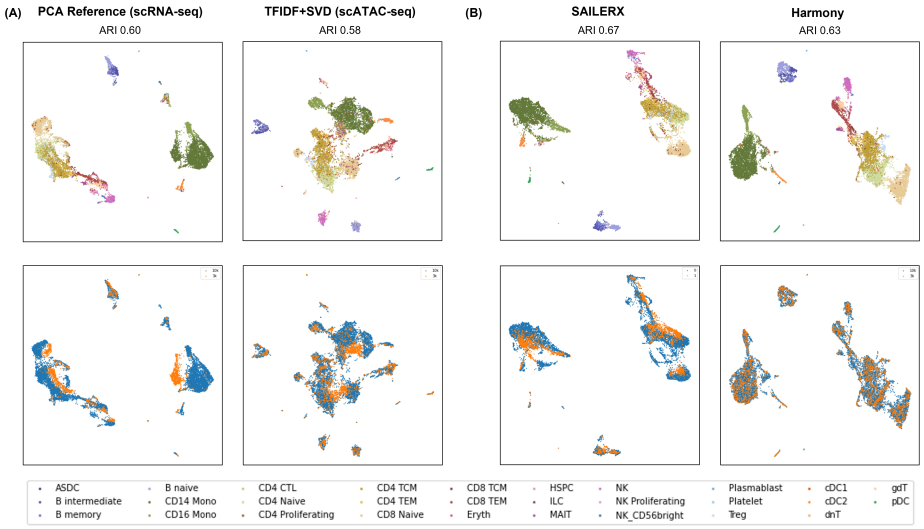

Fig. S7: Results on batch effect correction on PBMC 10k and 3k datasets. (A) UMAP Visualizations of PCA (left) embedding on gene expression modality and TF-IDF + SVD (right) embedding on chromatin accessibility modality before batch effect corrections. (B) UMAP visualization of embeddings after batch effect correction. Top row: colored by cell types; Bottom row: colored by batches.

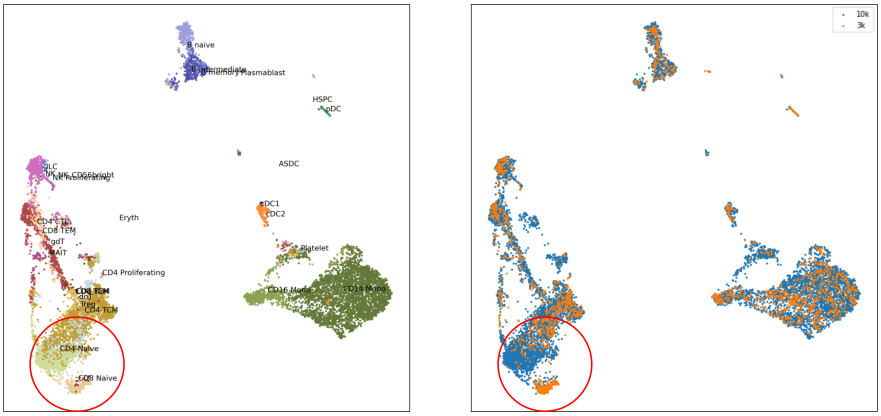

Fig. S8: UMAP visualizations of the embedding generated by SAILERX. Left: colored by cell types; Right: colored by batches.

6 Results on Motif Enrichment Analysis

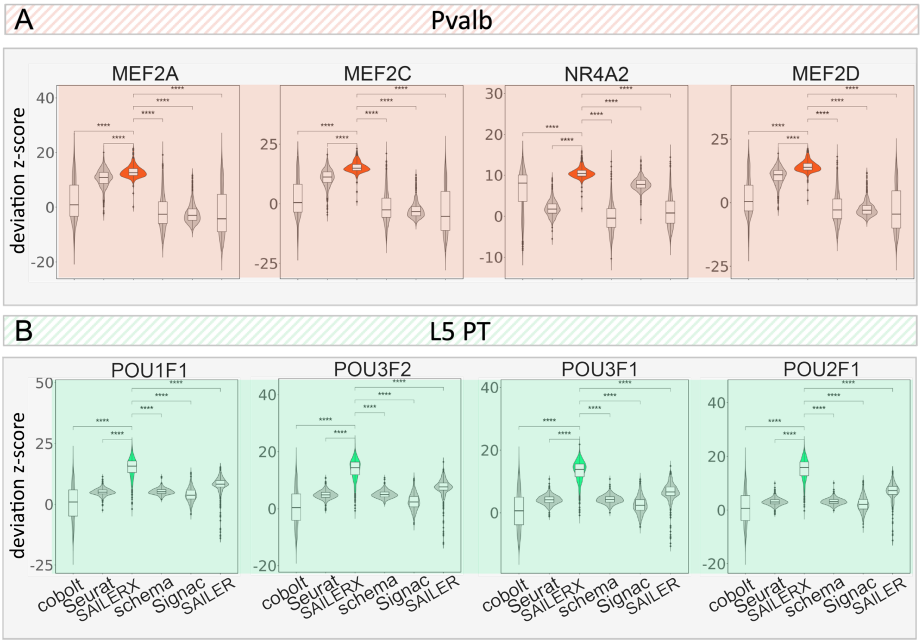

Fig. S9: Motif deviation z-scores on cells identified as (A) Pvalb and (B) L5 PT by different methods from SNARE-seq imputed data. The data is imputed through SAILERX. For each cell type, four enriched motifs are selected. Pairwise t-tests are performed between SAILERX and all other methods. Three-stars refers to differential significance between two methods (p-value less than 0.05).