Supplemental Materials for

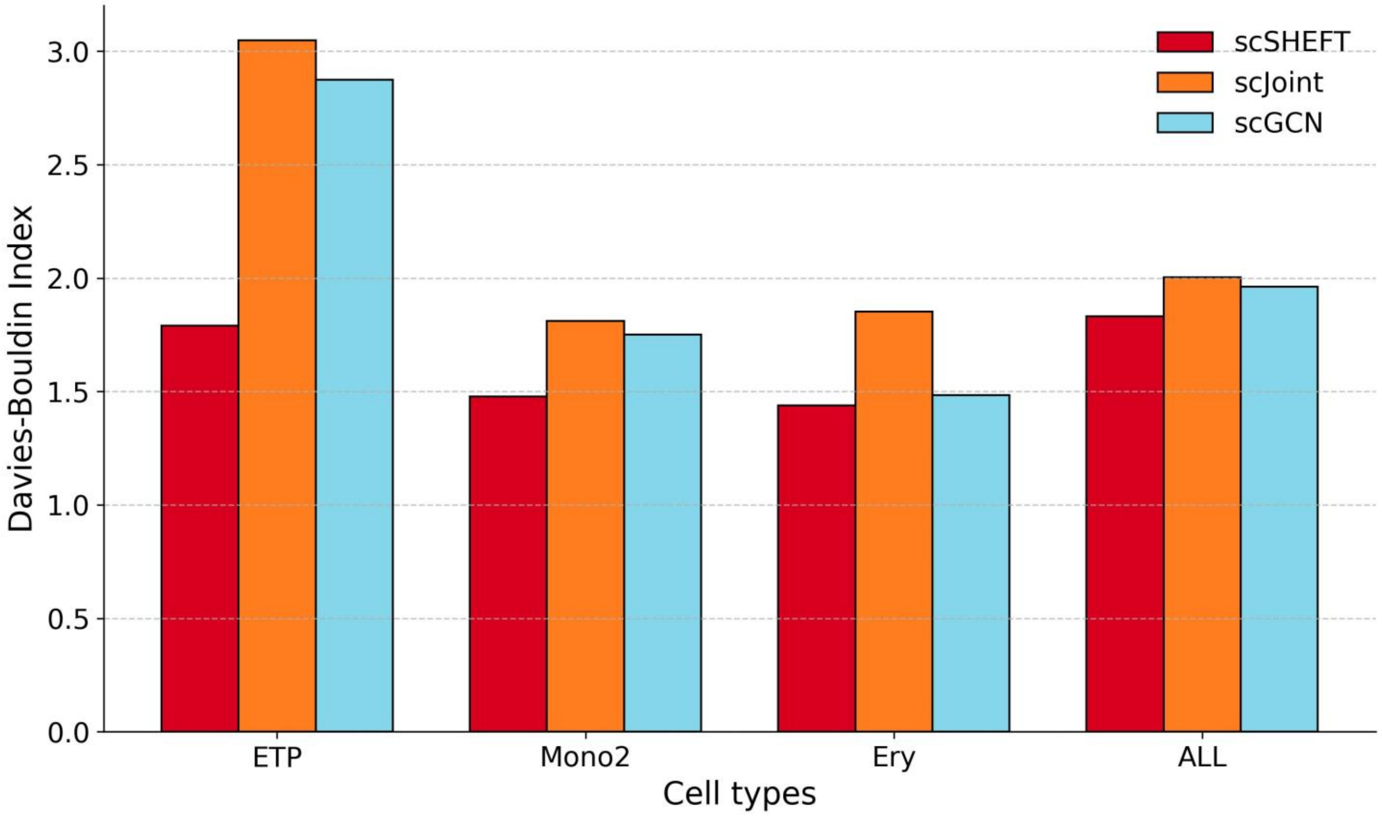
scSHEFT enables multi-omics label transfer from scRNA-seq to scATAC-seq through dual alignment

Supplemental Figures:

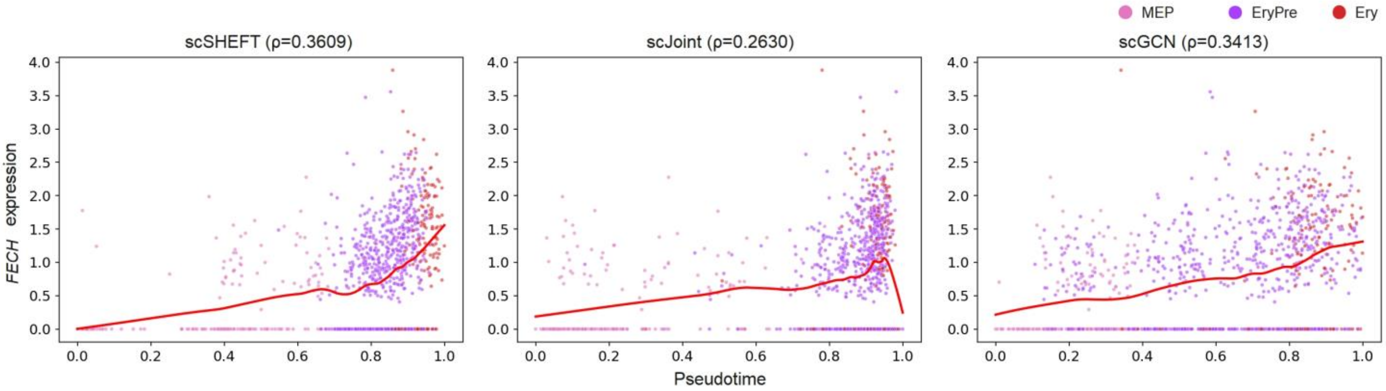
Supplemental Tables:

Supplemental Notes:

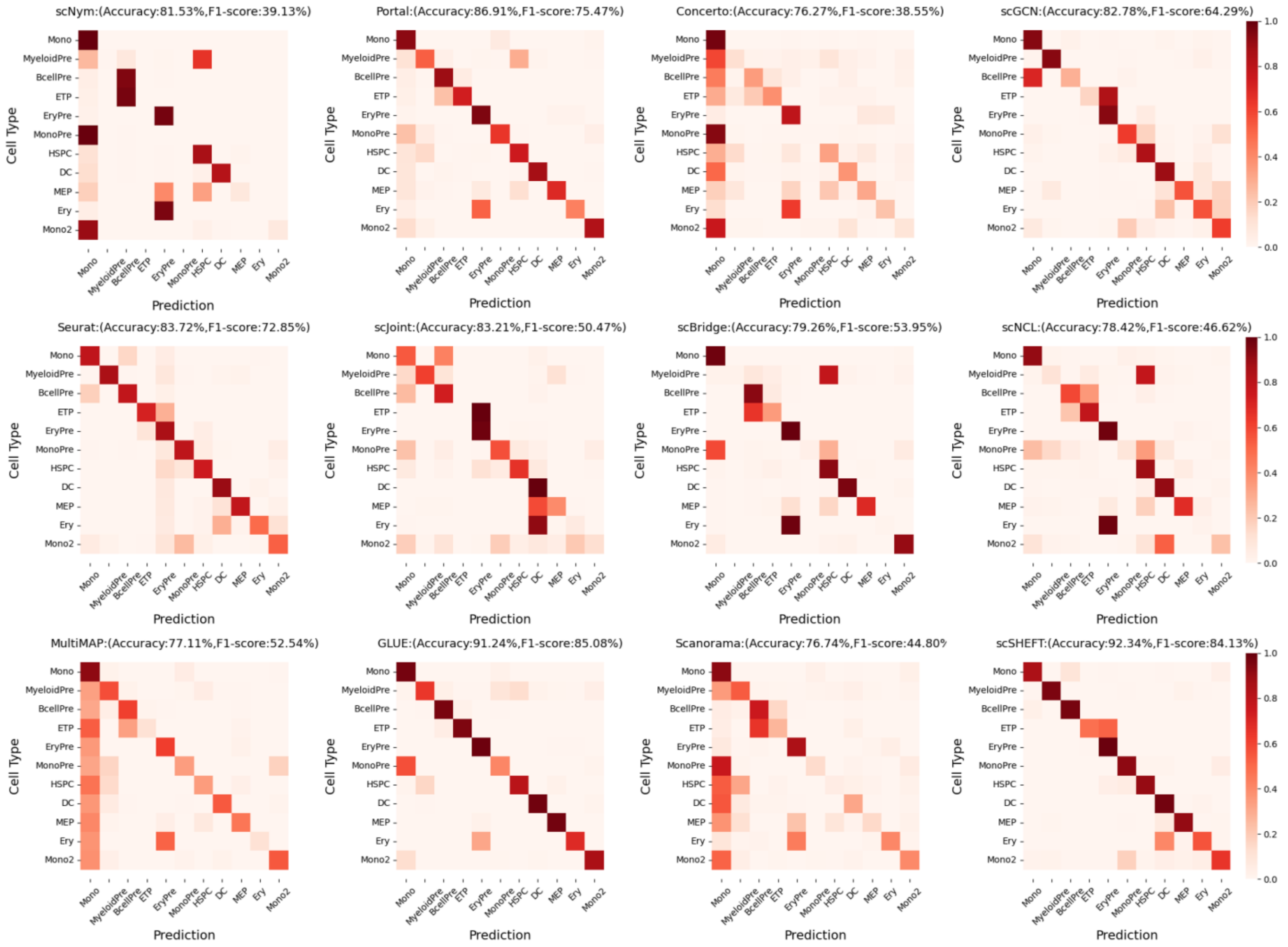
Supplemental Figures



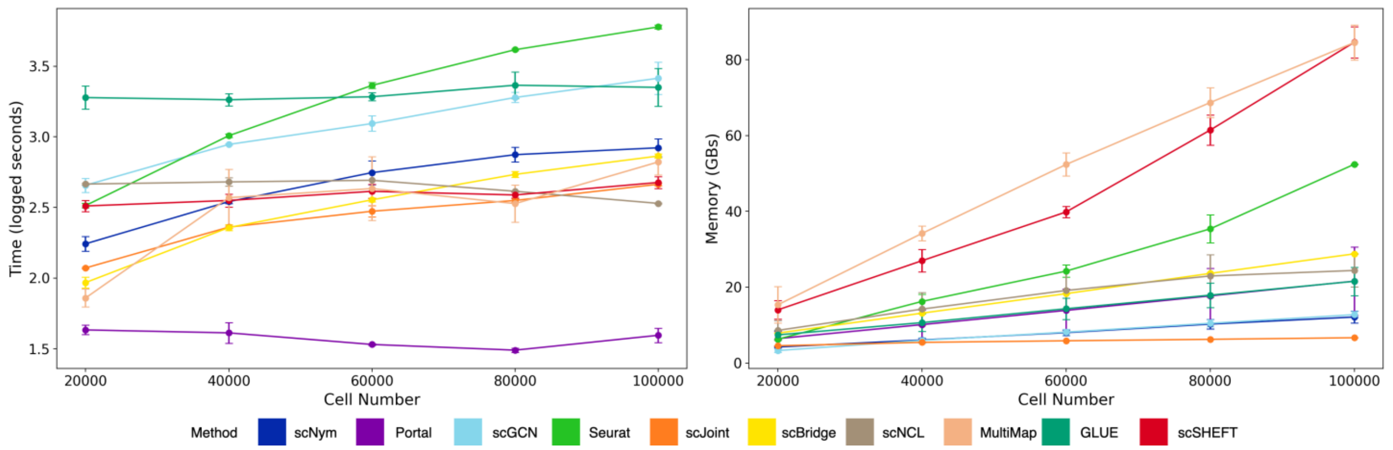
Supplemental Figure S1. Comparison of Davies-Bouldin Index (DBI) across different cell types (ETP, Mono2, Ery) and the entire T cell bone marrow dataset (ALL) using three methods (scSHEFT, scGCN, scJoint). The metrics are computed based on original cell embeddings to evaluate the clustering performance. Lower DBI values indicate better clustering performance.



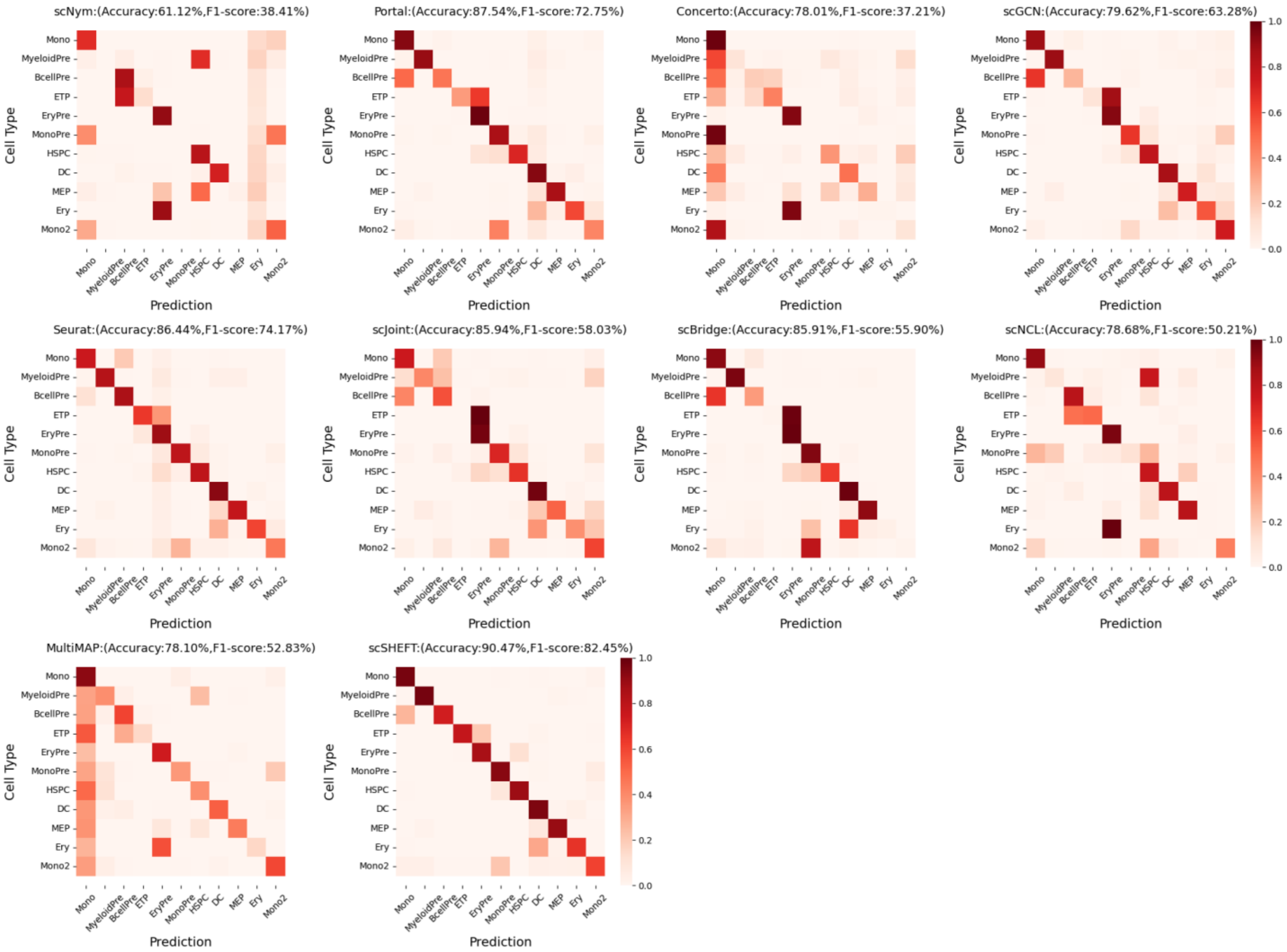
Supplemental Figure S2. *FECH* expression dynamics along diffusion-inferred pseudotime across methods. Dots represent individual cells colored by cell type. The red line denotes the LOWESS-fitted trend of *FECH* expression along pseudotime. The Spearman's correlation (ρ) between *FECH* expression and pseudotime quantifies trajectory continuity, with all correlations statistically significant (p < 1e-5). Higher correlation values reflect better capture of erythroid differentiation progression.



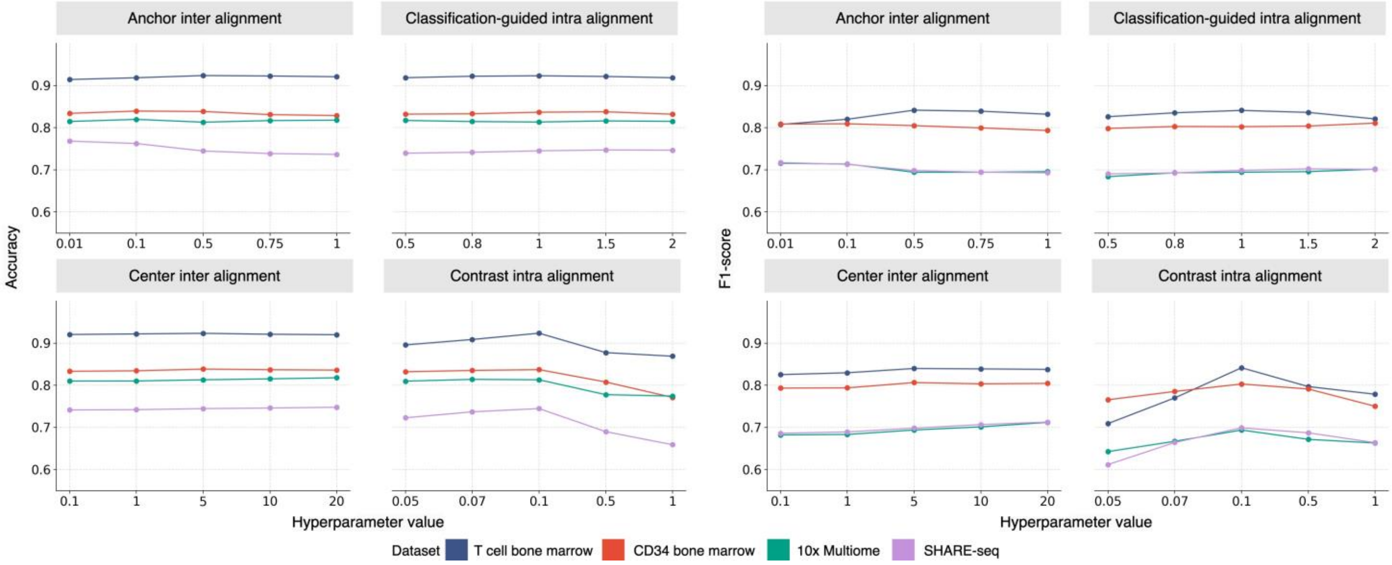
Supplemental Figure S3. Label transfer matrices comparing predicted labels with ground-truth annotations for scSHEFT and eleven baselines on the T cell bone marrow dataset. A clearer diagonal structure indicates better label transfer performance.



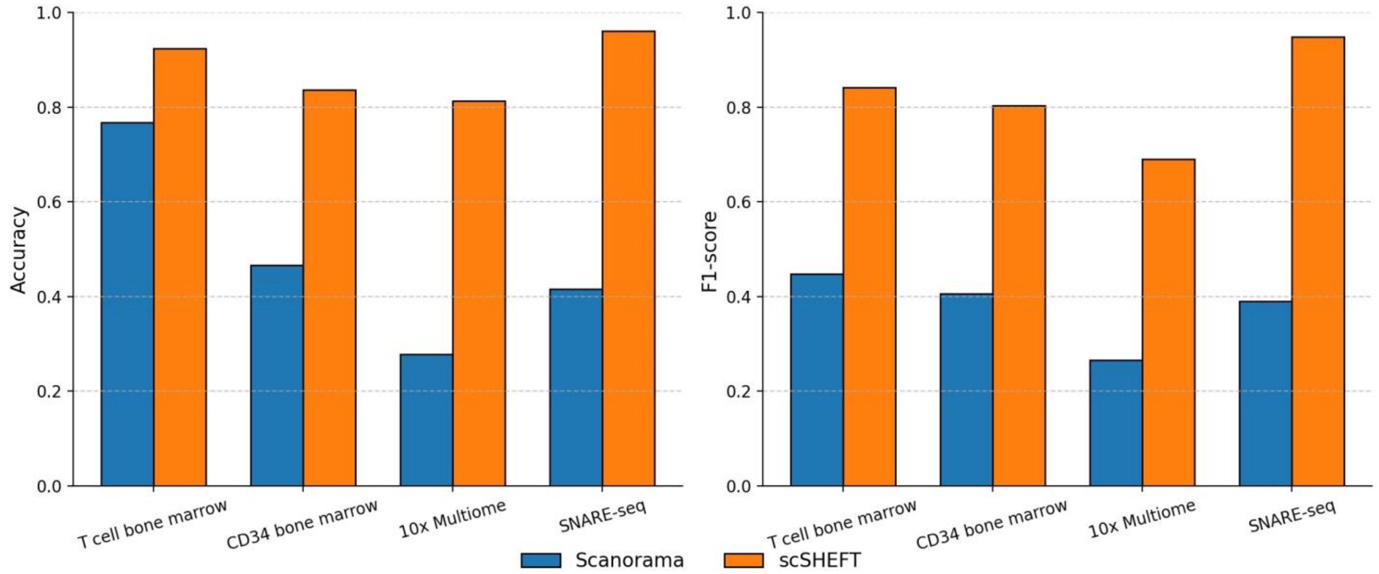
Supplemental Figure S4. The running time and memory usage of different methods on subsets of the PBMC dataset with 20,000 to 100,000 cells.



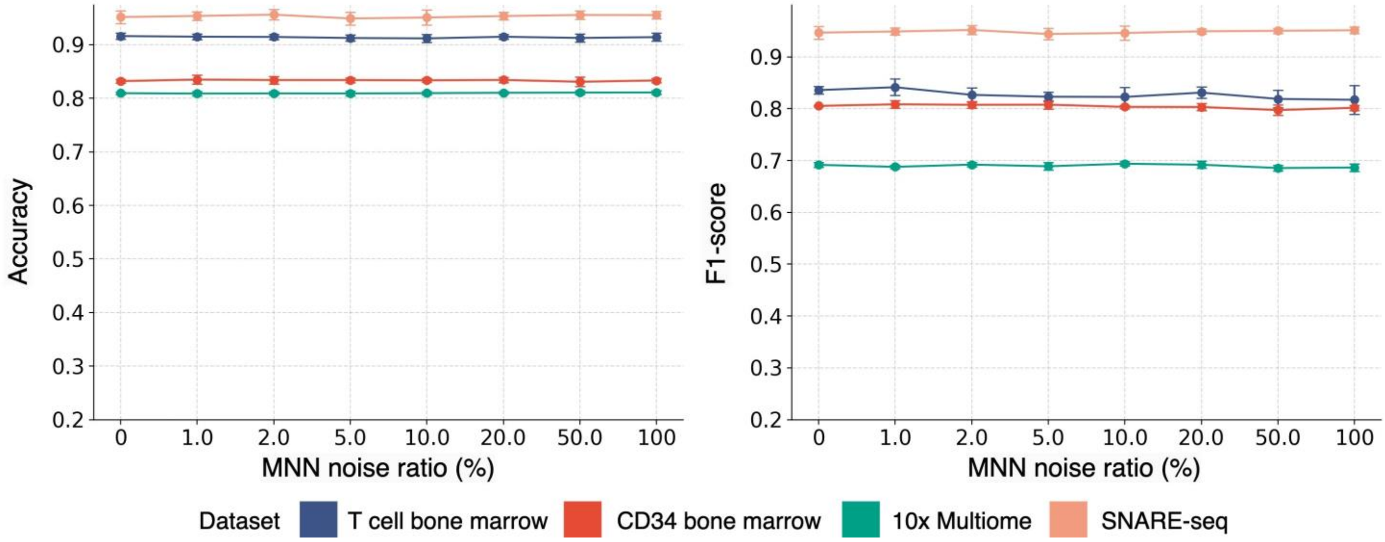
Supplemental Figure S5. Label transfer matrices comparing predicted labels with ground-truth annotations for scSHEFT and nine baselines under 50% dropout on scRNA-seq data. A clearer diagonal structure indicates better label transfer performance.



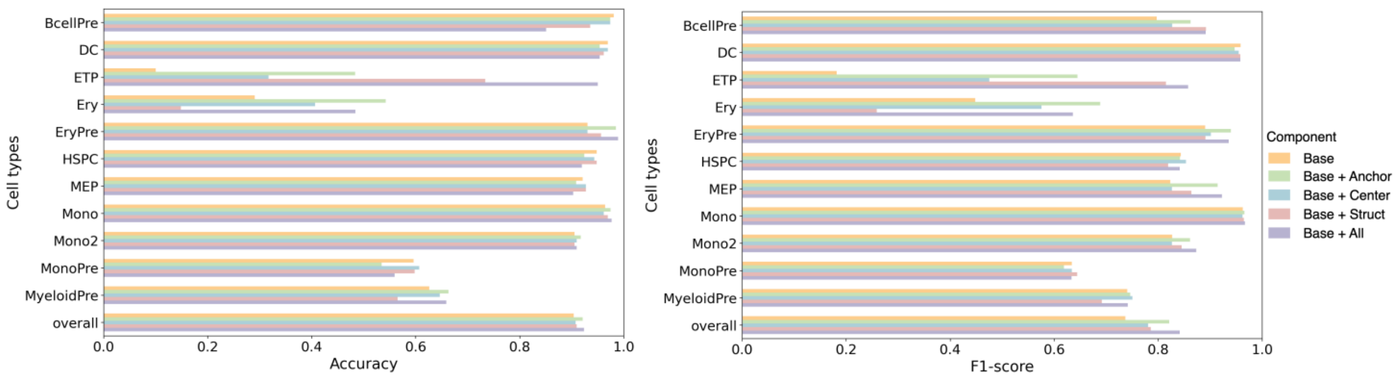
Supplemental Figure S6. Label transfer performance of scSHEFT under different hyperparameter settings.



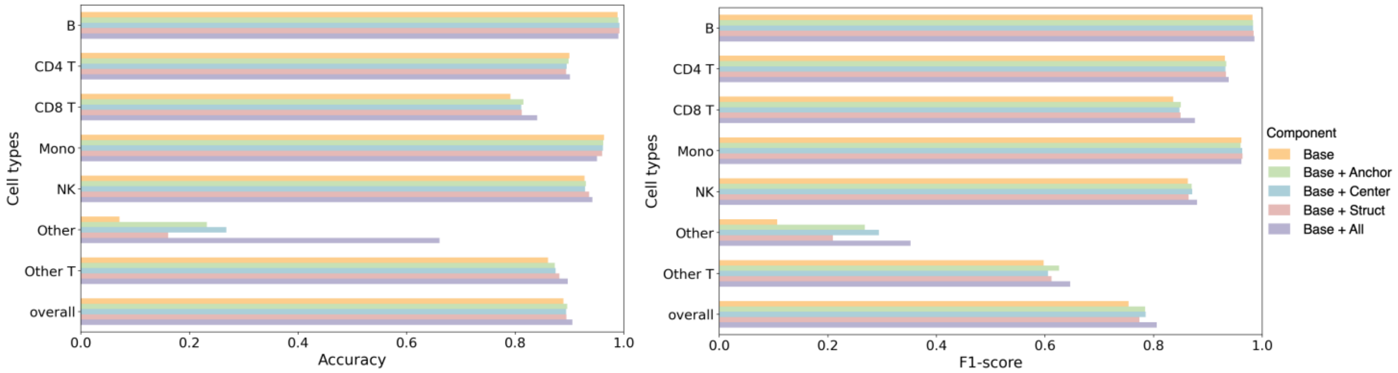
Supplemental Figure S7. Comparison of scSHEFT and the MNN-based baseline method Scanorama across datasets.



Supplemental Figure S8. Label transfer performance of scSHEFT with varying simulated MNN noise across different datasets (n=5 repeats with different model random seeds). Error bars indicate mean ± s.d.



Supplemental Figure S9. Cell-type-specific Accuracy and F1-score for ablation analysis on the T cell bone marrow (paired) dataset.



Supplemental Figure S10. Cell-type-specific Accuracy and F1-score for ablation analysis on the PBMC (unpaired) dataset.

	CrossEntropy loss	InfoNCE loss	Anchor loss	Center loss	Fusion embedding
Base	√	√	-	-	-
Base+Anchor	√	√	√	-	-
Base+Center	√	√	-	√	-
Base+Struct	√	√	-	-	√
Base+All	√	√	√	√	√

Supplemental Table S1. Overview of component inclusion across baseline and ablation variants.

Supplemental Notes

Baseline Methods

scNym: Python package scNym (v0.3.2) was used for all datasets. The gene expression matrix (GEM) from scRNA-seq and the gene activity score (GAS) from scATAC-seq were used as input, normalized (size_factor=1e6), and log-transformed before training. For the CITE-ASAP dataset, the preprocessed GEM/GAS matrix was concatenated with the log-normalized ADT matrix prior to training. Model training and cell type prediction were performed using the *scnym.api.scnym_api* function with default parameters.

Portal: Python package Portal (v1.0.2) was used for all datasets. The GEM from scRNA-seq and the GAS from scATAC-seq were used as input and preprocessed with the *portal.model.Model.preprocess* function using default parameters. For the CITE-ASAP dataset, the preprocessed GEM/GAS matrix was concatenated with the log-normalized ADT matrix before training. The model was then trained and evaluated with default settings. Based on the integrated latent representations of scRNA-seq and scATAC-seq, the *knn_classifier* function from Concerto's implementation was used to assign labels to scATAC-seq cells, providing both predicted labels and prediction confidence scores.

Concerto: Python package Concerto-reproducibility was used for all the datasets. The GEM from scRNA-seq and the GAS from scATAC-seq were used as input, and both were preprocessed using the *preprocessing_rna* function with *is_hvg* set to *False*. For the CITE-ASAP dataset, the preprocessed GEM/GAS matrix was concatenated with the ADT matrix before training. Concerto supports two approaches for label transfer: query-to-reference mapping and semi-supervised learning. We evaluated both and found that semi-supervised learning generally achieved better performance. The preprocessed GEM and GAS were saved in tfrecord format using the *concerto_make_tfrecord_supervised* function. Model training was performed with *concerto_train_inter_supervised_uda*, and prediction with *concerto_test_inter_supervised*, yielding integrated latent representations for scRNA-seq and scATAC-seq. Finally, cell type annotation for scATAC-seq was performed using the *knn_classifier* function (k=30), which provided both predicted labels and prediction confidence scores.

scGCN: Python package scGCN was used for all the datasets. Following the official tutorials, Seurat v4.3.0 was used to prepare input data for scGCN. Specifically, the *save_processed_data* function in *data_preprocess_utility.R* was used to preprocess raw GEM and GAS, and to construct intra- and inter-modality graphs. Because scGCN's strategy for selecting highly variable genes is time-consuming, we subsampled the scRNA-seq data to 10,000 cells (if the dataset contained more than 10,000 cells) during gene selection. For the CITE-ASAP dataset, the GEM/GAS matrix was concatenated with the ADT matrix before preprocessing. Model training and inference were performed using the default settings in the released code. For novel cell type detection, the *metrics* function in *data_preprocess_utility.R* was used to compute the entropy score (H) and enrichment score (E) for all ATAC cells, and the prediction confidence p was calculated as $E - H$.

Seurat: R package Seurat (v.4.3.0) was used for all the datasets. The GEM from scRNA-seq and the GAS from scATAC-seq were used as input and normalized using the *NormalizeData* function. For the CITE-ASAP dataset, the log-normalized GEM/GAS matrix was concatenated with the log-normalized ADT matrix. The *FindVariableFeatures* function (method = *"vst"*) was used to select the top 4,000 most variable genes from the scRNA-seq data. Anchors between scRNA-seq and scATAC-seq were identified using the *FindTransferAnchors* function with 'cca' reduction. Cell type annotations were then transferred from scRNA-seq to scATAC-seq using the *TransferData* function. Prediction confidence scores p were directly exported from the transfer results.

scJoint: Python package scJoint was used for all the datasets. The GEM from scRNA-seq and the GAS from scATAC-seq were binarized using zero thresholding. For the CITE-ASAP dataset, the binarized GEM/GAS matrix was concatenated with the log-transformed, normalized ADT matrix prior to training. Training parameters followed the configuration notes provided in the official GitHub repository. Prediction confidence scores p were calculated as described in the original paper.

scBridge: Python package scBridge was used for all the datasets. The GEM from scRNA-seq and the GAS from scATAC-seq were used as inputs. GEM was preprocessed by normalization, log transformation, and scaling, while GAS was processed with TF-IDF transformation and scaling, following the default scBridge procedure. For the CITE-ASAP dataset, the preprocessed GEM/GAS matrix was concatenated with the log-normalized ADT matrix before training.

Training parameters followed the configuration notes provided in the official GitHub repository. Prediction confidence scores p were calculated as described in the original paper.

scNCL: Python package scNCL was used for all the datasets. The GEM from scRNA-seq and the GAS from scATAC-seq were binarized using zero thresholding. For the CITE-ASAP dataset, the binarized GEM/GAS matrix was concatenated with the log-transformed, normalized ADT matrix before training. Training parameters followed the configuration notes provided in the official GitHub repository. Prediction confidence scores p were calculated as described in the original paper.

MultiMap: Python package MultiMap was used for all the datasets. The GEM from scRNA-seq, and both the GAS and peak count data from scATAC-seq, were used as inputs. GEM was preprocessed by normalization, log transformation, scaling, and principal component analysis (PCA), with results stored in .obsm['X_pca'] for downstream integration. Peak count data was processed with TF-IDF transformation followed by LSI dimensionality reduction, stored in .obsm['X_lsi'], as described in the MultiMap tutorial. These precomputed reductions were integrated using the *MultiMAP.Integration* function. The integrated latent representations were then used to annotate scATAC-seq cells using the *sklearn.neighbors.KNeighborsClassifier* function with neighborhood size *k* set to 5.

GLUE: Python Package GLUE (v 0.3.2) was used for all datasets. The GEM from scRNA-seq and the peak count data from scATAC-seq were used as input. GEM was preprocessed by normalization, log transformation, scaling, and principal component analysis (PCA). The scATAC-seq data was reduced to 100 dimensions using the *scglue.data.lsi* function. To construct a guidance graph of feature interactions, the *scglue.data.get_gene_annotation* function was used to supplement gene coordinate information from GTF files, and the *scglue.genomics.rna_anchored_guidance_graph* function was used to compute the guidance graph. The *scglue.model.configure_dataset* function was used to configure the training and test datasets with default parameters. Although the training configuration includes the *use_cell_type* option for supervised classification on scRNA-seq data, we found that enabling this option degraded performance, so it was not used. Before training, a subgraph was extracted from the guidance graph. Model training was performed using the *scglue.models.fit_SCGLUE* function with the extracted subgraph. Finally, the integrated latent representations were used to annotate scATAC-seq cells using the *sklearn.neighbors.KNeighborsClassifier* function with neighborhood size *k* set to 5.

Scanorama: Python Package Scanorama (v 1.7.4) was used for all datasets. The GEM from scRNA-seq and the GAS from scATAC-seq were used as inputs and concatenated, with a domain field added to distinguish the two modalities. Both datasets were preprocessed by normalization, log transformation, selection of the top 2,000 most variable genes, and principal component analysis (PCA). Data integration was performed using the *sce.pp.scanorama_integrate* function. Finally, integrated latent representations were used to annotate scATAC-seq cells using the *sklearn.neighbors.KNeighborsClassifier* function with neighborhood size k set to 5.

scSHEFT: The GEM from scRNA-seq, as well as the GAS and peak count data from scATAC-seq, were used as inputs. GEM and GAS were preprocessed by normalization, log transformation, and scaling. For the CITE-ASAP dataset, the binarized GEM/GAS matrix was concatenated with the log-transformed, normalized ADT matrix prior to training. To obtain low-dimensional representations of the raw scATAC-seq data, different approaches were used: for the CITE-ASAP dataset, the ADT matrix served as the low-dimensional representation of the ASAP data; for other datasets, following Seurat's tutorial, we computed dimension-reduced representations (LSI) of the raw scATAC-seq data by first calculating the TF-IDF matrix and then performing singular value decomposition (SVD) on the TF-IDF matrix.

Details of data preprocessing

scSHEFT accepts the GEM of scRNA-seq and the GAS and peak count data of scATAC-seq as the inputs. The GEM and GAS were preprocessed by normalization, log transformation and scaling. Note that for the CITE-ASAP dataset, the binarized GEM/GAS matrix and the log-transformed, normalized ADT matrix were concatenated before training. The detailed preprocessing steps for each dataset are elaborated below:

- **Human SNARE-seq dataset.** The gene expression, gene activity, and peak-by-cell matrices were downloaded from NCBI GEO accession number GSE126074. Additionally, the human v44 GTF files were obtained for downstream analysis. The GAS data was generated using Episcanpy v0.4.0, resulting in a dataset with 1,017 cells and 8,459 common genes for the analysis.

- **T cell bone marrow multiome dataset.** The T cell bone marrow multiome dataset is provided in GSE200046. For convenience, we downloaded the filtered and processed count matrices, including cell type annotations and ATAC fragment files from Zenodo (https://doi.org/10.5281/zenodo.6383269). Additionally, the human v44 GTF files were obtained for downstream analysis. The GAS data was generated using Episcanpy v0.4.0, resulting in a dataset with 7,439 cells and 13,916 common genes for analysis.

- **The CD34 bone marrow multiome dataset.** The CD34 bone marrow multiome dataset is provided in GSE200046. For convenience, we downloaded the filtered and processed count matrices, including cell type annotations and ATAC fragment files from Zenodo (https://doi.org/10.5281/zenodo.6383269). Additionally, the human v44 GTF files were obtained for downstream analysis. The GAS data was generated using Episcanpy v0.4.0, resulting in a dataset with 6,881 cells and 12,066 common genes for analysis.

- **The CITE-ASAP dataset.** The original CITE-seq data, ASAP-seq data, and fragments file are provided in GSE15647838. For convenience, we downloaded the preprocessed data provided in https://github.com/SydneyBioX/scJoint/blob/main/data.zip, which contains 4,644 CITE-seq and 4,506 ASAP-seq cells of 7 common types.

- **Mouse SHARE-seq skin dataset.** The gene expression, gene activity, and peak-by-cell matrices were downloaded from NCBI GEO accession number GSE140203. For convenience, we downloaded the preprocessed data provided in https://scglue.readthedocs.io/en/latest/data.html (referred to Ma-2020). Additionally, the mouse vM33 GTF files were obtained for downstream analysis. The GAS data was generated using Episcanpy v0.4.0, resulting in a dataset with 32,231 cells and 16,375 common genes for analysis.

- **The 10x Multiome dataset.** The gene expression, peak-by-cell matrix, and fragments file were downloaded from https://www.10xgenomics.com/cn/datasets/fresh-frozen-lymph-node-with-b-cell-lymphoma-14-k-sorted-nuclei-1-standard-1-0-0. Additionally, the human v44 GTF files were obtained for downstream analysis. Following the settings of scJoint, we first removed cell types with a proportion of less than approximately 1%. The GAS data was generated using Episcanpy v0.4.0, resulting in a dataset with 11,363 cells and 7,667 common genes for analysis.

- **The PBMC COVID-19 vaccine dataset.** The gene expression matrix, peak-by-cell matrix, and fragments file were downloaded from https://zenodo.org/records/8240488. We parsed the .rds files to extract the data and metadata, and converted them into h5ad format. Additionally, the human v44 GTF file was obtained for downstream analysis. Following the settings of scJoint, we first removed cell types with a proportion of less than approximately 1%. Gene activity scores (GAS) were generated using Episcanpy v0.4.0, resulting in a dataset containing 111,351 scRNA-seq cells, 77,810 scATAC-seq cells, and 15,233 common genes for analysis.