## Supplementary Figures

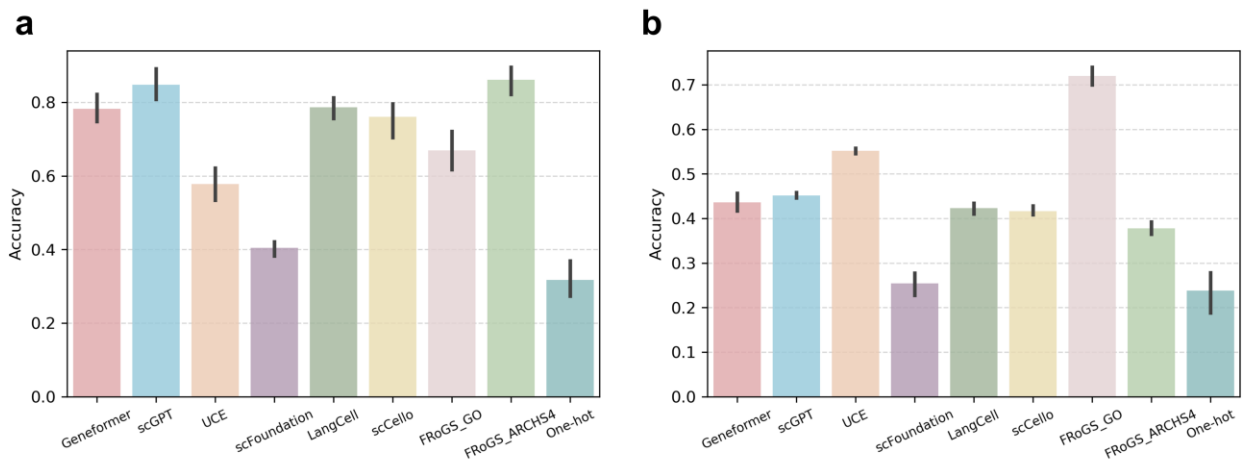



**Fig. S1. Benchmarking results on gene-level tasks.** a, Classification of tissue-specific genes. b, Classification of GO terms.

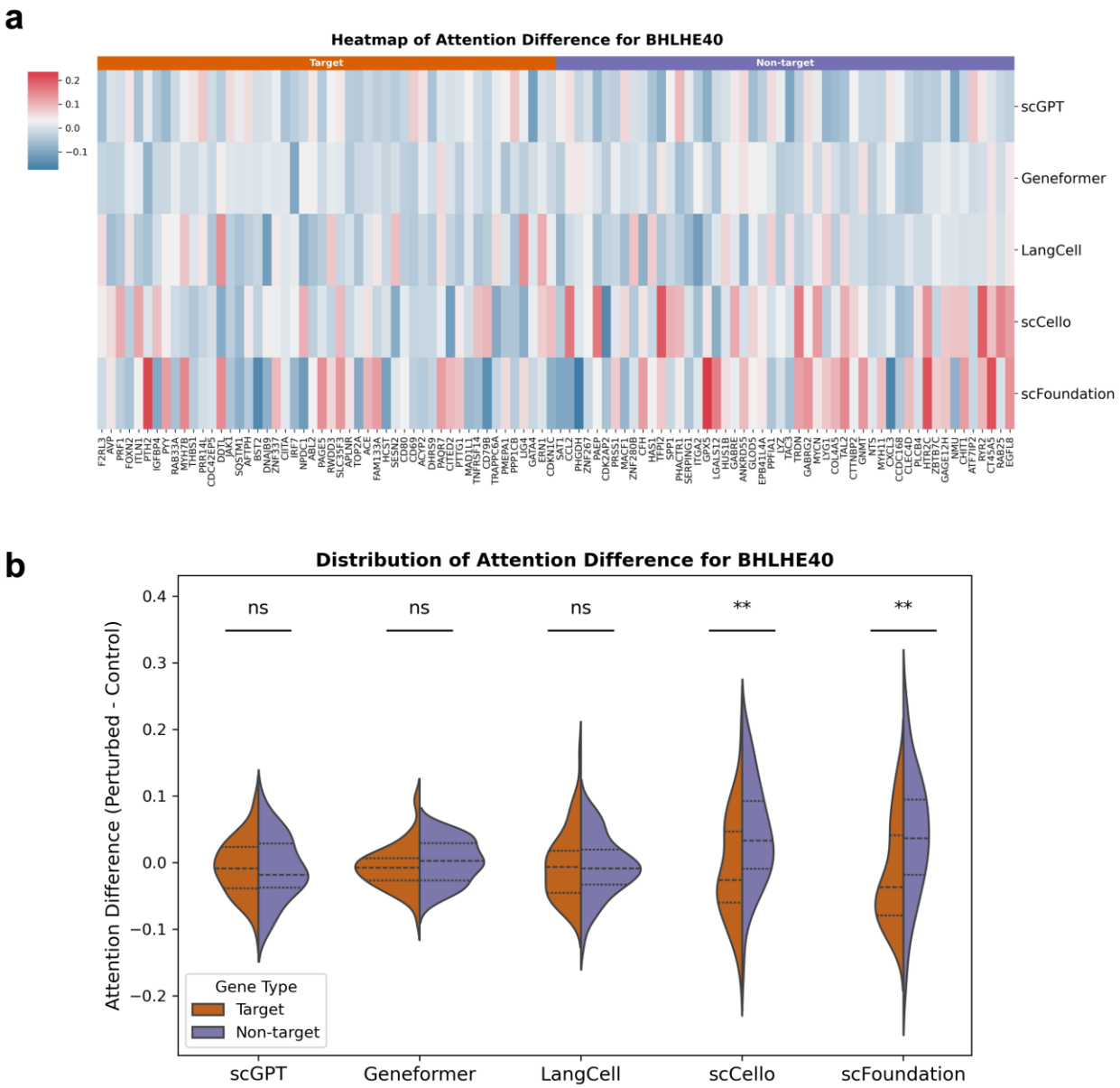



**Fig. S2. Visualization of attention difference (Perturbed - Control) from 50 target genes and 50 non-target genes to BHLHE40 using a heatmap (a) and violin plots (b).** The significance level was set as: ***, $p \leq 0.001$; **, $0.001 < p \leq 0.01$; *, $0.01 < p \leq 0.05$; ns, $p > 0.05$.

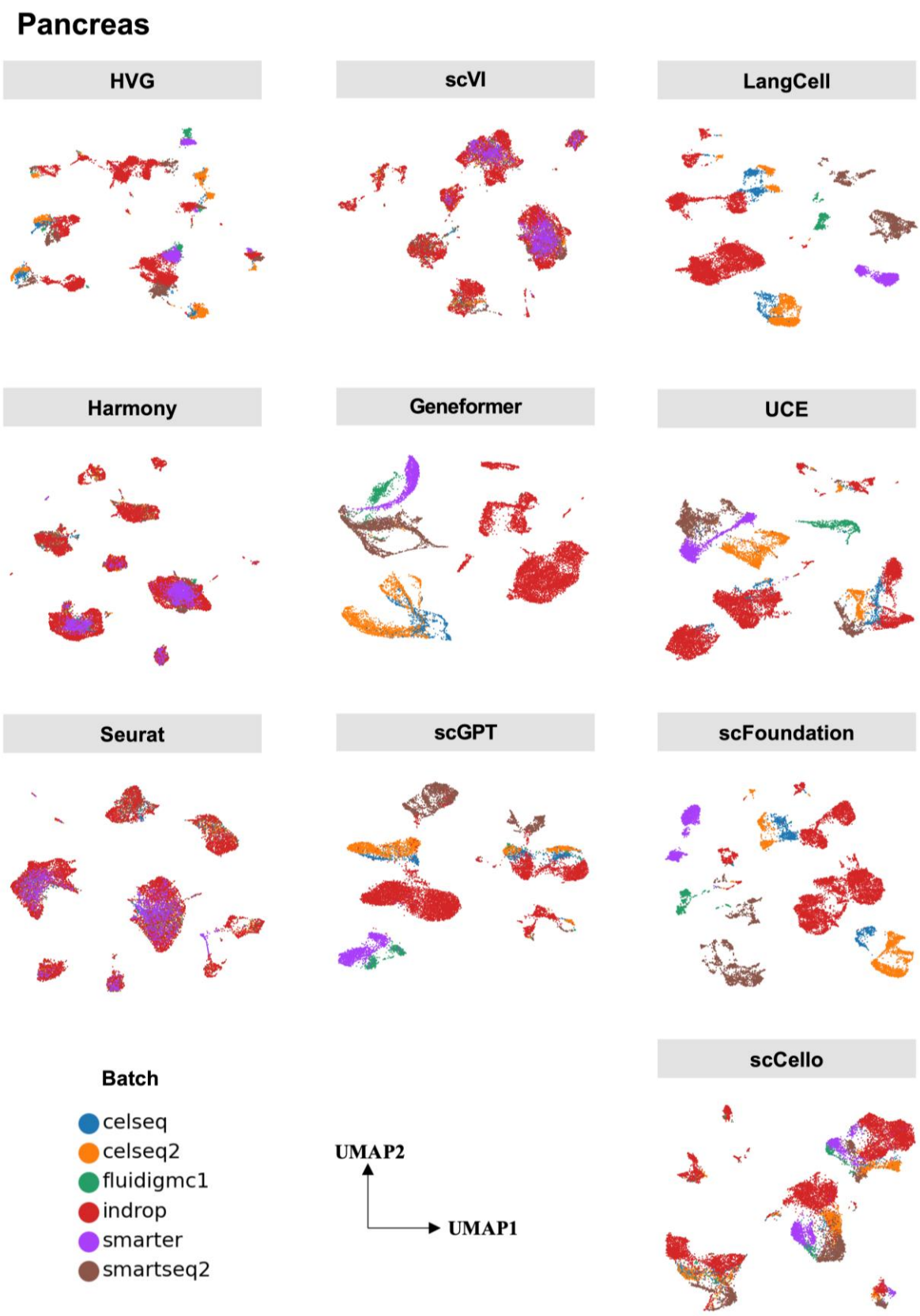



Fig. S3. UMAP visualization of batch integration results on the Pancreas dataset colored by batch labels.

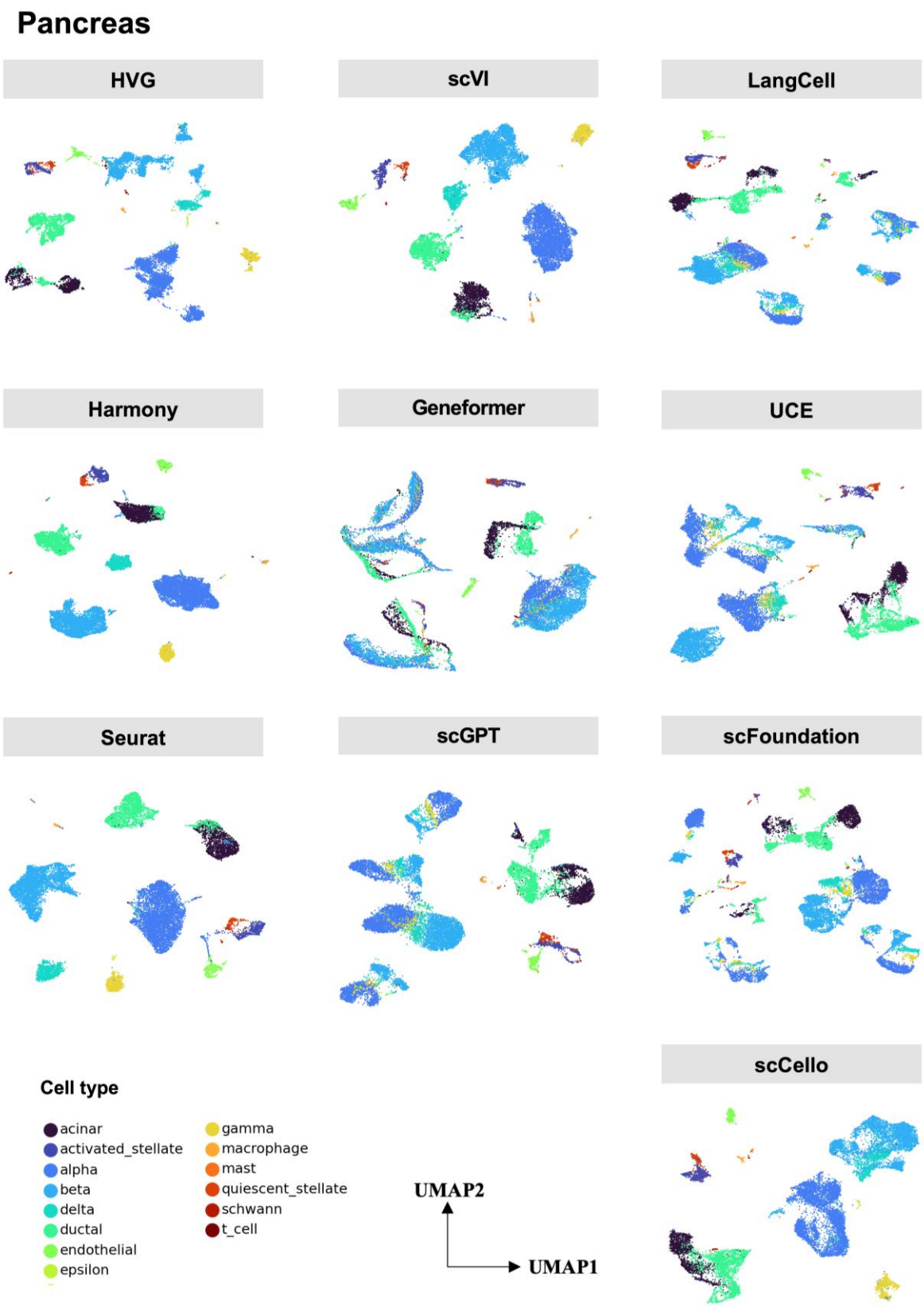



Fig. S4. UMAP visualization of batch integration results on the Pancreas dataset colored by cell types.

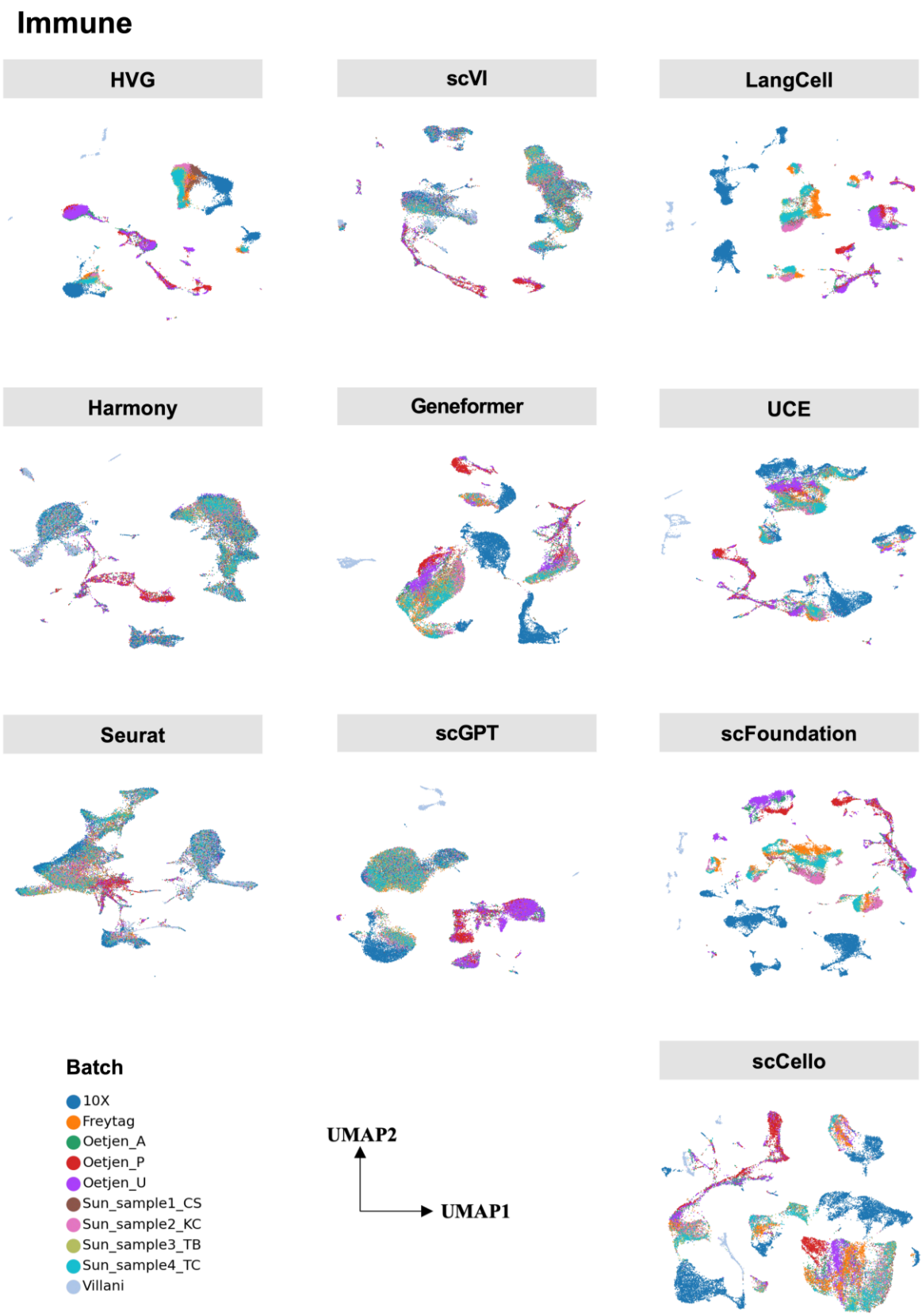



Fig. S5. UMAP visualization of batch integration results on the Immune (human) dataset colored by batch labels.

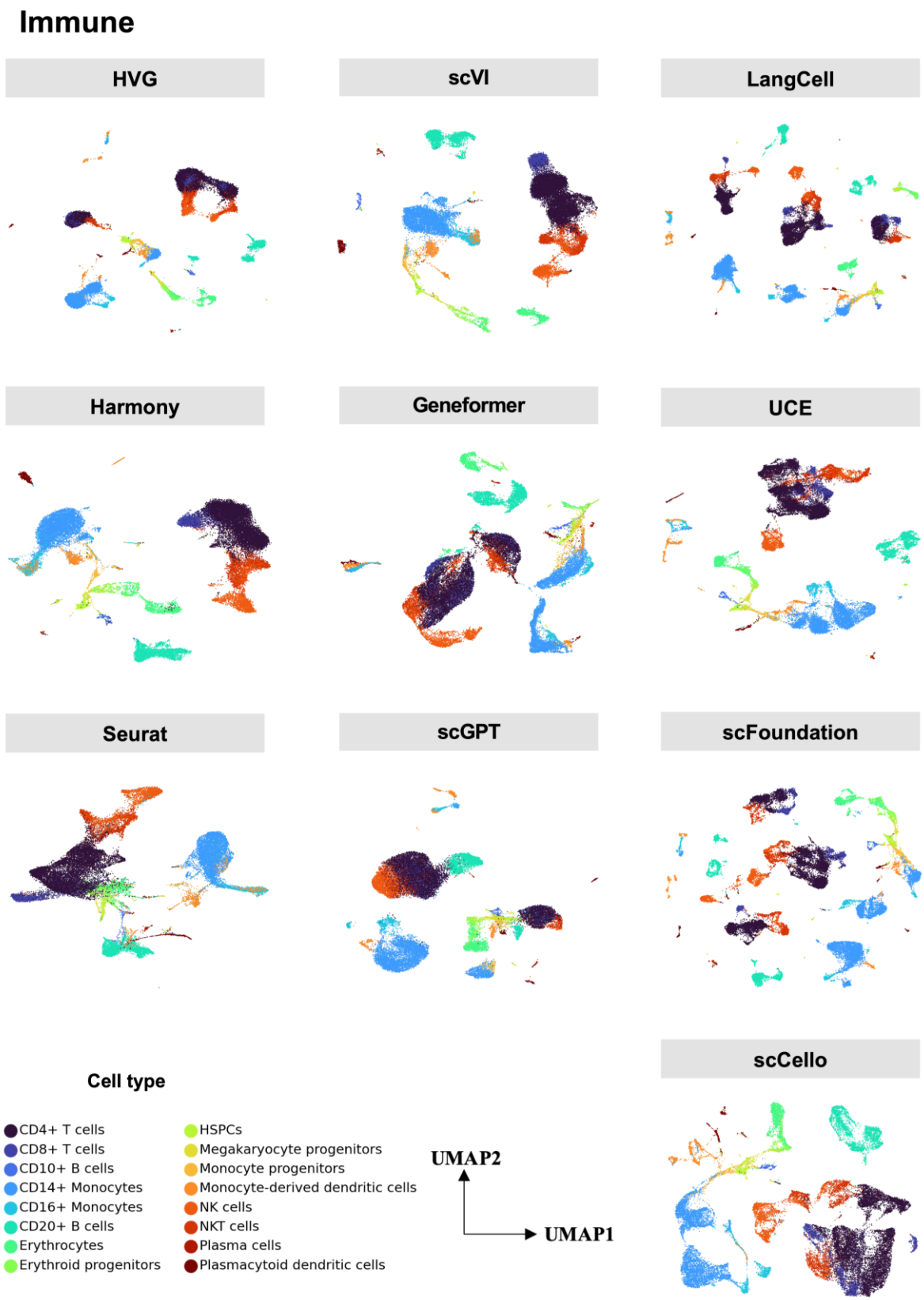



Fig. S6. UMAP visualization of batch integration results on the Immune (human) dataset colored by cell types.

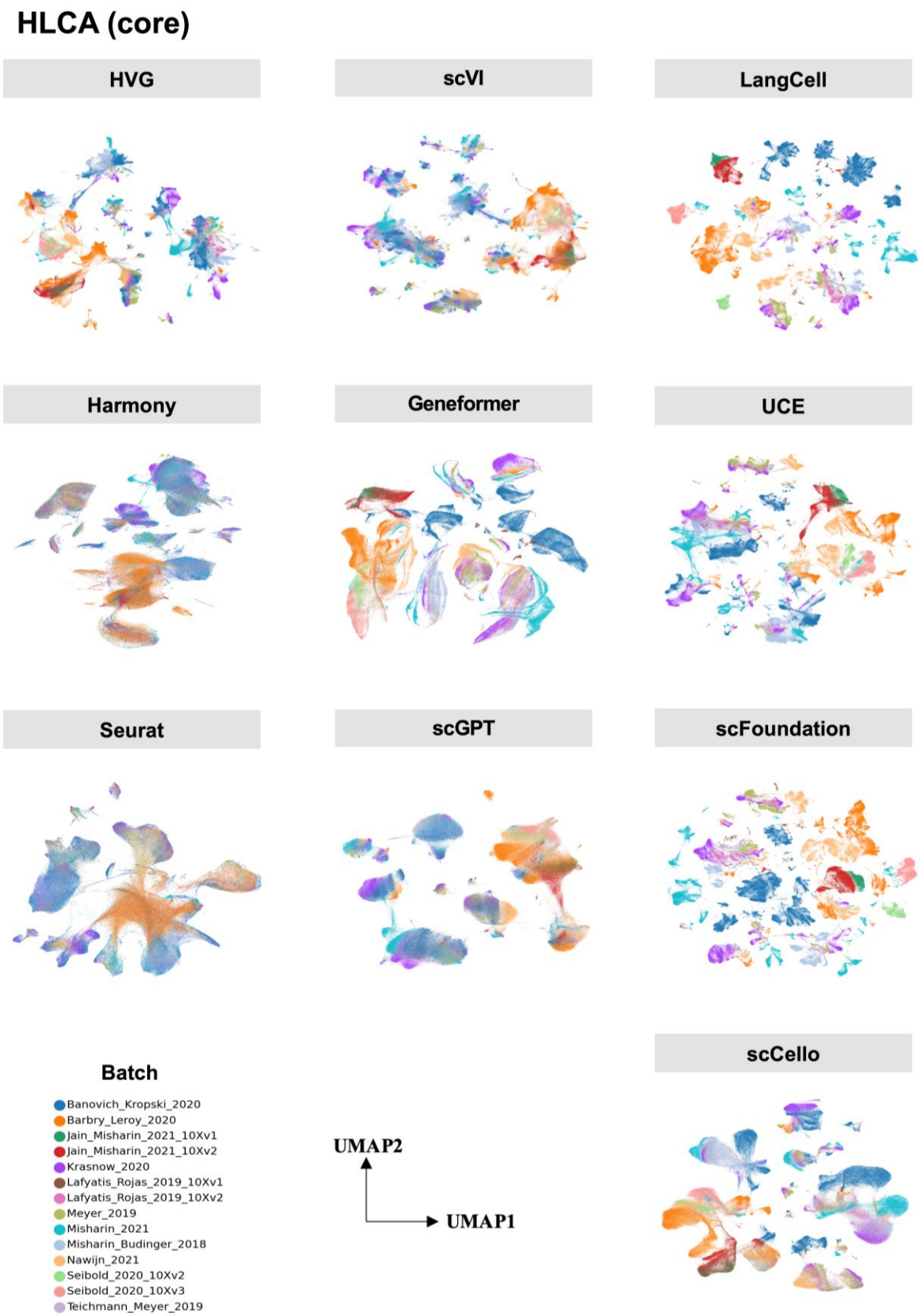



Fig. S7. UMAP visualization of batch integration results on the HLCA (core) dataset colored by batch labels.

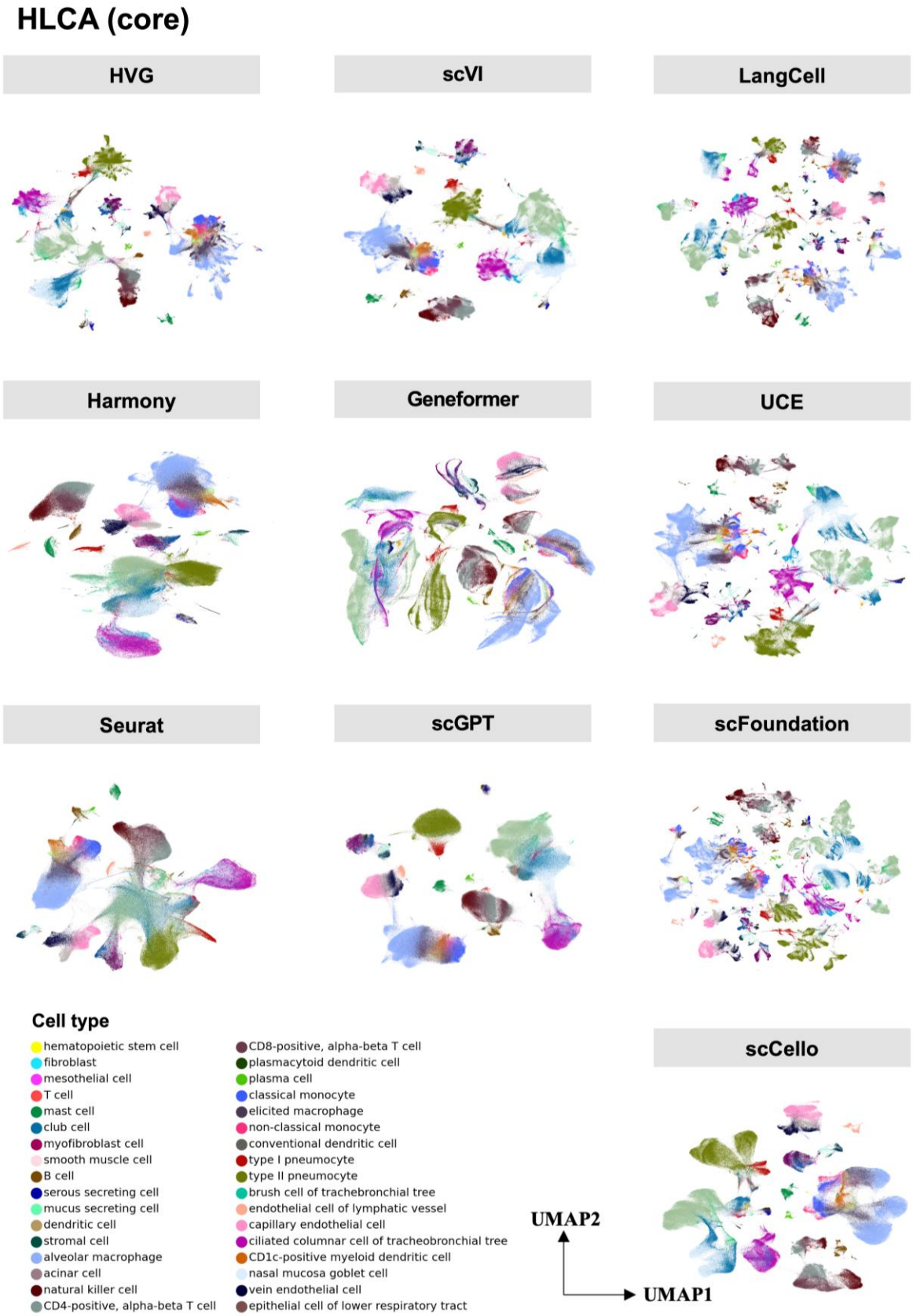



**Fig. S8. UMAP visualization of batch integration results on the HLCA (core) dataset colored by cell types.**

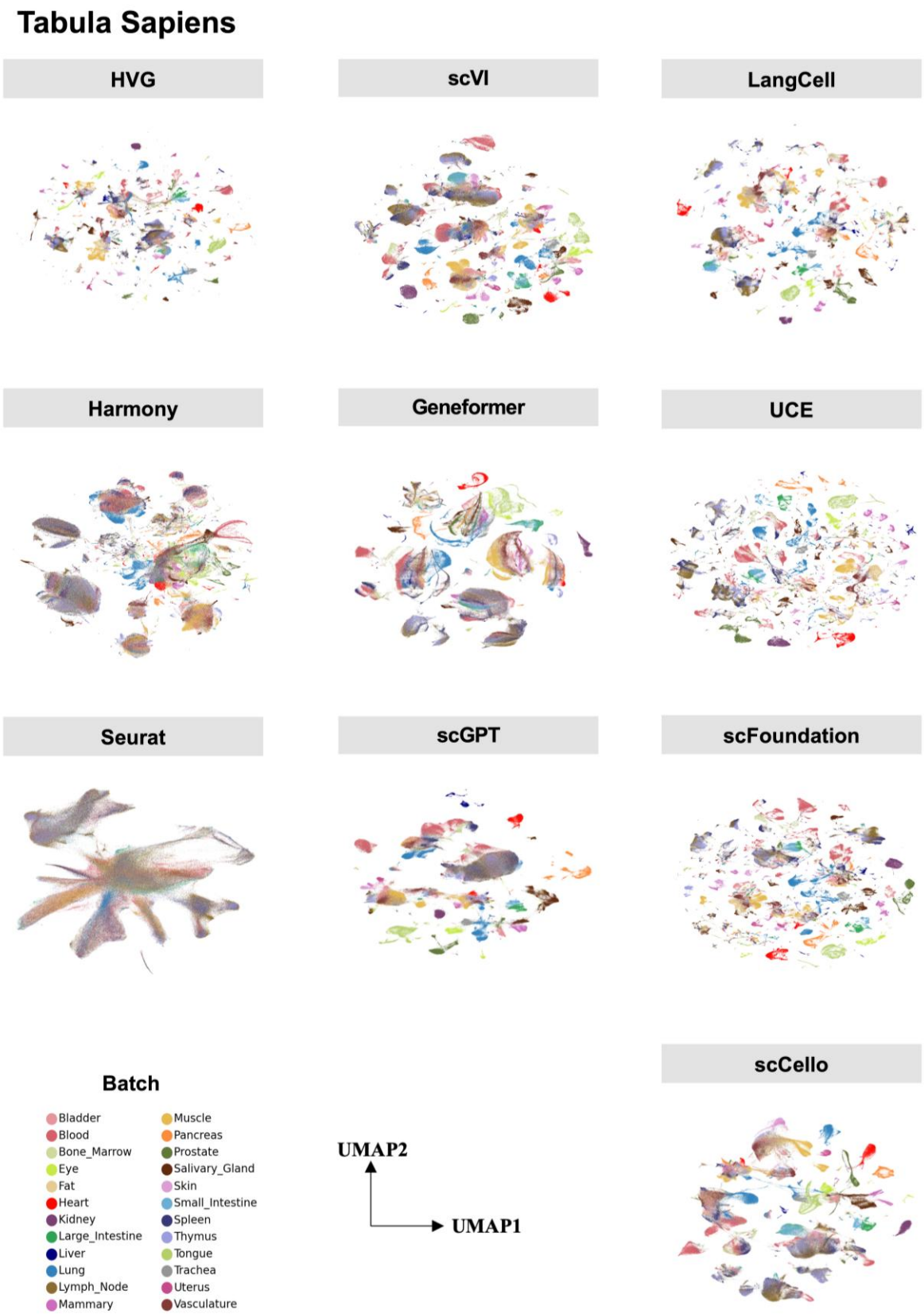



**Fig. S9. UMAP visualization of batch integration results on the Tabula Sapiens dataset colored by batch labels.**

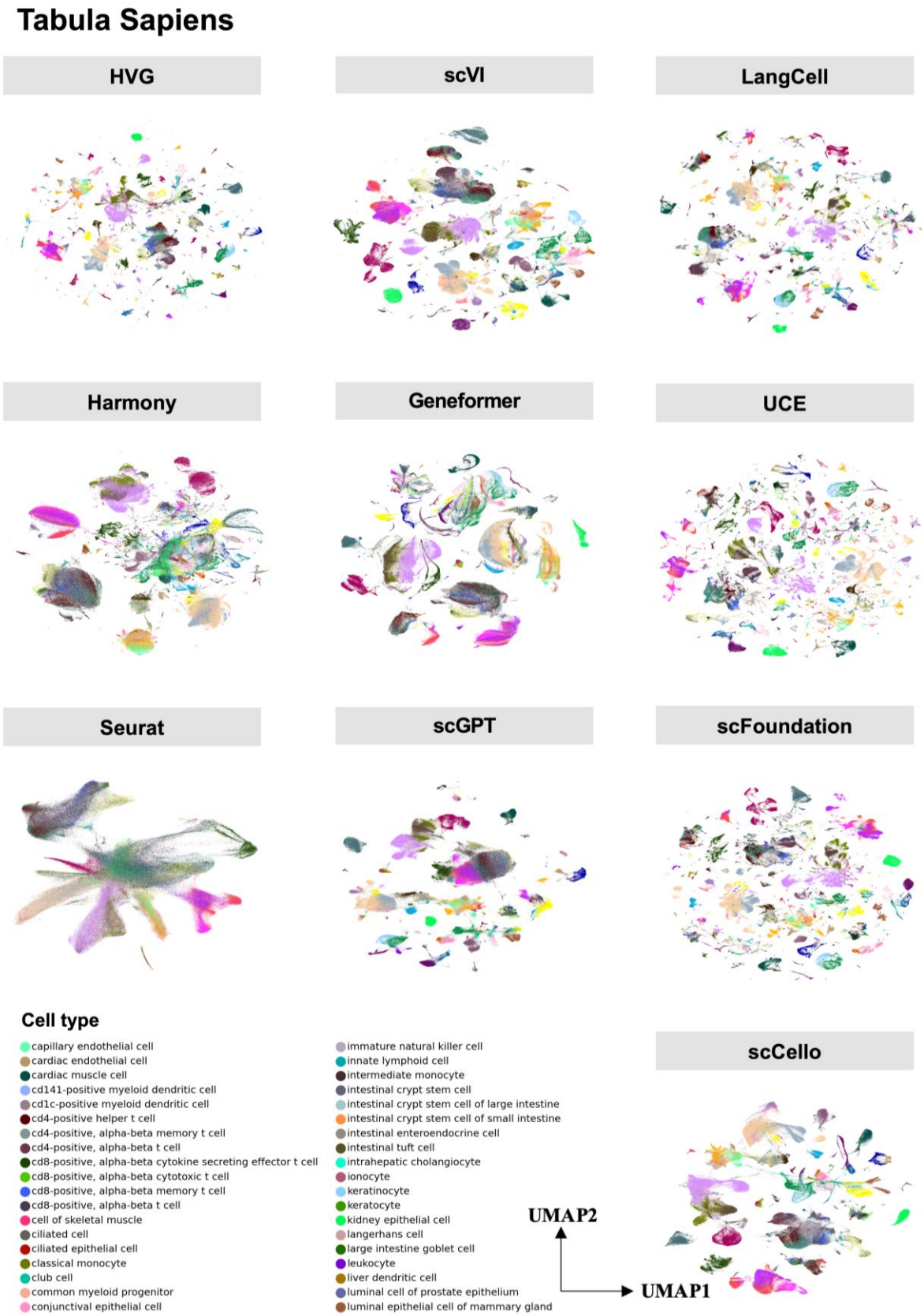



Fig. S10. UMAP visualization of batch integration results on the Tabula Sapiens dataset colored by cell types.

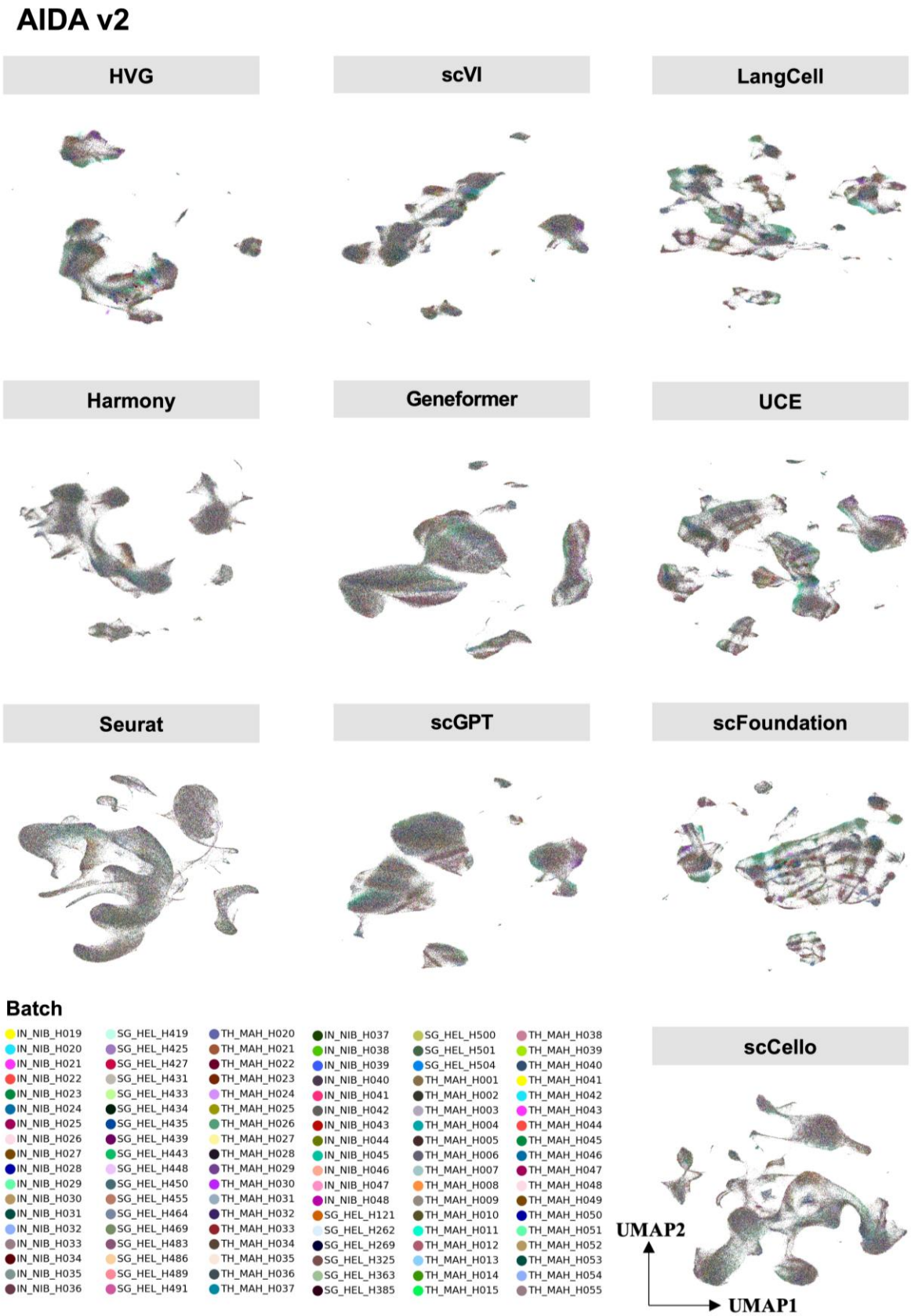



**Fig. S11. UMAP visualization of batch integration results on the AIDA v2 dataset colored by batch labels.**

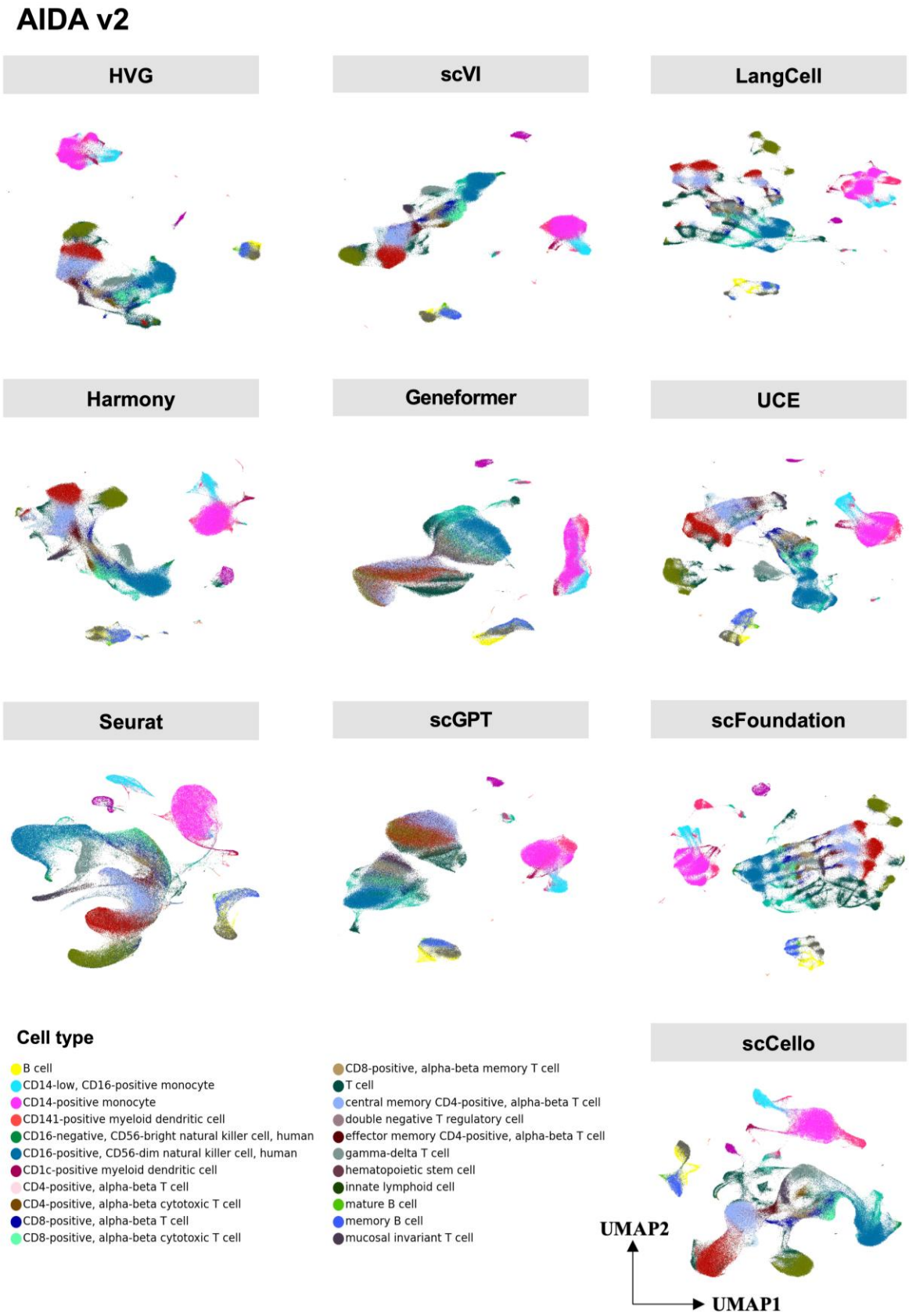
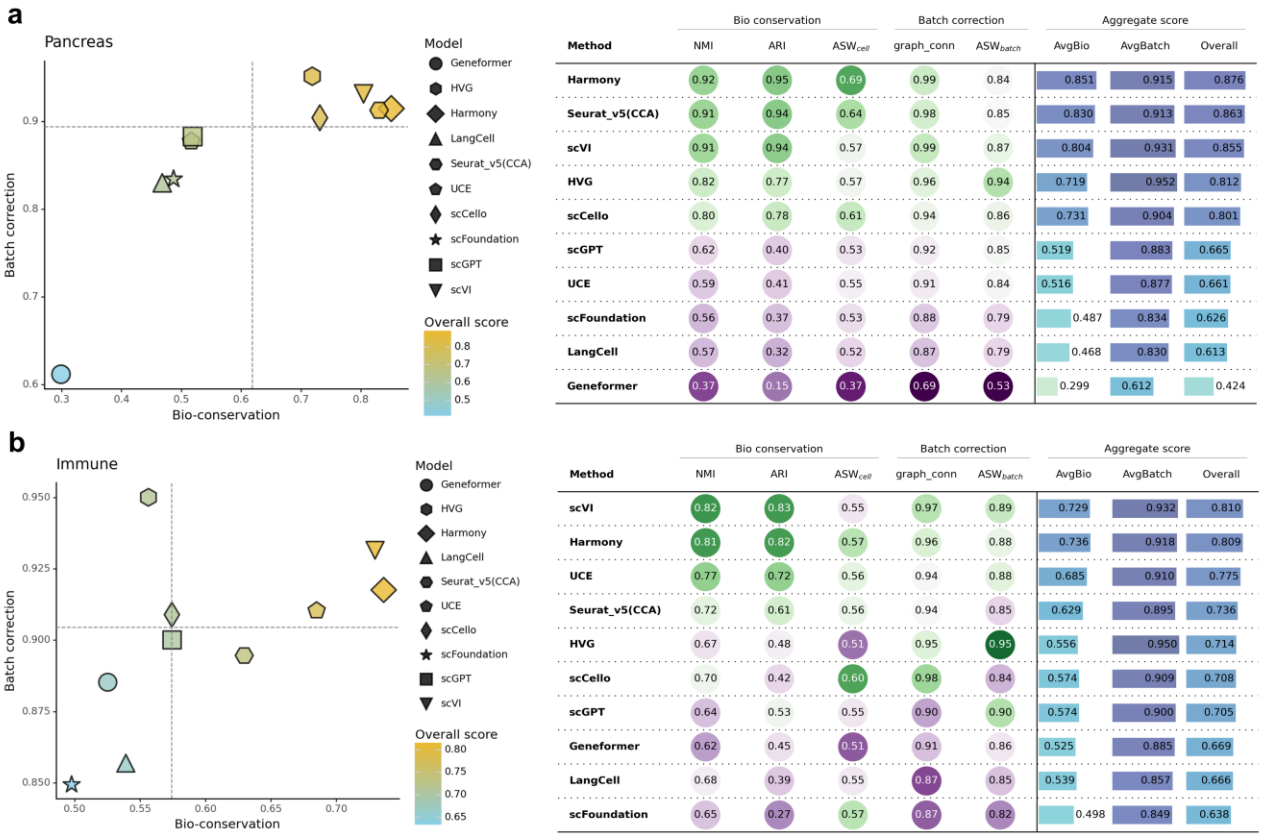



Fig. S12. UMAP visualization of batch integration results on the AIDA v2 dataset colored by cell types.




| Method | Bio conservation | | | Batch correction | | Aggregate score | | |
|---|---|---|---|---|---|---|---|---|
| | NMI | ARI | $ASW_{cell}$ | graph_conn | $ASW_{batch}$ | AvgBio | AvgBatch | Overall |
| **Harmony** | 0.92 | 0.95 | 0.69 | 0.99 | 0.84 | 0.851 | 0.915 | 0.876 |
| **Seurat_v5(CCA)** | 0.91 | 0.94 | 0.64 | 0.98 | 0.85 | 0.830 | 0.913 | 0.863 |
| **scVI** | 0.91 | 0.94 | 0.57 | 0.99 | 0.87 | 0.804 | 0.931 | 0.855 |
| **HVG** | 0.82 | 0.77 | 0.57 | 0.96 | 0.94 | 0.719 | 0.952 | 0.812 |
| **scCello** | 0.80 | 0.78 | 0.61 | 0.94 | 0.86 | 0.731 | 0.904 | 0.801 |
| **scGPT** | 0.62 | 0.40 | 0.53 | 0.92 | 0.85 | 0.519 | 0.883 | 0.665 |
| **UCE** | 0.59 | 0.41 | 0.55 | 0.91 | 0.84 | 0.516 | 0.877 | 0.661 |
| **scFoundation** | 0.56 | 0.37 | 0.53 | 0.88 | 0.79 | 0.487 | 0.834 | 0.626 |
| **LangCell** | 0.57 | 0.32 | 0.52 | 0.87 | 0.79 | 0.468 | 0.830 | 0.613 |
| **Geneformer** | 0.37 | 0.15 | 0.37 | 0.69 | 0.53 | 0.299 | 0.612 | 0.424 |

| Method | Bio conservation | | | Batch correction | | Aggregate score | | |
|---|---|---|---|---|---|---|---|---|
| | NMI | ARI | $ASW_{cell}$ | graph_conn | $ASW_{batch}$ | AvgBio | AvgBatch | Overall |
| **scVI** | 0.82 | 0.83 | 0.55 | 0.97 | 0.89 | 0.729 | 0.932 | 0.810 |
| **Harmony** | 0.81 | 0.82 | 0.57 | 0.96 | 0.88 | 0.736 | 0.918 | 0.809 |
| **UCE** | 0.77 | 0.72 | 0.56 | 0.94 | 0.88 | 0.685 | 0.910 | 0.775 |
| **Seurat_v5(CCA)** | 0.72 | 0.61 | 0.56 | 0.94 | 0.85 | 0.629 | 0.895 | 0.736 |
| **HVG** | 0.67 | 0.48 | 0.51 | 0.95 | 0.95 | 0.556 | 0.950 | 0.714 |
| **scCello** | 0.70 | 0.42 | 0.60 | 0.98 | 0.84 | 0.574 | 0.909 | 0.708 |
| **scGPT** | 0.64 | 0.53 | 0.55 | 0.90 | 0.90 | 0.574 | 0.900 | 0.705 |
| **Geneformer** | 0.62 | 0.45 | 0.51 | 0.91 | 0.86 | 0.525 | 0.885 | 0.669 |
| **LangCell** | 0.68 | 0.39 | 0.55 | 0.87 | 0.85 | 0.539 | 0.857 | 0.666 |
| **scFoundation** | 0.65 | 0.27 | 0.57 | 0.87 | 0.82 | 0.498 | 0.849 | 0.638 |

**Fig. S13. Detailed batch integration performance on the Pancreas dataset (a) and Immune (human) dataset (b) assessed by scIB metrics.** The dash lines indicate the median score across all evaluated models.

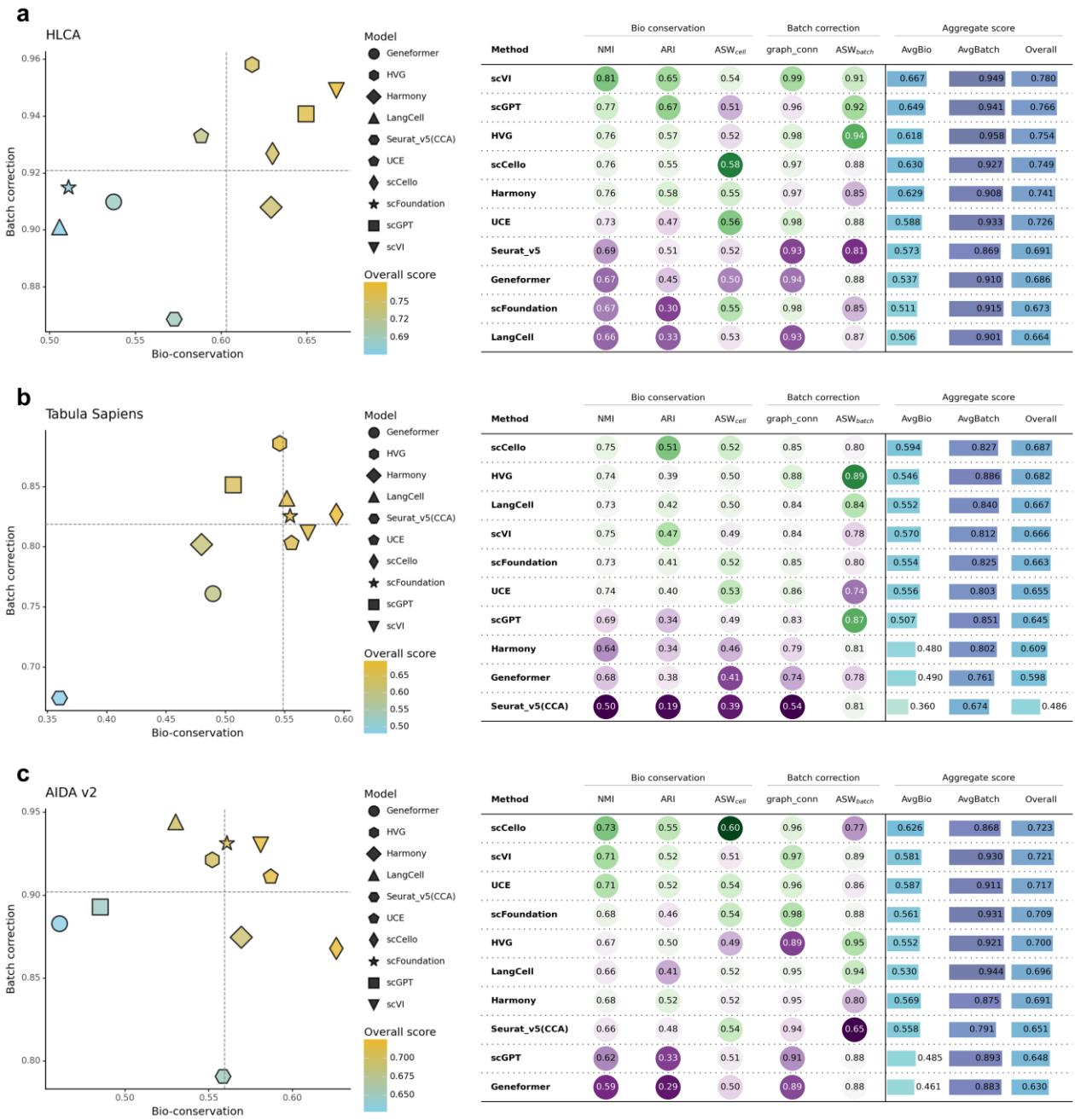



**Fig. S14. Detailed batch integration performance on the HLCA (core) dataset (a), Tabula Sapiens dataset (b) and AIDA v2 dataset (c) assessed by scIB metrics.** The dash lines indicate the median score across all evaluated models.

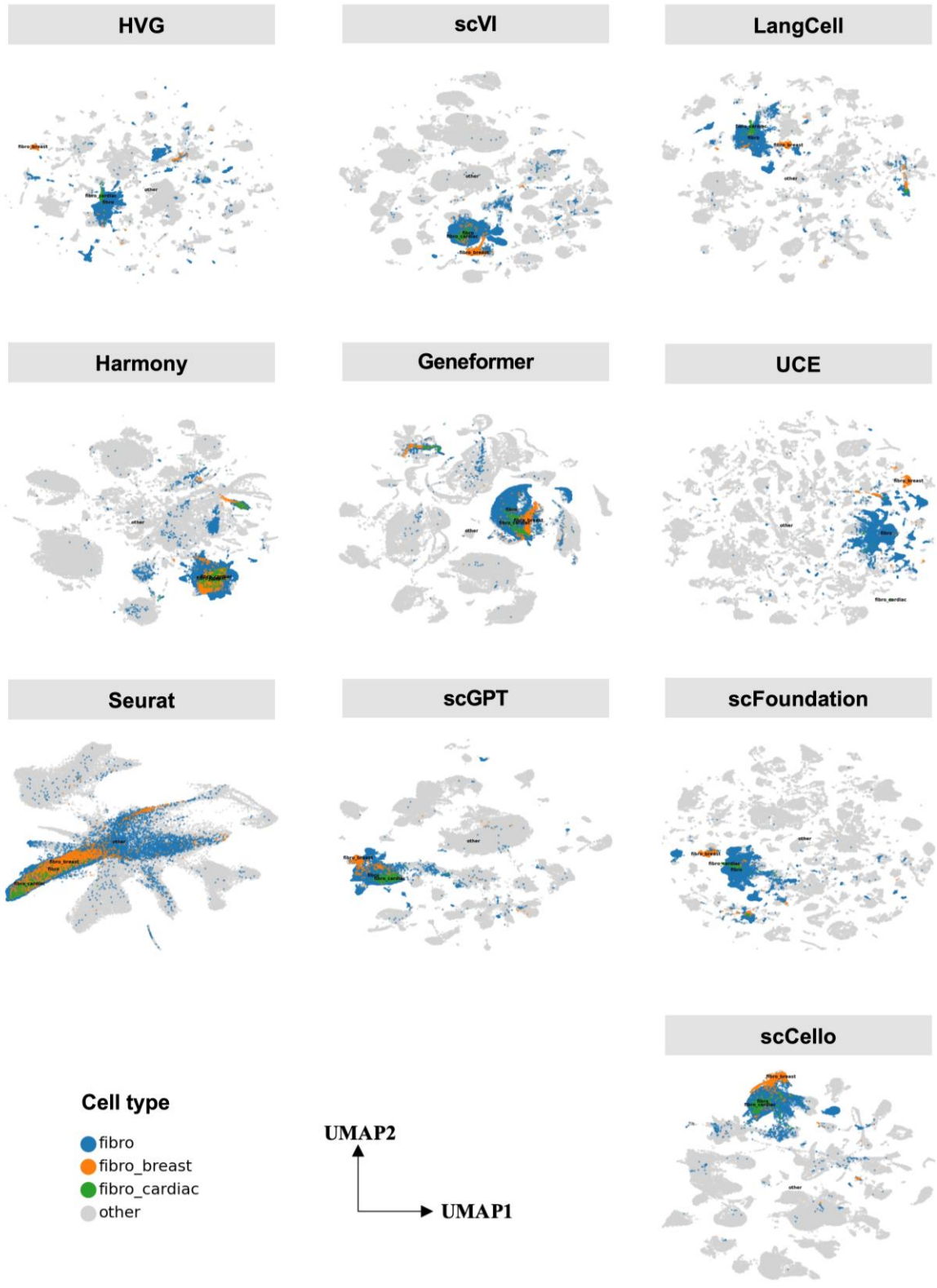
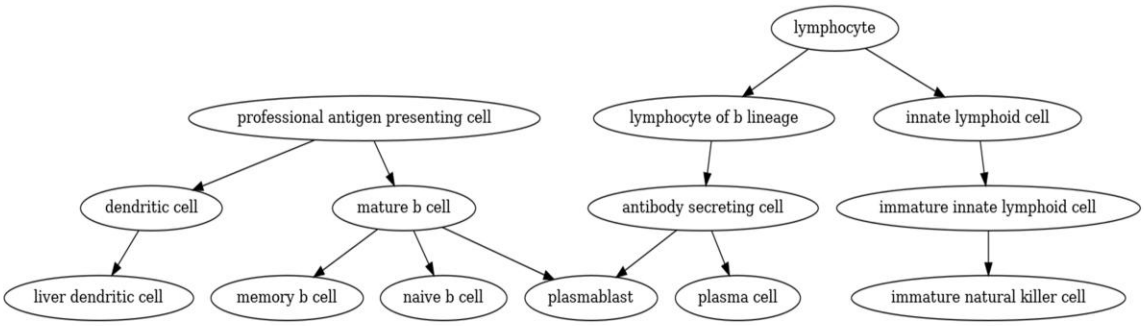



**Fig. S15. UMAP plots of batch integration results on the Tabula Sapiens dataset, focusing on fibroblast subtypes and down-sampled (10%) other cell types.**

a




| Cell ontology id | Cell name | DAG distance | PCA distance | OntoRWR distance |
|---|---|---|---|---|
| CL:0000786 | plasma cell | 0.0 | 0.000000 | 0.000000 |
| CL:0000980 | plasmablast | 2.0 | 0.364187 | 3.022473 |
| CL:0000787 | memory b cell | 4.0 | 0.464566 | 5.066945 |
| CL:0000788 | naive b cell | 4.0 | 0.504570 | 4.768262 |
| CL:0000823 | immature natural killer cell | 6.0 | 0.792220 | 5.111935 |
| CL:2000055 | liver dendritic cell | 6.0 | 0.143257 | 5.153073 |

b

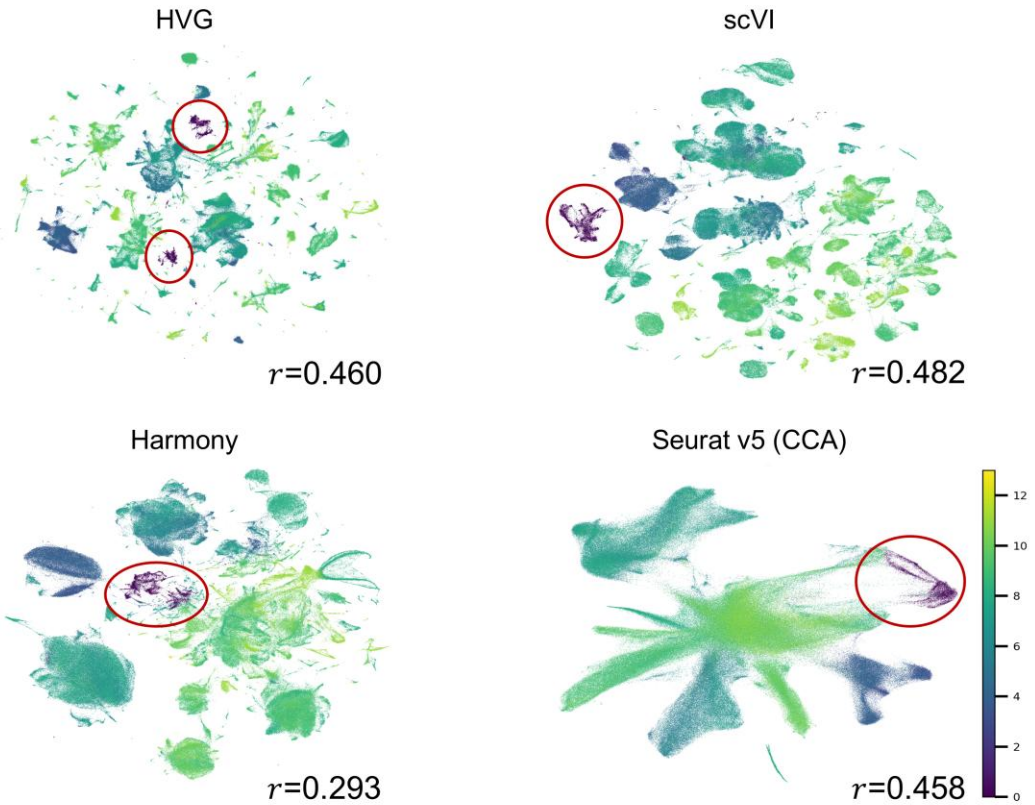



**Fig. S16. A case study on plasma cell.** a, Comparison of the PCA distance and OntoRWR distance to DAG distance in the cell ontology graph. b, UMAP plots of cell embeddings from the Tabula Sapiens dataset provided by different models, colored by the shortest path to the target plasma cell based on the cell ontology graph (DAG distance). The red circles refer to the plasma cells. The $r$ score reported in

each subplot is the plasma cell-specific Pearson correlation coefficient between its neighborhood affinities defined in two graphs, where the first graph is derived from the provided embeddings, and the other is the reference graph derived from performing random walk with restart on the text similarity-weighted cell ontology graph (i.e., the scGraph-OntoRWR score).

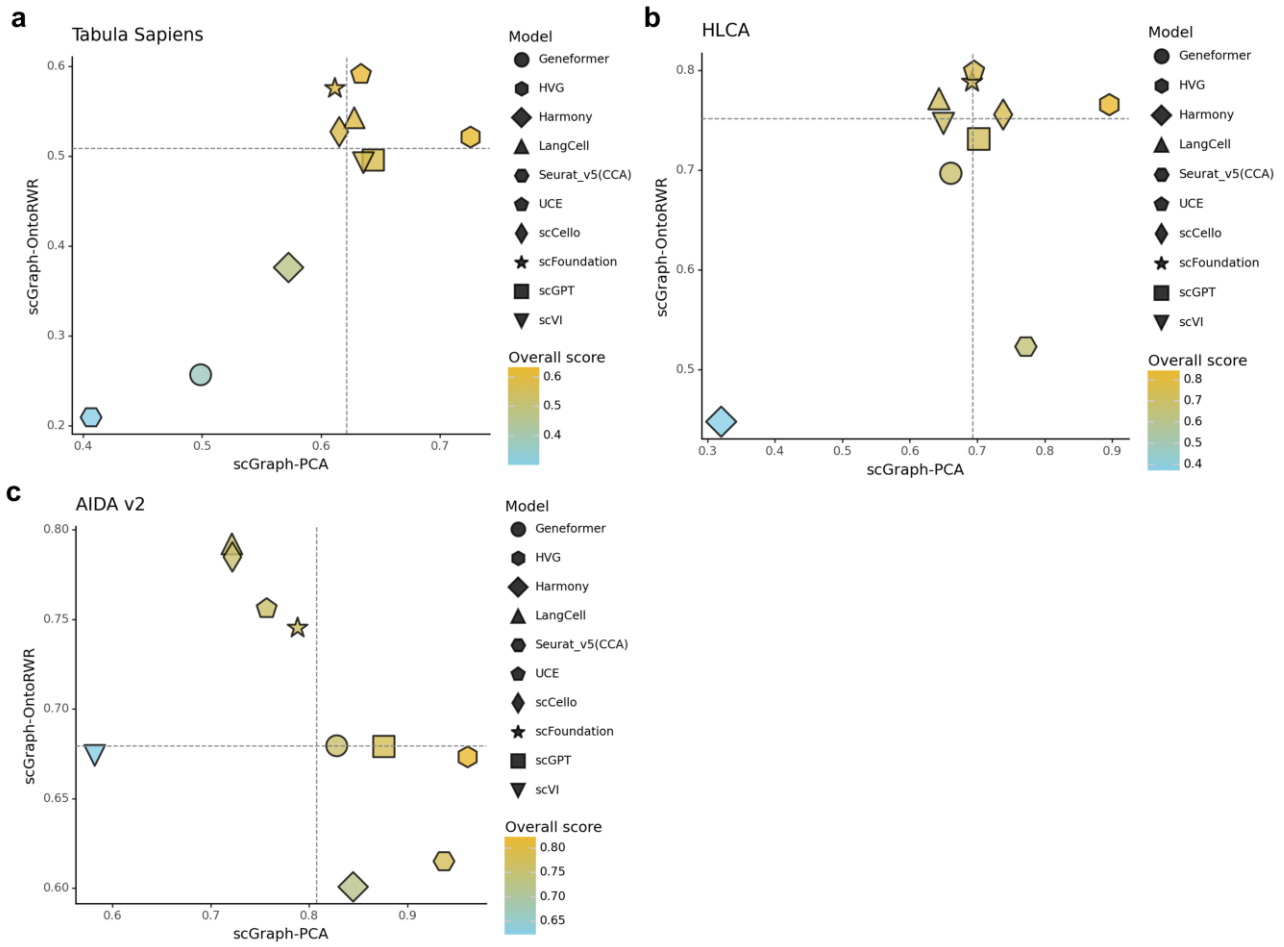



**Fig. S17. Batch integration performance on the Tabula Sapiens dataset (a), HLCA (core) dataset (b), and AIDA v2 dataset (c) assessed by scGraph metrics.** The dash lines indicate the median score across all evaluated models.

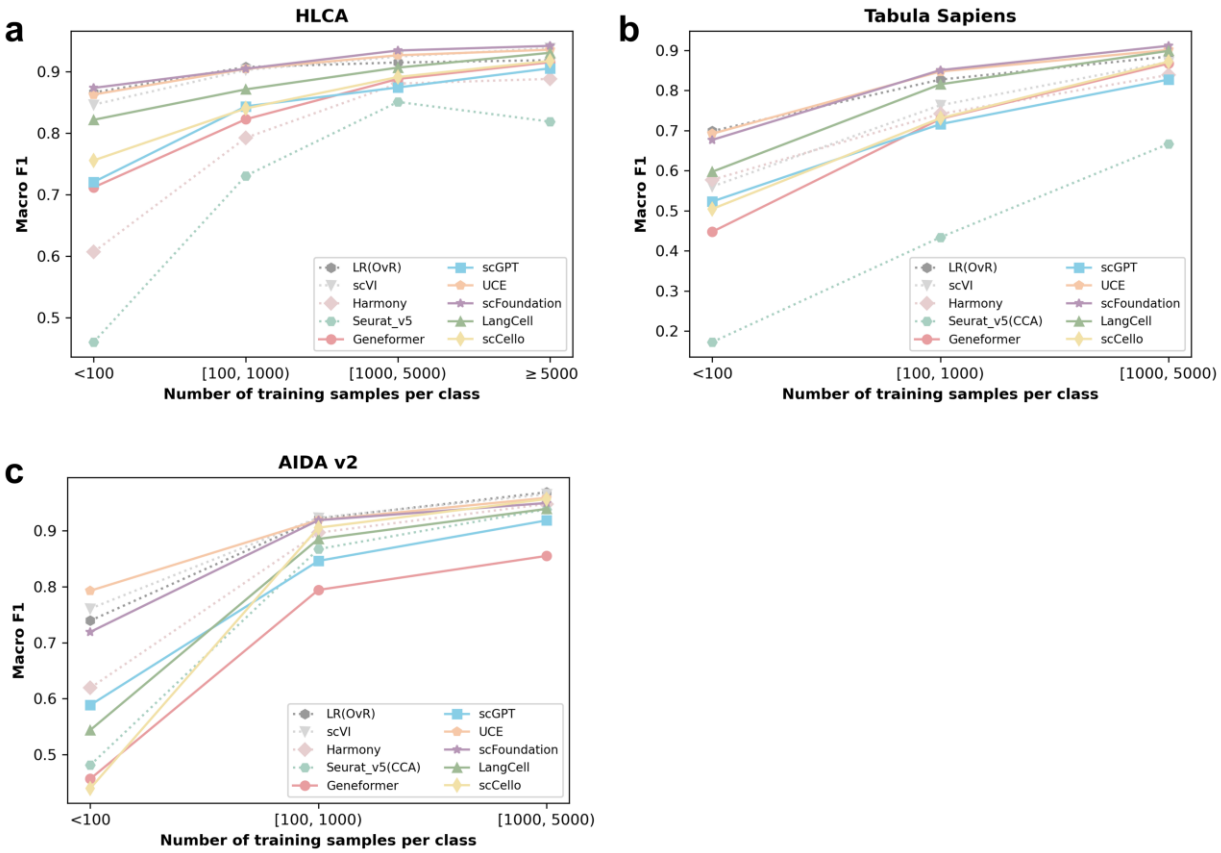



Fig. S18. Macro-F1 scores of cell type annotation grouped by number of training samples.

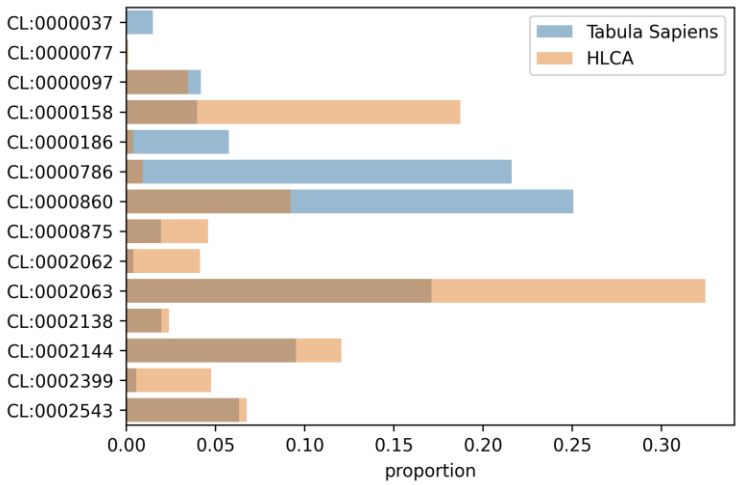



Fig. S19. Distribution of 14 overlapped cell types in the Tabula Sapiens and HLCA datasets.

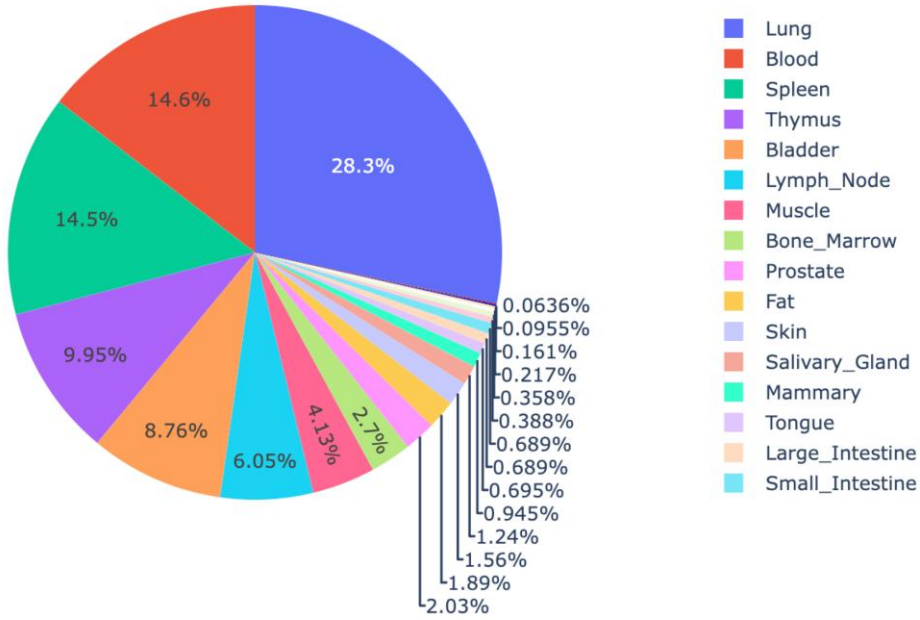



**Fig. S20. Tissue composition of the Tabula Sapiens test set used for cross-dataset validation.**

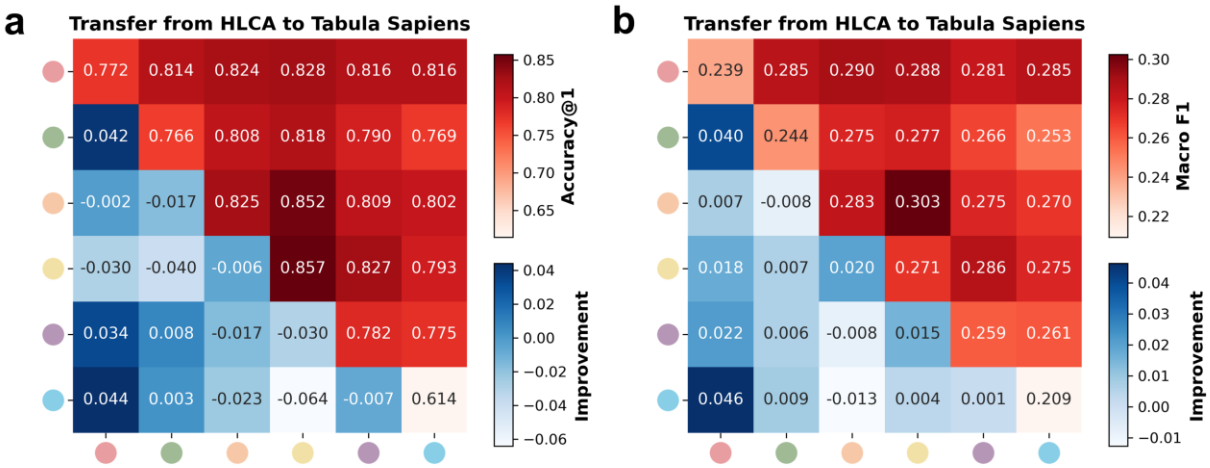



**Fig. S21. The model complementary analysis in the setting of transferring from HLCA to Tabula Sapiens dataset.** The ensemble performance and absolute improvement over the best component are colored by red and blue, respectively.

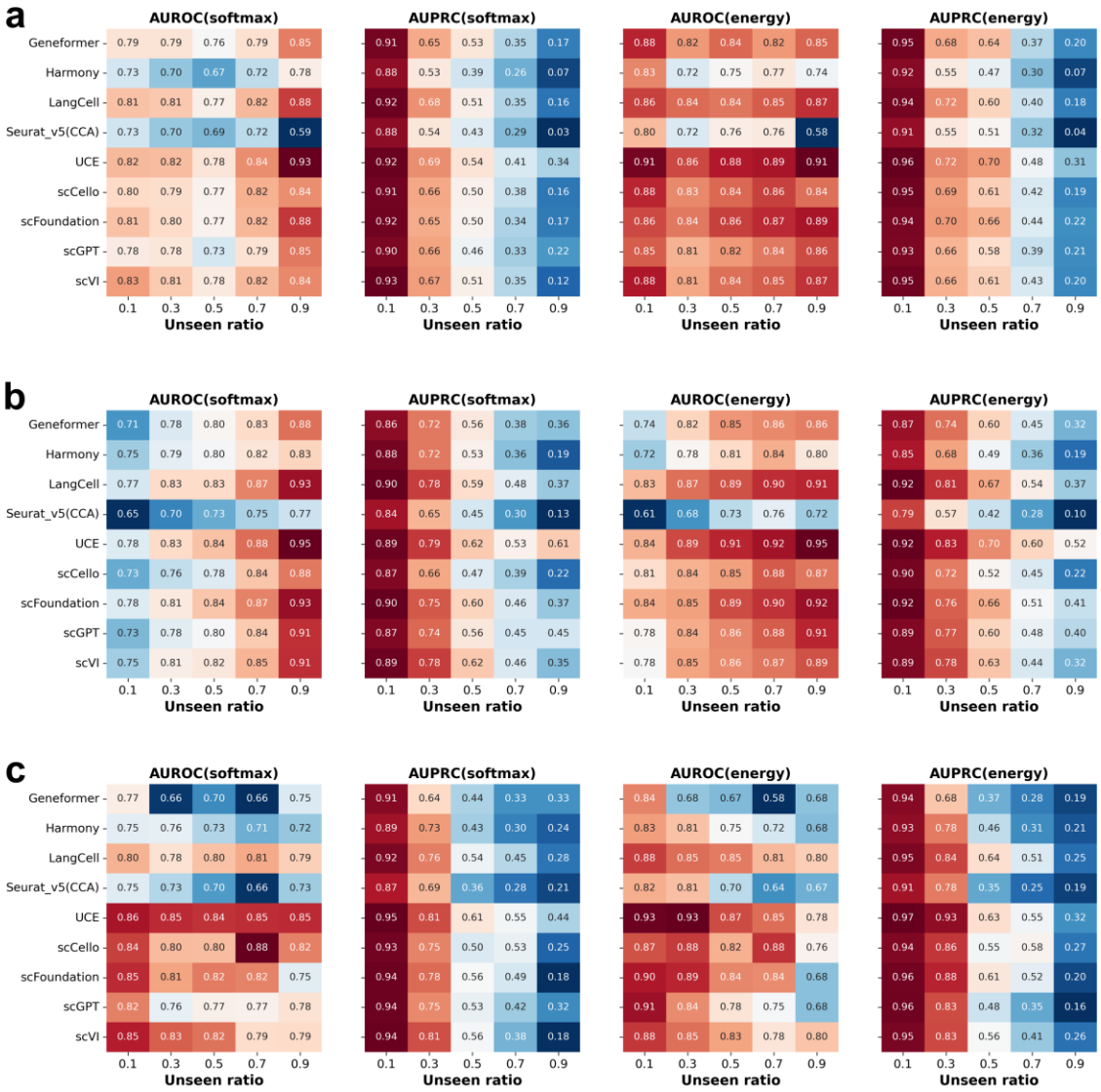



**Fig. S22. Heatmap of the model performance in distinguishing seen from unseen cell types on the HLCA (core) dataset (a). Tabula Sapiens dataset (b), and AIDA v2 dataset (c).** Specifically, the seen cell types are treated as positive (in-distribution), while the unseen cell types are treated as negative (out-of-distribution).

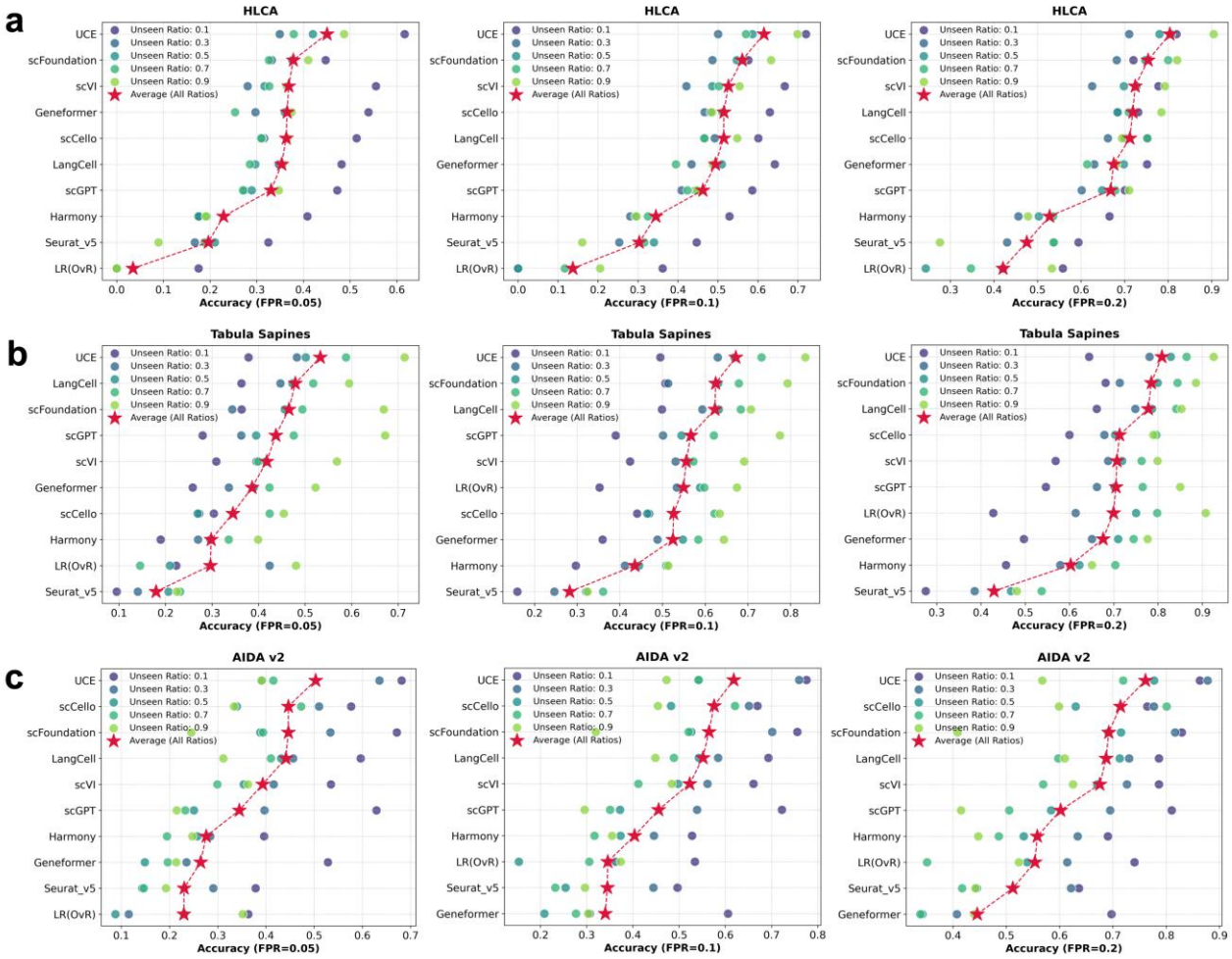



**Fig. S23. Cell type annotation accuracy with different ratios of novel cell types on the HLCA dataset (a), Tabula Sapiens dataset (b) and AIDA v2 dataset (c).** Accuracy is reported at fixed False Positive Rates (FPR) of 0.05, 0.1, and 0.2.

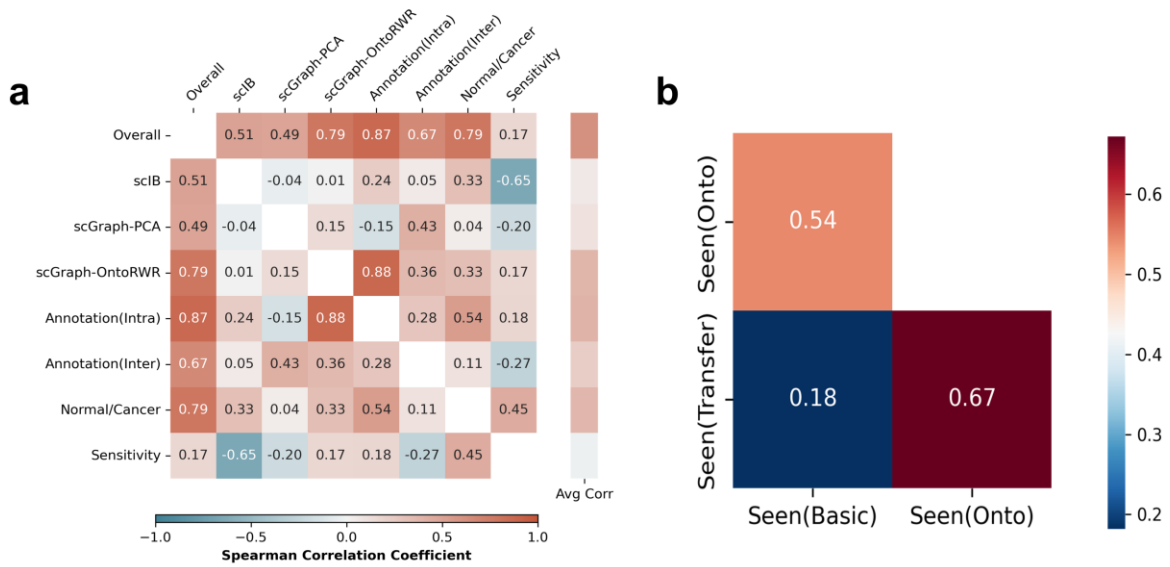



**Fig. S24. Analysis of task correlations**. a, Heatmap of Spearman correlations between model rankings across batch integration metrics and downstream tasks. b, Heatmap of Spearman correlations between model rankings across cell type annotation scenarios.

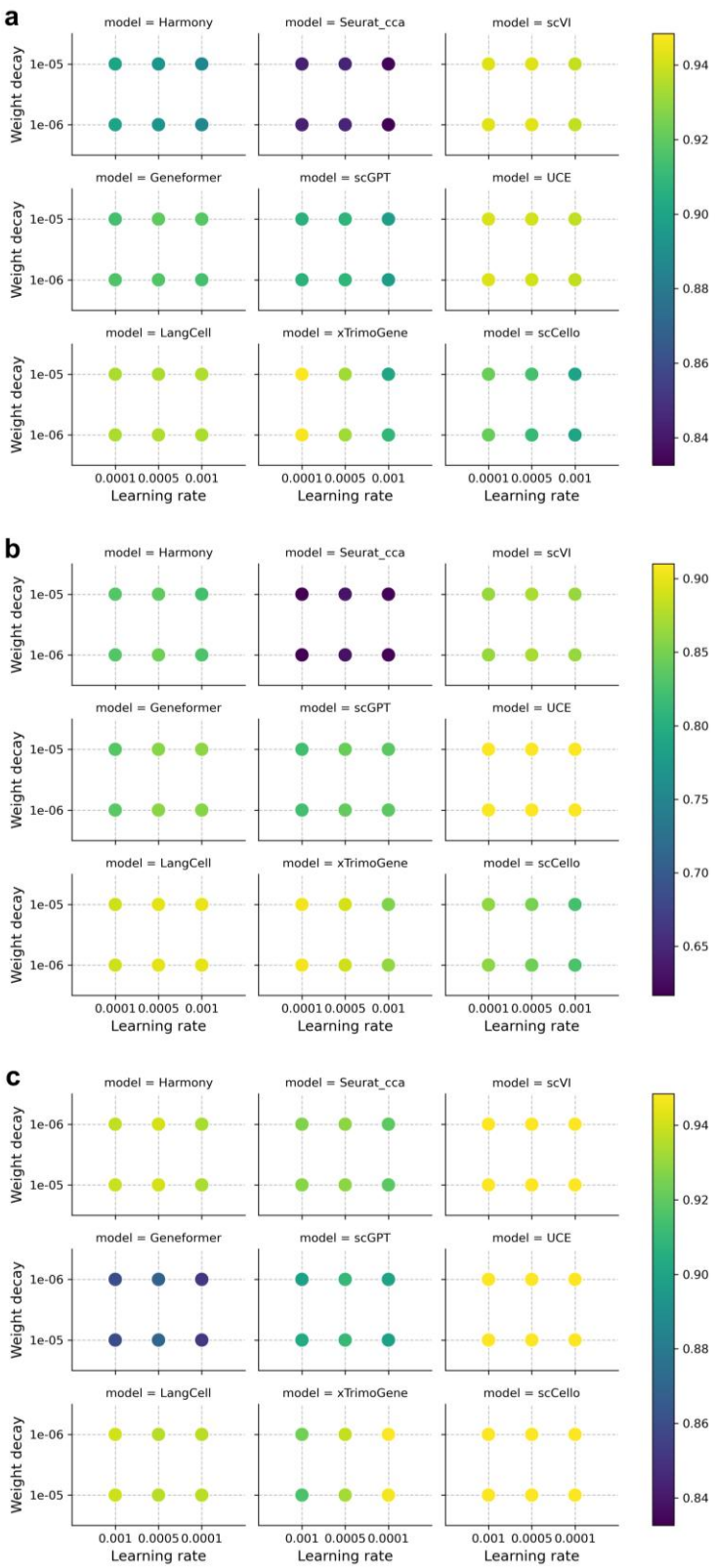



**Fig. S25. Hyperparameter search for embedding-based OnClass models.** Model performance under different weight decay and learning rate on the HLCA dataset (a), Tabula Sapiens dataset (b), and AIDA v2 dataset (c).

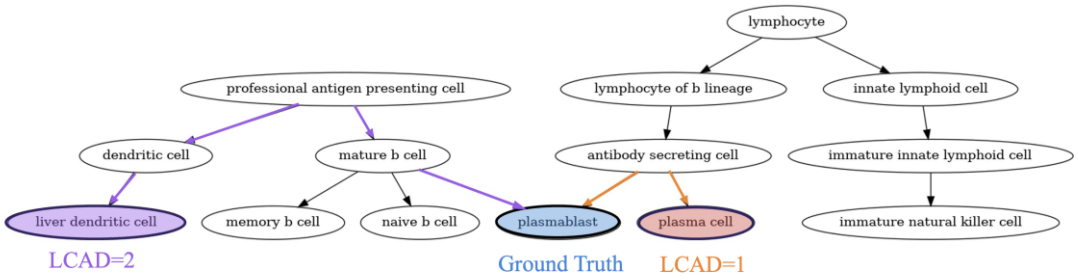



**Fig. S26. Illustration of the calculation of lowest common ancestor distance (LCAD).** When the ground truth label is plasmablast, the lowest common ancestor of plasma cell and ground truth is antibody secreting cell, thus the LCAD is 1 (in earthy yellow). Similarly, the lowest common ancestor of liver dendritic cell and ground truth is professional antigen presenting cell, thus the LCAD is 2 (in purple).

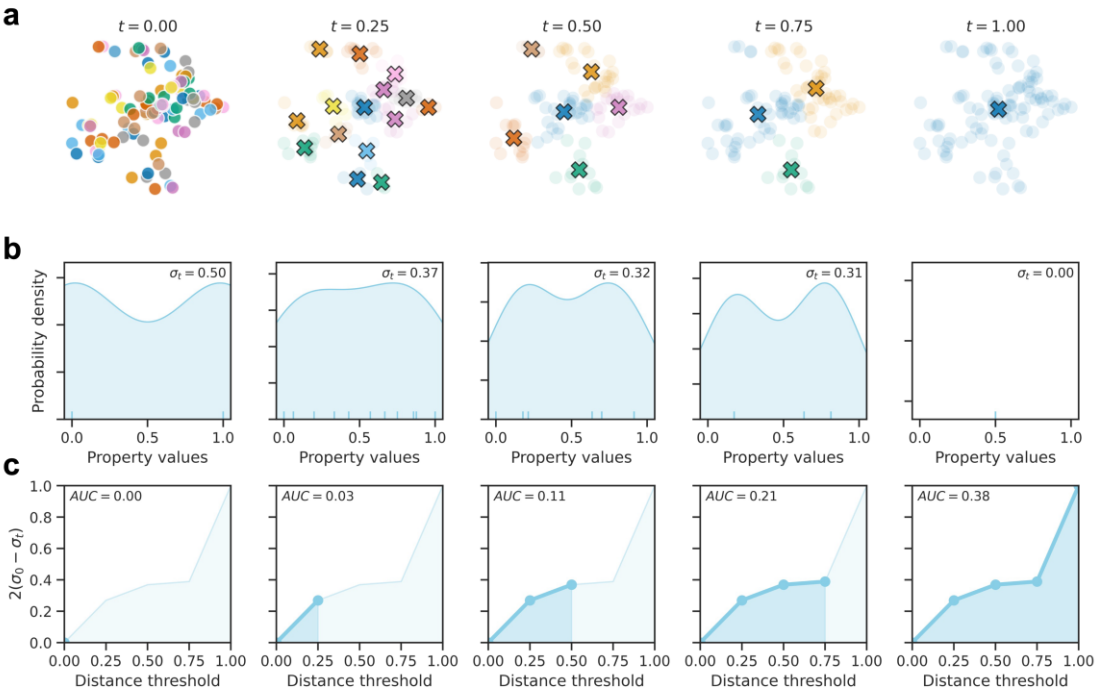



**Fig. S27. An example of the calculation of roughness index (ROGI).** (a) Data clustering with threshold $t$ ranging from 0 to 1. Each cluster is represented by a *prototype*, whose property is defined as the average property of all data points within the cluster. (b) The weighted standard deviation $\sigma_t$ of the property distribution across all *cluster prototypes* captures the property dispersion at different threshold $t$. (c) ROGI is measured as the area under the curve (AUC) of $2(\sigma_0 - \sigma_t)$ versus $t$.