Supplementary Materials:

Effectively preserving biological variations in multi-batch and multi-condition single-cell data integration

Qingbin Zhou[1, †], Tao Ren[2, 3, †], Fan Yuan[4], Jiating Yu[5], Jiacheng Leng[6], Jiahao Song[1], Duanchen Sun[1, 7, *], and Ling-Yun Wu[2, 3, *]

1 School of Mathematics, Shandong University, Jinan 250100, China;
2 State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China;
3 School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China;
4 School of Mathematics and Information Science, Yantai University, Yantai 264005, China;
5 School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing 210044, China;
6 Zhejiang Lab, Hangzhou 311121, China;
7 Shandong Key Laboratory of Cancer Digital Medicine, Jinan 250033, China;
† These authors contributed equally: Qingbin Zhou, Tao Ren.
* These authors jointly supervised this work: Duanchen Sun, Ling-Yun Wu. Email: dcsun@sdu.edu.cn, lywu@amss.ac.cn

Contents

1. Dataset descriptions

Alzheimer's disease dataset. The Alzheimer's disease (AD) snRNA-seq dataset used for benchmarking was downloaded from the online website (http://adsn.ddnetbio.com)[1]. This dataset contains 11,884 cells from 12 donors, categorized into disease and healthy conditions, with each condition containing three batches. We did not consider the cells with unknown cell types, and the preprocessed data contains 11,884 cells with 2,000 highly variable genes (HVG). Additionally, we used an external AD bulk RNA-seq dataset to test the effectiveness of our signature. This bulk dataset includes 87 AD and 74 control samples and was downloaded from the Gene Expression Omnibus (GEO; accession number: GSE5281). The raw count matrix was processed using the standard pipeline provided by Seurat[2] to obtain a normalized expression matrix for downstream analysis.

COVID-19 dataset. The COVID-19 raw and processed data used for benchmarking was downloaded from the GEO (accession number: GSE149689)[3]. We excluded cells with unknown cell types. The preprocessed dataset contains 17 patients, 58,199 cells, and five conditions: healthy donors (n = 4), hospitalized patients with severe influenza (n = 5), and patients with severe (n = 4), mild (n = 3), and asymptomatic COVID-19 (n = 1). Except for COVID-19 patients C3, C6, and C7, who were sampled twice, all other patients were sampled only once, making a total of 20 batches for this COVID-19 dataset.

Diabetes dataset. The diabetes dataset used in this study was downloaded from the online website (https://hpap.pmacs.upenn.edu/analysis). The dataset contains 222,077 human islet cells obtained from 67 donors within the Human Pancreas Analysis Program (HPAP)[4, 5]. There are four conditions for the collected diabetes dataset: non-diabetic, type 1 diabetes autoantibody positive (AAb+), type 1 diabet (T1D), and type 2 diabetes (T2D).

PBMC atlas. The COVID-19 peripheral blood mononuclear cell (PBMC) atlas[6] used in this study was downloaded from the online website (https://figshare.com/projects/scPoli_data/155018). The original data contains six conditions, and we only kept the 'COVID-19' and 'Normal' samples, which accounted for the vast majority of the cells. We filtered the cells with unknown cell types, and the final atlas contains 6,740,034 PBMC cells from 2,162 samples, representing cells from 25 datasets, two conditions, and five sequencing protocols.

Supplementary Table 7. Detailed information of datasets used in this study.

Dataset	#Cell	#HVG	#Cell type	#Condition	#Batch
AD dataset	11,884	2,000	6	2	6
COVID-19 dataset	58,199	2,000	13	5	20
Diabetes dataset	222,077	3,000	10	4	67
PBMC atlas	6,740,034	10,000	12	2	25

2. Neural network architecture of scFLASH

scFLASH consisted of four neural network units: encoder E, decoder D, batch classifier B, and condition classifier C (Supplementary Table 8). All units used multilayer perceptron with ReLU activation functions (except for the last layer of D). During the forward pass, both B and C used the SoftMax function to calculate the predicted probabilities for each class.

Supplementary Table 8. Detailed information of neural network architecture in scFLASH.

Unit	Architecture	Activation Function
Encoder E	(#HVG +#Batch) → 128 → 32	ReLU (first layer)
Decoder D	(32 + #Batch) → 128 → #HVG	ReLU (hidden layer) Tanh (final layer)
Batch classifier B	32 → 100 → 100 → #Batch	ReLU (all hidden layers) SoftMax (forward pass)
Condition classifier C	8 → 50 → #Condition	ReLU (hidden layer) SoftMax (forward pass)

3. Evaluation metrics

In this section, we described the evaluation metrics used to benchmark the integration performances of different methods.

Biological conservation metrics:

Adjusted Rand Index (ARI): ARI quantifies the clustering accuracy, taking values ranging from -1 to 1, with higher values indicating a better clustering[7]. In order to calculate the ARI, we compared the cell type labels with the clustering results and then scaled the values from 0 to 1.

Normalized Mutual Information (NMI): NMI evaluates the similarity between the cell type labels and the clustering results. We calculated the NMI using the cell type labels with the clustering results.

Batch correction metrics:

integration Local Inverse Simpson's Index (iLISI): iLISI measures the variability (i.e., diversity or evenness) between batches[8]. We calculated an iLISI score for each cell, ranging from 1 to B (where B represents the number of batches), indicating perfect separation and perfect mixing. We scaled iLISI scores to the range of 0 to 1 and used the median as the final metric.

Average silhouette width across batches (ASW_batch): The silhouette width measures the relationship between a cell's within-cluster distances and between-cluster distances to the closest cluster[9]. For the batch mixing metric, we considered the absolute silhouette width $|s_i|$ on batch labels per cell i, where 0 indicates well-mixed batches, and any deviation from 0 indicates a batch effect.

To compute the ASW_batch score, we averaged the $1 - |s_i|$ within each cell type label, ensuring that higher scores represent better batch mixing, and then averaged across all cell type labels[10]:

$$\text{ASW\_batch} = \frac{1}{|M|}\sum_{j \in M}\frac{1}{|C_j|}\sum_{i \in C_j}(1 - |s_i|)$$

where C_j is the set of cells with the cell type label j and M is the set of unique cell types. The range of ASW_batch is [0, 1], with 1 representing ideal batch mixing and 0 indicating strongly separated batches.

Condition conservation metrics:

Average silhouette width across condition (ASW_cond): ASW_cond was calculated using the classical definition of ASW on condition categories and scaled to a value between 0 and 1 using $\text{ASW\_cond} = (\text{ASW\_cond\_raw} + 1)/2$, with a higher value indicating better protection of condition differences.

F1 score of the k-nearest neighbor classifier (cond_knn): We trained a *k*-nearest neighbor classifier (*k*=15) for all cells using the condition categories as labels. After this, we computed the F1 score using the predicted and true condition categories. We used the F1 score as cond_knn (range from 0 to 1), and a higher score indicates more separable cells from different conditions.

4. Configurations of competing methods

We compared scFLASH with ten state-of-the-art methods (BBKNN[11], Harmony[8], Scanorama[12], scVI[13], scDREAMER[14], Seurat[2], scINSIGHT[15], scMerge2[16], scDisInFact[17], and scDisco[18]). All methods were implemented following their pipelines described in their GitHub repositories with the default parameters (Supplementary Table 9).

Supplementary Table 9. Configurations of the competing methods.

Methods	Principle	Configuration	Github (https://github.com/)
BBKNN	KNN graph integration	1.3.9	Teichlab/bbknn
Harmony	PCA + clustering-based correction	0.0.10	immunogenomics/harmony
Scanorama	SVD + Mutual nearest neighbors	1.7.4	brianhie/scanorama
scVI	Conditional variational autoencoder	0.17.3	scverse/scvi-tools
scDREAMER	VAE with adversarial classifier	NA	Zafar-Lab/scDREAMER
Seurat	CCA + Mutual nearest neighbors	4.4.0	satijalab/seurat
scINSIGHT	Non-negative matrix factorization	0.1.4	Vivianstats/scINSIGHT
scMerge2	Factor analysis + pseudo-bulk construction	1.18.0	SydneyBioX/scMerge
scDisInFact	MMD+VAE-based disentanglement	0.1.0	ZhangLabGT/scDisInFact
scDisco	VAE with condition domain adaptation	1.0	Victory-LRJ/scDisco

5. Ablation study

We conducted ablation studies to evaluate the contributions of the penalized condition classifier and the batch classifier, as well as the entropy loss designed in our penalized condition classifier. The results were summarized in Supplementary Table 10.

We found that scFLASH without the batch classifier (scFLASH-woBC) outperformed scFLASH without the condition classifier (scFLASH-woCond) in condition conservation scores. In contrast, scFLASH-woCond performed better in batch correction scores. Besides, scFLASH with the entropy loss could improve the effectiveness of the condition classifier and thus contribute to the overall integration performance.

In summary, all components in scFLASH are indispensable. When used together, scFLASH effectively unifies batch correction and condition preservation, providing a balanced and reliable framework for multi-batch and multi-condition integration.

Supplementary Table 10. The quantitative evaluations of biological conservation, batch correction, and condition conservation for ablation studies.

Dataset	Model	Biological conservation	Batch correction	Condition conservation	Total
Alzheimer's disease dataset	scFLASH	0.956	0.681	0.701	0.779
	scFLASH-woCond*	0.954	0.738	0.582	0.758
	scFLASH-woBC**	0.957	0.581	0.764	0.767
	scFLASH-woEN***	0.947	0.677	0.706	0.776
COVID-19 dataset	scFLASH	0.735	0.533	0.619	0.630
	scFLASH-woCond*	0.746	0.558	0.581	0.628
	scFLASH-woBC**	0.648	0.484	0.729	0.621
	scFLASH-woEN***	0.719	0.530	0.625	0.624

*scFLASH without the condition classifier

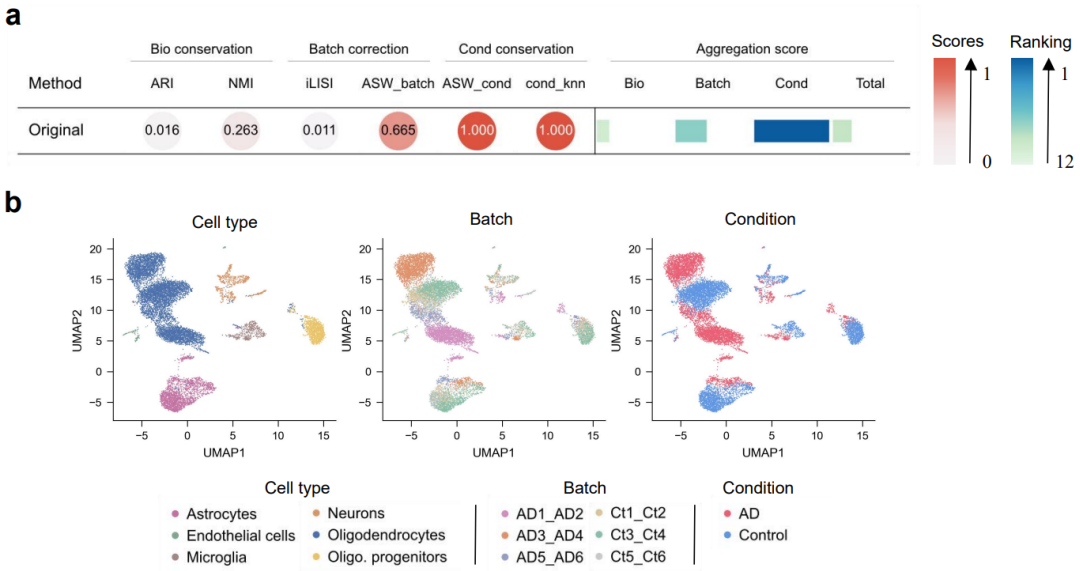
**scFLASH without the batch classifier

***scFLASH without the entropy loss

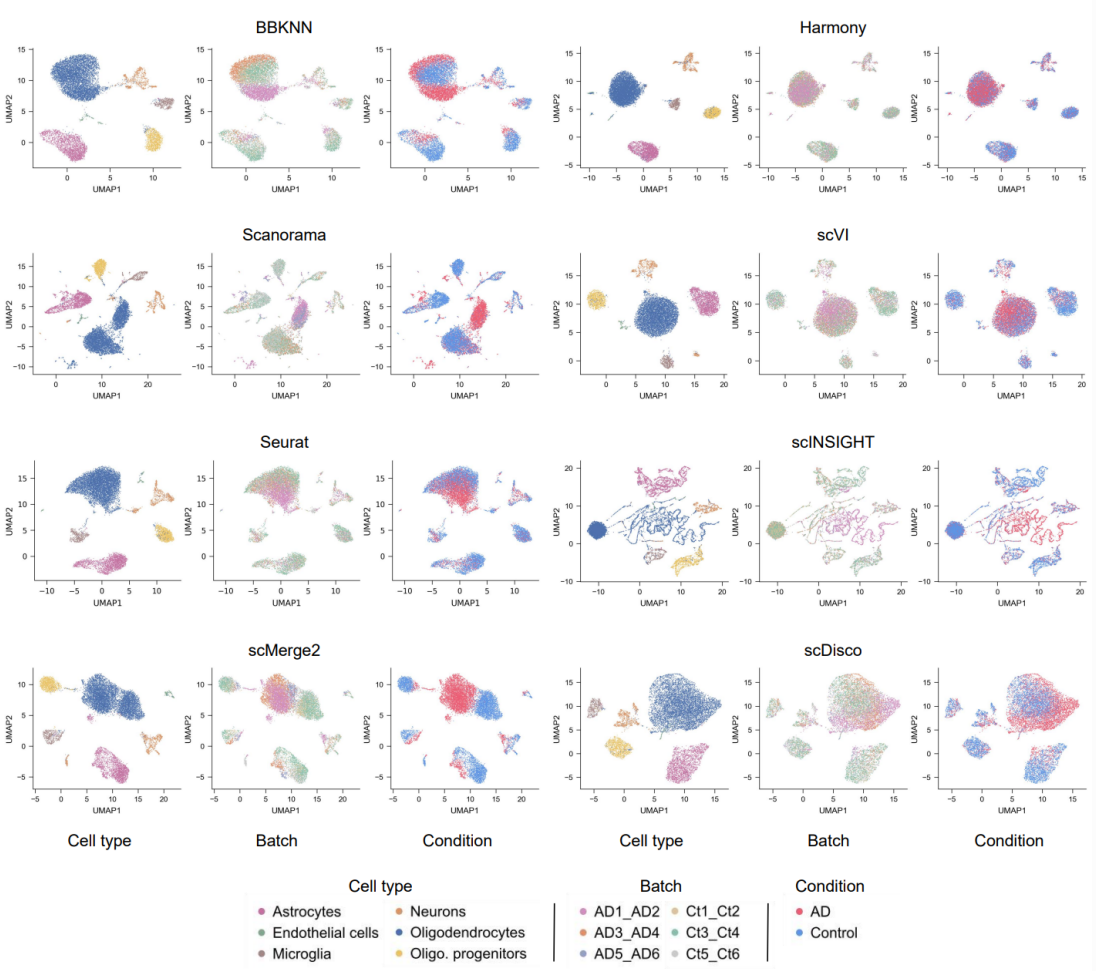
6. Supporting results for main Fig. 2

In this section, we provided supporting results for benchmarking the Alzheimer's disease dataset. Supplementary Fig. 1 showed that the biological preservation and batch correction scores were low for the original unintegrated data, which indicates prominent batch effects. These observations can lead to exaggerated phenotypic differences and impair the accurate representation of cellular heterogeneity. Supplementary Fig. 2 displayed UMAP visualizations for the Alzheimer's disease dataset after integration by BBNKK, Harmony, Scanorama, scVI, Seurat, scINSIGHT, scMerge2, and scDisco.

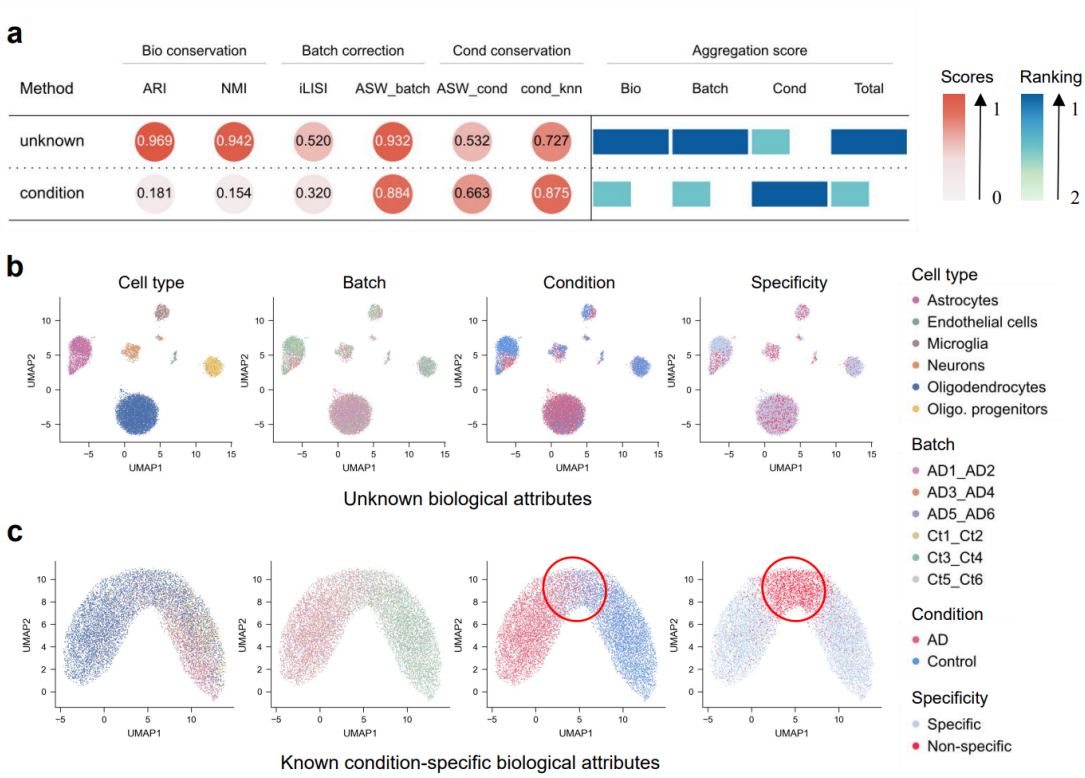
To evaluate scFLASH's disentangling capability, we evaluated the integration performances using the latent known condition-specific attributes and unknown biological attributes. Supplementary Fig. 3a showed that these two embeddings captured different perspectives for biological conservation, batch correction, and condition conservation. Supplementary Fig. 3b and 3c presented UMAP visualizations of latent space embeddings. Notably, the condition-specific cells identified by scFLASH effectively delineated the boundary between healthy and diseased states, as highlighted by red ellipses in Supplementary Fig. 3c.



Supplementary Fig. 1. a, Original metrics for the unintegrated Alzheimer's disease dataset. **b,** UMAP visualizations of the unintegrated Alzheimer's disease dataset. The cells were annotated and colored by cell types (left), batches (middle), and conditions (right).



Supplementary Fig. 2. UMAP visualizations of latent space embeddings for the Alzheimer's disease dataset after integration by eight different methods. The cells were annotated and colored by cell types (left), batches (middle), and conditions (right).

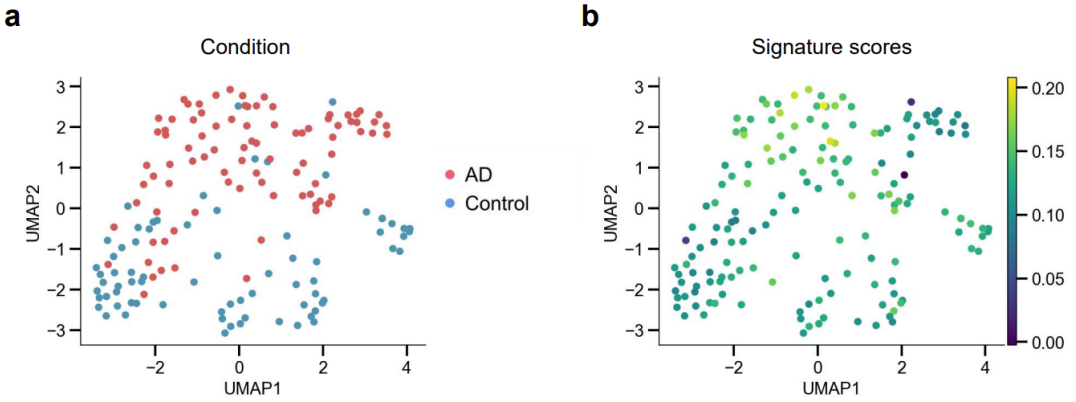


Supplementary Fig. 3. a, The quantitative metrics for unknown biological attributes (unknown) and known condition-specific biological attributes (condition), distinguished by scFLASH in the Alzheimer's disease dataset. **b-c,** UMAP visualizations of Alzheimer's disease dataset using (**b**) unknown biological attributes and (**c**) known condition-specific biological attributes. The cells were annotated and colored by cell types, batches, conditions, and specificities.

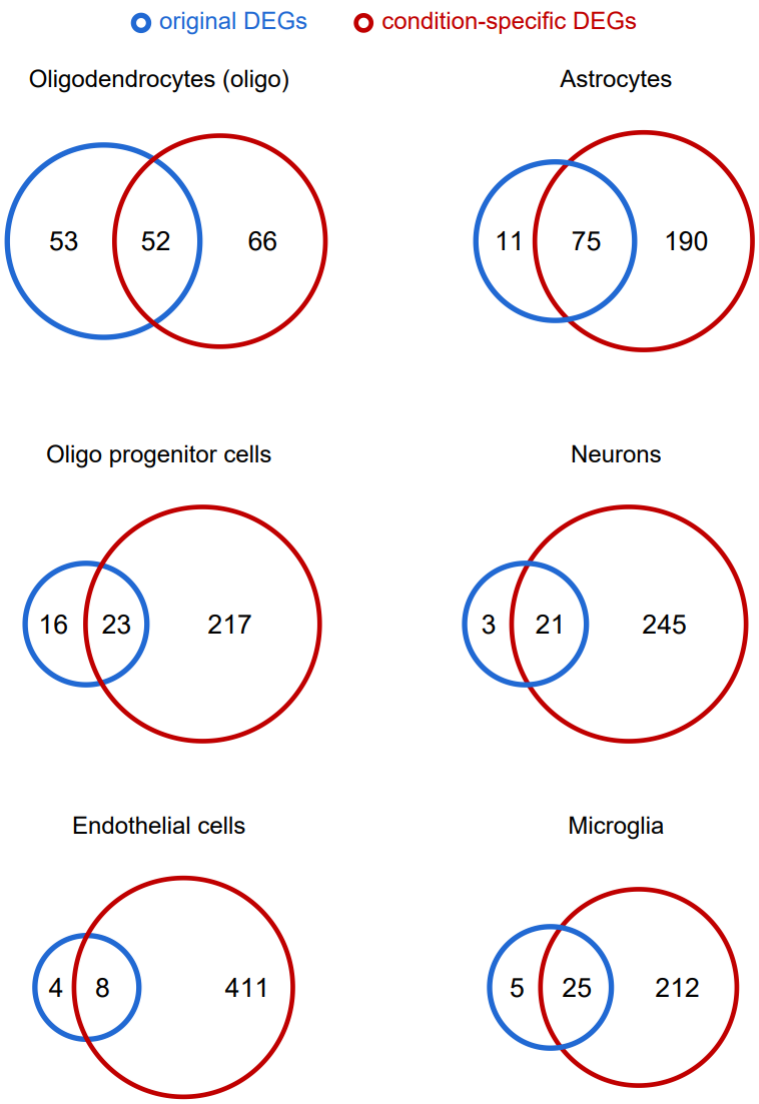
7. Supporting results for main Fig. 3

In this section, we offered supporting materials for the analysis results of scFLASH on Alzheimer's disease (AD) dataset as supplements to Fig. 3 in the main text.

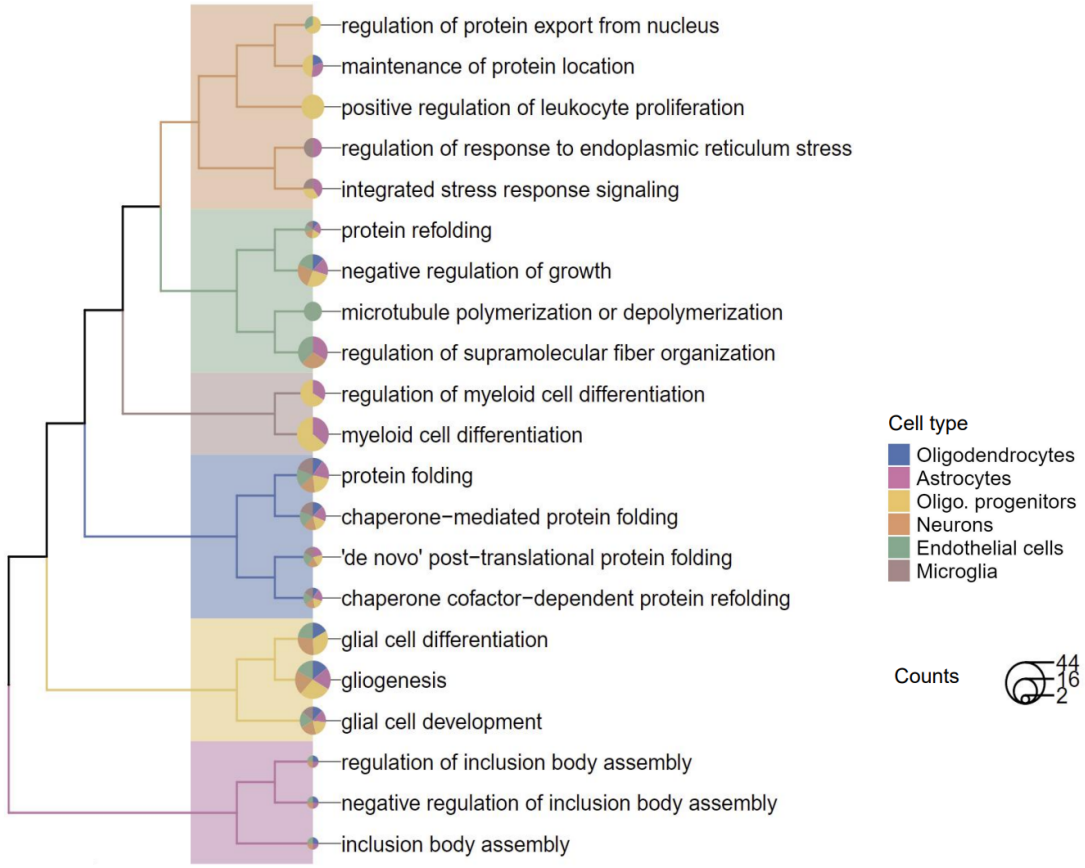
Supplementary Fig. 4 showed the UMAP visualizations for GSE5281 samples. We found that the AD patients had relatively higher signature scores. Supplementary Fig. 5 indicated that scFLASH identified more AD-upregulated DEGs when directly comparing condition-specific cells in oligodendrocytes, astrocytes, oligo progenitor cells, neurons, endothelial cells, and microglia. Supplementary Fig. 6 presented functional enrichment analysis results across different cell types, revealing enriched protein misfolding-related pathways in AD cells. Supplementary Fig. 7 used volcano plots to show that scFLASH identified *LINGO1* as an upregulated DEG across multiple cell types. Supplementary Fig. 8 and Supplementary Fig. 9 both demonstrated that scFLASH-corrected expression profiles can improve cell type annotation, reduce noise, and successfully detect weak biological signals compared to other methods.



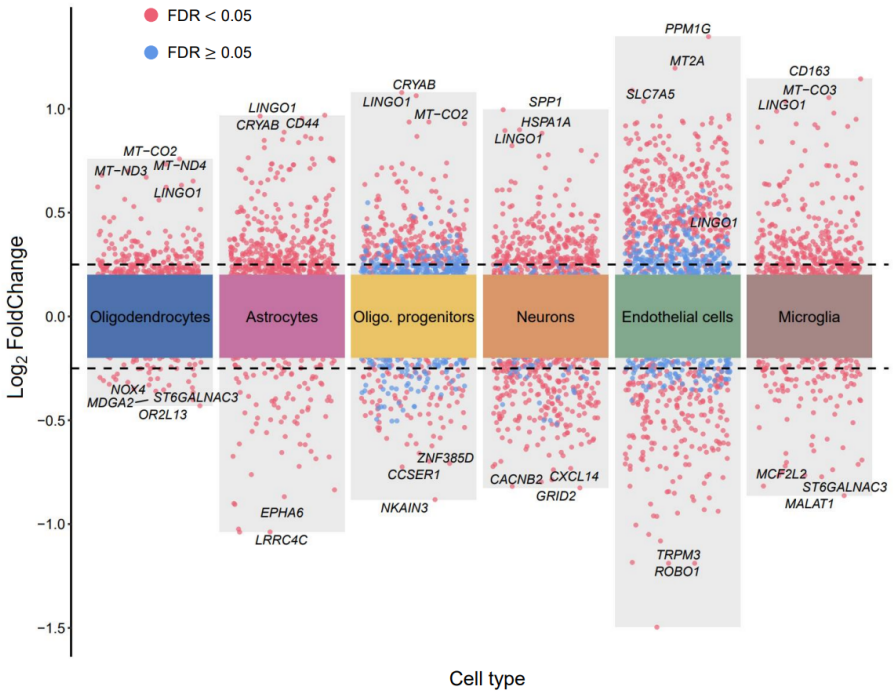
Supplementary Fig. 4. UMAP visualizations for GSE5281 samples. The bulk samples were annotated and colored by (**a**) conditions and signature scores (**b**).



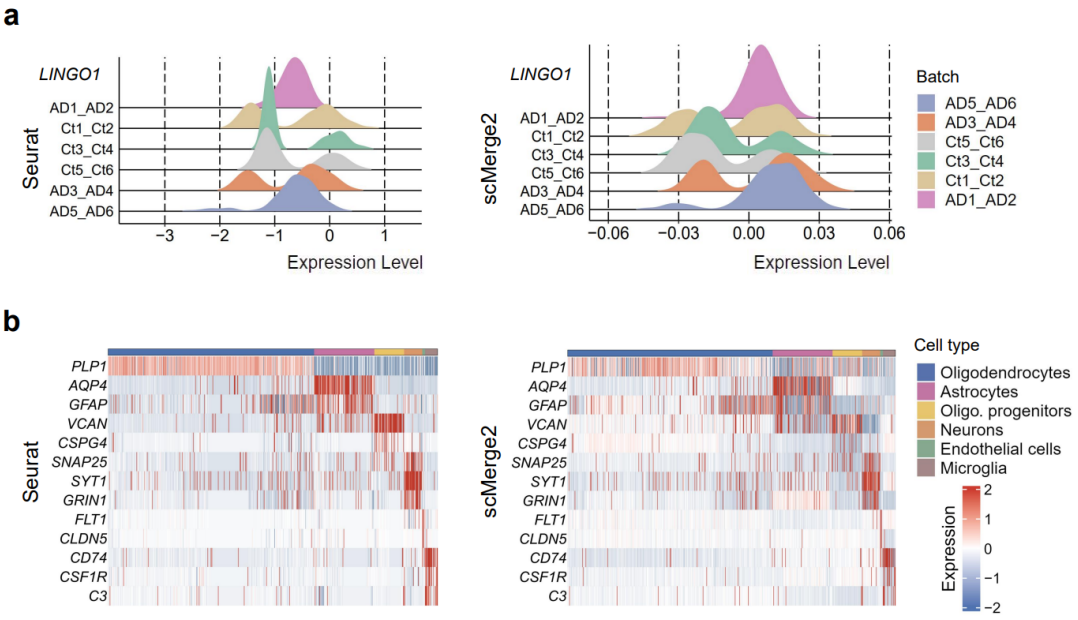
Supplementary Fig. 5. Venn diagrams show overlaps between upregulated DEGs identified using the original cells and the condition-specific cells for Oligodendrocytes (oligo), Astrocytes, Oligo progenitor cells, Neurons, Endothelial cells, and Microglia.



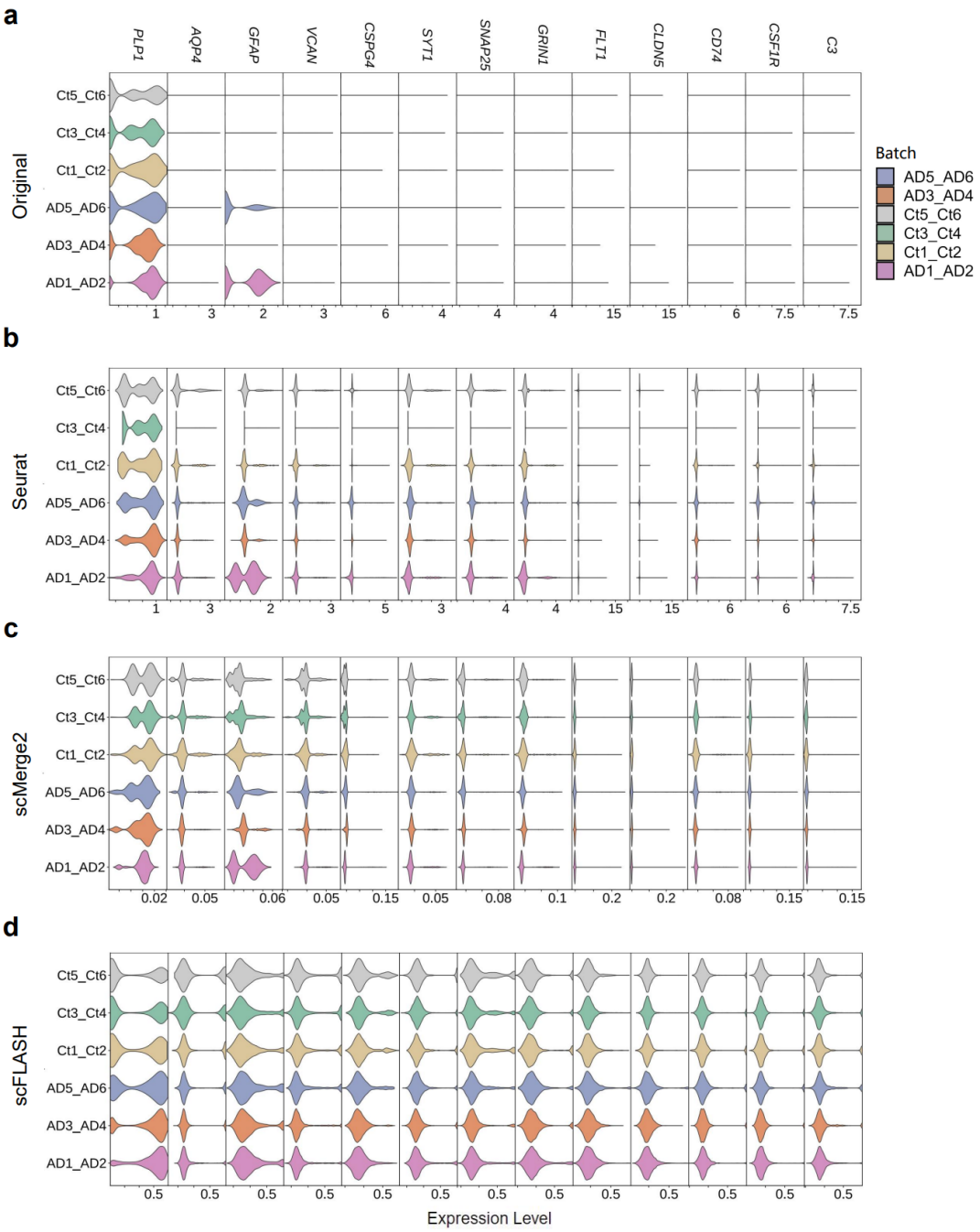
Supplementary Fig. 6. Hierarchical clustering of enriched pathways across different cell types. The size of each pie chart represents the number of genes in the corresponding pathway, and the segment in the pie chart represents the proportion of enriched genes from a particular cell type.



Supplementary Fig. 7. Differential gene expression analysis revealed up- and down-regulated genes in all six cell types. The points with FDR less than 0.05 and larger than 0.05 are shown in red and blue, respectively. The two horizontal dashed lines represent ±0.25 log-transformed fold changes in gene expression.



Supplementary Fig. 8. a, Ridgeline plots show the expression of *LINGO1* across different batches corrected by Seurat (left) and scMerge2 (right). **b,** Heat maps of known marker gene expressions corrected by Seurat (left) and scMerge2 (right).

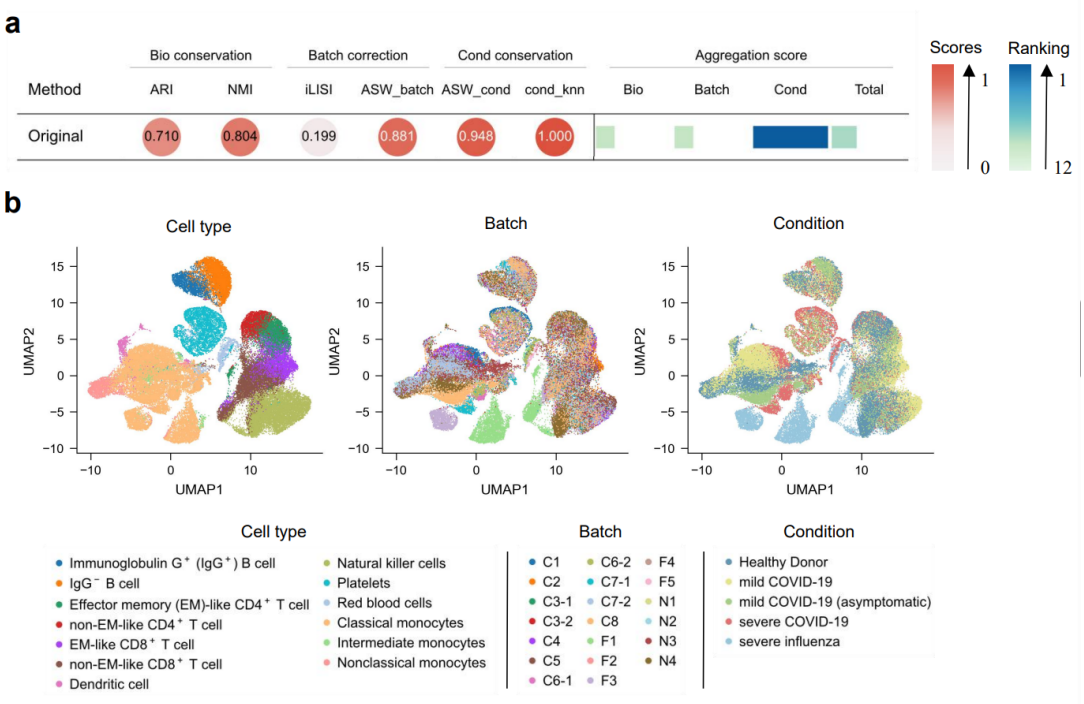


Supplementary Fig. 9. Violin plots display the expression levels of known marker genes corrected by (**a**) original, (**b**) Seurat, (**c**) scMerge2, and (**d**) scFLASH.

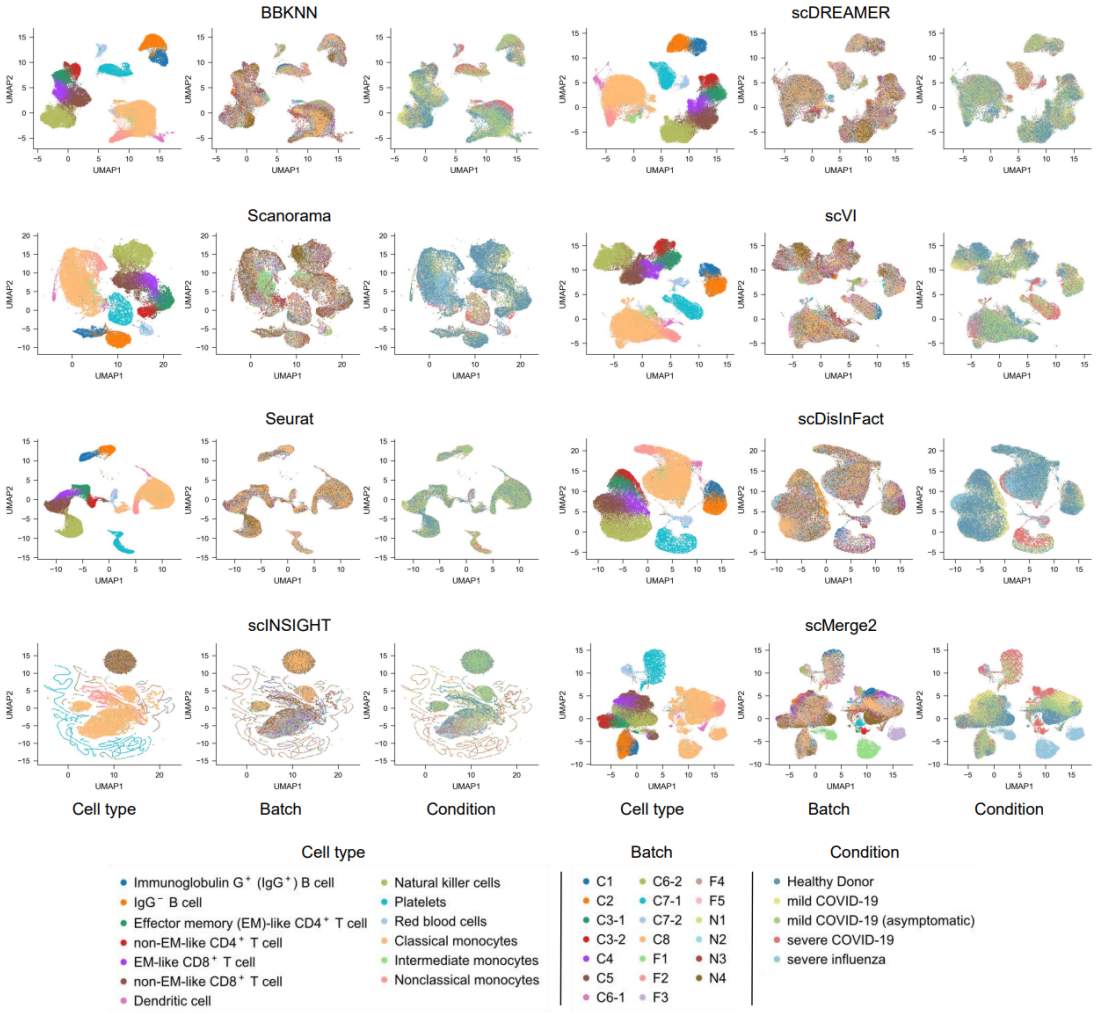
8. Supporting results for main Fig. 4

In this section, we provided supporting results for benchmarking the COVID-19 dataset. Supplementary Fig. 10 showed that the original gene expression data exhibited clustering driven by cell types. However, technical variations between batches, such as classical monocytes from F1 and F3 in the influenza condition, led to batch-specific clustering rather than reflecting real biological differences. Supplementary Fig. 11 displayed UMAP visualizations for the COVID-19 dataset after integration by BBKNN, scDREAMER, Scanorama, scVI, Seurat, scDisInFact, scINSIGHT, and scMerge2.

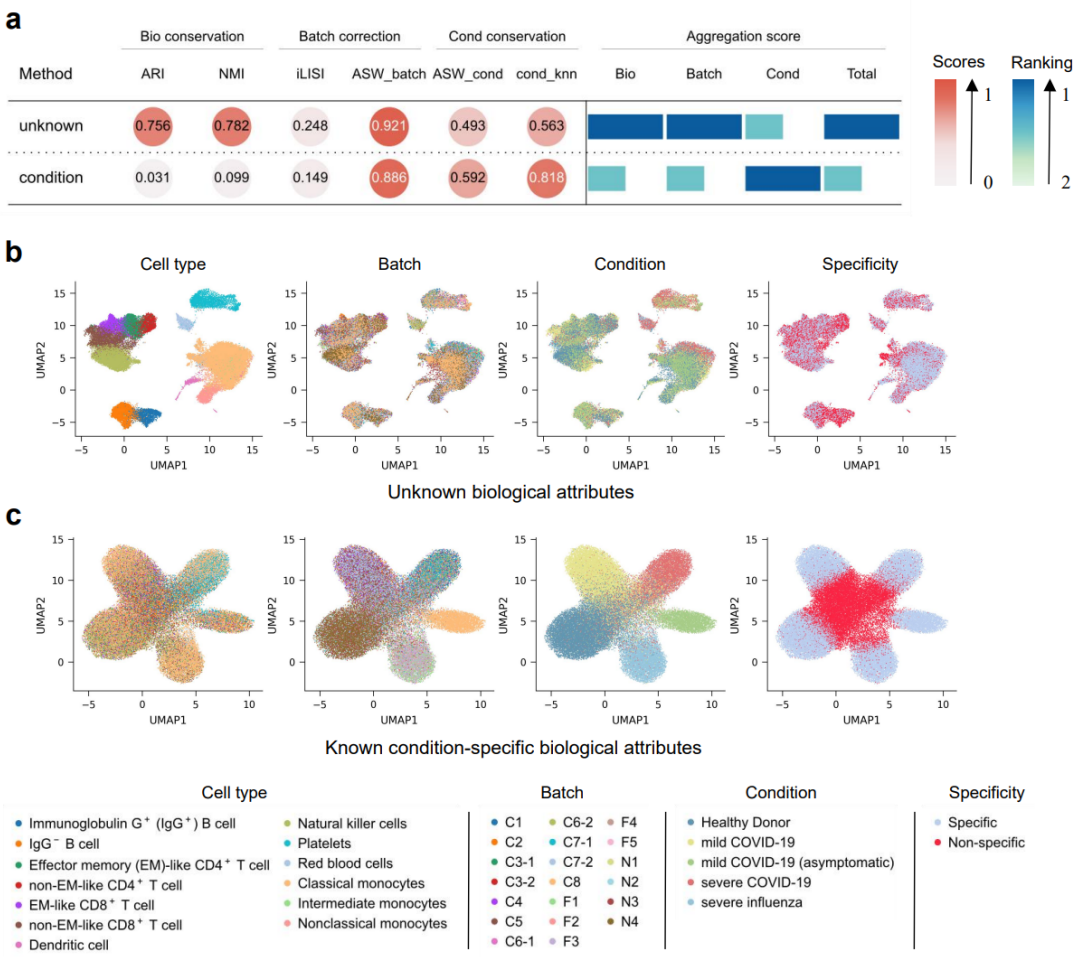
To evaluate scFLASH's disentangling capability, we evaluated the integration performances using the latent known condition-specific attributes and unknown biological attributes. Supplementary Fig. 12a showed that these two embeddings captured different perspectives for biological conservation, batch correction, and condition conservation. Supplementary Fig. 12b and 12c presented UMAP visualizations of latent space embeddings. Notably, even though the COVID-19 dataset has multiple conditions, the condition-specific cell subpopulations identified by scFLASH effectively characterized different states, as showed in Supplementary Fig. 12c.



Supplementary Fig. 10. a, Original metrics for the unintegrated COVID-19 dataset. **b,** UMAP visualizations of the unintegrated COVID-19 dataset. The cells were annotated and colored by cell types (left), batches (middle), and conditions (right).



Supplementary Fig. 11. UMAP visualizations of latent space embeddings for the COVID-19 dataset after integration by eight different methods. The cells were annotated and colored by cell types (left), batches (middle), and conditions (right).

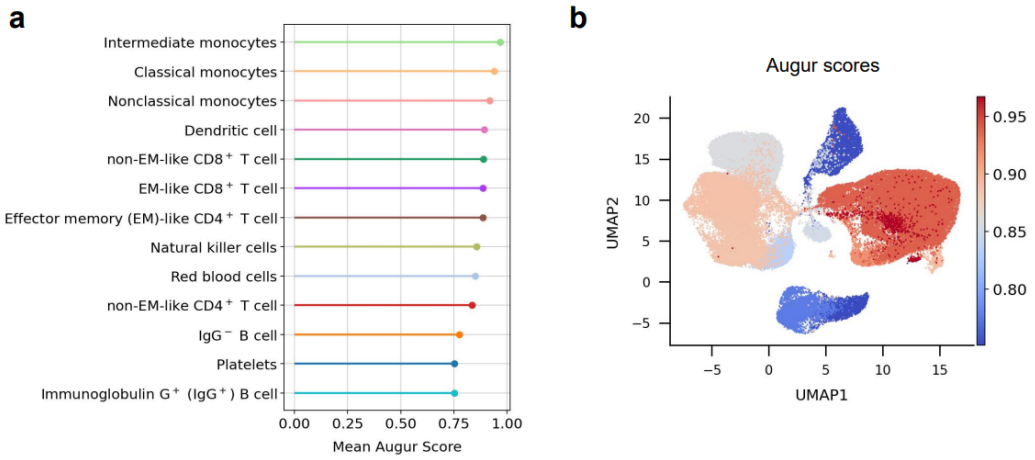


Supplementary Fig. 12. a, The quantitative metrics for unknown biological attributes (unknown) and known condition-specific biological attributes (condition), distinguished by scFLASH in the COVID-19 dataset. **b-c,** UMAP visualizations of COVID-19 dataset using (**b**) unknown biological attributes and (**c**) known condition-specific biological attributes. The cells were annotated and colored by cell types, batches, conditions, and specificities.

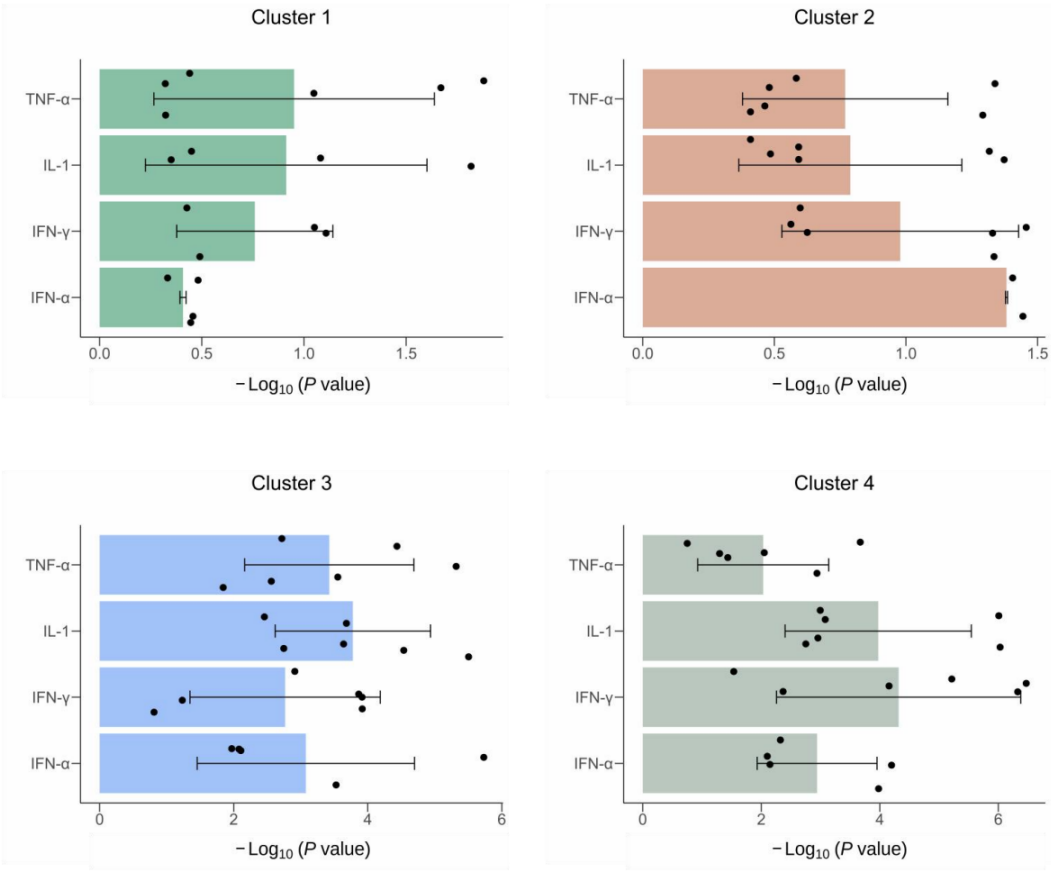
9. Supporting results for main Fig. 5

In this section, we showed supporting materials for the analysis results of scFLASH on COVID-19 datasets as supplements to Fig. 5 in the main text, characterizing the transcriptional identities of classical monocytes under different conditions.

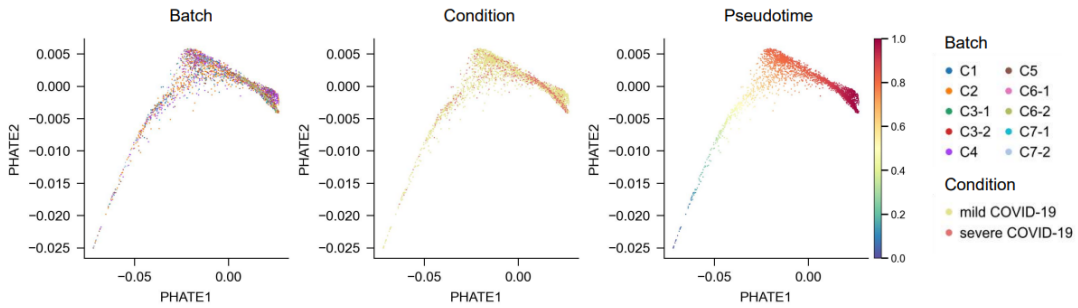
To predict condition-specific prioritizations of cell types, we used the Augur[19] algorithm in the pertpy Python package[20]. Specifically, the expression profiles corrected by scFLASH were adjusted to be non-negative by subtracting the minimum value, serving as the input. The “random_forest_classifier” was selected as the estimator parameter, and Augur was run with default settings. Supplementary Fig. 13 showed that the classical monocytes had the highest Augur score in condition-specific prioritizations. Supplementary Fig. 14 listed the enriched inflammation-related pathways for different gene clusters. Besides, Supplementary Fig. 15 showed that scFLASH arranged CD4 T cells along the severity of disease (mild and severe COVID-19), revealing a connection between pseudotime and different disease states.



Supplementary Fig. 13. a, Lollipop plot of cell type condition-specific prioritizations calculated by Augur. **b,** UMAP visualization using latent space embeddings after integration by scFLASH. The cells were colored by Augur scores.



Supplementary Fig. 14. Enrichment bar plots of cytokine-responsive gene sets from LINCS using DEGs from different clusters. The bars represent the standard errors.

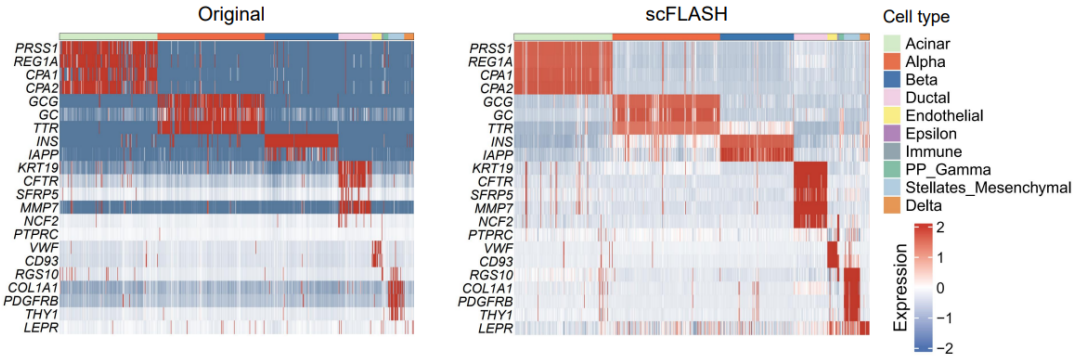


Supplementary Fig. 15. PHATE visualizations of CD4 T cells integrated by scFLASH. The cells were annotated and colored by batches (left), conditions (middle), and pseudotime (right).

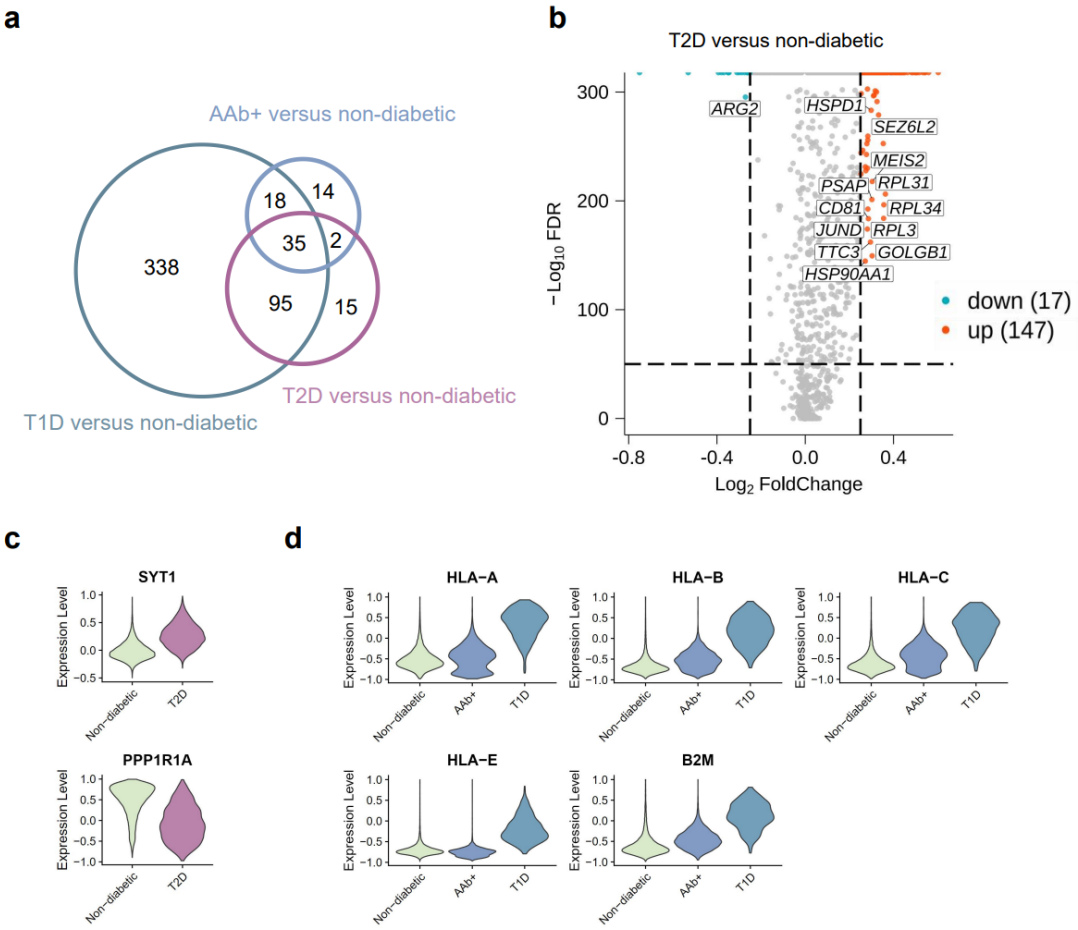
10. Supporting results for main Fig. 6

In this section, we provided some necessary results to support that scFLASH can remove batch effects from different donors while preserving conditional information in diabetes.

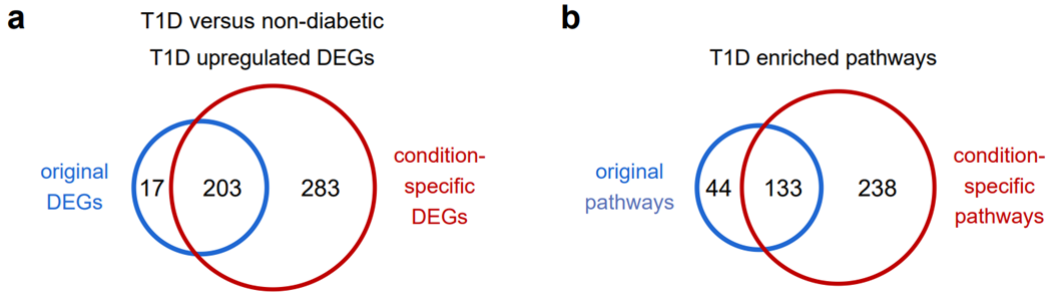
Supplementary Fig. 16 showed that compared to the original data, the expression profiles corrected by scFLASH better captured intercellular variations and reduced the noise. Supplementary Fig. 17 displayed several upregulated genes in T1D and T2D versus non-diabetic individuals. Some of these DEGs exhibited minimal changes in AAb+ patients. Supplementary Fig. 18 indicated that scFLASH-derived condition-specific cells could help to identify more biologically meaningful T1D upregulated DEGs and GO terms. Besides, Supplementary Fig. 19 showed the pseudotime inference results in T2D beta cells.



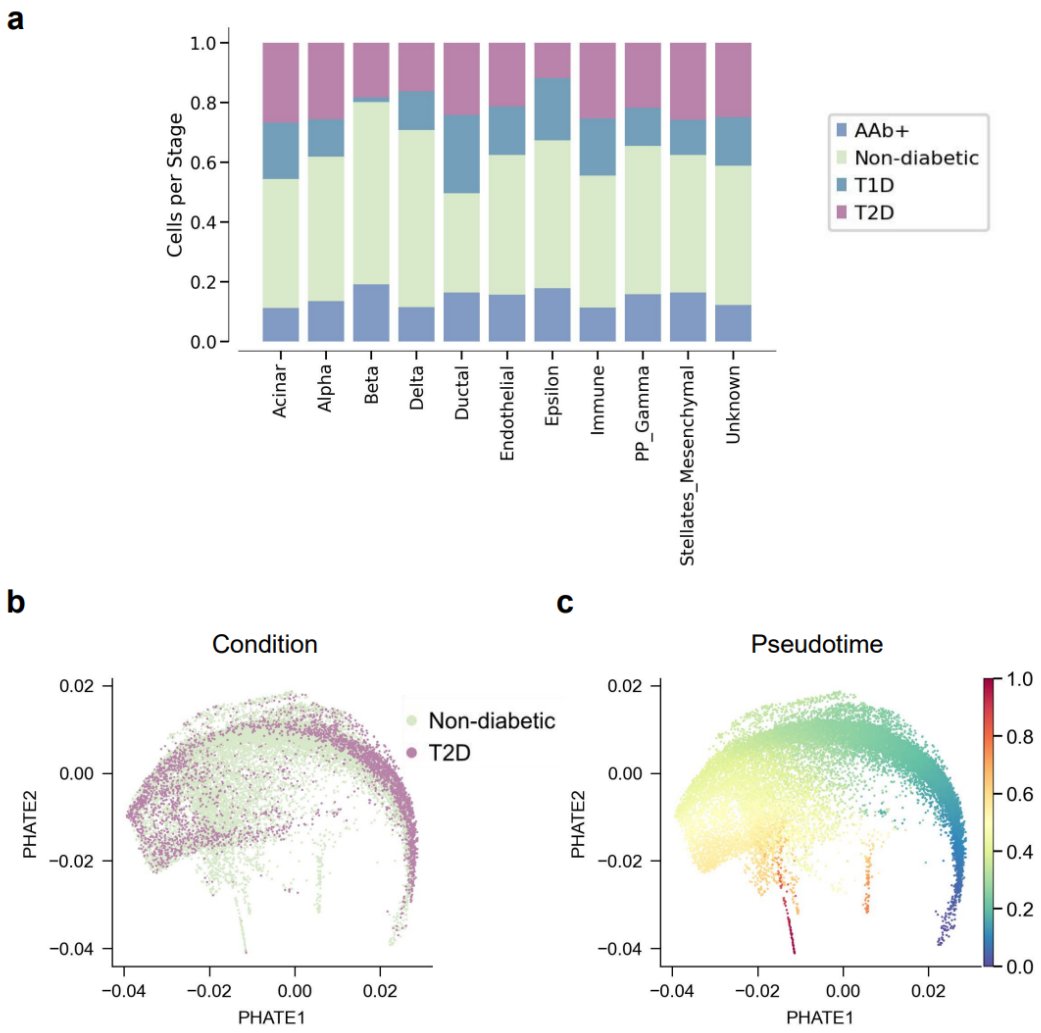
Supplementary Fig. 16. Heat maps of known marker genes using the original expressions (left) and the corrected expressions by scFLASH (right).



Supplementary Fig. 17. a, Venn diagram show overlaps between upregulated DEGs of AAb+, T1D, and T2D compared to non-diabetic. **b,** Volcano plot of DEGs in T2D versus non-diabetic. The two vertical dashed lines represent ±0.25 log-transformed fold changes in gene expression, and the horizontal dashed line denotes an FDR cutoff of 1e-50. The FDR was the adjusted *P* value calculated by the Wilcoxon rank-sum test. **c-d,** Violin plots of expression levels of selected DEGs in (**c**) T2D and (**d**) T1D, respectively.



Supplementary Fig. 18. a-b, Venn diagrams show overlaps between T1D upregulated DEGs (left) and enriched GO terms (right) identified using the original cells and the condition-specific cells.



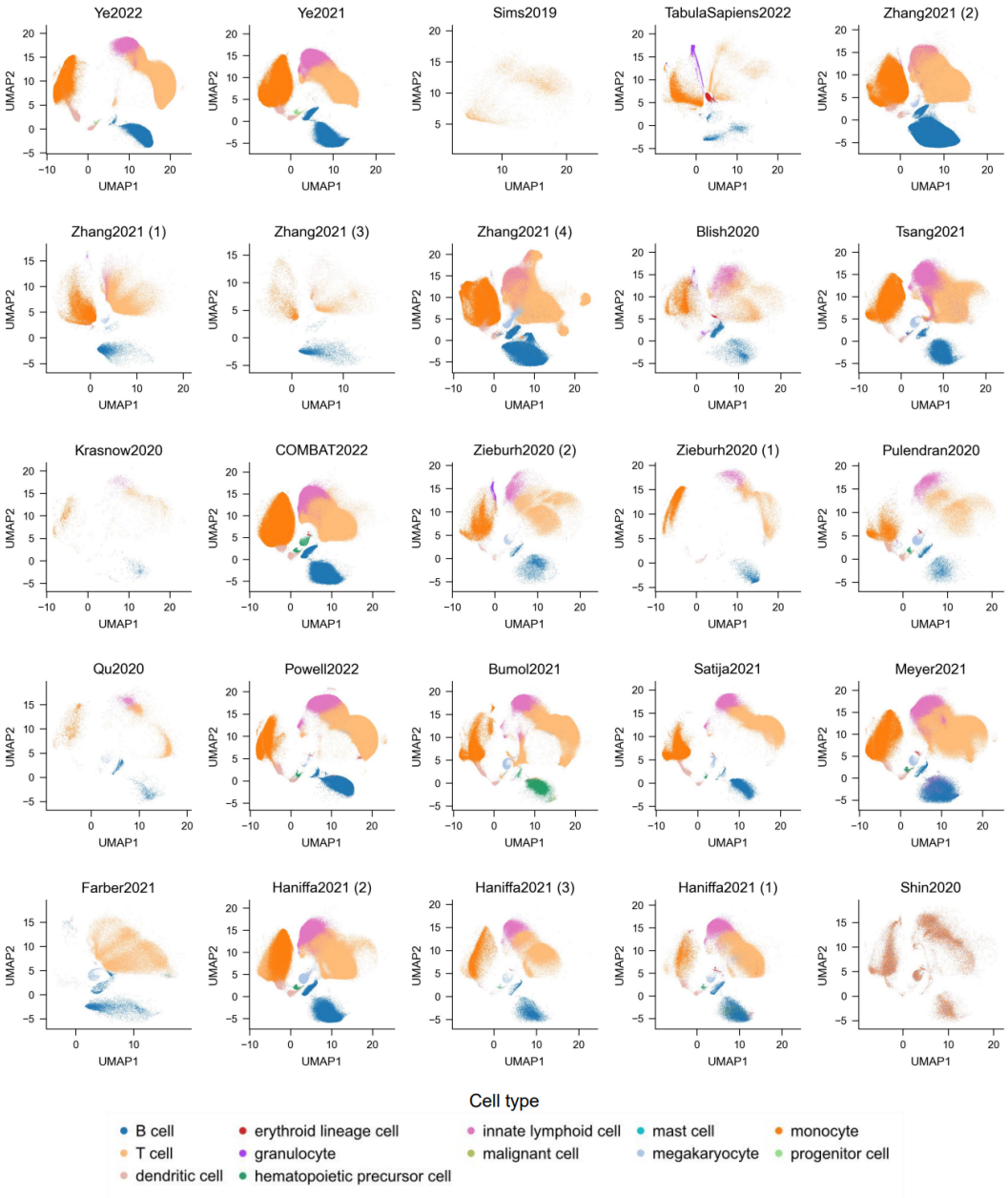
Supplementary Fig. 19. a, Proportions of cells from different conditions in a specific cell type. **b-c,** PHATE visualizations of control and T2D beta cells integrated by scFLASH. The cells were annotated and colored by (**b**) conditions and (**c**) pseudotime.

11. Supporting results for main Fig. 7

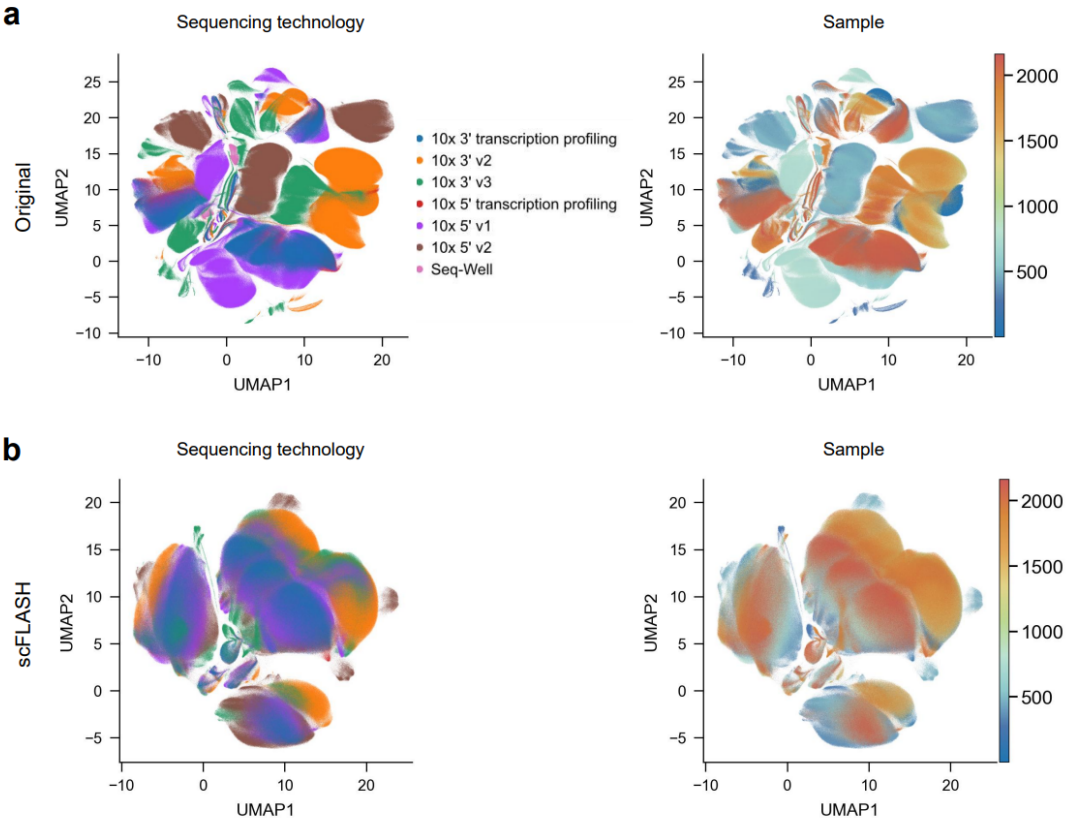
In this section, we presented supporting findings that illustrate the effectiveness of scFLASH in mitigating batch effects across various studies, sequencing technologies, and samples. We also demonstrated scFLASH's applicability on large-scale integration tasks.

Supplementary Fig. 20 used the UMAP visualizations to show that scFLASH successfully separated cell types within each study and aligned similar cell clusters across various studies. Supplementary Fig. 21 showed that scFLASH is a practical integration approach to mitigate batch effects originating from sequencing technologies and samples. Supplementary Fig. 22 indicated that scFLASH effectively distinguished the phenotypic differences among studies such as Zhang2021, Tsang2021, and Meyer2021. Supplementary Fig. 23 displayed the UMAP visualizations of pseudo-bulk profiles for each sample, consistent with the enrichment analysis results for monocyte populations.

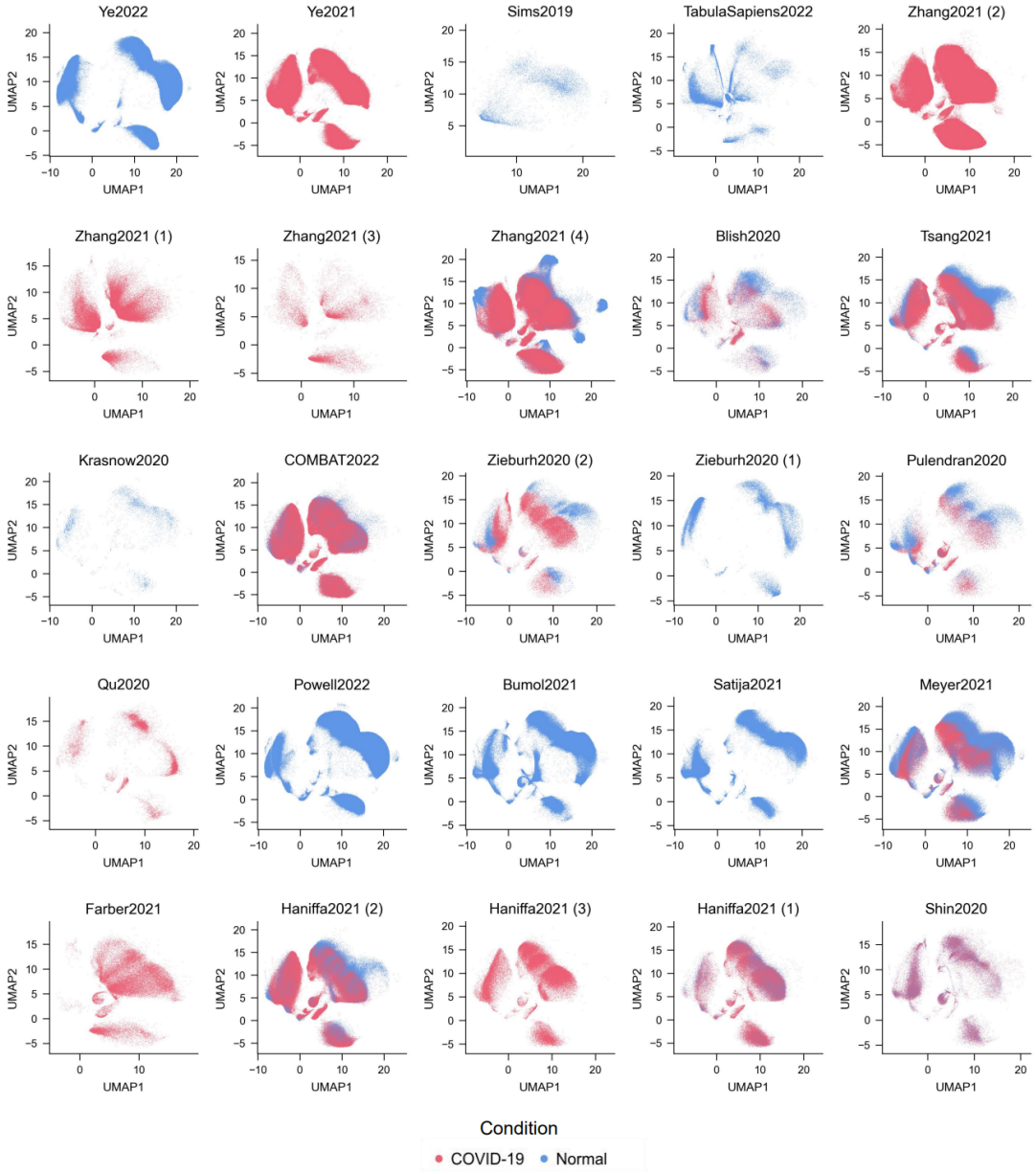
Last, we reported the runtime of scFLASH on the PBMC atlas, with the number of cells ranging from 10,000 to 100,000. The evaluations were performed on a Linux system with an x86_64 CPU, leveraging an A100-PCIE-40GB GPU and operating with CUDA version 11.6. Supplementary Fig. 24 showed that the scFLASH had comparable runtimes with other integration methods and could be executed on large-scale integration tasks efficiently.



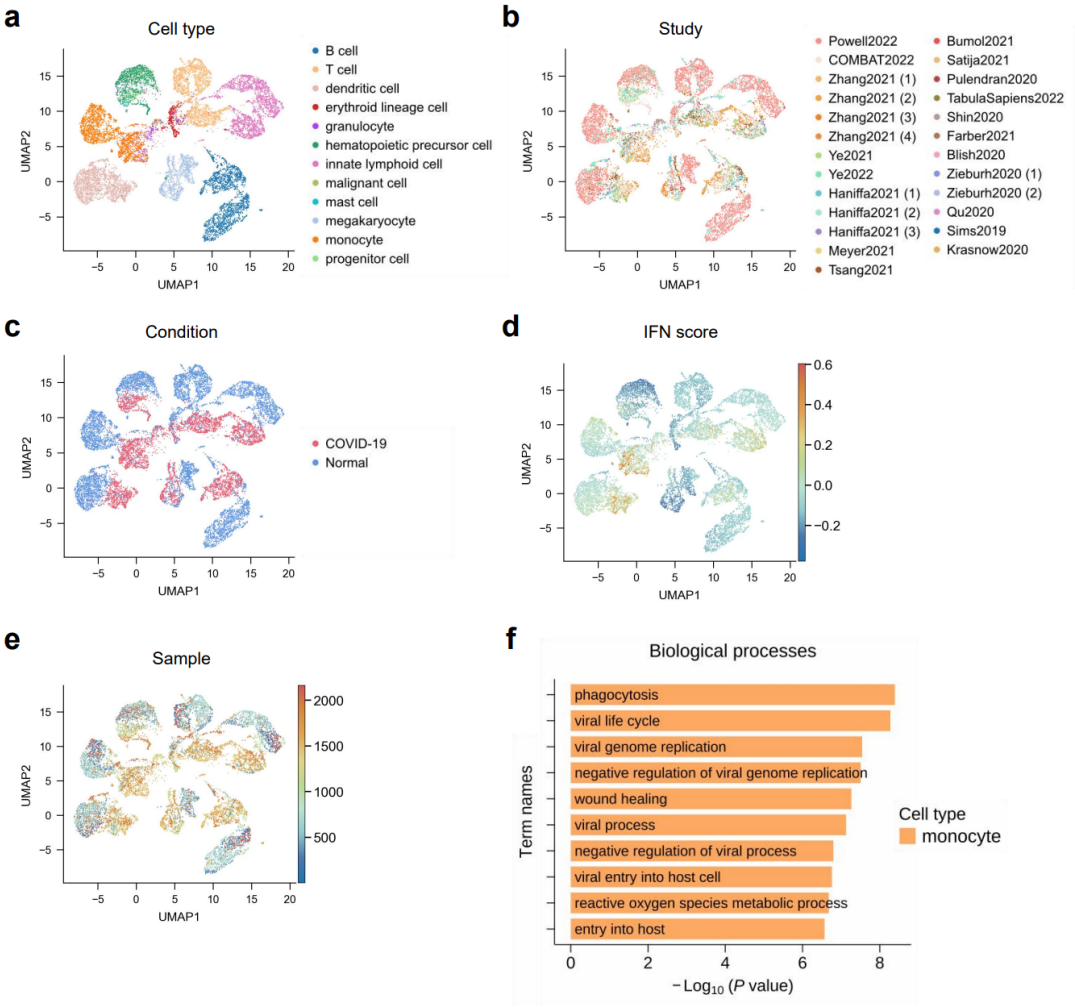
Supplementary Fig. 20. UMAP visualizations using latent space embeddings integrated by scFLASH. The cells were colored by cell types and faceted by studies.



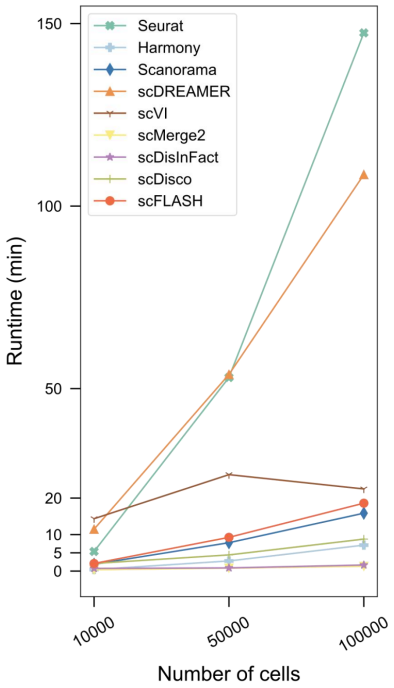
Supplementary Fig. 21. a, UMAP visualizations of the unintegrated PBMC atlas. The cells were annotated and colored by sequencing technologies (left) and samples (right). **b,** UMAP visualizations using latent space embeddings integrated by scFLASH. The cells were annotated and colored by sequencing technologies (left) and samples (right).



Supplementary Fig. 22. UMAP visualizations using latent space embeddings integrated by scFLASH. The cells were colored by conditions and faceted by studies.



Supplementary Fig. 23. a-e, UMAP visualizations of pseudo-bulk profiles for each sample. The cells were annotated and colored by (**a**) cell types, (**b**) studies, (**c**) conditions, (**d**) IFN scores, and (**e**) samples. **f,** Bar plots of enriched pathways using COVID-19 upregulated DEGs in monocytes.



Supplementary Fig. 24. The runtime comparisons of scFLASH and other methods.

References

1. Grubman, A. et al. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* **22**, 2087-2097 (2019).
2. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821 (2019).
3. Lee, J.S. et al. Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. *Sci. Immunol.* **5** (2020).
4. Shapira, S.N., Naji, A., Atkinson, M.A., Powers, A.C. & Kaestner, K.H. Understanding islet dysfunction in type 2 diabetes through multidimensional pancreatic phenotyping: The Human Pancreas Analysis Program. *Cell Metab.* **34**, 1906-1913 (2022).
5. Kaestner, K.H., Powers, A.C., Naji, A., Consortium, H. & Atkinson, M.A. NIH Initiative to Improve Understanding of the Pancreas, Islet, and Autoimmunity in Type 1 Diabetes: The Human Pancreas Analysis Program (HPAP). *Diabetes* **68**, 1394-1402 (2019).
6. De Donno, C. et al. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nat. Methods.* **20**, 1683-1692 (2023).
7. Rand, W.M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**, 846-850 (1971).
8. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods.* **16**, 1289-1296 (2019).
9. Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53-65 (1987).
10. Luecken, M.D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods.* **19**, 41-50 (2022).
11. Polanski, K. et al. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964-965 (2020).
12. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685-691 (2019).
13. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods.* **15**, 1053-1058 (2018).
14. Shree, A., Pavan, M.K. & Zafar, H. scDREAMER for atlas-level integration of single-cell datasets using deep generative model paired with adversarial classifier. *Nat. Commun.* **14**, 7781 (2023).
15. Qian, K., Fu, S., Li, H. & Li, W.V. scINSIGHT for interpreting single-cell gene expression from biologically heterogeneous data. *Genome Biol.* **23**, 82 (2022).
16. Lin, Y., Cao, Y., Willie, E., Patrick, E. & Yang, J.Y.H. Atlas-scale single-cell multi-sample multi-condition data integration using scMerge2. *Nat. Commun.* **14**, 4272 (2023).
17. Zhang, Z., Zhao, X., Bindra, M., Qiu, P. & Zhang, X. scDisInFact: disentangled learning for integration and prediction of multi-batch multi-condition single-cell RNA-sequencing data. *Nat. Commun.* **15**, 912 (2024).

18. Liu, R., Qian, K., He, X. & Li, H. Integration of scRNA-seq data by disentangled representation learning with condition domain adaptation. *BMC Bioinformatics* **25**, 116 (2024).
19. Skinnider, M.A. et al. Cell type prioritization in single-cell data. *Nat. Biotechnol.* **39**, 30-34 (2021).
20. Heumos, L. et al. Pertpy: an end-to-end framework for perturbation analysis. Preprint at bioRxiv https://doi.org/10.1101/2024.1108.1104.606516 (2024).