Additional details

Basic framework for scRNA-seq data

Assuming that the scRNA-seq data is divided into reference data and target data, they can come from the same scRNA-seq dataset or different scRNA-seq datasets. Reference data is recorded as $\mathcal{D}_r = \{(x_i^r, y_i^r)_{i=1}^{n_r}\}$ and target data is recorded as $\mathcal{D}_t = \{(x_i^t)_{i=1}^{n_t}\}$. The label sets of reference and target data are denoted as $\mathcal{C}_r$ and $\mathcal{C}_t$, respectively. To begin, the preprocessed data matrix is denoted as X, where x_{ij} $(1 \leq i \leq n_r + n_t, 1 \leq j \leq m)$ represents the expression level of the i-th cell on the j-th gene and m represents the number of genes.

First, considering the discrete, sparse, and large variance characteristics of scRNA-seq data, we use the zero-inflated negative binomial (ZINB) distribution to model this gene expression pattern [1], that is:

$$p_{ZINB}(x_{ij}^*|\pi_{ij}, \mu_{ij}, \theta_{ij}) = \pi_{ij}\delta_{x_{ij}^*=0} + (1-\pi_{ij})\times \frac{\Gamma(x_{ij}^* + \theta_{ij})}{\Gamma(x_{ij}^*+1)\Gamma(\theta_{ij})} \times (\frac{\theta_{ij}}{\theta_{ij}+\mu_{ij}})^{\theta_{ij}} \times (\frac{\mu_{ij}}{\theta_{ij}+\mu_{ij}})^{x_{ij}^*}. \quad (1)$$

Among them, x_{ij}^* represents the raw read counts of the i-th cell on the j-th gene. π_{ij}, μ_{ij}, θ_{ij} represent the zero-inflated parameters, mean parameters, and dispersion parameters, respectively, and they constitute the parameters to be estimated for the model.

Due to the complex interaction between genes, these three sets of parameters are not independent of each other but fall into a low-dimensional manifold. Therefore, we use the DCA model to estimate the parameters, and at the same time, to approximate the manifold, to effectively reduce the dimension and denoise the scRNA-seq data [2]. Specifically, let $f_e(x) : R^m \rightarrow R^d$ be the encoder function that maps the cells into the low-dimensional embedding space and gets the embedding representation $z = f_e(x)$. Similarly, let $f_d(x) : R^d \rightarrow R^m$ be the decoder function and get the reconstructed variable $x_r = f_d(z)$. Then we use the reconstruct variable x_r to estimate the parameters:

$$\hat{\pi} = sigmoid(w_\pi' x_r);\ \hat{\theta} = exp(w_\theta' x_r);\ \hat{\mu} = exp(w_\mu' x_r) \quad (2)$$

where w_π, w_θ, w_μ are the corresponding weights. Given the parameters, we can assume that the conditional distribution of the reconstructed data is independent, so we can use the negative log-likelihood of ZINB distribution [3, 4] as the first loss function:

$$\mathcal{L}_{zinb} = -\sum_{i=1}^{n_r+n_t}\sum_{j=1}^{m} p(x_{ij}^*|\hat{\pi}_{ij}, \hat{\mu}_{ij}, \hat{\theta}_{ij}). \quad (3)$$

Using data reconstruction as another kind of regularization can help reveal the global probabilistic structure [5].

As shown in previous work (scNAME) [5], the ZINB-based denoising network is less capable of capturing the correlations across genes. Inspired by the recent progress in semi-supervised learning, we use the same data augmentation strategy as in scNAME to generate different gene expression matrices. Specifically, we first construct two auxiliary matrices: a binary mask matrix B that samples from the Bernoulli distribution and a shuffled expression matrix X' obtained by randomly shuffling the original data within each feature column. Then the augmented data matrix $\tilde{X}$ can be generated as:

$$\tilde{X} = B \odot X' + (1-B) \odot X, \quad (4)$$

where $\odot$ represents element-wise multiplication. After $\tilde{X}$ passes through the denoising autoencoder network, we can get the estimation value $\hat{B}$ of mask matrix B. To account for the dependencies of genes, the binary cross-entropy loss is applied to train the model, that is,

$$\mathcal{L}_{mask} = -\sum_{i=1}^{n_r+n_t}\sum_{j=1}^{m}(B_{ij}\log\hat{B}_{ij} + (1-B_{ij})\log(1-\hat{B}_{ij})). \quad (5)$$

To sum up, the overall loss of our basic framework is

$$\mathcal{L}_{den} = \mathcal{L}_{zinb} + \mathcal{L}_{mask} \quad (6)$$

Basic framework for spatial transcriptomics data

We start with the graph construction process. Assuming that the spatial coordinates of cells are $\{s_i^r\}_{i=1}^{n_r}$ and $\{s_i^t\}_{i=1}^{n_t}$ for reference and target data, respectively. We first calculate the Euclidean distances d_{ij} for each pair of cells (s_i, s_j) from the same region, and the edge is generated if $d_{ij} < \kappa$, where κ is a tunable threshold. We set $\kappa = 50$ by default. This graph construction step is independent of the subsequent method and can be changed as long as the constructed graph meaningfully reflects spatial similarities between cells. In this way, we can obtain the reference graph $\mathcal{G}_r$ and the target graph $\mathcal{G}_t$ for reference and target data, respectively. The node features correspond to gene expressions of cells.

Then we apply the encoder function to map cells from both graphs into a joint embedding space that captures spatial and molecular similarities between the cells. The cell embedding encoder function f_θ is parameterized by a learnable graph convolutional neural network (GCN) [6]. The encoder function f_θ generates the low-dimensional cell embedding z. Specifically, the encoder function contains one fully connected layer followed by the nonlinear activation function,

$$h_i^{(1)} = \phi(W^{(0)}h_i^{(0)} + b^{(0)}), \quad (7)$$

where $h_i^{(k)}$ is the hidden state of node s_i in k-th layer of the neural network and $k = 0, 1$. W is a parameter matrix, b is bias vector and ϕ denotes nonlinear activation function. The hidden state $h_i^{(0)}$ in layer 0 is set to the node feature x_{s_i}, i.e., a gene expression vector. The rectified linear unit (ReLU) is used as the activation function ϕ : ReLU$(\cdot) = \max(0, \cdot)$. Then we use a graph convolutional layer to enable message passing among nearby cells,

$$h_i^{(2)} = W_0^{(1)}h_i^{(1)} + \sum_{j\in\mathcal{N}^i} W_1^{(1)}h_j^{(1)}, \quad (8)$$

where $\mathcal{N}^i$ denotes neighborhood on node s_i. The final embedding of node s_i is $z_i = h_i^{(2)}$. On top of the encoder function, we add two parallel linear layers parameterized by the reference and target prototypes to achieve cell annotation and clustering. The implementation of this part is the same as that of scRNA-seq data.

Implementation of baseline methods.

The details of the eight comparative baselines are summarized in Table S1, including three clustering methods and five

annotation methods. To ensure the fairness of the comparison, we run the algorithm at their default settings to get the results. When the scRNA-seq algorithms are applied to spatial transcriptome data, we do not need to use the information of the spatial coordinate set.

Real dataset information.

Biological and statistical information of the five scRNA-seq datasets and one spatial transcriptomics dataset we used in intra-data annotation experiments are shown in Table S2, and the five groups of scRNA-seq datasets and one group of spatial transcriptomics dataset we used in cross-data annotation experiments are shown in Table S3. During the experiments, unless otherwise noted, we split the whole dataset into common cell types and private cell types according to Table S4.

Additional results

Specific values. In the text, we present the results on each experimental dataset in the form of radar charts, and here we also present the specific values of these results in Table S5, Table S6, and Table S7 for the convenience of readers to consult and compare. It can be seen that our approach is competitive in all three kinds of accuracy for both the scRNA-seq dataset and the spatial transcriptome dataset.

Statistical analysis. To prove the consistency and stability of the results of our method, we report their standard deviation values. Corresponding to Table S5, Table S6, and Table S7, the standard deviations of three-run results are within the interval (0.5, 2.5) for scBOL, which fluctuates relatively little. We also conduct the significance test of the improvements in results. Specifically, for scRNA-seq data, we choose the best clustering baseline scCNC and the best annotation baseline MARS to perform the one-sided pairwise t-test with scBOL on the overall accuracy. Corresponding to Table S5 and Table S6, the p-values are 2.4e-6 (scBOL vs scCNC) and 2.1e-08 (scBOL vs MARS), demonstrating that the improvement is significant. For spatial transcriptomics data, we also take a one-sided pairwise t-test between STELLAR and scBOL on the overall accuracy. Corresponding to Table S7, the p-value is 2.4e-3 (scBOL vs STELLAR), which validates that our method is significantly better than STELLAR.

Robustness analysis. For a better illustration, we provide detailed tables of the results of robustness analysis on the tested datasets. For scRNA-seq data, Table S8 shows the results for the three kinds of accuracy, i.e., annotation accuracy, clustering accuracy, and overall accuracy, with the varying of novel cell type number $|\mathcal{C}_n|$. From the table, we can see that no matter which accuracy we used, the performance of scBOL is always stable and excellent without being affected by changes in $|\mathcal{C}_n|$. In contrast, the three kinds of accuracy of the other methods all show relatively large fluctuations, which validates the robustness of scBOL to the novel cell type number. Table S9 shows the results for the three kinds of accuracy with the varying ratios of labeled data. we can conclude that the performance of scBOL is satisfactory and stable even when encountering dramatic changes in the ratio of labeled data. However, there is a certain degree of oscillation with all other competitive methods. Besides, our method is the best performer among all methods, validating its effectiveness and practicality. For spatial transcriptomics data, we give the results of the three kinds of accuracy with the different labeled ratios in Table S10. It can be seen that scBOL is consistently better than the other baselines regardless of the labeled ratios.

In alignment with the concept of the labeled ratio, we introduce a "novel ratio", defined as the fraction of sampled target private cells arising from previously unidentified cell types. This metric indirectly reflects the challenge associated with uncovering new cell types. By default, the novel ratio is established at 1.0 in the primary text. To examine the robustness of varied methods in response to alterations in the novel ratio, controlled experiments were executed utilizing the Quake 10x and Quake Smart-seq2 datasets. The novel ratio was varied incrementally, adopting values of 0.2, 0.4, 0.6, 0.8, and 1.0 during experimentation. Data presented in Figure S1 indicate that as the novel ratio increases—corresponding to a higher proportion of target private cells—the precision of scBOL in classifying novel cell types displays a marginal ascent, while its accuracy for all cell types experiences a minimal decline. This suggests that an augmentation in the quantity of novel cells may exert a subtle deleterious effect on the capability to identify established cell types. In comparison, other methodologies exhibit more pronounced fluctuations in accuracy pertaining to both known and novel cell types upon varying the novel ratio. ItClust, notably, demonstrates this sensitivity, which serves to accentuate scBOL's relative stability. Despite scDECL's consistent performance across known and novel accuracy, it fails to capitalize on the informative potential of reference data, resulting in accuracy levels that lag significantly behind alternative approaches. Furthermore, scBOL maintains superior overall accuracy irrespective of the novel ratio fluctuations, affirming the method's dominance and resilience when confronted with an increasing presence of target private cells. For enhanced clarity, an exhaustive tabulation of the results has been provided. The conclusions drawn from the tabular data in Table S11 align with those observed in the corresponding line graph, reinforcing the findings.

Low abundance experiment. Given that cell types of newfound interest typically present in low abundance under real-world conditions, we opted for the Quake 10x Limb Muscle and Quake 10x Mammary Gland as our reference and target datasets, respectively, for the current investigation. Within these datasets, prevalent cell populations such as T cells, B cells, and macrophages were designated as common cell types, while rarer populations including stromal cells, luminal epithelial cells, endothelial cells, and basal cells were categorized as novel cell types due to their smaller sample sizes. To evaluate the efficacy of our algorithm, scBOL, we conducted a comparative analysis with three alternative algorithms. The performance metrics for these algorithms were as follows: ItClust achieved 89.6 accuracy on known cell types, 45.9 on novel cell types, and an overall accuracy of 75.4; scNym scored 98.9 on known types, 38.1 on novel types, and 55.4 overall; and scArches attained 88.4, 55.1, and 75.0 across the respective categories. In contrast, scBOL outperformed these algorithms with impressive scores of 99.1 on known cell types, 93.3 on novel cell types, and an overall accuracy of 96.8. This denotes a significant advantage of scBOL in accurately identifying both common and novel cell types, despite the latter's reduced prevalence, underscoring its applicability in a variety of practical contexts. Additionally, we incorporated a Sankey diagram to provide a more visually intuitive comparison of scBOL's predictive proficiency relative to that of the other algorithms. As exemplified in Figure S2, we can deduce that scBOL not only precisely identifies cells belonging to infrequent

novel classes but also aligns a vast array of common cell types with corresponding cells in the reference dataset. In contrast, competing methods failed to match this performance. For instance, ItClust pooled most novel cell types into a single cluster. Meanwhile, scNym predominantly misclassified novel cells as known macrophages while incorrectly segregating some T cells into an additional novel group. Furthermore, scArches wrongly grouped basal cells and luminal epithelial cells into a collective novel cluster. Overall, our findings reinforce the robustness and superiority of our method in detecting low-abundance cell types within heterogeneous datasets.

Additional spatial transcriptomic data. Our method has proven to be versatile, showing applicability to data sourced from a myriad of spatial transcriptomic technologies, encompassing those that rely on in situ hybridization, sequencing, and imaging modalities. Within the body of the text, we have focused our analysis on two datasets hailing from the CODEX multiplexed imaging technique. Furthering our investigations, we have expanded our experimental scope to encompass additional datasets derived from in situ hybridization and sequencing technologies, namely those curated by Lohoff, utilizing seqFISH, and Chen, through Stereo-seq. Their fundamental characteristics are detailed in Table S2. Our comparative studies, involving scBOL and five alternative methods, are showcased in Figure S3. The comprehensive metrics are tabulated in Table S12. These results underscore the superior and consistent performance of scBOL across both datasets. For a granular and graphical examination of the annotation outcomes, we have leveraged both Sankey diagrams and UMAP plots to illustrate the methodologies' efficacies on the Lohoff dataset. The Sankey diagram, depicted in Figure S4A, reveals that scBOL adeptly categorizes extant cell types and uncovers new ones with efficacy not mirrored by other methods, which often conflate disparate cell types into a single cluster. This finding is further corroborated by the UMAP visualization in Figure S4B, where scBOL distinctly segregates various cell types within the embedding space—a feat not replicated by competing algorithms. In summation, the exemplary performance of scBOL across diverse spatial transcriptomic datasets not only underscores its robustness but also its potential for broad adoption in the field.

Differential gene expression analysis. Our methodology is capable of assigning cluster labels to target private cells; however, the biological relevance of these clusters remains to be substantiated. To corroborate the capability of scBOL in unveiling novel cell types, we conducted marker gene identification in the context of a biological analysis, which involved classification and clustering delineation. This process entailed leveraging gene expression matrices alongside both predicted and true labels to isolate differentially expressed genes (DEGs), utilizing the Scanpy software suite [7]—a step critical for pinpointing marker genes within each cluster. Subsequently, we ascertained the degree of concordance between the top 100 DEGs of both the true clusters and the predicted ones. The proportional overlap, divided by 100, furnished a quantitative measure of similarity, thereby indicating the biological pertinence of the novel cell types identified by our method. This experimental approach was applied to two scRNA-seq datasets: the Vento Smart-seq2 dataset served as the reference, and the Vento 10x dataset as the target. Comparative analysis of the outcomes, as depicted in Figure S5, demonstrates that scBOL efficaciously discerned four novel cell types enriched with biological significance, which are also discernibly distinct from previously characterized cell types. In contrast, while scArches and scNym exhibited moderate performance in classifying known cell types, they were less effective in accurately identifying the four novel cell types, particularly with regard to the natural killer cells and trophoblast cells.

Running time experiment. To investigate the association between cell count and the computational time required for model training, we employed the widely recognized Splatter package [8] to simulate datasets comprising 5k, 10k, 25k, 50k, and 100k cells. Each dataset was composed of 10 balanced cell types, encompassing a total of 5000 genes. The simulation settings included a dropout median (dropout.mid) of 0.5, a dropout shape (dropout.shape) parameter set to -1, and a differential expression factor scale (de.facScale) of 0.2. To enhance the robustness of our findings, we produced three distinct datasets for each specified size by initiating the simulation process with varying random seeds. Furthermore, we partitioned the 10 cell types into two subgroups: the first five were designated as known cell types and the last five as novel cell types. Within the known category, we equally split the cells into reference and target data subsets. Subsequent model computations were performed on a consistent hardware setup, utilizing a cluster machine equipped with 2 Tesla A100 GPUs. During these computational experiments, we meticulously documented both the run-time duration and the overall accuracy. Our analysis entailed averaging the metrics across the separate datasets corresponding to identical cell counts. These aggregated results were systematically tabulated and illustrated in Figure S6 and Table S13, respectively. Inspection of these results revealed a notable pattern: with increasing cell numbers, the runtime exhibited divergent trends across the different methodologies tested. Notably, only MARS and scBOL demonstrated significant improvements in overall accuracy, with scBOL consistently outperforming other methods. Except for scCNC and scDECL, scBOL demonstrated a trend that closely approximated a linear increase, much in alignment with four other methods. Contrastingly, scCNC and scDECL's computational times surged disproportionately beyond a specific dataset threshold. In the context of computational efficiency, scBOL performed admirably; it was slower compared to scNym, ItClust, and MARS but showed a superior speed relative to scArches, scDECL, and scCNC.

Hyperparameter sensitivity. We first investigate the robustness of annotation results when using different temperature parameter τ settings (0.1, 0.2, 1.0, 5.0, 10.0) on Quake 10x and Quake Smart-seq2 datasets. Figure S7(a) shows that the performance of scBOL is quite robust no matter the variation of τ, which indicates the stability of our method. Meanwhile, we also examine the robustness of scBOL by artificially changing the value of sample ratio α, which varies in the range of [10, 15, 20, 25, 30]. From Figure S7(b), it is easy to conclude that the overall accuracy of scBOL changes little and is relatively stable in the face of the variation of α.

References

1. Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.
2. Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications*,

10(1):1–14, 2019.

3. Kelvin KW Yau, Kui Wang, and Andy H Lee. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal: journal of mathematical methods in biosciences*, 45(4):437–452, 2003.
4. Liang Chen, Qiuyan He, Yuyao Zhai, and Minghua Deng. Single-cell rna-seq data semi-supervised clustering and annotation via structural regularized domain adaptation. *Bioinformatics*, 37(6):775–784, 2021.
5. Hui Wan, Liang Chen, and Minghua Deng. scname: neighborhood contrastive clustering with ancillary mask estimation for scrna-seq data. *Bioinformatics*, 38(6):1575–1583, 2022.
6. Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*, pages 1–6. Ieee, 2017.
7. F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5, 2018.
8. Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.
9. Hai-Yun Wang, Jian-Ping Zhao, Chun-Hou Zheng, and Yan-Sen Su. sccnc: a method based on capsule network for clustering scrna-seq data. *Bioinformatics*, 2022.
10. Yanglan Gan, Yuhan Chen, Guangwei Xu, Wenjing Guo, and Guobing Zou. Deep enhanced constraint clustering based on contrastive learning for scrna-seq data. *Briefings in Bioinformatics*, 24(4):bbad222, 2023.
11. Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications*, 13(1):1739, 2022.
12. Maria Brbić, Marinka Zitnik, Sheng Wang, Angela O Pisco, Russ B Altman, Spyros Darmanis, and Jure Leskovec. Mars: discovering novel cell types across heterogeneous single-cell experiments. *Nature methods*, 17(12):1200–1206, 2020.
13. Jian Hu, Xiangjie Li, Gang Hu, Yafei Lyu, Katalin Susztak, and Mingyao Li. Iterative transfer learning with neural network for clustering and cell type classification in single-cell rna-seq analysis. *Nature machine intelligence*, 2(10):607–618, 2020.
14. Jacob C Kimmel and David R Kelley. scnym: Semi-supervised adversarial neural networks for single cell classification. *bioRxiv*, 2020.
15. Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, et al. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40(1):121–130, 2022.
16. Maria Brbić, Kaidi Cao, John W Hickey, Yuqi Tan, Michael P Snyder, Garry P Nolan, and Jure Leskovec. Annotation of spatially resolved single-cell data with stellar. *Nature Methods*, 19(11):1411–1418, 2022.
17. Junyue Cao, Jonathan S Packer, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667, 2017.
18. Tabula Muris Consortium et al. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367–372, 2018.
19. Daniel E Wagner, Caleb Weinreb, Zach M Collins, James A Briggs, Sean G Megason, and Allon M Klein. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 360(6392):981–987, 2018.
20. Amit Zeisel, Hannah Hochgerner, Peter Lönnerberg, Anna Johnsson, Fatima Memic, Job Van Der Zwan, Martin Häring, Emelie Braun, Lars E Borm, Gioele La Manno, et al. Molecular architecture of the mouse nervous system. *Cell*, 174(4):999–1014, 2018.
21. Yury Goltsev, Nikolay Samusik, Julia Kennedy-Darling, Salil Bhate, Matthew Hale, Gustavo Vazquez, Sarah Black, and Garry P Nolan. Deep profiling of mouse splenic architecture with codex multiplexed imaging. *Cell*, 174(4):968–981, 2018.
22. T Lohoff, S Ghazanfar, A Missarova, N Koulena, N Pierson, JA Griffiths, ES Bardot, C-HL Eng, RCVg Tyser, R Argelaguet, et al. Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. *Nature biotechnology*, 40(1):74–85, 2022.
23. Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185(10):1777–1792, 2022.
24. Mauro J Muraro, Gitanjali Dharmadhikari, Dominic Grün, Nathalie Groen, Tim Dielen, Erik Jansen, Leon Van Gurp, Marten A Engelse, Francoise Carlotti, Eelco Jp De Koning, et al. A single-cell transcriptome atlas of the human pancreas. *Cell systems*, 3(4):385–394, 2016.
25. Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems*, 3(4):346–360, 2016.
26. Roser Vento-Tormo, Mirjana Efremova, Rachel A Botting, Margherita Y Turco, Miquel Vento-Tormo, Kerstin B Meyer, Jong-Eun Park, Emily Stephenson, Krzysztof Polański, Angela Goncalves, et al. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature*, 563(7731):347–353, 2018.
27. Adam L Haber, Moshe Biton, Noga Rogel, Rebecca H Herbst, Karthik Shekhar, Christopher Smillie, Grace Burgin, Toni M Delorey, Michael R Howitt, Yarden Katz, et al. A single-cell survey of the small intestinal epithelium. *Nature*, 551(7680):333–339, 2017.
28. Lindsey W Plasschaert, Rapolas Žilionis, Rayman Choo-Wing, Virginia Savova, Judith Knehr, Guglielmo Roma, Allon M Klein, and Aron B Jaffe. A single-cell atlas of the airway epithelium reveals the cftr-rich pulmonary ionocyte. *Nature*, 560(7718):377–381, 2018.
29. Daniel T Montoro, Adam L Haber, Moshe Biton, Vladimir Vinarsky, Brian Lin, Susan E Birket, Feng Yuan, Sijia Chen, Hui Min Leung, Jorge Villoria, et al. A revised airway epithelial hierarchy includes cftr-expressing ionocytes. *Nature*, 560(7718):319–324, 2018.

Table S1. Summary of eight baseline methods for comparison.

	Method	Year	Programming	Download URL
Clustering	scCNC [9]	2022	Python	https://github.com/WHY-17/scCNC
	scDECL [10]	2023	Python	https://github.com/DBLABDHU/scDECL
	STAGATE [11]	2022	Python	https://github.com/QIFEIDKN/STAGATE
Annotation	MARS [12]	2020	Python	https://github.com/snap-stanford/mars
	ItClust [13]	2020	Python	https://github.com/jianhuupenn/ItClust
	scNym [14]	2020	Python	https://www.github.com/calico/scnym
	scArches [15]	2022	Python	https://github.com/theislab/scarches
	STELLAR [16]	2022	Python	https://github.com/snap-stanford/stellar

Table S2. Real datasets information without batch effect in intra-data annotation experiments.

Dataset	Organ	Platform	Cell types	Cells	Reference
Cao	Atlas	sci-RNA-seq	16	30960	[17]
Quake_10x	Atlas	10x	36	53587	[18]
Quake_Smart-seq2	Atlas	Smart-seq2	45	41526	[18]
Wagner	Atlas	inDrop	14	34348	[19]
Zeisel	Nervous System	10x	17	110704	[20]
Hubmap	Intestine	CODEX	6	110633	[21]
Lohoff	Embryo	seqFISH	6	11529	[22]
Chen	Embryo	Stereo-seq	8	38746	[23]

Table S3. Real datasets information with batch effect in inter-data annotation experiments.

Domain	Dataset	Organ	Platform	Cell types	Cells	Reference
Reference	Muraro	Pancreas	Smart-seq2	4	1724	[24]
Target	Baron_human	Pancreas	inDrop	8	8451	[25]
Reference	Vento-Tormo Smart-seq2	Placenta	Smart-seq2	4	4310	[26]
Target	Vento-Tormo 10x	Placenta	10x	8	54976	[26]
Reference	Mammary Smart-seq2	Mammary Gland	Smart-seq2	4	2405	[18]
Target	Mammary 10x	Mammary Gland	10x	7	4481	[18]
Reference	Haber largecell	Small Intestine	10x	3	10396	[27]
Target	Haber region	Small Intestine	10x	6	11665	[27]
Reference	Plasschaert	Trachea	inDrop	2	6152	[28]
Target	Montoro 10x	Trachea	10x	4	7006	[29]
Reference	Tonsil	Tonsil	CODEX	3	173968	[21]
Target	BE	Barretts Esophagus	CODEX	6	45958	[21]

Table S4. The split information of datasets used in our intra-data and inter-data cell type identification experiments.

	Cao	Quake1	Quake2	Wagner	Zeisel	Hubmap
$\lvert\mathcal{C}_s\rvert$	6	12	15	5	3	3
$\lvert\bar{\mathcal{C}}_r\rvert$	4	12	15	4	0	0
$\lvert\bar{\mathcal{C}}_t\rvert$	6	12	15	5	3	3
	Muraro (R) Baron_human (T)	**Vento Smart-seq2 (R) Vento 10x (T)**	**Plasschaert (R) Montoro 10x (T)**	**Mammary Smart-seq2 (R) Mammary 10x (T)**	**Haber largecell (R) Haber region (T)**	**Tonsil (R) BE (T)**
$\lvert\mathcal{C}_s\rvert$	4	4	2	4	3	3
$\lvert\bar{\mathcal{C}}_r\rvert$	0	0	0	0	0	0
$\lvert\bar{\mathcal{C}}_t\rvert$	4	4	2	3	3	3

Table S5. Performance comparison between the various baselines on five scRNA-seq datasets in the intra-data setting. Quake1 and Quake2 refer to Quake 10x and Quake Smart-seq2, respectively.

	Cao			**Quake1**			**Quake2**			**Wagner**			**Zeisel**		
	known	novel	overall	known	novel	overall	known	novel	overall	known	novel	overall	known	novel	overall
scCNC [9]	55.8	42.8	28.7	84.4	65.3	68.5	58.0	35.3	38.5	83.1	**57.4**	59.8	64.9	79.1	70.1
scDECL [10]	51.4	41.6	26.3	31.7	45.9	26.1	22.0	30.3	25.9	32.5	48.8	35.0	55.8	69.4	49.5
MARS [12]	92.7	58.2	63.1	96.4	49.8	67.5	88.1	78.9	78.3	78.0	53.0	54.8	89.8	87.0	83.9
ItClust [13]	3.4	45.0	48.0	54.3	43.3	53.2	10.9	62.0	65.9	29.4	30.8	36.0	32.1	69.3	63.7
scNym [14]	**98.5**	63.1	61.2	**98.5**	48.1	53.7	95.3	69.6	65.9	93.9	44.6	43.5	**99.3**	62.2	62.4
scArches [15]	78.0	45.4	57.7	90.3	57.3	70.1	64.0	56.1	58.1	65.2	41.0	46.8	73.1	63.2	63.6
scBOL	96.5	**74.6**	**77.4**	98.0	**65.8**	**77.5**	**96.2**	**82.1**	**82.5**	**94.9**	54.7	**62.5**	96.5	**91.7**	**89.1**

Table S6. Performance comparison between the various baselines on five groups of scRNA-seq datasets in the inter-data setting. "R" represents the reference data and "T" refers to the target data.

	Muraro (R)			**Vento Smart-seq2 (R)**			**Plasschaert (R)**			**Mammary Smart-seq2 (R)**			**Haber largecell (R)**		
	Baron_human (T)			**Vento 10x (T)**			**Montoro 10x (T)**			**Mammary 10x (T)**			**Haber region (T)**		
	known	novel	overall	known	novel	overall	known	novel	overall	known	novel	overall	known	novel	overall
scCNC [9]	75.0	40.8	61.1	83.4	47.1	43.7	79.7	73.1	73.0	92.4	65.5	76.2	62.7	69.4	55.9
scDECL [10]	36.7	38.4	29.9	42.8	46.4	22.9	59.9	72.9	54.9	93.7	92.0	83.9	83.6	63.8	39.3
MARS [12]	79.5	**82.3**	80.0	94.5	78.6	83.8	88.6	**94.5**	89.1	81.5	97.5	86.9	57.1	75.1	68.2
ItClust [13]	80.9	56.4	69.2	64.3	75.0	58.2	90.1	75.1	83.2	36.8	70.5	67.2	53.4	58.2	56.4
scNym [14]	88.2	55.5	63.9	**98.1**	70.4	80.6	**96.1**	77.7	83.1	95.1	48.6	49.8	**95.8**	44.4	51.2
scArches [15]	89.3	52.8	80.9	83.4	66.8	75.2	91.4	67.4	85.3	62.0	55.5	59.0	72.3	51.7	59.6
scBOL	**96.1**	81.9	**91.8**	95.7	**97.0**	**93.3**	95.6	88.7	**93.1**	**95.8**	**98.8**	**97.6**	90.9	**77.4**	**79.7**

Table S7. Performance comparison between the various baselines on two spatial transcriptomics datasets, i.e., Hubmap and TonsilBE from CODEX multiplexed imaging technology.

	Hubmap (CODEX)			**TonsilBE (CODEX)**		
	known	novel	overall	known	novel	overall
MARS [12]	**97.5**	60.3	58.4	95.0	45.4	67.4
scNym [14]	66.9	64.8	63.1	7.1	66.4	65.9
scArches [15]	9.4	49.6	47.0	19.0	34.6	40.6
STAGATE [11]	10.9	69.8	60.4	4.9	59.1	50.8
STELLAR [16]	92.5	72.9	68.8	**95.2**	60.4	76.1
scBOL	93.1	**96.3**	**95.8**	92.5	**96.9**	**94.3**

Table S8. Performance comparison for different novel cell type numbers with labeled ratio 0.5 across various methods on two scRNA-seq datasets.

	Quake 10x			Quake 10x			Quake 10x			Quake 10x			Quake 10x		
	Known=32, novel=4			Known=25, novel=11			Known=18, novel=18			Known=11, novel=25			Known=4, novel=32		
Methods	Known	Novel	Overall	Known	Novel	Overall	Known	Novel	Overall	Known	Novel	Overall	Known	Novel	Overall
MARS [12]	81.8	55.1	77.4	90.0	47.4	69.0	92.1	52.8	68.9	97.7	57.8	67.8	98.4	65.2	69.1
scNym [14]	95.8	75.2	81.8	93.5	51.0	61.1	98.4	52.8	60.8	**99.0**	43.4	46.7	**99.0**	30.6	31.9
scArches [15]	85.9	61.8	81.9	88.0	58.2	72.5	88.3	56.6	69.1	78.2	52.3	59.8	87.0	42.7	49.3
ItClust [13]	23.3	26.1	25.4	53.3	43.3	43.7	70.5	47.3	52.3	73.3	34.0	55.3	42.7	51.6	74.1
scCNC [9]	76.6	**80.7**	65.9	79.8	53.1	62.2	85.0	49.8	61.3	81.2	53.0	58.1	73.4	46.9	56.0
scDECL [10]	17.0	78.7	14.6	22.3	47.2	22.4	24.8	38.9	21.3	32.3	27.8	19.2	61.5	19.6	16.8
scBOL	**98.4**	60.8	**91.8**	**98.4**	**69.5**	**83.2**	**98.7**	**69.6**	**78.5**	88.2	**65.3**	**71.9**	93.9	**67.7**	**71.4**
	Quake Smart-seq2			Quake Smart-seq2			Quake Smart-seq2			Quake Smart-seq2			Quake Smart-seq2		
	Known=40, novel=5			Known=31, novel=14			Known=22, novel=23			Known=13, novel=32			Known=4, novel=41		
Methods	Known	Novel	Overall	Known	Novel	Overall	Known	Novel	Overall	Known	Novel	Overall	Known	Novel	Overall
MARS [12]	78.0	63.8	74.8	79.6	80.2	74.8	80.3	**70.6**	69.2	89.3	**70.3**	69.9	91.0	72.0	72.6
scNym [14]	**95.3**	63.2	82.3	**95.4**	69.4	64.9	**96.9**	59.2	56.4	96.3	57.7	54.2	**99.3**	29.6	29.4
scArches [15]	73.5	65.2	68.9	74.3	60.6	64.7	72.3	54.7	57.2	55.2	50.7	49.9	41.2	38.2	37.6
ItClust [13]	9.3	49.3	67.4	12.4	68.5	65.9	32.7	55.5	49.4	30.7	44.1	40.2	7.4	23.2	22.1
scCNC [9]	39.8	48.8	34.7	56.8	31.7	34.1	65.0	40.8	39.0	59.1	21.7	23.8	86.6	26.6	25.9
scDECL [10]	11.2	30.7	9.8	12.1	31.4	15.9	15.6	21.8	14.3	45.1	23.8	20.0	85.8	39.4	37.5
scBOL	91.1	**71.2**	**84.4**	92.3	**81.3**	**79.5**	94.8	**70.6**	**73.0**	**98.6**	68.9	**70.0**	98.7	**73.2**	**73.9**

Table S9. Performance comparison for different labeled ratios across various methods on tested two scRNA-seq datasets.

	Quake 10x			Quake 10x			Quake 10x			Quake 10x			Quake 10x		
	ratio=0.1			ratio=0.3			ratio=0.5			ratio=0.7			ratio=0.9		
Methods	Known	Novel	Overall	Known	Novel	Overall	Known	Novel	Overall	Known	Novel	Overall	Known	Novel	Overall
MARS [12]	88.8	63.3	77.1	91.0	58.8	74.8	92.1	52.8	68.9	91.4	54.3	65.7	67.3	50.7	52.5
scNym [14]	94.0	52.0	68.2	98.4	54.7	64.0	98.4	52.8	60.8	93.8	53.7	59.1	88.9	62.3	60.2
scArches [15]	86.7	54.4	73.2	90.4	59.1	74.9	88.3	56.6	69.1	86.1	57.8	66.1	73.5	52.4	54.3
ItClust [13]	71.2	49.1	58.9	75.5	51.1	62.1	70.5	47.3	52.3	42.3	41.4	45.0	26.3	39.1	35.7
scCNC [9]	85.5	48.1	65.5	85.3	41.4	60.8	85.0	49.8	61.3	86.5	50.2	56.3	86.4	54.4	55.2
scDECL [10]	40.5	56.0	34.7	24.9	38.9	18.0	24.8	38.9	21.3	24.9	38.9	26.0	24.7	38.9	33.5
scBOL	**98.0**	**69.4**	**82.8**	**98.6**	**71.7**	**81.6**	**98.7**	**69.6**	**78.5**	**98.3**	**65.4**	**72.3**	**98.0**	**70.3**	**70.2**
	Quake Smart-seq2			Quake Smart-seq2			Quake Smart-seq2			Quake Smart-seq2			Quake Smart-seq2		
	ratio=0.1			ratio=0.3			ratio=0.5			ratio=0.7			ratio=0.9		
Methods	Known	Novel	Overall	Known	Novel	Overall	Known	Novel	Overall	Known	Novel	Overall	Known	Novel	Overall
MARS [12]	81.0	**72.4**	74.7	79.1	**76.3**	74.9	80.3	**70.6**	69.2	78.5	72.2	71.3	74.0	67.7	67.2
scNym [14]	85.4	58.8	59.0	86.8	65.5	59.2	**96.9**	59.2	56.4	**97.2**	49.1	49.8	**90.4**	53.5	52.4
scArches [15]	71.4	53.6	60.5	71.2	49.8	57.3	72.3	54.7	57.2	64.7	54.5	53.5	39.9	54.9	52.8
ItClust [13]	48.8	67.2	57.8	43.4	55.1	50.7	32.7	55.5	49.4	5.8	55.2	49.0	0.0	47.3	44.6
scCNC [9]	65.3	43.1	42.2	63.7	32.1	35.3	65.0	40.8	39.0	89.6	70.9	72.1	81.7	71.7	70.9
scDECL [10]	41.8	55.3	37.9	15.5	21.8	12.5	15.6	21.8	14.3	15.5	21.8	16.6	15.8	21.8	19.8
scBOL	**94.5**	70.0	**75.3**	**95.3**	73.1	**76.3**	94.7	**70.6**	**73.0**	96.3	**73.1**	**74.0**	88.0	**75.5**	**70.5**

Table S10. Performance comparison for different labeled ratios across various methods on tested two spatial transcriptomics datasets.

	Hubmap ratio=0.3			Hubmap ratio=0.4			Hubmap ratio=0.5			Hubmap ratio=0.6			Hubmap ratio=0.7		
Methods	**Known**	**Novel**	**Overall**	**Known**	**Novel**	**Overall**	**Known**	**Novel**	**Overall**	**Known**	**Novel**	**Overall**	**Known**	**Novel**	**Overall**
MARS [12]	80.8	73.6	75.1	79.4	71.7	73.1	**97.5**	60.3	58.4	**96.8**	58.1	59.4	95.2	60.1	59.4
scNym [14]	66.7	62.2	60.8	66.8	63.5	62.0	66.9	64.8	63.1	67.3	65.9	64.3	68.9	62.5	60.1
scArches [15]	1.0	48.7	44.3	1.0	52.4	49.6	9.4	49.6	47.0	0.0	70.7	64.7	1.0	54.4	51.7
STAGATE [11]	23.4	62.6	54.3	10.8	70.0	59.0	10.9	69.8	60.4	11.2	69.7	62.0	11.4	69.9	64.0
STELLAR [16]	**95.0**	77.6	81.3	**95.5**	79.9	73.9	92.5	72.9	68.8	96.3	82.6	78.0	**98.1**	67.0	65.3
scBOL	**95.0**	**85.5**	**87.3**	90.8	**96.8**	**95.7**	93.1	**96.3**	**95.8**	95.9	**89.1**	**91.6**	92.4	**97.0**	**96.6**
	TonsilBE ratio=0.3			**TonsilBE ratio=0.4**			**TonsilBE ratio=0.5**			**TonsilBE ratio=0.6**			**TonsilBE ratio=0.7**		
Methods	**Known**	**Novel**	**Overall**	**Known**	**Novel**	**Overall**	**Known**	**Novel**	**Overall**	**Known**	**Novel**	**Overall**	**Known**	**Novel**	**Overall**
MARS [12]	82.5	58.0	67.3	90.4	43.2	66.0	95.0	45.4	67.4	82.1	52.0	63.7	75.9	56.6	65.3
scNym [14]	6.8	66.7	66.3	6.8	66.5	66.1	7.1	66.4	65.9	7.1	66.5	66.0	17.3	51.8	53.7
scArches [15]	14.0	62.6	56.9	16.7	57.2	48.1	19.0	34.6	40.6	18.1	40.6	40.3	15.5	63.2	49.3
STAGATE [11]	5.2	37.2	40.7	5.4	37.1	39.6	4.9	59.1	50.8	5.2	34.4	37.4	5.3	56.3	49.0
STELLAR [16]	**96.5**	56.0	75.6	**95.0**	76.1	83.3	**95.2**	60.4	76.1	94.1	63.6	76.5	94.9	64.8	77.3
scBOL	92.4	**92.7**	**92.5**	94.7	**96.5**	**95.6**	92.5	**96.9**	**94.3**	**95.2**	**94.3**	**94.7**	**95.9**	**93.7**	**94.8**

Table S11. Performance comparison for different novel ratios across various methods on tested two scRNA-seq datasets.

	Quake 10x ratio=0.2			Quake 10x ratio=0.4			Quake 10x ratio=0.6			Quake 10x ratio=0.8			Quake 10x ratio=1.0		
Methods	**Known**	**Novel**	**Overall**	**Known**	**Novel**	**Overall**	**Known**	**Novel**	**Overall**	**Known**	**Novel**	**Overall**	**Known**	**Novel**	**Overall**
MARS [12]	86.3	61.5	80.1	88.9	61.2	78.1	90.8	58.3	75.8	93.3	59.4	75.5	92.1	52.8	68.9
scNym [14]	98.5	57.2	82.5	**98.7**	51.6	71.5	**98.7**	48.7	66.4	**98.6**	52.9	65.7	**98.4**	52.8	60.8
scArches [15]	93.4	46.6	83.5	92.1	49.7	77.1	90.7	50.5	72.5	89.4	50.8	69.0	88.3	56.6	69.1
ItClust [13]	84.0	46.3	81.9	76.2	45.0	70.7	75.5	54.5	67.6	64.3	48.5	60.2	70.5	47.3	52.3
scCNC [9]	83.6	50.9	72.4	87.9	48.8	69.9	83.6	42.4	60.5	80.5	47.6	60.0	85.0	49.8	61.3
scDECL [10]	24.8	39.7	19.9	24.8	38.7	16.7	24.8	38.6	16.3	24.8	39.2	18.4	24.8	38.9	21.3
scBOL	**98.7**	**65.3**	**88.0**	97.7	**63.6**	**81.9**	97.1	**64.5**	**79.8**	97.4	**68.6**	**79.6**	96.9	**69.6**	**78.5**
	Quake Smart-seq2 ratio=0.2			**Quake Smart-seq2 ratio=0.4**			**Quake Smart-seq2 ratio=0.6**			**Quake Smart-seq2 ratio=0.8**			**Quake Smart-seq2 ratio=1.0**		
Methods	**Known**	**Novel**	**Overall**	**Known**	**Novel**	**Overall**	**Known**	**Novel**	**Overall**	**Known**	**Novel**	**Overall**	**Known**	**Novel**	**Overall**
MARS [12]	78.1	75.0	75.9	79.7	**73.7**	74.0	82.0	72.6	72.8	79.1	72.9	71.6	80.3	**70.6**	69.2
scNym [14]	**96.7**	57.9	69.5	**96.8**	52.7	60.2	**96.6**	62.1	59.5	**96.8**	50.4	53.3	**96.9**	59.2	56.4
scArches [15]	83.1	53.3	70.6	77.8	50.7	62.0	77.7	57.8	63.4	74.0	52.8	57.2	72.3	54.7	57.2
ItClust [13]	24.6	53.8	65.6	60.3	60.1	59.9	50.5	50.3	50.0	37.8	67.7	59.2	32.7	55.5	49.4
scCNC [9]	66.4	42.7	52.8	65.9	42.0	48.6	65.0	40.9	42.3	65.2	40.5	41.1	64.3	40.8	39.0
scDECL [10]	15.5	21.8	11.2	15.5	22.0	9.5	15.5	21.8	11.6	15.5	21.9	13.2	15.6	21.8	14.3
scBOL	95.2	**69.3**	**77.1**	94.9	73.1	**75.8**	94.2	**74.7**	**75.0**	93.6	**73.6**	**74.4**	94.7	**70.6**	**73.0**

Table S12. Performance comparison between the various baselines on another two spatial transcriptomics datasets, i.e., Lohoff from seqFISH technology and Chen from Stereo-seq technology.

	Lohoff (seqFISH)			Chen (Stereo-seq)		
	known	novel	overall	known	novel	overall
MARS [12]	79.0	63.3	53.0	52.5	48.8	36.5
scNym [14]	58.1	50.4	51.3	67.2	38.8	38.6
scArches [15]	48.5	71.1	62.0	46.1	45.4	44.8
STAGATE [11]	42.1	47.5	45.7	12.8	58.2	43.4
STELLAR [16]	**98.4**	84.7	79.5	35.2	61.1	50.9
scBOL	96.4	**93.7**	**94.8**	**83.9**	**69.1**	**71.2**

Table S13. Time-consuming (second) and performance (overall accuracy) comparison for different data sizes across various methods.

	size = 5k		size = 10k		size = 25k		size = 50k		size = 100k	
Methods	**Time**	**Overall**	**Time**	**Overall**	**Time**	**Overall**	**Time**	**Overall**	**Time**	**Overall**
scCNC [9]	9104	50.5	32170	55.4	163152	54.2	509184	52.7	1578470	54.8
scDECL [10]	12037	13.7	22038	14.2	53095	13.6	211591	13.5	222072	13.4
ItClust [13]	420	38.2	1480	34.9	1876	36.0	3592	34.9	7972	36.9
MARS [12]	1248	58.4	2304	84.0	4182	95.5	7893	97.8	12438	98.5
scNym [14]	**342**	38.5	**534**	39.3	**1032**	39.7	**1614**	40.8	**3024**	41.1
scArches [15]	2798	37.5	5903	36.7	12620	36.3	28408	35.9	56764	39.3
scBOL	1785	**63.5**	3367	**87.9**	8564	**98.3**	19913	**99.6**	37558	**99.9**

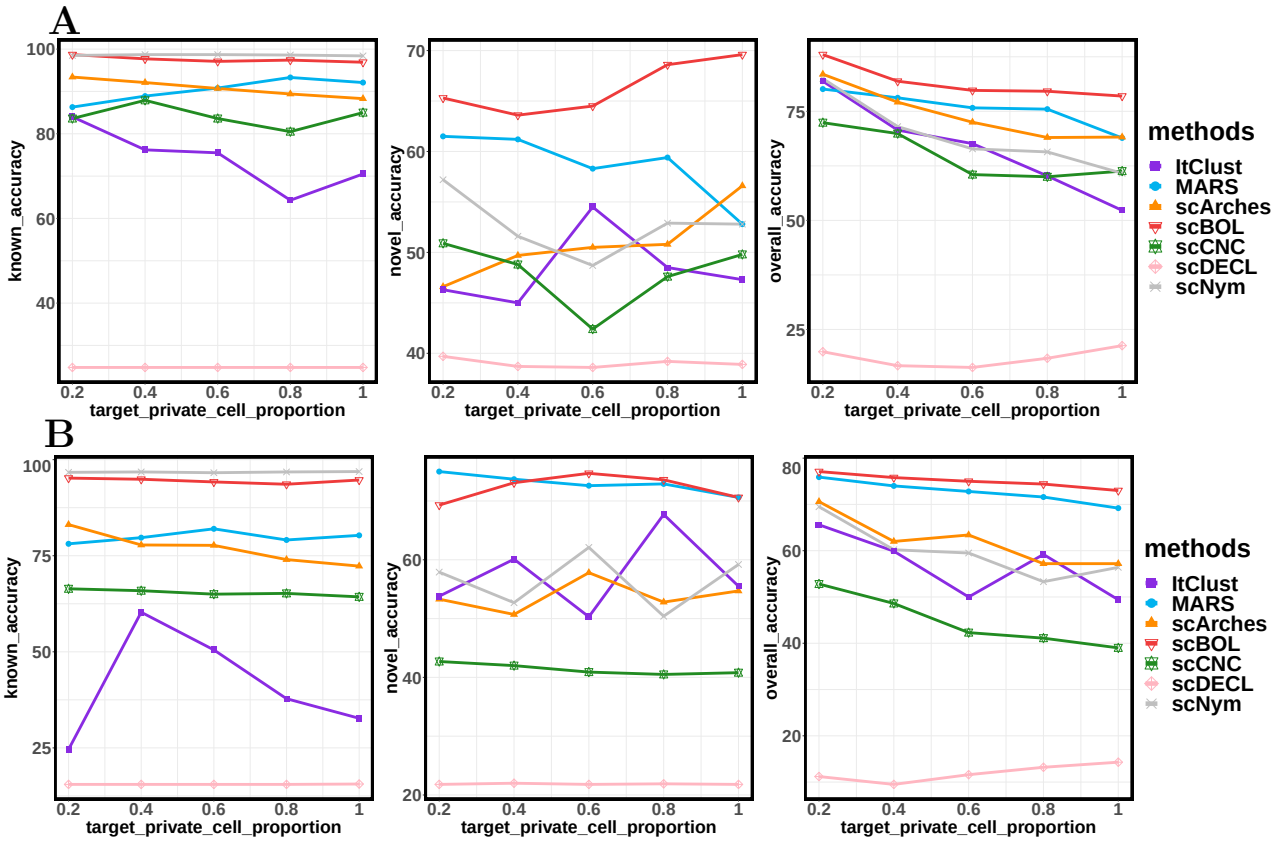


Fig. S1. Robustness analysis. A. The trend of known accuracy, novel accuracy, and overall accuracy with varying proportions of target private cells sampled in Quake 10x datasets respectively. **B.** The trend of known accuracy, novel accuracy, and overall accuracy with varying proportions of target private cells sampled in Quake Smart-seq2 datasets respectively.

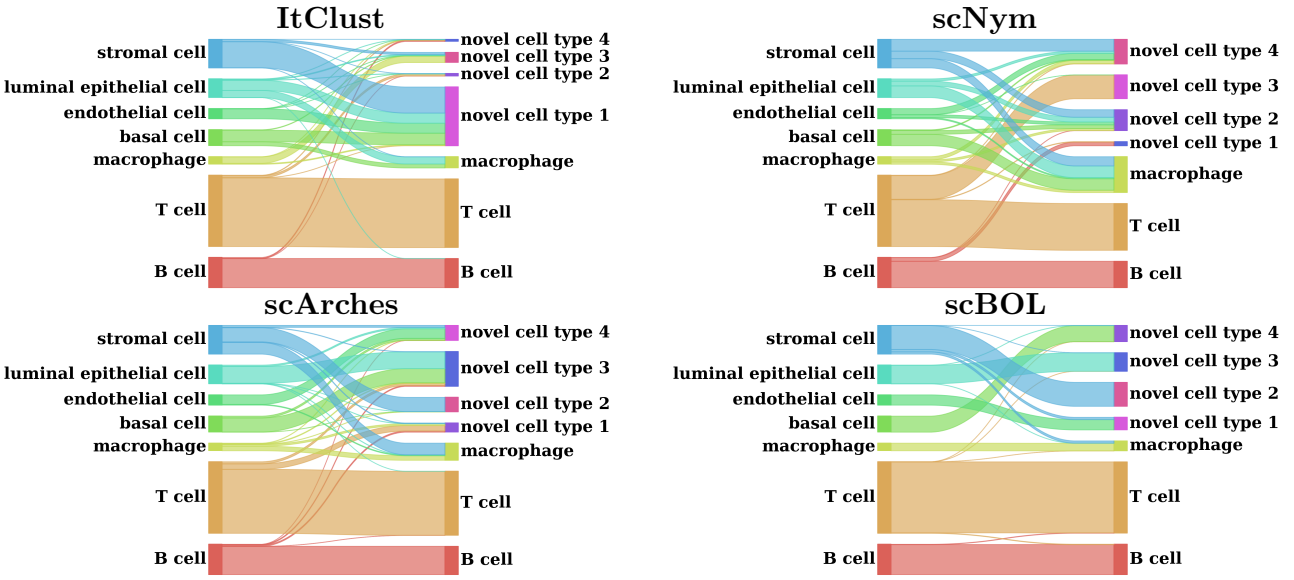


Fig. S2. Mapping relationship among prediction results of scBOL and other three methods via Sankey plots for the experiment where Quake 10x Limb Muscle is the reference data and Quake 10x Mammary Gland is the target data.

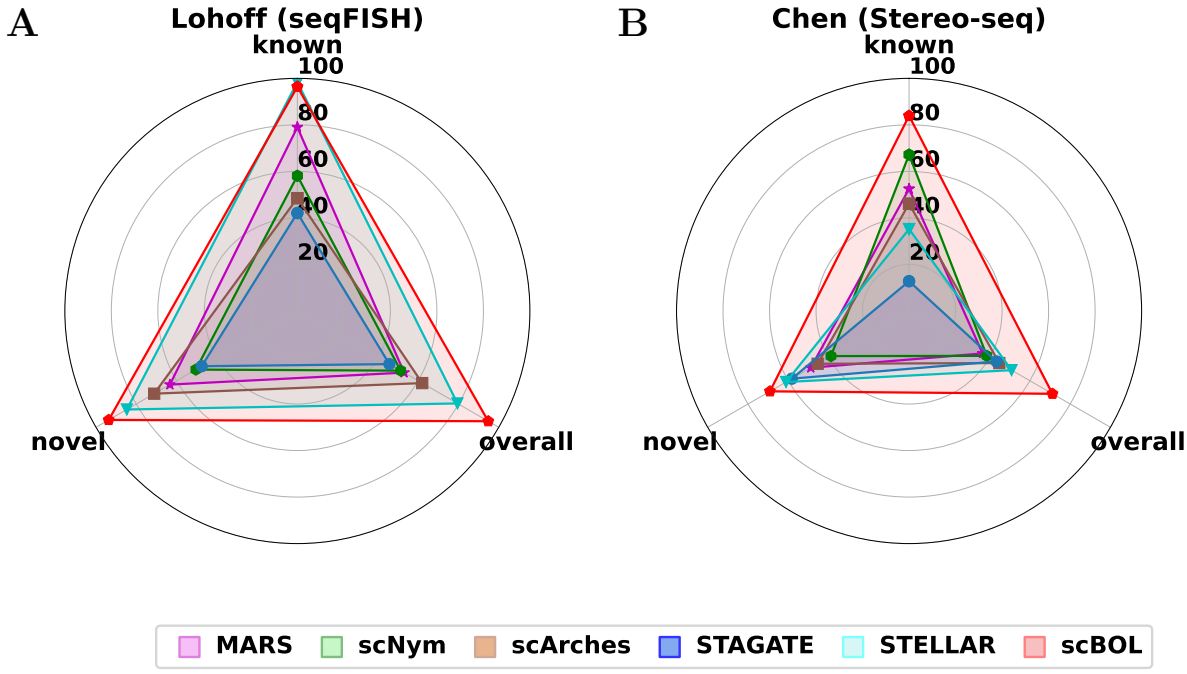


Fig. S3. Radar plot of scBOL and other five competing methods measured by three kinds of accuracy. A. Radar plots on the Lohoff dataset obtained by seqFISH. **B.** Radar plots on the Chen dataset obtained by Stereo-seq.

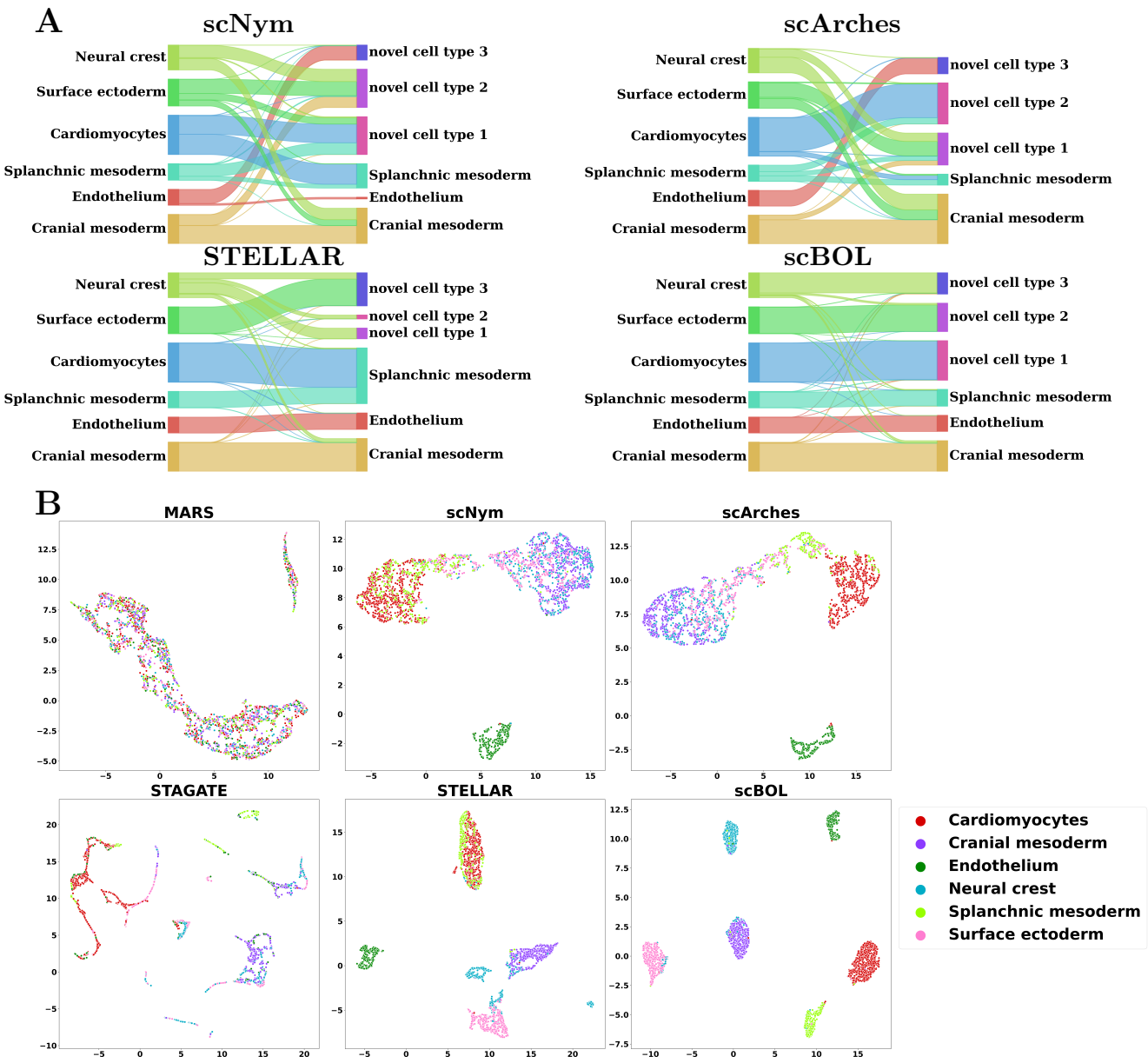


Fig. S4. Intra-data experiments on the Lohoff dataset. A. Mapping relationship among prediction results of scBOL and other three methods via Sankey plots. **B.** Visualization plots via UMAP calculated using the latent representations of scBOL and other five methods colored by cell types.

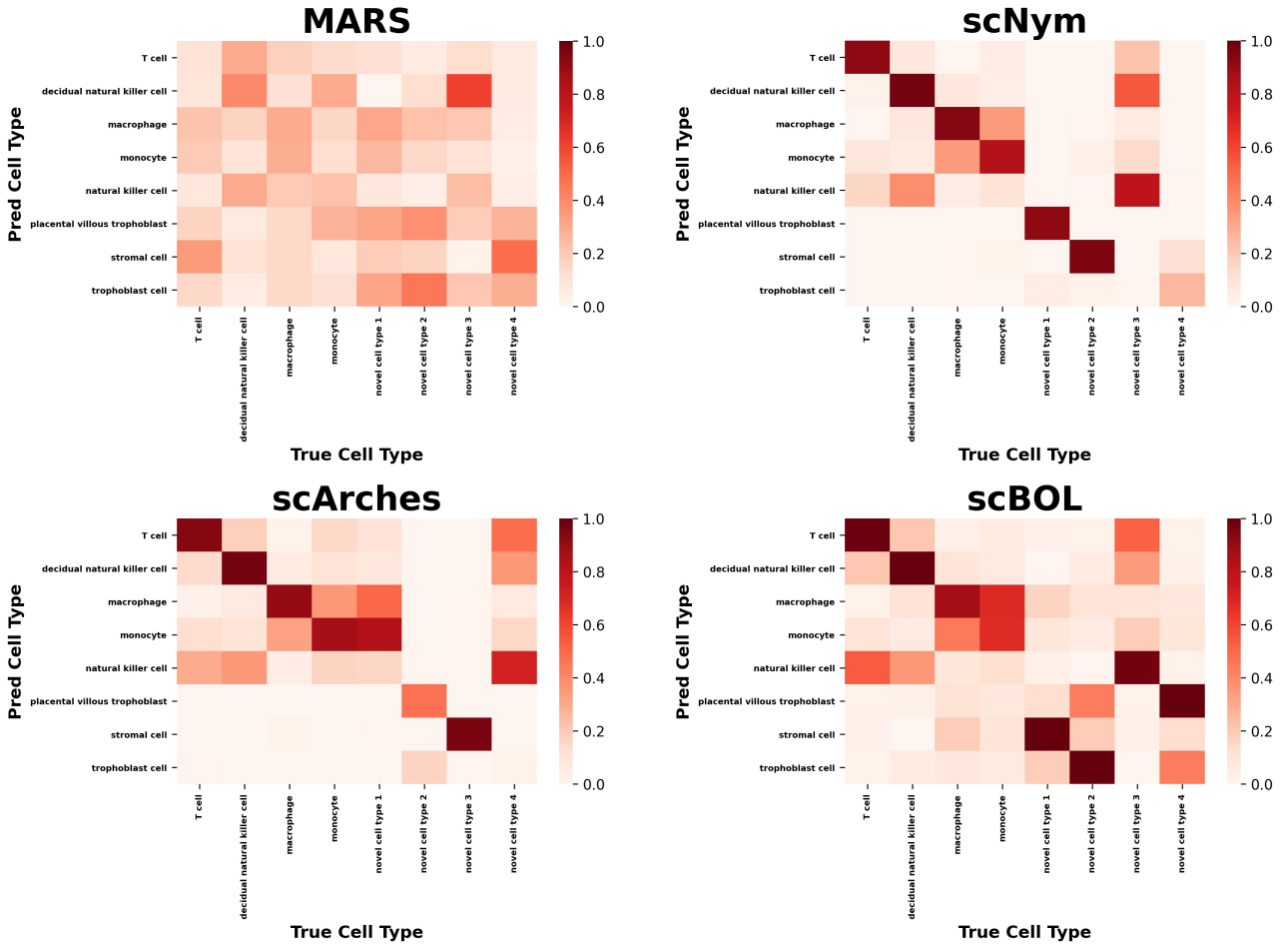


Fig. S5. Differential gene expression analysis: the similarity heatmap based on marker gene sets identified by ground-truth cell types and prediction cell types of each method.

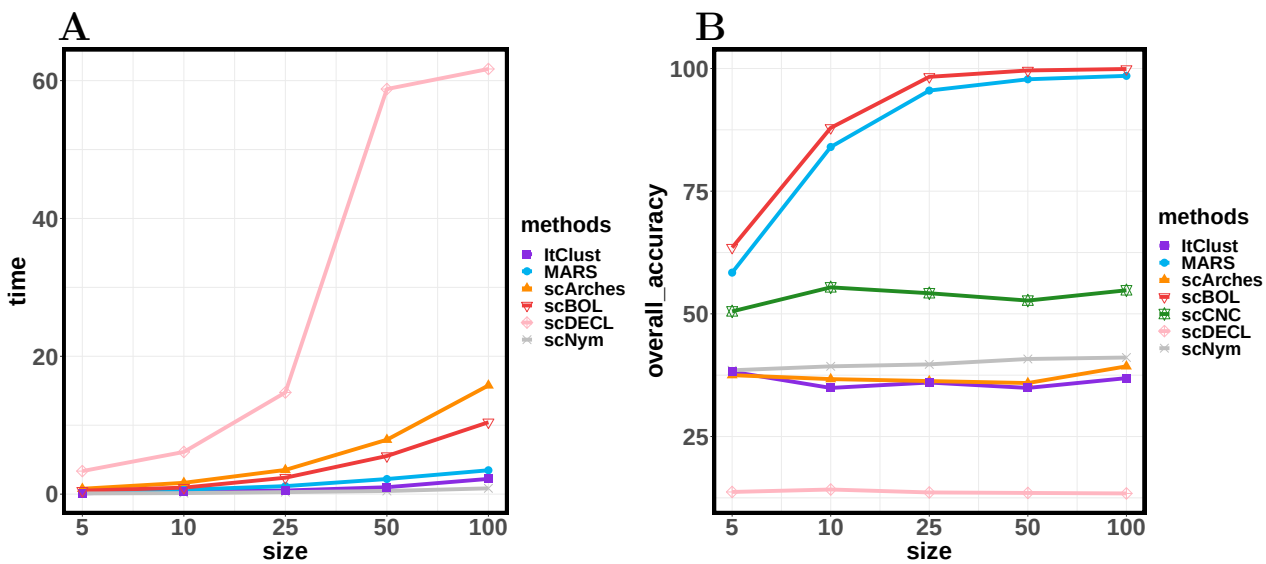
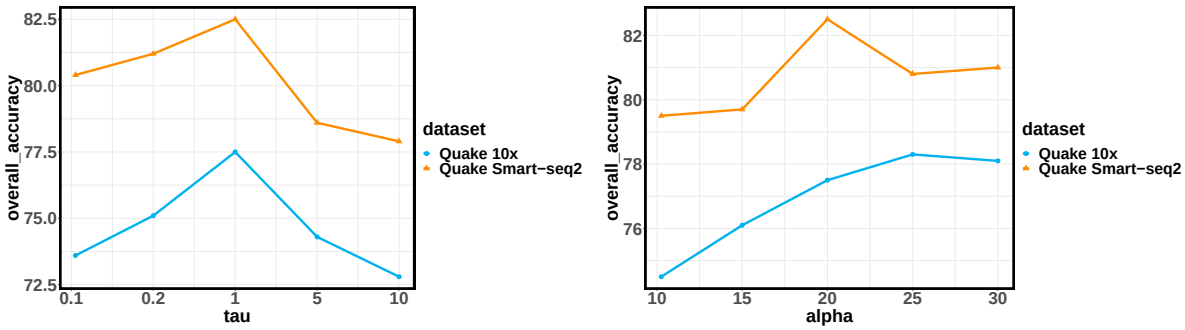


Fig. S6. A. The trend of time-consuming (hour) with respect to the change in data sizes (K). Since the running time of scCNC is much longer than the other methods, we do not draw its line to make the line graph clearer. **B.** The trend of overall accuracy with respect to the change in data sizes.



(a) τ value (b) α value

Fig. S7. Hyperparameter sensitivity of scBOL to the changes of τ and α on Quake 10x, Quake Smart-seq2 datasets.