SwarmMAP: Swarm Learning for Decentralized Cell Type Annotation in Single Cell Sequencing Data

Oliver Lester Saldanha[1†], Vivien Goepp[4†], Kevin Pfeiffer[1], Hyojin Kim[2], Jie Fu Zhu[1], Rafael Kramann[4], Sikander Hayat[4*], Jakob Nikolas Kather[1,2,3*]

[1] Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Fetscherstraße 74, Dresden, 01307, Saxony, Germany .

[2] Department of Medicine I, Faculty of Medicine and University Hospital Carl Gustav Carus, Technical University Dresden Fetscherstraße 74, Dresden, 01307, Saxony, Germany .

[3] Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Im Neuenheimer Feld 460, Heidelberg, 69120, Baden-Wuerttemberg, Germany .

[4] Department of Medicine 2, RWTH Aachen University, Medical Faculty, Pauwelsstrasse 30, Aachen, 52074, North Rhine-Westphalia, Germany .

*Corresponding author(s). E-mail(s): shayat@ukaachen.de; jakob_nikolas.kather@tu-dresden.de;
Contributing authors: oliverlestersaldanha25@gmail.com; vgoepp@ukaachen.de; kevin.pfeiffer@tu-dresden.de; genehaus@gmail.com; jeffzhu6969@gmail.com; rkramann@ukaachen.de;
†These authors contributed equally to this work.

Abstract

Rapid technological advancements have made it possible to generate single-cell data at a large scale. Several laboratories around the world can now generate single-cell transcriptomic data from different tissues. Unsupervised clustering, followed by annotation of the cell type of the identified clusters, is a crucial step in single-cell analyses. However, there is no consensus on the marker genes to use for annotation, and cell-type annotation is currently mostly done by manual inspection of marker genes, which is irreproducible, and poorly scalable. Additionally, patient-privacy is also a critical issue with human datasets. There is a critical need to standardize and automate cell-type annotation across datasets in a privacy-preserving manner. Here, we developed SwarmMAP that uses Swarm Learning to train machine learning models for cell-type classification based on single-cell sequencing data in a decentralized way. SwarmMAP does not require any exchange of raw data between data centers. SwarmMAP has a F1-score of 0.93, 0.98, and 0.88 for cell type classification in human heart, lung, and breast datasets, respectively. Swarm Learning-based models yield an average performance of **0.907** which is on par with the performance achieved by models trained on centralized data (*p*-val=**0.937**, Mann-Whitney *U* Test). We also find that increasing the number of datasets increases cell-type prediction accuracy and enables handling higher cell-type diversity. Together, these findings demonstrate that Swarm Learning is a viable approach to automate cell-type annotation. SwarmMAP is available at https://github.com/hayatlab/SwarmMAP.

Keywords: Swarm Learning, Single-cell RNA Transcriptomics, Cell Type Annotation, Classification, Decentralized Learning

1 Introduction

Recent technological advances in single-cell sequencing have led to a plethora of scientific discoveries improving our understanding of human tissue and diseases [1], including COVID [2, 3], lung [4], cardiovascular [5–7], renal diseases [8], and cancer [9–11] at single-cell resolution. Typical single-cell analysis pipelines use unsupervised clustering, followed by cell-type annotation of identified clusters based on the expression level and specificity of selected marker genes [12]. Cell-type annotation is still primarily a manual effort in which subject experts review marker genes per cluster to annotate cell types. There is no consensus on marker genes and their importance for cell type annotation yet. This reduces reproducibility as the selection of marker genes and their importance for annotation is dependent on the expert annotating the data and can vary from person to person. Furthermore, with increasing amounts of data emerging from different labs, this approach is not scalable or transferable. Additionally, secure data sharing and maintaining data privacy is a critical issue while working with human patient data [13].

To leverage the full potential of multiple studies, tools to generate standardized cell-type annotation while maintaining data privacy are needed to unify and compare data across studies. Furthermore, to increase reproducibility, scalability, and limit individual bias, manual annotation of cell clusters should be increasingly replaced by machine learning models that automatically assign individual cells to a cell type [14–17]. Some tools have also been developed to map new unannotated data to a reference data set [18–20]. However, it is challenging to train a universal machine learning model to classify cell types based on individual single-cell sequencing datasets due to the underlying technical batch effects. Moreover, generalizable machine learning models need to be trained on large, multi-centric, and diverse datasets to account for this variability. The usual procedure to create such datasets is centralized data collection. This requires multiple participating institutions to send their data to a single location. Such data transfer can create practical, legal, and even ethical problems and is often a rate-limiting step to train machine learning models in biology and medicine [21].

Swarm learning is a computational technique to co-train machine learning models at multiple institutions in a decentralized way, without exchanging underlying data[22]. Swarm learning does not require a central coordinator of the network and thus avoids monopolization of resources and machine learning models [23]. In medical image and computational pathology analysis, Swarm learning has been shown to enable a high performance of machine learning models, which is on par with models trained in a centralized way [23–25]. Ultimately, Swarm learning could enable training of machine learning models in a massively parallelized way, increasing the resilience of the training process and democratizing access to the resulting models.

Here, we show that Swarm learning can be efficiently applied to train machine learning models for cell type classification based on single cell sequencing data. We evaluate this on human data from multiple organs generated by different research centers. Our tool, SwarmMAP, shows high accuracy for cell-type classification in a privacy-preserved setting where patient data is not shared among users. SwarmMAP enables comparative analyses across datasets, enabling novel discoveries in single-cell datasets while maintaining patient privacy.

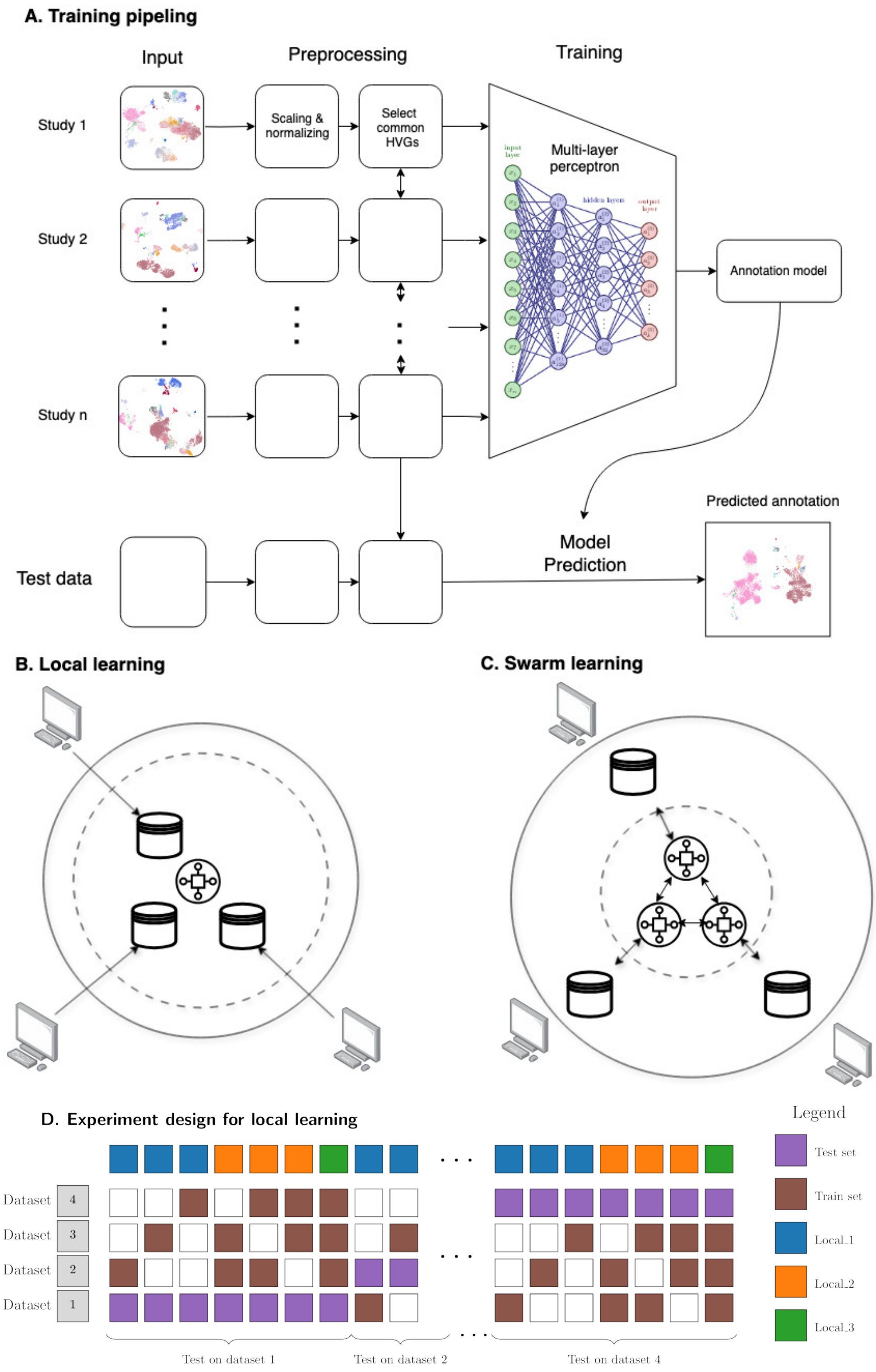


Fig. 1: Overview of the SwarmMAP workflow. (A) Pipeline for training our annotation classifier: scaling and normalization is done independently for each study, except for selecting the highly variable genes (HVGs), where the union of the top 2000 HVGs across studies is used; (B) Local learning (LL) framework, where all the data is shared to learn a single model (dashed inner circle); (C) Swarm learning (SL) framework, where no data is shared and decentralized sharing of models' parameters is enable by a blockchain (dashed inner circle); (D) Experimental design for local learning (LL): for each organ, all combinations of train and test sets are used. The settings where 1, 2, or 3 training sets are used are termed Local_1, Local_2, and Local_3. For Swarm leaning (SL), all 3 training sets are used in a data privacy-preserving manner.

2 Materials and Methods

2.1 Overview of the workflow

SwarmMAP is a Swarm learning based method to classify cell-types in single-cell transcriptomic data. It is trained in a supervised learning manner for each organ. We access the utility of SwarmMAP in both local learning (LL) and Swarm learning (SL) settings using the same data pre-processing pipeline, and classifier is used to compare performance (Figure 1A). In LL, a single model is trained using a common training dataset (Figure 1B) while in SL, each agent keeps its data and model private (Figure 1C), as a blockchain-enabled Swarm learning framework allows each agent's model to learn from the other models. To assess model performance, the predicted annotation is compared with the annotation provided by the authors in the test dataset in a study-specific manner.

2.2 Dataset description

SwarmMAP is trained and tested on single-cell transcriptomic data from human heart, lung, and breast (Supplementary Table 1) totaling 284 donors and 1,956,243 cells. Each collection consists of 4 separate studies. Lung and breast collections are taken from the human lung cell atlas [4] and breast atlas [9], respectively. The heart collection is created from individual datasets, where the cell-types provided by individual studies have been standardized. Four studies were selected from each collection, with varying sample sizes and cell composition (S12). Training and testing is performed at two annotation levels: cell types and cell subtypes, which are used as the ground truth for the classifier. The datasets and their cell-types are represented in Supplementary Figures S1 and S2 using UMAP [26].

2.3 Preparation of labels from cell type annotation

We perform the classification of cell labels at two annotation levels: a coarse annotation level, "cell types", and a finer annotation level, "cell subtypes". Each annotation level is trained independently using its own model.

All datasets included in this study come with their own annotations by the respective authors. For the heart atlas, the level 1 annotation ("Annotation_1") was used as cell types (14 cell types). The level 1 annotation ("Subclustering") contained 65 cell subtypes, which are too specific for cell type annotation. Thus, the subtypes were grouped together to arrive at 17 subtypes of cells. The "Subclustering" labels were manually matched to their closest cell ontology terms [27]. Then, hierarchical clustering was performed across the 4 studies to measure distance between subtypes and the closest subtypes are merged together, the names being set to their respective common ancestors in the cell ontology. For heart, we obtained 17 cell subtypes from 65 initial "Subclustering" labels.

For lung and breast atlases, the same merging process was used, except that cell ontology terms were already available. For the lung atlas, the initial "cell_type" label with 50 categories was gradually merged to obtain 24 cell subtypes, and then further to obtain 17 cell types. For the breast atlas, the initial "cell_type" label with 26 categories was gradually merged to obtain 22 cell subtypes and 14 cell types.

The resulting annotations have the number of classes comparable between the heart, lung, and breast collections: 14, 17, and 14, respectively, for cell types, and 17, 24, and 22, respectively, for cell subtypes. The hierarchical clustering of the final annotations are provided in dendrograms (Supplementary Figure S7) and the correspondence between cell types and cell subtypes is represented in a Sankey diagram (Supplementary Figure S5).

2.4 Data preprocessing for supervised learning

For each data collection, we apply the following preprocessing steps before training.

- **Suspension type**: Suspension types (single-cell or single-nuclei) are filtered to ensure that all observations in every data collection have the same suspension type. The heart collection consists of single nuclei, and the lung and breast collections consist of single cell data.
- **Tissue**: Data are also filtered with respect to the tissue. For the heart, only cells from the left ventricle are selected. For the lung, only cells from the lung parenchyma are selected. The breast collection has no tissue specification.
- **Quality Control**: Standard quality control (QC) of counts is performed and cells deemed outliers are filtered out. Four QC metrics are used: log1p value of total counts, log1p value of number of genes by counts, percentage of counts in the top 20 genes, and percentage of reads mapped to mitochondrial genes. For each metric, each cell whose metric is smaller or greater than a margin of five times the median absolute deviation around the median value is set as an outlier. Furthermore, all cells with more than 8 percent of reads mapped to mitochondrial genes are also set as outliers.
- **Feature selection**: For each study, the 2000 most highly variable genes (HVGs) are computed. The choice of 2000 features constitutes a trade-off between model simplicity and model complexity. The effect of the number of HVGs is reported in Supplementary Figure S9. The set of features is then defined as the union of these genes in all studies (3516, 3566, and 3174 features for heart, lung, and breast, respectively). Consequently, in the SL setting, each model is trained after sharing the union of all other agents' HVGs. We consider that sharing this information constitutes no privacy breach concerning the expression data (see section 4). Moreover, the initial number of HVGs used can have an effect on the model performance and is the result of a trade-off between model complexity and generalizability. Supplementary Figure S9 displays the F1 scores for each cell type for several number of HVGs: 200, 500, 1000, and 2000. For the heart and lung, performance increases continually as the number of HVGs increases, but with diminishing returns. For breast collection, interestingly, the number of HVGs showed no effect on performance. Overall, these results suggest that a higher number of HVGs, is beneficial for classification. However, this comes at the cost of computing time and model complexity, we choose the standard value of 2000 HVGs throughout [28].

The final dataset consists of 1,196,647 (1,117,502), 243,031, 516,565 cells in heart (heart subtype), lung and breast collections, respectively (see S12) from 144, 51, and 89 donors (see Supplementary Table 2), respectively. Finally, raw counts are normalized (10,000 counts per cell) and scaled using the log(1+x) transform.

2.5 Experimental design

The experiment design for local learning (LL) is detailed in Figure 1D. Each column represents an experiment and the experiment design (28 experiments in total) and is applied to each organ separately. Each experiment only uses one dataset as a test set. Then, for each choice of test set, all combinations of training sets are used, that is, training on 1 ("Local_1", 12 experiments), 2 ("Local_2", 12 experiments), and 3 ("Local_3", 4 experiments) datasets. This design compares the classification performance as a function of the number of cells, averaging out the difference in cell count between datasets. For Swarm learning (SL), only the four experiments designs with 3 training sets are used, using all combinations of test set.

2.6 Machine learning classifier model

Briefly, the classifier is trained in both the local and Swarm learning setups separately. In the local learning (LL) setup, classification performance is evaluated when training on 1, 2, or 3 datasets, where the fourth dataset is used for testing. These simulation settings are called Local_1, Local_2, and Local_3, respectively. All possible combinations of training and testing datasets are used. In the Swarm learning (SL) setup, 3 datasets are used for training and one for testing. For each combination of train and test datasets, the validation set is obtained by splitting the training data into train and validation sets.

The model used for classification is a multi-layer perceptron (MLP) classifier. We use two fully connected inner layers with 128 and 32 neurons, respectively. We use a tanh activation function and a cross-entropy loss. Optimization is performed using the Adam optimizer. After hyperparameter fine-tuning using cross-validation on the heart dataset with cell types as labels, the following parameters are used: a learning rate of 1e-3 and no weight decay; a batch size of 128; 100 training epochs. The following alternative configurations were tested:

- using dropout for regularization (dropout rates of 0.25 and 0.5);
- using weight decay with the `AdamW` optimizer (value between 1e-3 and 1e-7);
- using weighted class resampling to counterweight the class imbalance; and
- including inverse class proportions as weights in the loss function;
- using different batch sizes (32, 64, or 256).

These did not produce a significant improvement in classification.

2.7 Swarm Learning

Swarm learning (SL) allows decentralized collaborative training of machine learning (ML) models on multiple physically distinct computing systems (peers) [29]. Here, we implemented SL using three separate peers, representing the institutions. Each peer independently trained a machine learning model on its proprietary dataset, with no raw data shared between peers. During training, model weights and biases were exchanged in multiple synchronization events (sync events). These sync events occurred at the end of each synchronization interval, defined as a fixed number of training batches. At each sync event, the model weights were averaged, and training was resumed at each peer using the updated parameters. To account for differences in sizes of the datasets, we applied a weighted SL approach, where the contributions of each peer were scaled by a weighting factor proportional to the size of its dataset. Motivated by previous studies on pathological and radiology data [23, 30]. This approach ensured balanced contributions from peers with varying dataset sizes, where larger datasets are not overrepresented in the final model. After completing all training epochs, a final round of model merging is performed, providing all peers with a unified model. The Hewlett Packard Enterprise (HPE) SL framework was utilized, which consists of five major components: the ML node, the SL node, the Swarm Network (SN) node, identity management, and HPE license management. The ML node defines the ML model and access to the data. Where the SL node process handles the parameter sharing, while the SN node process manages peer communication. To manage global model state information and enable decentralized parameter merging, an Ethereum blockchain (https://ethereum.org) was employed. Unlike traditional federated learning, SL does not rely on a central server; instead, smart contracts facilitate the selection of peers for parameter merging. All processes were executed in Docker containers. A detailed description of this process and instructions for reproduction can be found under https://github.com/KatherLab/swarm-learning-hpe/tree/dev_single_cell.

2.8 Comparison with state-of-the-art cell type classifiers

The Swarm learning framework can be applied using any machine learning model, the only constraint being that no data are shared between agents. Thus, SwarmMAP can be built using a variety of classifiers. There have been many approaches to cell type classification in single-cell RNA sequencing data. Garnett [31] uses marker genes already curated to annotate cells using a regularized multinomial linear classifier. ACTINN [32] uses an MLP classifier with 3 hidden layers (100, 50 and 25 neurons, respectively). Celltypist [14] uses L2 regularized logistic regression. Supervised Contrastive Learning for Single Cell (SCLSC) [33] employs contrastive learning to learn an embedding representation for cell types and a KNN classifier to annotate cells. devCellPy is a machine learning-enabled pipeline for automated annotation of complex multilayered single-cell transcriptomic data, based on the XGBoost classifier. scTab [34] introduces a feature-attention-based classifier model for single cell transcriptomic data based on TabNet, a deep learning classifier for tabular data [35]. The classification performance of TabNet was compared with several models, including XGBoost [36] and MLP. [34] found that the feature attention-based classifier model outperformed the other models in the context of large-scale and curated datasets (significant differences in the macro F1 score), but with minor differences in F1 values (0.83 for scTab, 0.81 for XGBoost, 0.80 for MLP).

However, scTab is trained on very large datasets with between 10^3 and 10^6 cells per cell type, while SwarmMAP is considering the more challenging setting where some cell types have cell counts of the order of 10^2, and 10^1 for some cell types. Since in general settings, XGBoost is preferred over TabNet [37] and is considered to be more flexible for classification task, we chose to compare the performance of our MLP classifier with XGBoost. Specifically, XGBClassifier classifier was used from the XGBoost Python package, with default parameters. XGBoost was compared to MLP for cell type classification in the LL framework, on the same data (3 organs, see Section 2.4 and with the same experiment design (Local_1, Local_2, and Local_3, see Section 2.5). MLP compares slightly favorably to XGBoost while being faster to train by a factor of 2 to 4 (see Supplementary Figure S3). Thus, SwarmMAP is built using an MLP classifier.

2.9 Data and code availability

The 12 datasets are publicly available from CellxGene[1] and the Broad Institute Single Cell portal[2]. The download links are provided in Supplementary Table 3. The processed datasets will be made available on Zenodo upon publication of the manuscript. The SwarmMAP method and the code to reproduce the results in this study are available at https://github.com/hayatlab/SwarmMAP.

3 Results

3.1 Machine learning-based cell type prediction in multiple datasets

The average weighted F1 score for main cell-type classification in heart datasets are 0.947, 0.957, and 0.958 when training on 1, 2 or 3 (Local_3) datasets, respectively. The corresponding values for cell subtype classification are 0.961, 0.968, and 0.972 respectively (Figure 2). Similar values are obtained from the lung and breast datasets (Figure 2). Here, weighted F1 score (which averages the F1 score for each class weighted by the support of the class) was used as the main classification metric. Results from micro F1 score (based on global true positives, false negatives, and false positives) and the macro F1 score (unweighted average of the F1 score for each class) are also provided in Supplementary Figure S4.

Mann–Whitney *U* tests were performed to compare F1 scores between the different simulation settings. Despite the increase in mean performance as more datasets are used for training, the differences are not statistically significant, owing to the small sample size of the simulation runs (12, 12, and 4 respectively). Moreover, some cell types are hard to classify (see Section 3.1.2), making the distribution of F1 scores more dispersed. This is especially true for breast data, in which classification is harder (see Section 3.2.3). However, there is an increase in the averaged F1 scores, especially for heart subtypes, lung main cell types, and subtypes. The respective average scores of Local_1, Local_2, and Local_3 are 0.961, 0.968, and 0.972 for heart subtypes; 0.978, 0.981, and 0.982 for lung types; and 0.945, 0.954, and 0.958 for lung subtypes (see Supplementary Table 4 for weighted F1 scores).

[1] https://cellxgene.cziscience.com/datasets

[2] https://singlecell.broadinstitute.org/single_cell

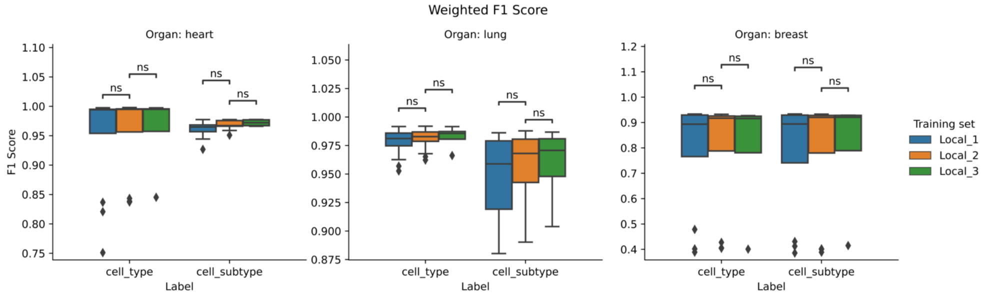


Fig. 2: Classification performance of the local learning framework. Weighted F1 scores are averaged over all simulation runs. F1 score distributions for classifying main cell types and subtypes using Local_1 (training on 1 dataset), Local_2 (training on 2 datasets), and Local_3 (training on 3 datasets) setting in local learning.

3.1.1 Classification performance increases as more datasets are used for training

3.1.2 Classification performances vary greatly between cell-types.

Model performance (F1-score) per cell type varies between different cell types (Figure 3). The corresponding results for cell subtypes are represented in Supplementary Figure S11. In particular, for heart data, “Epicardium” cells and “Ischemic cells (Myocardial infarction)” are difficult to classify, in line with the scarcity of these classes, their uneven distribution among datasets (see cell count barplots in Figure S8), and the difficulty to define ischemic cells biologically (see UMAP representation in Figure S1). In addition, the ischemic cells is a collection of cells from different lineages including cardiomyocytes, epithelial, etc. Thus, they do not have well-defined marker genes and are thus inherently difficult to classify. For the lung data, all cell types are well classified, except “respiratory basal cells”, which are classified as “epithelial cells”. For the breast data, classification is more challenging, with four cell types with F1 score below 0.5: mature alpha-beta T cells, mature B cells, naive thymus-derived CD4-positive, alpha-beta T cells, and capillary endothelial cells. Some other cell types like endothelial tip cells and macrophages have a high disparity in classification performance between simulation runs (see Section 3.2.3).

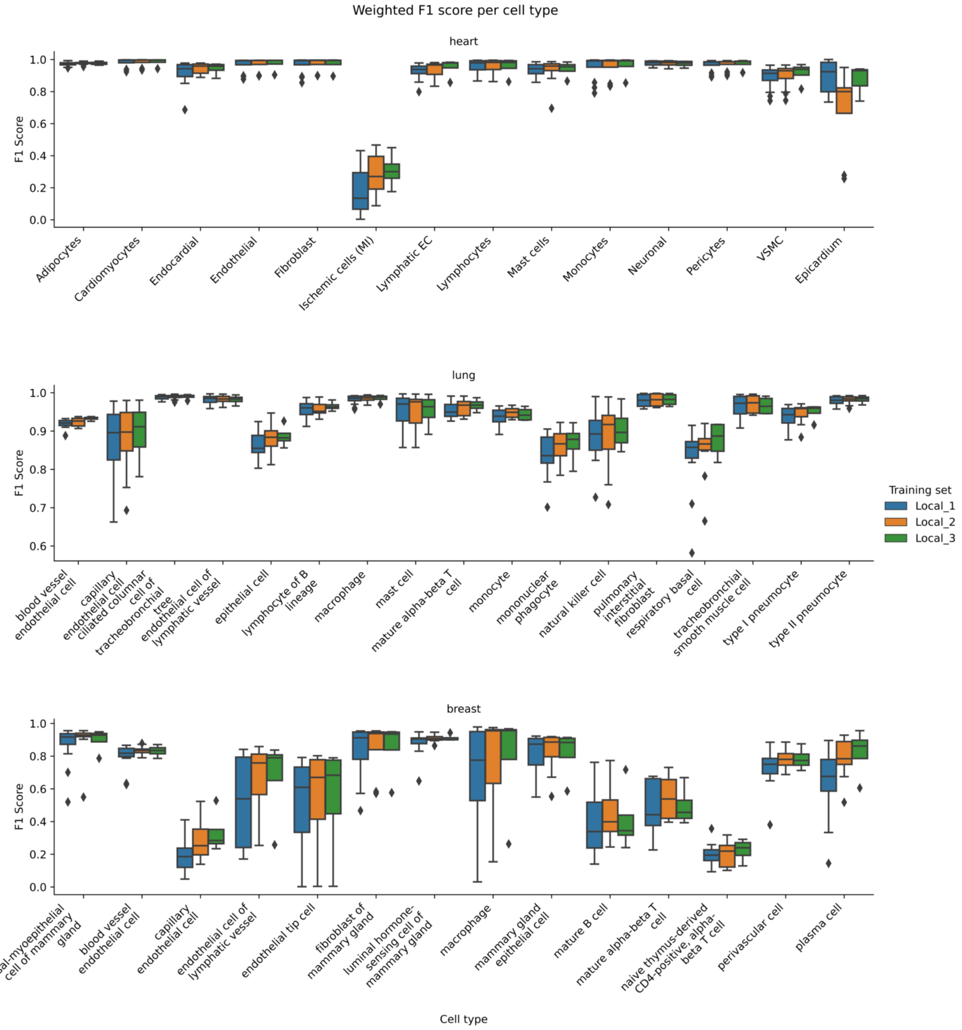
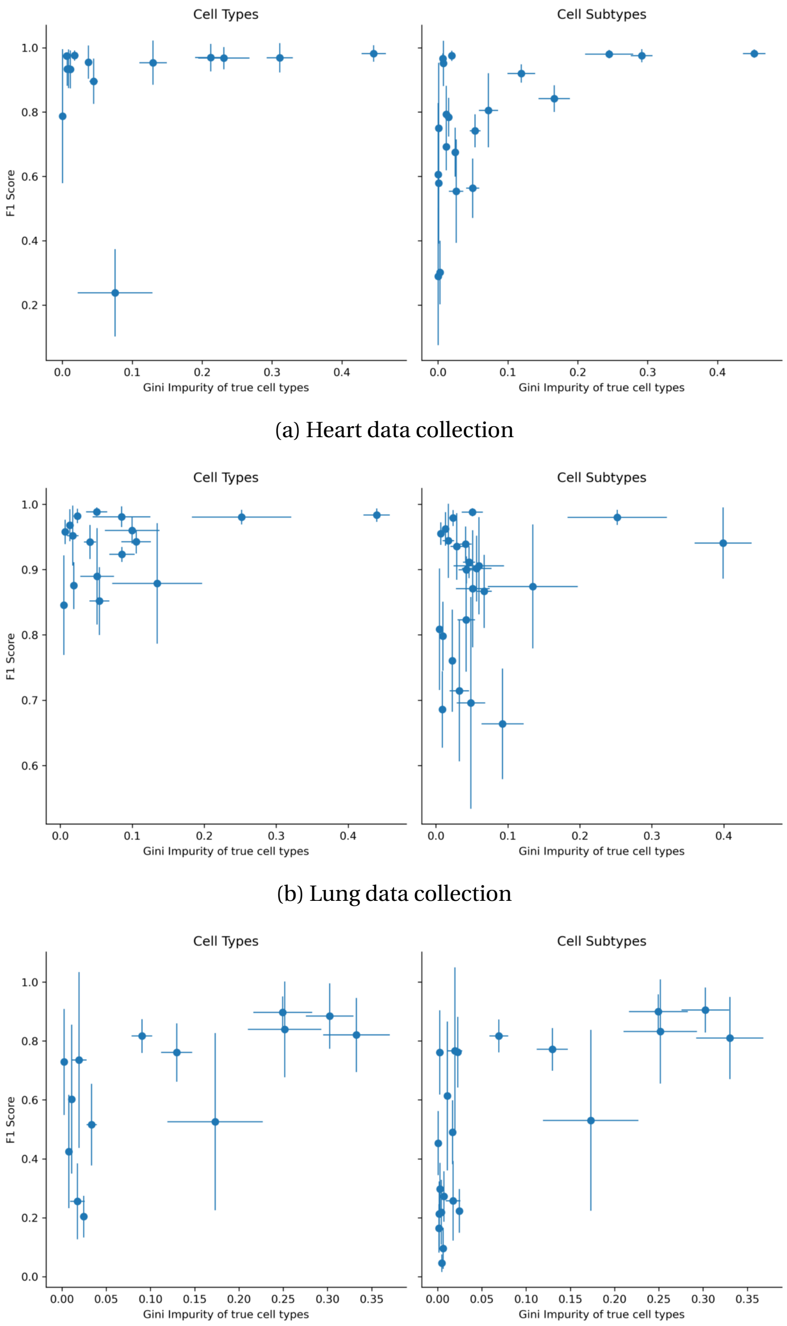


Fig. 3: Classification performance of the local learning framework. F1 scores are averaged over all simulation runs for each cell type. Overall, there is a notable disparity in classification accuracy between cell types.

3.1.3 Cell type prediction performance improves as cell count increases

To investigate that increasing the sample size of a cell type improves its classification accuracy, we evaluate the link between classification performance and the rarity of cell types. For each cell type and averaged over all studies, we compute the Gini impurity index, which is a measure of the rarity of the cell type (higher values are rarer classes). Then we computed the F1 score for each cell type and each study, averaged over all simulation runs (28 runs, combining Local_1, Local_2, and Local_3 together). Figure 4 represents the

F1 score as a function of the Gini impurity index of cell types. The results show that the classification performance increases as the rarity of the cell types increases.



(a) Heart data collection

(b) Lung data collection

(c) Breast data collection

Fig. 4: Classification performance as a function of the rarity of the cell types. The x-axis represents the Gini impurity index of true cell types (higher values are rarer classes). The y-axis represents the F1 score. Each dot is a cell type and the values are averaged over all studies for the Gini indices and all simulation runs (Local_1, Local_2, Local_3) for the F1 scores.

Organ	Classification level	Intercept	Slope	p-value of slope	Adjusted R^2
Heart	Cell type	-0.007	0.131	0.531	-0.047
Lung	Cell type	-0.585	0.718	0.211	0.042
Breast	Cell type	-0.104	0.344	**0.013**	0.368
Heart	Cell subtype	-0.143	0.291	**0.018**	0.224
Lung	Cell subtype	-0.104	0.194	0.300	0.006
Breast	Cell subtype	-0.052	0.249	**0.001**	0.408

Table 1: Parameters of the linear regression models fitting the F1 score as a function of cell type rarity. Significant p-values are in bold.

To quantify this, a linear model was fitted to the data for each organ and level (considering the mean values as independent samples and discarding the estimated confidence intervals), and the results are reported in Table 1. All linear models have an estimated positive slope, with a significant p-value for the 3 cases: the breast collection (both cell types and subtypes) and the heart collection for cell subtypes. This confirms that the classification performance increases as the rarity of the cell types increases.

3.2 Swarm learning performs on par with centralized models

3.2.1 Classification performance across cell types using Swarm learning

The SL classifier is compared to the LL classifier trained on 3 studies ("Local_3"). Figure 5 shows the weighted F1 score in all cell types and subtypes for LL and SL. The SL setting is directly compared to the corresponding LL setting Local_3, while Local_1 and Local_2 are also included for comparison purposes. The difference in distribution between SL and LL F1 scores is non-significant (p-value < 0.05) in all settings using a two-sided Mann-Whitney test. When comparing Local_3 with SL in each organ, the values are 0.958 versus 0.934, and 0.972 versus 0.966 for the heart; 0.982 versus 0.982, and 0.958 versus 0.958 for the lung; 0.970 versus 0.809, and 0.796 versus 0.808 for the breast datasets, respectively. The numeric prediction accuracy values for all settings, as well as their confidence intervals, are reported in Supplementary Table 4.

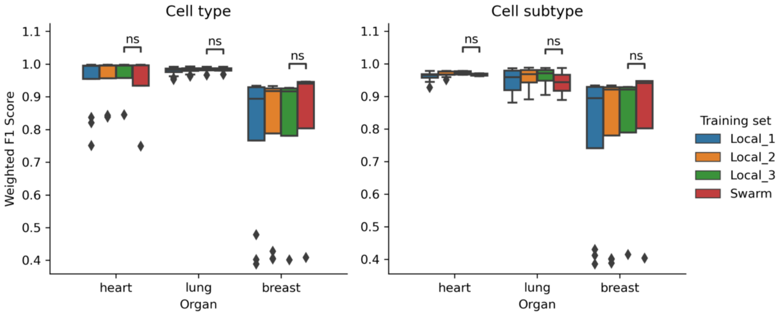


Fig. 5: Weighted F1 score across all cell types and subtypes for LL and SL settings. Significance levels are provided by Mann-Whitney tests. Overall, the performance of local and Swarm learning approaches are comparable, showing that there is no performance loss even when training is done without sharing datasets and models in the Swarm learning setting.

3.2.2 Classification performance per cell type using Swarm learning

The classification accuracy shows largely high accuracy for most cell types across the three organs as shown in the normalized confusion matrices (Figure 6 and Supplementary Figure S6). For heart, SL performs on par with LL for all cell types except Ischemic cells, Epicardium, and, to a lesser extent, Lymphatic ECs. Epicardium cells are present in low quantity and their small sample sizes (217, 0, 9, and 61) hinders efficient classification, which is reflected in lower SL performance, while “Ischemic cells” and “Lymphatic ECs” are difficult to differentiate using expression profiles. For lung, SL performs as well as LL for all cell types except respiratory basal cells, which are classified mainly as epithelial cells. This is explained by the fact that this cell type is present in small numbers (262, 245, 50, and 11, the smallest sample size for this collection). Finally, for the breast datasets, the SL classifier performs overall as well as LL. It performs on par or outperforms slightly for well-classified cell types (basal-myoepithelial cells of mammary gland, endothelial cell of lymphatic vessel, fibroblast of mammary gland, liminal hormone-sensing cell of mammary gland, macrophage, mammary gland epithelial cell, mature alpha-beta T cell, perivascular cell), it outperforms for blood vessel endothelial cells, and it slightly under-performs for some cell types which are already poorly classified by LL (capillary endothelial cell, endothelial tip cell, mature B cell, naive thymus-derived CD4-positive, alpha-beta T cell, and plasma cell). Overall, as more data are added to the collections, the “rare” cell types are better classified by LL, and thus also by SL. For LL, blood vessel endothelial cells are predicted as endothelial cells. Furthermore, several cell types are predicted as mammary gland epithelial cell: mature B cell, and two T cell types (mature alpha-beta T cell and naive thymus-derived CD4-positive, alpha-beta T cell). The very specific cell type naive thymus-derived CD4-positive, alpha-beta T cell has prediction scattered over 5 cell types. In LL setting, endothelial tip cells are predicted as mammary gland epithelial cell. For SL, in comparison, some cell types have their prediction performances significantly degraded: capillary endothelial cells, mature B cells, naive thymus-derived CD4-positive, alpha-beta

T cells (which is a difficult case, even for Local), and plasma cells, which were well classified in the Local_3 setting.

Visualization of the classification prediction score is available in Supplementary Figure S13.

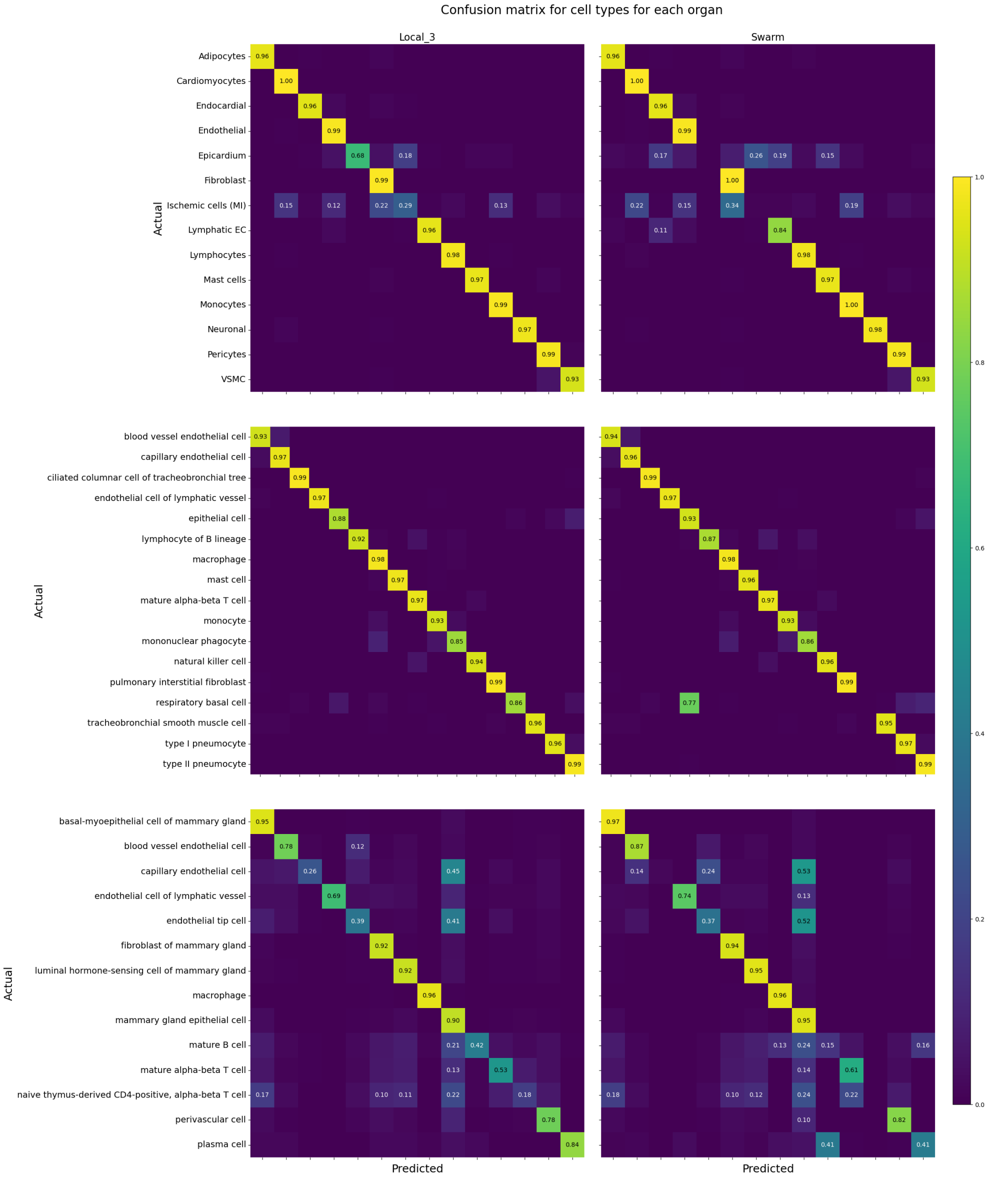


Fig. 6: Confusion matrices for classifying cell types using Local_3 (left) versus Swarm_3 (right) classifier on heart (top), lung (center), and breast (bottom) datasets. The accuracies are are averaged over all simulation runs and normalized by row.

3.2.3 Cell subtype classification is challenging

Cell subtypes are defined as cell-states in a given cell lineage or main cell type that are closely related to each other and share biological properties with other cell subtypes identified within the main cell types.

For heart, in LL setting, the only cell subtype that is misclassified in the relative majority of the time is dendritic cells, which are predicted to be macrophages 47% of the time. In SL, this cell subtype is correctly classified < 10% of the time and is predicted as macrophages a (absolute) majority of the time. The other misclassified subtypes are: CD8-positive, alpha-beta regulatory T cell and T cell, jointly; circulating angiogenic cell and endothelial cells, jointly; lymphocytes, as alpha-beta regulatory T cells, T cells, and natural killer cells, and plasma cells as natural kill cells, and myofibroblast cells as fibroblasts. For SL setting, 7 cell subtypes have an average accuracy < 10% (B cells, CD8-positive, alpha-beta regulatory T cells, circulating angiogenic cells, dendritic cells, lymphocytes, megakaryocytes, and plasma cells), and another subtype has an accuracy < 50% (myofibroblast cells). All of these subtypes also have low prediction accuracy in LL, with the notable addition of B cells (0.84% accuracy in LL), classified by SL mostly as monocyte and natural killer cells. All other cell subtypes with accuracy greater than 70% in LL had similar or better accuracy in SL.

In the lung datasets, in LL setting, 13% of CD1c-positive myeloid dendritic cells were predicted as lung macrophages; 16% of CD4-positive, alpha-beta T cells were predicted as and CD8-positive, alpha-beta T cells; 14% of elicited macrophages as alveolar macrophages; 12% of epithelial cell of alveolus of lung as epithelial cell of lower respiratory tract and 18% as type II pneumocytes; 15% of non-classical monocytes as classical monocytes; and 11% of pulmonary artery endothelial cells as capillary endothelial cells. In SL, the six aforementioned cell subtypes show a drop in accuracy, with epithelial cells of alveolus of lung having the largest drop in accuracy from 65% to 13%. Notably, most subtypes well classified in LL are also well classified in SL, with respiratory basal cells standing out, having an accuracy of 82% in local learning down to below 10% in Swarm learning.

In breast datasets, only seven subtypes have an accuracy over 80% in LL: basal-myoepithelial cells of mammary gland, fibroblasts of mammary gland, luminal adaptive secretory precursor cell of mammary gland, macrophages, plasma cells, and vein endothelial cells. Of the 11 subtypes which are classified with > 50% accuracy in LL, 10 are classified with an increased or equal accuracy (the outlier is mature NK T cells, which drop from an accuracy of 50% to < 10%). On the opposite side, of the 11 subtypes which are classified with < 50% accuracy in LL, 10 are classified with decreased accuracy in SL (the outlier is CD8-positive, alpha-beta memory T cells, which increase from 47% to 60%). Furthermore, for breast in LL, 6 subtypes are misclassified as luminal adaptive secretory precursor cells of mammary gland in a relative majority of cases: Tc1 cells (20%), capillary endothelial cells (41%), class switched memory B cells (25%), mammary gland epithelial cells (misclassified in an absolute majority of cases, 51%), naive thymus-derived CD4-positive, alpha-beta T cells (21%), and unswitched memory B cells (15%). This could be explained by a difficulty in finding expression signatures which differentiate well these subtypes from the biomarkers of luminal adaptive secretory precursor cells of mammary gland. These results are carried over to SL, in which the misclassification rates increase in these 6 subtypes.

Taken together, these results highlight that SL performs worse than LL when the cell subtypes are poorly classified in LL and performs better when the cell subtypes are sufficiently well classified by LL.

4 Discussion

Our study shows that there were no significant differences in the precision of cell type prediction between the Swarm model and the model trained on all combined data. This suggests that using the Swarm model approach is efficient method for predicting cell types in large-scale single-cell transcriptomics datasets while maintaining data privacy. However, challenges remain for predicting cell-types that are not clearly differentiable. This could be due to low sample count, lineage that is a mixture of multiple cell-types e.g. ischemic cells, over- or under- clustering, or closely related cell-types. Figures 6 and S6 highlight *(i)* the challenge in annotating cell types which are present in small numbers and *(ii)* the higher accuracy of annotating cell types which are well-defined and present in higher numbers.

In heart datasets tested here, cluster annotated as ischemic cells was difficult to classify in both LL and SL settings. This is due to ischemic cells being a aggregated cluster comprising of several lineages including cardiomyocytes, fibroblasts and endothelial cells. Additionally, lymphatic endothelial cells are misclassified as endocardial cells. Both cell-types are closely related and share marker genes [38]. For lung datasets, performance is good throughout classes for LL, except for “respiratory basal cells” (cell ontology id “CL:0002633”), which are classified as “epithelial cells” (cell ontology if “CL:0000066”). This is biologically sensible, as the former is a subtype of the latter in the Cell Ontology. Likewise, SL performs well across cell types except for the latter case, in which 77% of respiratory basal cells are misclassified as epithelial cells. For breast datasets, cell types including blood vessel endothelial cells, fibroblasts of mammary gland, luminal hormone-sensing cells of mammary gland, and mammary gland epithelial cells, that are well classified in LL setting, are also classified correctly in the SL setting.

The cell subtype classification task is harder than the prediction of the main cell type as different cell states can have similar marker genes (see Supplementary Figure 4). In general, the classification is fairly accurate across subtypes for the lung (1 cell subtype out of 24 with below 70% accuracy), followed by the heart (5 of 21 cell subtypes below 70% accuracy) and finally the breast (13 of 22 cell types below 70% accuracy).

This study highlights the potential of Swarm learning to build scalable and privacy-preserving models from single-cell transcriptomic data. The next step will consist in training hierarchical models for all annotation levels. Using the cell ontology as *a priori* information, hierarchical classifiers [39] can be used to annotate cells at the correct depth in the ontology. This approach will benefit from fine-grained annotated datasets which are now increasingly available to resolve automated fine-grained annotation. Another extension of this work is the building of a cross-organ model for annotation. Since many cell types are present in different organs, pooling datasets across organs will directly increase the available training size in terms of cell count per cell type, without hindering prediction of organ-specific cell types. Finally, as Swarm learning is applicable to any classifier, it can be applied to multi-omics data, leveraging different types of biological information (surface protein, chromatic accessibility, etc.) to gain a deeper insight into not only cell types, but also cell states.

Funding. This work was funded by RWTH Aachen START (ID 692308), CRU344 and Leducq Immuno-Fib HF seed grant to SH. JNK is supported by the German Cancer Aid (DECADE, 70115166), the German Federal Ministry of Education and Research (PEARL, 01KD2104C; CAMINO, 01EO2101; TRANSFORM LIVER, 031L0312A; TANGERINE, 01KT2302 through ERA-NET Transcan; Come2Data, 16DKZ2044A; DEEP-HCC, 031L0315A), the German Academic Exchange Service (SECAI, 57616814), the European Union's Horizon Europe and innovation programme (ODELIA, 101057091; GENIAL, 101096312), the European Research Council (ERC; NADIR, 101114631), the National Institutes of Health (EPICO, R01 CA263318) and the National Institute for Health and Care Research (NIHR, NIHR203331) Leeds Biomedical Research Centre.

Conflict of interest. SH is a consultant for Turbine AI, co-founder and shareholder of Sequantrix GmbH and has recieved research funding from AskBio and Novo Nordisk. JNK declares consulting services for Bioptimus, France; Owkin, France; DoMore Diagnostics, Norway; Panakeia, UK; AstraZeneca, UK; Mindpeak, Germany; and MultiplexDx, Slovakia. Furthermore, he holds shares in StratifAI GmbH, Germany, Synagen GmbH, Germany; has received a research grant by GSK; and has received honoraria by AstraZeneca, Bayer, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, BMS, Roche, Pfizer, and Fresenius. No other competing financial interests are declared by any of the remaining authors.

Acknowledgments. The authors gratefully acknowledge Tore Bleckwehl for providing the standardized heart atlas that was used in this study.

Author contribution. Conceptualization: OLS, JNK, SH; Methodology: VG, KP, OLS, HK, JFZ, JNK, SH; Software: VG, KP; Formal analysis: VG, KP; Resources: JNK, RK, SH; Data Curation: VG, HK, KP; Writing - Original Draft: VG, KP, SH; Writing - Review & Editing: VG, OLS, KP, RK, JNK, SH; Visualization: VG, KP; Supervision: JNK, SH; Project administration: OLS, SH; Funding acquisition: JNK, SH;

References

[1] Rozenblatt-Rosen, O., Stubbington, M.J.T., Regev, A., Teichmann, S.A.: The Human Cell Atlas: From vision to reality. Nature **550**(7677), 451–453 (2017) https://doi.org/10.1038/550451a

[2] Schreibing, F., Hannani, M.T., Kim, H., Nagai, J.S., Ticconi, F., Fewings, E., Bleckwehl, T., Begemann, M., Torow, N., Kuppe, C., Kurth, I., Kranz, J., Frank, D., Anslinger, T.M., Ziegler, P., Kraus, T., Enczmann, J., Balz, V., Windhofer, F., Balfanz, P., Kurts, C., Marx, G., Marx, N., Dreher, M., Schneider, R.K., Saez-Rodriguez, J., Costa, I., Hayat, S., Kramann, R.: Dissecting CD8+ T cell pathology of severe SARS-CoV-2 infection by single-cell immunoprofiling. Frontiers in Immunology **13** (2022) https://doi.org/10.3389/fimmu.2022.1066176

[3] Stephenson, E., Reynolds, G., Botting, R.A., Calero-Nieto, F.J., Morgan, M.D., Tuong, Z.K., Bach, K., Sungnak, W., Worlock, K.B., Yoshida, M., Kumasaka, N., Kania, K., Engelbert, J., Olabi, B., Spegarova, J.S., Wilson, N.K., Mende, N., Jardine, L., Gardner, L.C.S., Goh, I., Horsfall, D., McGrath, J., Webb, S., Mather, M.W., Lindeboom, R.G.H., Dann, E., Huang, N., Polanski, K., Prigmore, E., Gothe, F., Scott, J., Payne, R.P., Baker, K.F., Hanrath, A.T., Schim van der Loeff, I.C.D., Barr, A.S., Sanchez-Gonzalez, A., Bergamaschi, L., Mescia, F., Barnes, J.L., Kilich, E., de Wilton, A., Saigal, A., Saleh, A., Janes, S.M., Smith, C.M., Gopee, N., Wilson, C., Coupland, P., Coxhead, J.M., Kiselev, V.Y., van Dongen, S., Bacardit, J., King, H.W., Rostron, A.J., Simpson, A.J., Hambleton, S., Laurenti, E., Lyons, P.A., Meyer, K.B., Nikolić, M.Z., Duncan, C.J.A., Smith, K.G.C., Teichmann, S.A., Clatworthy, M.R., Marioni, J.C., Göttgens, B., Haniffa, M.: Single-cell multi-omics analysis of the immune response in COVID-19. Nature Medicine **27**(5), 904–916 (2021) https://doi.org/10.1038/s41591-021-01329-2

[4] Sikkema, L., Ramírez-Suástegui, C., Strobl, D.C., Gillett, T.E., Zappia, L., Madissoon, E., Markov, N.S., Zaragosi, L.-E., Ji, Y., Ansari, M., Arguel, M.-J., Apperloo, L., Banchero, M., Bécavin, C., Berg, M., Chichelnitskiy, E., Chung, M.-i., Collin, A., Gay, A.C.A., Gote-Schniering, J., Hooshiar Kashani, B., Inecik, K., Jain, M., Kapellos, T.S., Kole, T.M., Leroy, S., Mayr, C.H., Oliver, A.J., von Papen, M., Peter, L., Taylor, C.J., Walzthoeni, T., Xu, C., Bui, L.T., De Donno, C., Dony, L., Faiz, A., Guo, M., Gutierrez, A.J., Heumos, L., Huang, N., Ibarra, I.L., Jackson, N.D., Kadur Lakshminarasimha Murthy, P., Lotfollahi, M., Tabib, T., Talavera-López, C., Travaglini, K.J., Wilbrey-Clark, A., Worlock, K.B., Yoshida, M., van den Berge, M., Bossé, Y., Desai, T.J., Eickelberg, O., Kaminski, N., Krasnow, M.A., Lafyatis, R., Nikolic, M.Z., Powell, J.E., Rajagopal, J., Rojas, M., Rozenblatt-Rosen, O., Seibold, M.A., Sheppard, D., Shepherd, D.P., Sin, D.D., Timens, W., Tsankov, A.M., Whitsett, J., Xu, Y., Banovich, N.E., Barbry, P., Duong, T.E., Falk, C.S., Meyer, K.B., Kropski, J.A., Pe'er, D., Schiller, H.B., Tata, P.R., Schultze, J.L., Teichmann, S.A., Misharin, A.V., Nawijn, M.C., Luecken, M.D., Theis, F.J.: An integrated cell atlas of the lung in health and disease. Nature Medicine **29**(6), 1563–1577 (2023) https://doi.org/10.1038/s41591-023-02327-2

[5] Bleckwehl, T., Babler, A., Tebens, M., Maryam, S., Nyberg, M., Bosteen, M., Halder,

M., Shaw, I., Fleig, S., Pyke, C., et al.: Encompassing view of spatial and single-cell rna sequencing renews the role of the microvasculature in human atherosclerosis. Nature Cardiovascular Research, 1–19 (2024)

[6] Kuppe, C., Ramirez Flores, R.O., Li, Z., Hayat, S., Levinson, R.T., Liao, X., Hannani, M.T., Tanevski, J., Wünnemann, F., Nagai, J.S., Halder, M., Schumacher, D., Menzel, S., Schäfer, G., Hoeft, K., Cheng, M., Ziegler, S., Zhang, X., Peisker, F., Kaesler, N., Saritas, T., Xu, Y., Kassner, A., Gummert, J., Morshuis, M., Amrute, J., Veltrop, R.J.A., Boor, P., Klingel, K., Van Laake, L.W., Vink, A., Hoogenboezem, R.M., Bindels, E.M.J., Schurgers, L., Sattler, S., Schapiro, D., Schneider, R.K., Lavine, K., Milting, H., Costa, I.G., Saez-Rodriguez, J., Kramann, R.: Spatial multi-omic map of human myocardial infarction. Nature **608**(7924), 766–777 (2022) https://doi.org/10.1038/s41586-022-05060-x

[7] Kanemaru, K., Cranley, J., Muraro, D., Miranda, A.M.A., Ho, S.Y., Wilbrey-Clark, A., Patrick Pett, J., Polanski, K., Richardson, L., Litvinukova, M., Kumasaka, N., Qin, Y., Jablonska, Z., Semprich, C.I., Mach, L., Dabrowska, M., Richoz, N., Bolt, L., Mamanova, L., Kapuge, R., Barnett, S.N., Perera, S., Talavera-López, C., Mulas, I., Mahbubani, K.T., Tuck, L., Wang, L., Huang, M.M., Prete, M., Pritchard, S., Dark, J., Saeb-Parsy, K., Patel, M., Clatworthy, M.R., Hübner, N., Chowdhury, R.A., Noseda, M., Teichmann, S.A.: Spatially resolved multiomics of human cardiac niches. Nature **619**(7971), 801–810 (2023) https://doi.org/10.1038/s41586-023-06311-1

[8] Kuppe, C., Ibrahim, M.M., Kranz, J., Zhang, X., Ziegler, S., Perales-Patón, J., Jansen, J., Reimer, K.C., Smith, J.R., Dobie, R., Wilson-Kanamori, J.R., Halder, M., Xu, Y., Kabgani, N., Kaesler, N., Klaus, M., Gernhold, L., Puelles, V.G., Huber, T.B., Boor, P., Menzel, S., Hoogenboezem, R.M., Bindels, E.M.J., Steffens, J., Floege, J., Schneider, R.K., Saez-Rodriguez, J., Henderson, N.C., Kramann, R.: Decoding myofibroblast origins in human kidney fibrosis. Nature **589**(7841), 281–286 (2021) https://doi.org/10.1038/s41586-020-2941-1

[9] Reed, A.D., Pensa, S., Steif, A., Stenning, J., Kunz, D.J., Porter, L.J., Hua, K., He, P., Twigger, A.-J., Siu, A.J.Q., Kania, K., Barrow-McGee, R., Goulding, I., Gomm, J.J., Speirs, V., Jones, J.L., Marioni, J.C., Khaled, W.T.: A single-cell atlas enables mapping of homeostatic cellular shifts in the adult human breast. Nature Genetics **56**(4), 652–662 (2024) https://doi.org/10.1038/s41588-024-01688-9

[10] Khaliq, A.M., Erdogan, C., Kurt, Z., Turgut, S.S., Grunvald, M.W., Rand, T., Khare, S., Borgia, J.A., Hayden, D.M., Pappas, S.G., Govekar, H.R., Kam, A.E., Reiser, J., Turaga, K., Radovich, M., Zang, Y., Qiu, Y., Liu, Y., Fishel, M.L., Turk, A., Gupta, V., Al-Sabti, R., Subramanian, J., Kuzel, T.M., Sadanandam, A., Waldron, L., Hussain, A., Saleem, M., El-Rayes, B., Salahudeen, A.A., Masood, A.: Refining colorectal cancer classification and clinical stratification through a single-cell atlas. Genome Biology **23**(1), 113 (2022) https://doi.org/10.1186/s13059-022-02677-z

[11] Kang, J., Lee, J.H., Cha, H., An, J., Kwon, J., Lee, S., Kim, S., Baykan, M.Y., Kim, S.Y.,

An, D., Kwon, A.-Y., An, H.J., Lee, S.-H., Choi, J.K., Park, J.-E.: Systematic dissection of tumor-normal single-cell ecosystems across a thousand tumors of 30 cancer types. Nature Communications **15**(1), 4067 (2024) https://doi.org/10.1038/s41467-024-48310-4

[12] Luecken, M.D., Theis, F.J.: Current best practices in single-cell RNA-seq analysis: A tutorial. Molecular Systems Biology **15**(6), 8746 (2019) https://doi.org/10.15252/msb.20188746

[13] Lohmöller, J., Scheiber, J., Kramann, R., Wehrle, K., Hayat, S., Pennekamp, J.: sce (match): Privacy-preserving cluster matching of single-cell data (2024)

[14] Domínguez Conde, C., Xu, C., Jarvis, L.B., Rainbow, D.B., Wells, S.B., Gomes, T., Howlett, S.K., Suchanek, O., Polanski, K., King, H.W., Mamanova, L., Huang, N., Szabo, P.A., Richardson, L., Bolt, L., Fasouli, E.S., Mahbubani, K.T., Prete, M., Tuck, L., Richoz, N., Tuong, Z.K., Campos, L., Mousa, H.S., Needham, E.J., Pritchard, S., Li, T., Elmentaite, R., Park, J., Rahmani, E., Chen, D., Menon, D.K., Bayraktar, O.A., James, L.K., Meyer, K.B., Yosef, N., Clatworthy, M.R., Sims, P.A., Farber, D.L., Saeb-Parsy, K., Jones, J.L., Teichmann, S.A.: Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science **376**(6594), 5197 (2022) https://doi.org/10.1126/science.abl5197

[15] Hou, R., Denisenko, E., Forrest, A.R.R.: scMatch: A single-cell gene expression profile annotation tool using reference datasets. Bioinformatics **35**(22), 4688–4695 (2019) https://doi.org/10.1093/bioinformatics/btz292

[16] Xu, Y., Baumgart, S.J., Stegmann, C.M., Hayat, S.: MACA: Marker-based automatic cell-type annotation for single-cell expression data. Bioinformatics **38**(6), 1756–1760 (2022) https://doi.org/10.1093/bioinformatics/btab840

[17] Xu, Y., Kramann, R., McCord, R.P., Hayat, S.: MASI enables fast model-free standardization and integration of single-cell transcriptomics data. Communications Biology **6**(1), 1–14 (2023) https://doi.org/10.1038/s42003-023-04820-3

[18] Kang, J.B., Nathan, A., Weinand, K., Zhang, F., Millard, N., Rumker, L., Moody, D.B., Korsunsky, I., Raychaudhuri, S.: Efficient and precise single-cell reference atlas mapping with Symphony. Nature Communications **12**(1), 5890 (2021) https://doi.org/10.1038/s41467-021-25957-x

[19] Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M.I., Yosef, N.: Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Molecular Systems Biology **17**(1), 9620 (2021) https://doi.org/10/gs6dxt

[20] Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., *et al.*: Integrated analysis of multimodal single-cell data. Cell **184**(13), 3573–3587 (2021)

[21] Walker, C.R., Li, X., Chakravarthy, M., Lounsbery-Scaife, W., Choi, Y.A., Singh, R., Gürsoy, G.: Private information leakage from single-cell count matrices. Cell **187**(23), 6537–6549 (2024)

[22] Warnat-Herresthal, S., Schultze, H., Shastry, K.L., Manamohan, S., Mukherjee, S., Garg, V., Sarveswara, R., Händler, K., Pickkers, P., Aziz, N.A., *et al.*: Swarm learning for decentralized and confidential clinical machine learning. Nature **594**(7862), 265–270 (2021)

[23] Saldanha, O.L., Quirke, P., West, N.P., James, J.A., Loughrey, M.B., Grabsch, H.I., Salto-Tellez, M., Alwers, E., Cifci, D., Ghaffari Laleh, N., Seibel, T., Gray, R., Hutchins, G.G.A., Brenner, H., van Treeck, M., Yuan, T., Brinker, T.J., Chang-Claude, J., Khader, F., Schuppert, A., Luedde, T., Trautwein, C., Muti, H.S., Foersch, S., Hoffmeister, M., Truhn, D., Kather, J.N.: Swarm learning for decentralized artificial intelligence in cancer histopathology. Nature Medicine **28**(6), 1232–1239 (2022) https://doi.org/10.1038/s41591-022-01768-5

[24] Saldanha, O.L., Muti, H.S., Grabsch, H.I., Langer, R., Dislich, B., Kohlruss, M., Keller, G., Treeck, M., Hewitt, K.J., Kolbinger, F.R., *et al.*: Direct prediction of genetic aberrations from pathology images in gastric cancer with swarm learning. Gastric cancer **26**(2), 264–274 (2023)

[25] Cifci, D., Veldhuizen, G.P., Foersch, S., Kather, J.N.: Ai in computational pathology of cancer: improving diagnostic workflows and clinical outcomes? Annual Review of Cancer Biology **7**(1), 57–71 (2023)

[26] Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., Newell, E.W.: Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnology **37**(1), 38–44 (2019) https://doi.org/10.1038/nbt.4314

[27] Diehl, A.D., Meehan, T.F., Bradford, Y.M., Brush, M.H., Dahdul, W.M., Dougall, D.S., He, Y., Osumi-Sutherland, D., Ruttenberg, A., Sarntivijai, S., Van Slyke, C.E., Vasilevsky, N.A., Haendel, M.A., Blake, J.A., Mungall, C.J.: The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. Journal of Biomedical Semantics **7**(1), 44 (2016) https://doi.org/10.1186/s13326-016-0088-7 . Accessed 2025-01-08

[28] Zhao, R., Lu, J., Zhou, W., Zhao, N., Ji, H.: A systematic evaluation of highly variable gene selection methods for single-cell rna-sequencing. bioRxiv, 2024–08 (2024)

[29] Warnat-Herresthal, S., Schultze, H., Shastry, K.L., Manamohan, S., Mukherjee, S., Garg, V., Sarveswara, R., Händler, K., Pickkers, P., Aziz, N.A., Ktena, S., Tran, F., Bitzer, M., Ossowski, S., Casadei, N., Herr, C., Petersheim, D., Behrends, U., Kern, F., Fehlmann, T., Schommers, P., Lehmann, C., Augustin, M., Rybniker, J., Altmüller, J., Mishra, N., Bernardes, J.P., Krämer, B., Bonaguro, L., Schulte-Schrepping, J., De Domenico, E., Siever, C., Kraut, M., Desai, M., Monnet, B., Saridaki, M., Siegel, C.M., Drews, A., Nuesch-Germano, M., Theis, H., Heyckendorf, J., Schreiber, S., Kim-Hellmuth, S., Nattermann, J., Skowasch, D., Kurth, I., Keller, A., Bals, R., Nürnberg,

P., Rieß, O., Rosenstiel, P., Netea, M.G., Theis, F., Mukherjee, S., Backes, M., Aschenbrenner, A.C., Ulas, T., Breteler, M.M.B., Giamarellos-Bourboulis, E.J., Kox, M., Becker, M., Cheran, S., Woodacre, M.S., Goh, E.L., Schultze, J.L.: Swarm Learning for decentralized and confidential clinical machine learning. Nature **594**(7862), 265–270 (2021) https://doi.org/10.1038/s41586-021-03583-3

[30] Saldanha, O.L., Muti, H.S., Grabsch, H.I., Langer, R., Dislich, B., Kohlruss, M., Keller, G., van Treeck, M., Hewitt, K.J., Kolbinger, F.R., Veldhuizen, G.P., Boor, P., Foersch, S., Truhn, D., Kather, J.N.: Direct prediction of genetic aberrations from pathology images in gastric cancer with swarm learning. Gastric Cancer: Official Journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association **26**(2), 264–274 (2023) https://doi.org/10.1007/s10120-022-01347-0

[31] Pliner, H.A., Shendure, J., Trapnell, C.: Supervised classification enables rapid annotation of cell atlases. Nature Methods **16**(10), 983–986 (2019) https://doi.org/10.1038/s41592-019-0535-3

[32] Ma, F., Pellegrini, M.: ACTINN: automated identification of cell types in single cell RNA sequencing. Bioinformatics **36**(2), 533–538 (2020) https://doi.org/10.1093/bioinformatics/btz592 . Accessed 2025-01-10

[33] Heryanto, Y.D., Zhang, Y.-z., Imoto, S.: Predicting cell types with supervised contrastive learning on cells and their types. Scientific Reports **14**(1), 430 (2024) https://doi.org/10.1038/s41598-023-50185-2 . Publisher: Nature Publishing Group. Accessed 2025-01-06

[34] Fischer, F., Fischer, D.S., Biederstedt, E., Villani, A.-C., Theis, F.J.: Scaling Cross-Tissue Single-Cell Annotation Models. bioRxiv (2023). https://doi.org/10.1101/2023.10.07.561331

[35] Arik, S.O., Pfister, T.: TabNet: Attentive Interpretable Tabular Learning. arXiv (2020)

[36] Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD ’16, pp. 785–794. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785

[37] Shwartz-Ziv, R., Armon, A.: Tabular data: Deep learning is not all you need. Information Fusion **81**, 84–90 (2022) https://doi.org/10.1016/j.inffus.2021.11.011

[38] Ishii, Y., Langberg, J., Rosborough, K., Mikawa, T.: Endothelial cell lineages of the heart. Cell and tissue research **335**, 67–73 (2009)

[39] Silla, C.N., Freitas, A.A.: A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery **22**(1-2), 31–72 (2011) https://doi.org/10.1007/s10618-010-0175-9 . Accessed 2025-01-11

[40] Chaffin, M., Papangeli, I., Simonson, B., Akkad, A.-D., Hill, M.C., Arduini, A., Fleming, S.J., Melanson, M., Hayat, S., Kost-Alimova, M., Atwa, O., Ye, J., Bedi, K.C., Nahrendorf, M., Kaushik, V.K., Stegmann, C.M., Margulies, K.B., Tucker, N.R., Ellinor, P.T.: Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. Nature **608**(7921), 174–180 (2022) https://doi.org/10.1038/s41586-022-04817-8

[41] Litviňuková, M., Talavera-López, C., Maatz, H., Reichart, D., Worth, C.L., Lindberg, E.L., Kanda, M., Polanski, K., Heinig, M., Lee, M., Nadelmann, E.R., Roberts, K., Tuck, L., Fasouli, E.S., DeLaughter, D.M., McDonough, B., Wakimoto, H., Gorham, J.M., Samari, S., Mahbubani, K.T., Saeb-Parsy, K., Patone, G., Boyle, J.J., Zhang, H., Zhang, H., Viveiros, A., Oudit, G.Y., Bayraktar, O.A., Seidman, J.G., Seidman, C.E., Noseda, M., Hubner, N., Teichmann, S.A.: Cells of the adult human heart. Nature **588**(7838), 466–472 (2020) https://doi.org/10.1038/s41586-020-2797-4

[42] Reichart, D., Lindberg, E.L., Maatz, H., Miranda, A.M.A., Viveiros, A., Shvetsov, N., Gärtner, A., Nadelmann, E.R., Lee, M., Kanemaru, K., Ruiz-Orera, J., Strohmenger, V., DeLaughter, D.M., Patone, G., Zhang, H., Woehler, A., Lippert, C., Kim, Y., Adami, E., Gorham, J.M., Barnett, S.N., Brown, K., Buchan, R.J., Chowdhury, R.A., Constantinou, C., Cranley, J., Felkin, L.E., Fox, H., Ghauri, A., Gummert, J., Kanda, M., Li, R., Mach, L., McDonough, B., Samari, S., Shahriaran, F., Yapp, C., Stanasiuk, C., Theotokis, P.I., Theis, F.J., van den Bogaerdt, A., Wakimoto, H., Ware, J.S., Worth, C.L., Barton, P.J.R., Lee, Y.-A., Teichmann, S.A., Milting, H., Noseda, M., Oudit, G.Y., Heinig, M., Seidman, J.G., Hubner, N., Seidman, C.E.: Pathogenic variants damage cell composition and single cell transcription in cardiomyopathies. Science (New York, N.Y.) **377**(6606), 1984 (2022) https://doi.org/10.1126/science.abo1984

[43] Habermann, A.C., Gutierrez, A.J., Bui, L.T., Yahn, S.L., Winters, N.I., Calvi, C.L., Peter, L., Chung, M.-I., Taylor, C.J., Jetter, C., Raju, L., Roberson, J., Ding, G., Wood, L., Sucre, J.M.S., Richmond, B.W., Serezani, A.P., McDonnell, W.J., Mallal, S.B., Bacchetta, M.J., Loyd, J.E., Shaver, C.M., Ware, L.B., Bremner, R., Walia, R., Blackwell, T.S., Banovich, N.E., Kropski, J.A.: Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. Science Advances **6**(28), 1972 (2020) https://doi.org/10.1126/sciadv.aba1972

[44] Travaglini, K.J., Nabhan, A.N., Penland, L., Sinha, R., Gillich, A., Sit, R.V., Chang, S., Conley, S.D., Mori, Y., Seita, J., Berry, G.J., Shrager, J.B., Metzger, R.J., Kuo, C.S., Neff, N., Weissman, I.L., Quake, S.R., Krasnow, M.A.: A molecular cell atlas of the human lung from single-cell RNA sequencing. Nature **587**(7835), 619–625 (2020) https://doi.org/10.1038/s41586-020-2922-4

[45] Grant, R.A., Morales-Nebreda, L., Markov, N.S., Swaminathan, S., Querrey, M., Guzman, E.R., Abbott, D.A., Donnelly, H.K., Donayre, A., Goldberg, I.A., Klug, Z.M., Borkowski, N., Lu, Z., Kihshen, H., Politanska, Y., Sichizya, L., Kang, M., Shilatifard, A., Qi, C., Lomasney, J.W., Argento, A.C., Kruser, J.M., Malsin, E.S., Pickens, C.O., Smith, S.B., Walter, J.M., Pawlowski, A.E., Schneider, D., Nannapaneni, P., Abdala-Valencia, H., Bharat, A., Gottardi, C.J., Budinger, G.R.S., Misharin, A.V., Singer, B.D., Wunderink,

R.G.: Circuits between infected macrophages and T cells in SARS-CoV-2 pneumonia. Nature **590**(7847), 635–641 (2021) https://doi.org/10.1038/s41586-020-03148-w . Publisher: Nature Publishing Group. Accessed 2025-01-11

[46] Misharin, A.V., Budinger, G.R.S.: Targeting the Myofibroblast in Pulmonary Fibrosis. American Journal of Respiratory and Critical Care Medicine **198**(7), 834–835 (2018) https://doi.org/10.1164/rccm.201806-1037ED

[47] Murrow, L.M., Weber, R.J., Caruso, J.A., McGinnis, C.S., Phong, K., Gascard, P., Rabadam, G., Borowsky, A.D., Desai, T.A., Thomson, M., Tlsty, T., Gartner, Z.J.: Mapping hormone-regulated cell-cell interaction networks in the human breast at single-cell resolution. Cell Systems **13**(8), 644–6648 (2022) https://doi.org/10.1016/j.cels.2022.06.005

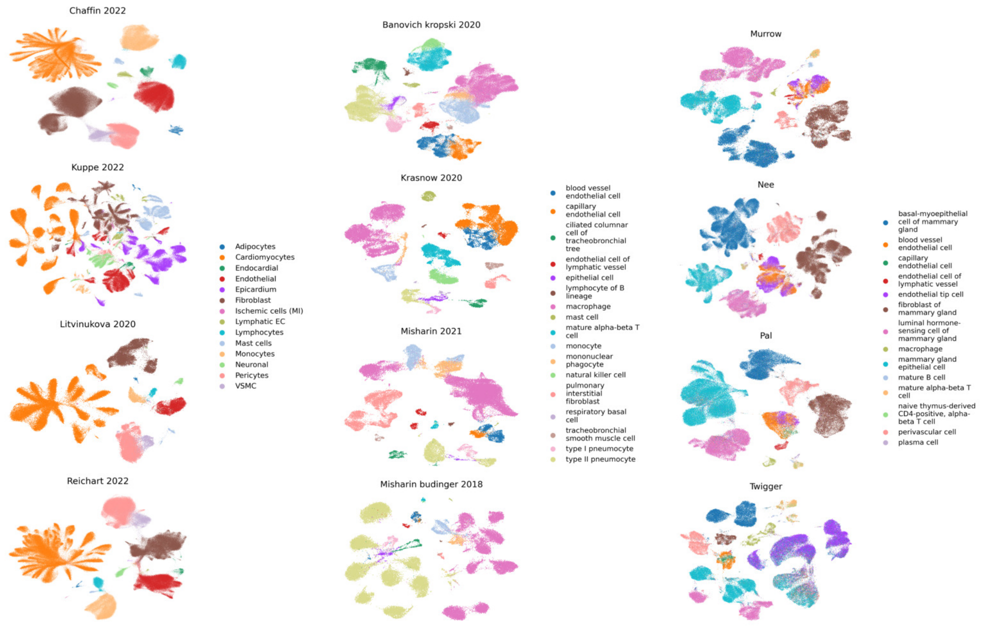
[48] Nee, K., Ma, D., Nguyen, Q.H., Pein, M., Pervolarakis, N., Insua-Rodríguez, J., Gong, Y., Hernandez, G., Alshetaiwi, H., Williams, J., Rauf, M., Dave, K.R., Boyapati, K., Hasnain, A., Calderon, C., Markaryan, A., Edwards, R., Lin, E., Parajuli, R., Zhou, P., Nie, Q., Shalabi, S., LaBarge, M.A., Kessenbrock, K.: Preneoplastic stromal cells promote BRCA1-mediated breast tumorigenesis. Nature Genetics **55**(4), 595–606 (2023) https://doi.org/10.1038/s41588-023-01298-x

[49] Pal, B., Chen, Y., Vaillant, F., Capaldo, B.D., Joyce, R., Song, X., Bryant, V.L., Penington, J.S., Di Stefano, L., Tubau Ribera, N., Wilcox, S., Mann, G.B., kConFab, Papenfuss, A.T., Lindeman, G.J., Smyth, G.K., Visvader, J.E.: A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. The EMBO Journal **40**(11), 107333 (2021) https://doi.org/10.15252/embj.2020107333

[50] Twigger, A.-J., Engelbrecht, L.K., Bach, K., Schultz-Pernice, I., Pensa, S., Stenning, J., Petricca, S., Scheel, C.H., Khaled, W.T.: Transcriptional changes in the mammary gland during lactation revealed by single cell sequencing of cells from human milk. Nature Communications **13**(1), 562 (2022) https://doi.org/10.1038/s41467-021-27895-0

Supplementary Information for the paper: “SwarmMAP: Swarm Learning for Decentralized Cell Type Annotation in Single Cell Sequencing Data”

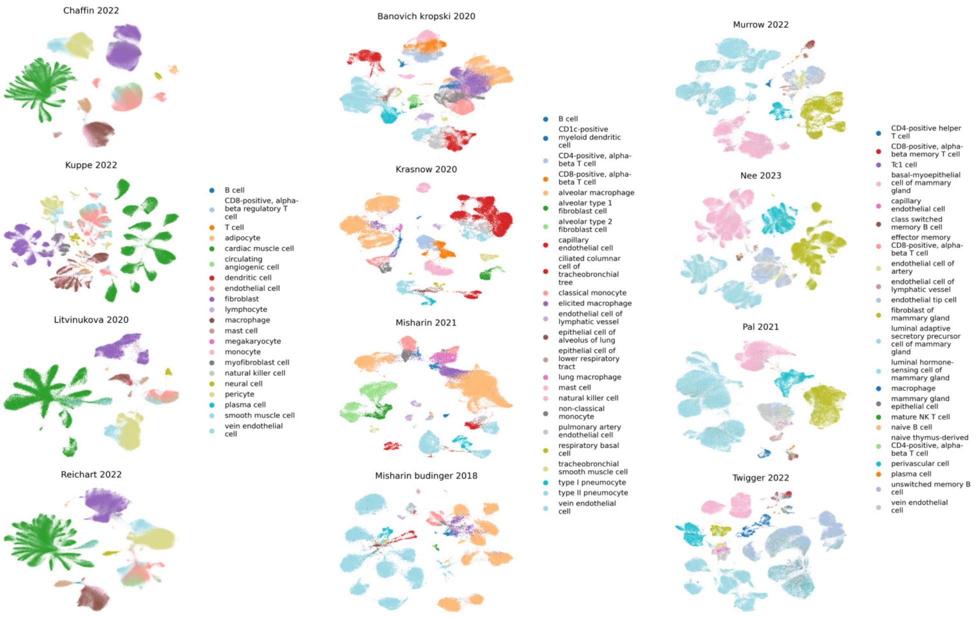
Supplementary Figure S1: UMAPs of all datasets used in this study colored by cell type



(a) Heart (b) Lung (c) Breast

Fig. S1: 2D Visualization of the dataset collections, colored by the manually-curated annotation of higher level ("cell types"). UMAP projections were computed independently for each study.

Supplementary Figure S2: UMAP of every dataset colored by cell subtype



(a) Heart (b) Lung (c) Breast

Fig. S2: UMAP representation of the cell types in the heart, lung, and breast collections showing the cell subtypes.

Supplementary Figure S3: Comparison of MLP and XGBoost classifiers

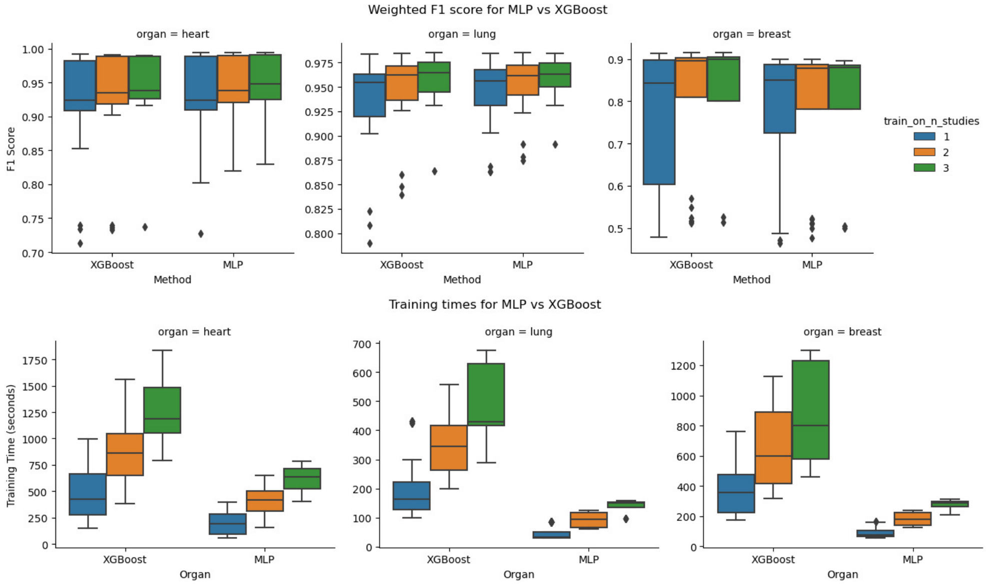


Fig. S3: Comparison of MLP and XGBoost classifiers in terms of weighted F1 score (top) and training time (bottom) at the cell type level in local learning.

Supplementary Figure S4: Weighted, macro, and micro F1 scores for local and swarm learning.

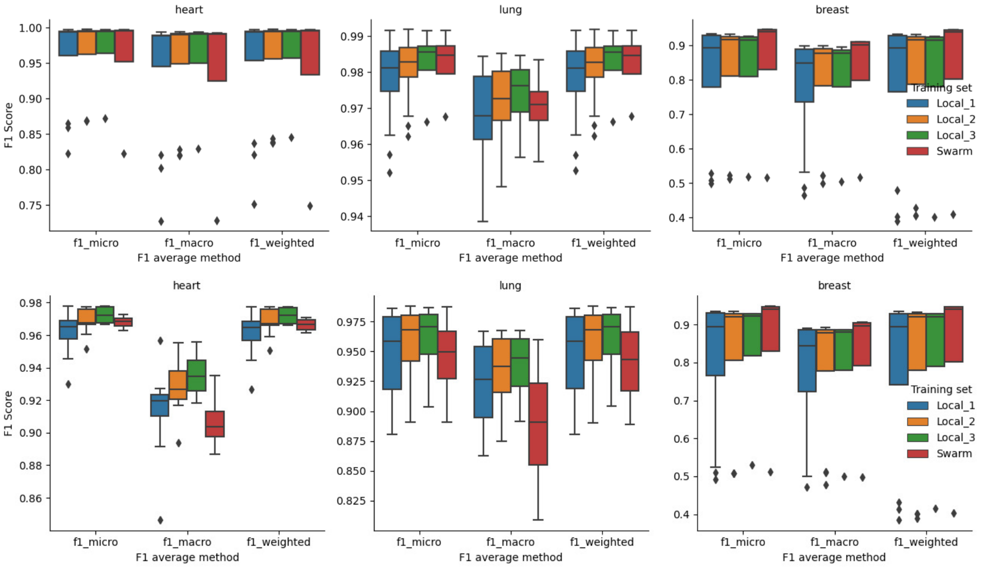


Fig. S4: Micro, macro, and weighted F1 scores for cell types (top) and cell subtypes (bottom) for LL and SL. Significance levels are not shown as no Mann-Whitney *U* tests are significant.

Supplementary Figure S5: Sankey plots between cell types and subtypes

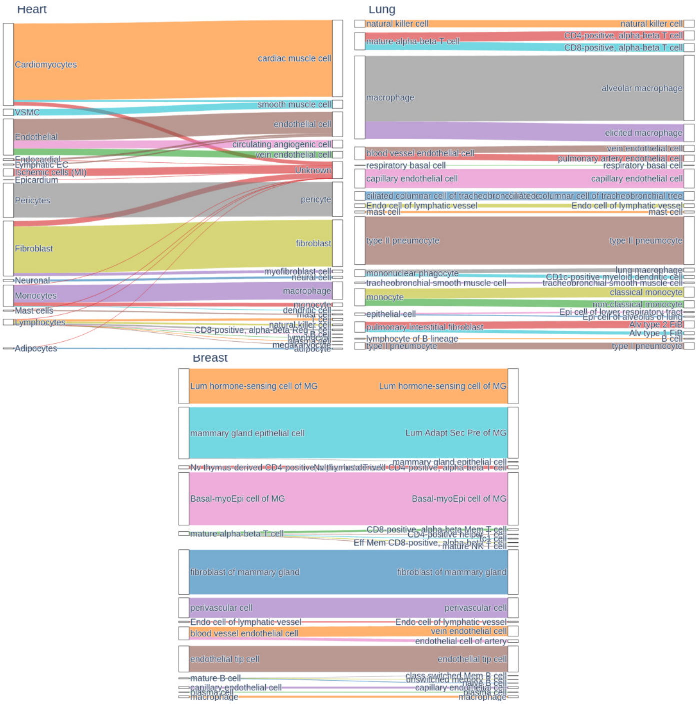


Fig. S5: Correspondence between cell types and subtypes for each organ.

Supplementary Figure S6: Confusion matrices for classifying cell subtypes

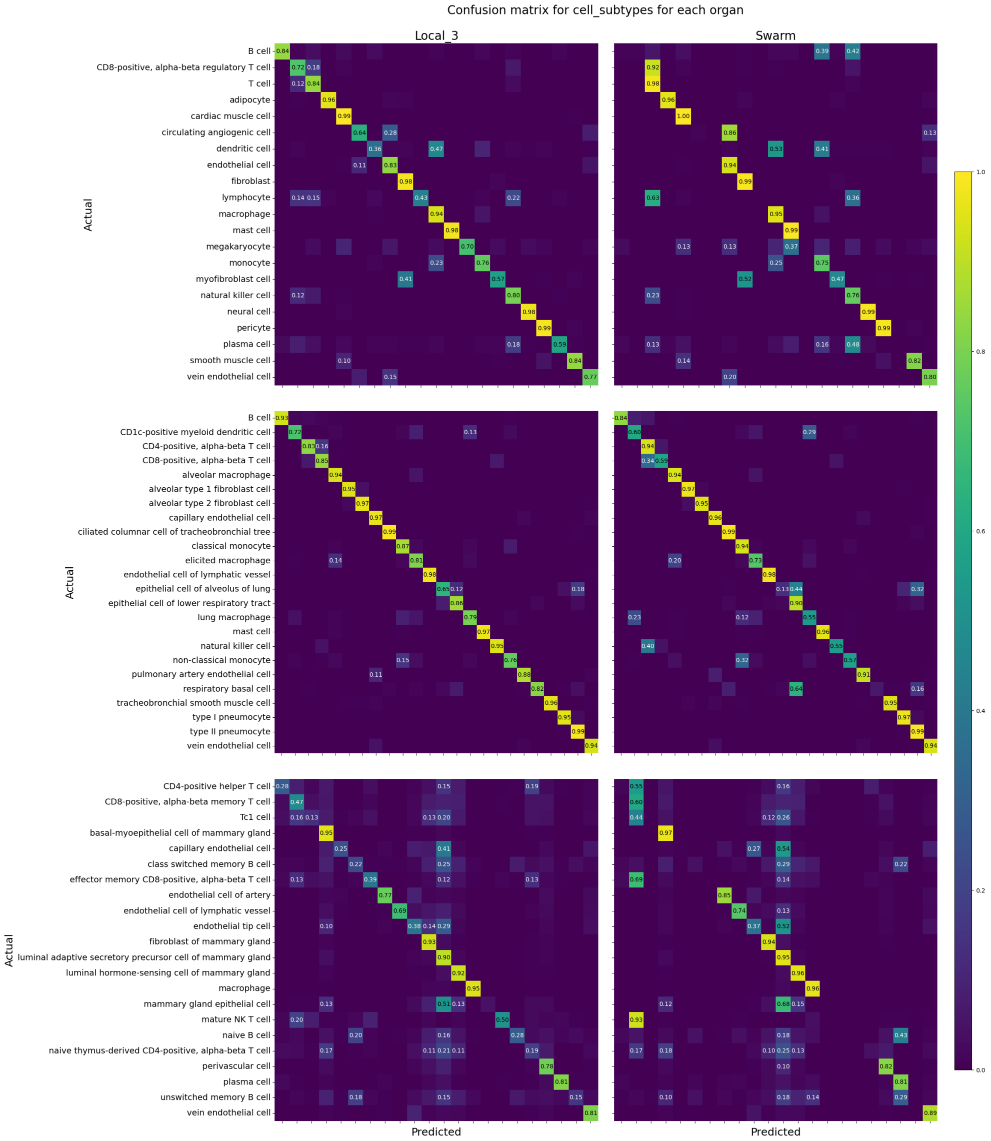


Fig. S6: Confusion matrices of Local_3 (left) versus Swarm (right) at cell subtype level for the heart (top), lung (center), and breast (bottom) datasets. The accuracies are averaged over all simulation runs and are normalized by row.

Supplementary Figure S7: Dendrograms of cell labels

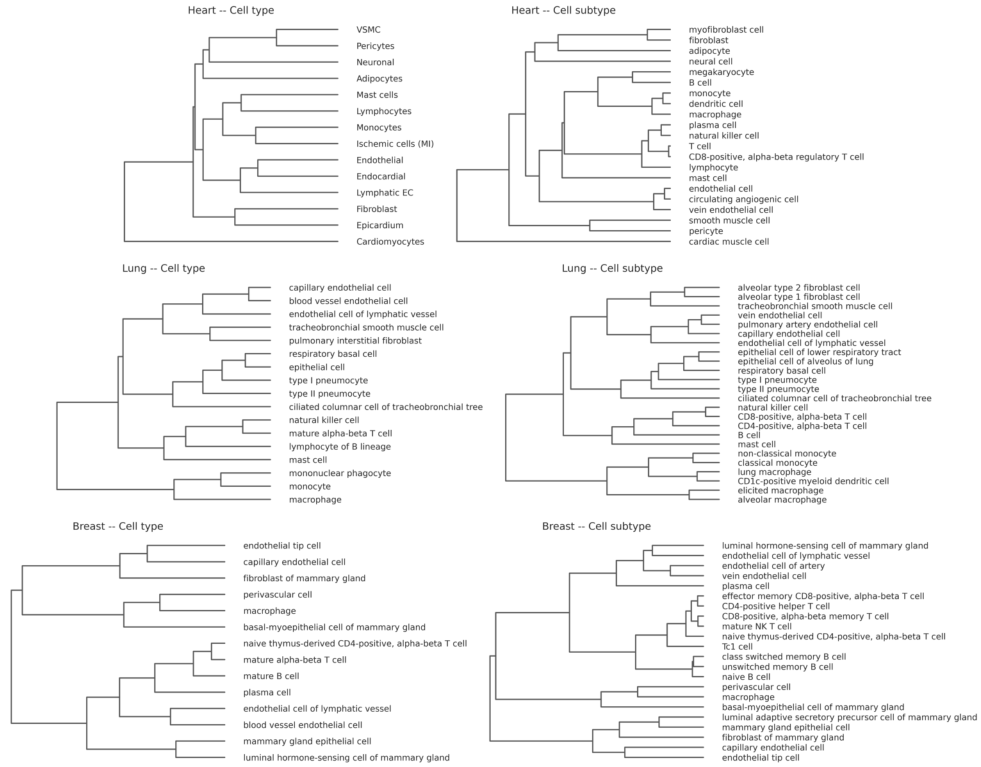

Heart -- Cell type
VSMC
Pericytes
Neuronal
Adipocytes
Mast cells
Lymphocytes
Monocytes
Ischemic cells (MI)
Endothelial
Endocardial
Lymphatic EC
Fibroblast
Epicardium
Cardiomyocytes
Heart -- Cell subtype
myofibroblast cell
fibroblast
adipocyte
neural cell
megakaryocyte
B cell
monocyte
dendritic cell
macrophage
plasma cell
natural killer cell
T cell
CD8-positive, alpha-beta regulatory T cell
lymphocyte
mast cell
endothelial cell
circulating angiogenic cell
vein endothelial cell
smooth muscle cell
pericyte
cardiac muscle cell
Lung -- Cell type
capillary endothelial cell
blood vessel endothelial cell
endothelial cell of lymphatic vessel
tracheobronchial smooth muscle cell
pulmonary interstitial fibroblast
respiratory basal cell
epithelial cell
type I pneumocyte
type II pneumocyte
ciliated columnar cell of tracheobronchial tree
natural killer cell
mature alpha-beta T cell
lymphocyte of B lineage
mast cell
mononuclear phagocyte
monocyte
macrophage
Lung -- Cell subtype
alveolar type 2 fibroblast cell
alveolar type 1 fibroblast cell
tracheobronchial smooth muscle cell
vein endothelial cell
pulmonary artery endothelial cell
capillary endothelial cell
endothelial cell of lymphatic vessel
epithelial cell of lower respiratory tract
epithelial cell of alveolus of lung
respiratory basal cell
type I pneumocyte
type II pneumocyte
ciliated columnar cell of tracheobronchial tree
natural killer cell
CD8-positive, alpha-beta T cell
CD4-positive, alpha-beta T cell
B cell
mast cell
non-classical monocyte
classical monocyte
lung macrophage
CD1c-positive myeloid dendritic cell
elicited macrophage
alveolar macrophage
Breast -- Cell type
endothelial tip cell
capillary endothelial cell
fibroblast of mammary gland
perivascular cell
macrophage
basal-myoepithelial cell of mammary gland
naive thymus-derived CD4-positive, alpha-beta T cell
mature alpha-beta T cell
mature B cell
plasma cell
endothelial cell of lymphatic vessel
blood vessel endothelial cell
mammary gland epithelial cell
luminal hormone-sensing cell of mammary gland
Breast -- Cell subtype
luminal hormone-sensing cell of mammary gland
endothelial cell of lymphatic vessel
endothelial cell of artery
vein endothelial cell
plasma cell
effector memory CD8-positive, alpha-beta T cell
CD4-positive helper T cell
CD8-positive, alpha-beta memory T cell
mature NK T cell
naive thymus-derived CD4-positive, alpha-beta T cell
Tc1 cell
class switched memory B cell
unswitched memory B cell
naive B cell
perivascular cell
macrophage
basal-myoepithelial cell of mammary gland
luminal adaptive secretory precursor cell of mammary gland
mammary gland epithelial cell
fibroblast of mammary gland
capillary endothelial cell
endothelial tip cell

Fig. S7: Ontology of cell types and subtypes obtained by hierarchical clustering of collections after dataset integration.

Supplementary Figure S8: Cell type composition per dataset

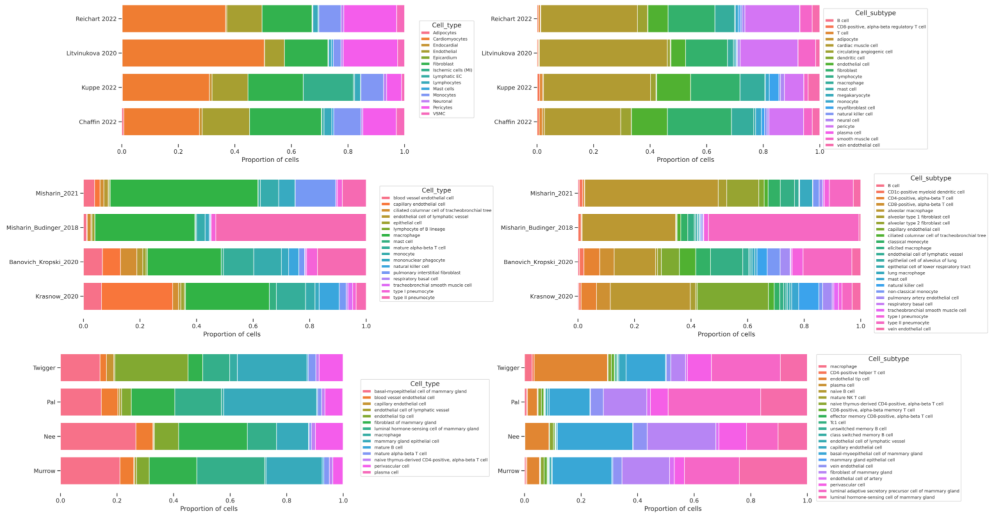
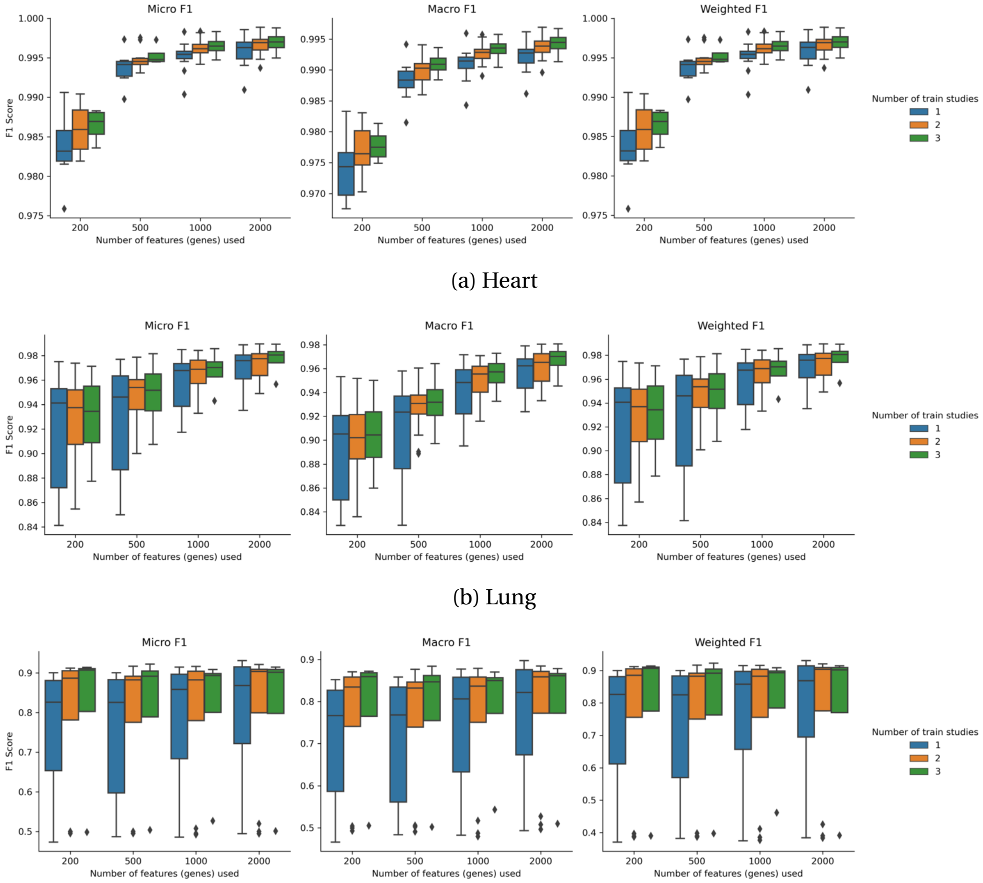


Fig. S8: Cell composition of all datasets, at both type and subtype levels.

Supplementary Figure S9: Effect of the number of HVGs



(a) Heart

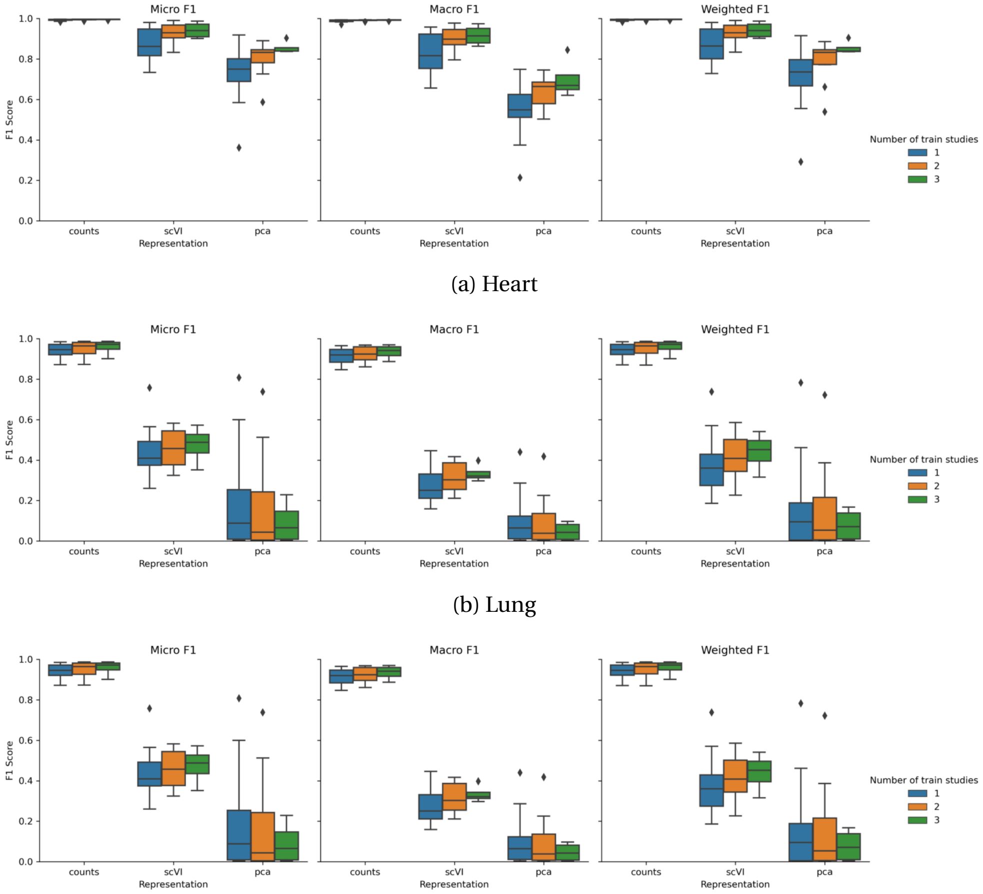
(b) Lung

(c) Breast

Fig. S9: Classification performance for various numbers of HVGs selected at preprocessing.

Supplementary Figure S10: Effect of using a low-dimensional embedding

Our classifier learns cell type compositions in the presence of batch effects between datasets without needing to account for these batch effects. This is possible because the feature space is the (normalized) counts. If a low-dimensional embedding were used instead, the classifier could not recover batch-effect agnostic decision boundaries. This is illustrated by the classification performance of the MLP classifier trained on (i) the normalized counts, (ii) the PCA of the counts using 50 PCs, and (iii) the embedding provided by scVI (Figure S10). The classification performance is significantly lower when using the low-dimensional embeddings, across organs and F1 scores average methods, with an even lower performance for the scVI embedding. This suggests that in the presence of batch effects, the low-dimensional embeddings do not capture the relevant information for cell type classification, and that the classifier is not able to recover the cell type composition in the presence of batch effects when using these embeddings.



(a) Heart

(b) Lung

(c) Breast

Fig. S10: Classification performance using different representations of the data.

Supplementary Figure S11: Local learning F1 scores per cell subtype

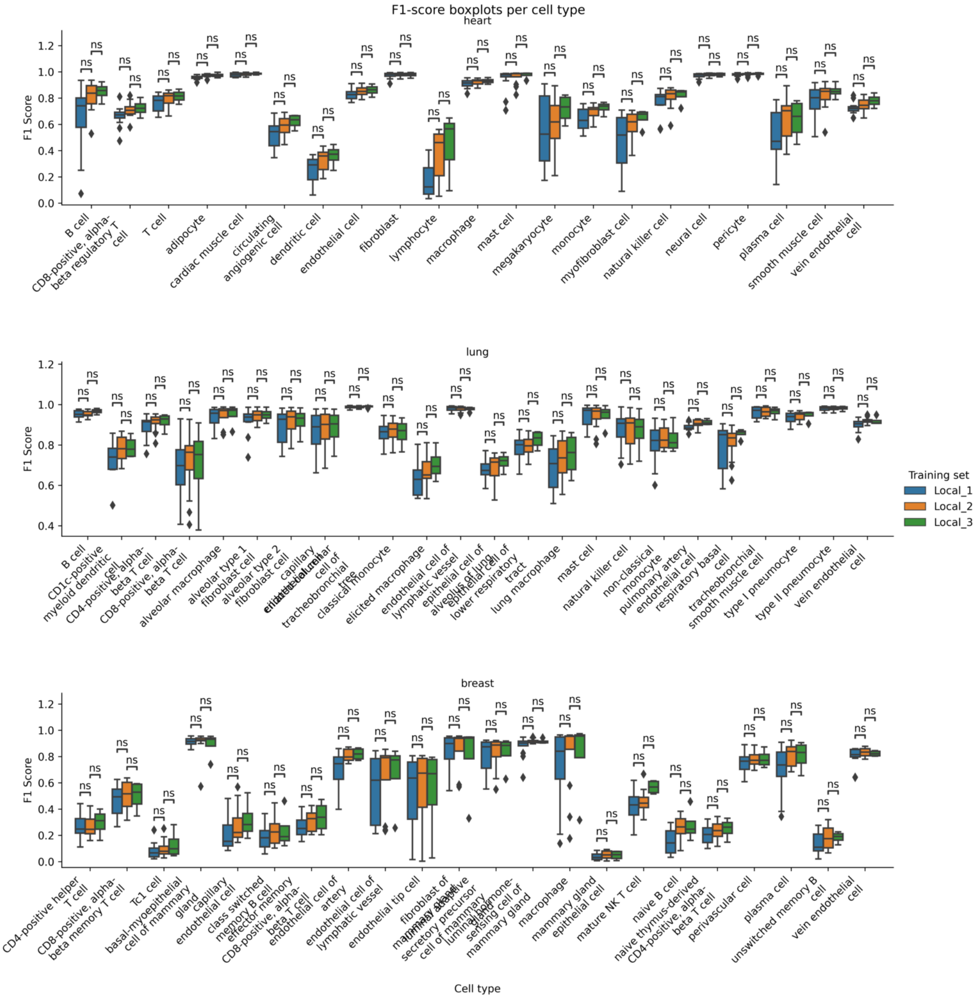


Fig. S11: F1 score for each cell subtype with LL when training on an increasing number of studies.

Supplementary Table 1: Composition of datasets by cell and cell types

Collection	Organ	Study	Number of cells	Number of cell types	Number of cell subtypes
1	Heart	Chaffin 2022 [40]	560696 (542216)	14	21
		Kuppe 2022 [6]	189349 (146325)	13	21
		Litvinukova 2020 [41]	84953 (79916)	14	21
		Reichart 2022 [42]	361649 (349045)	14	21
2	Lung [4]	Banovich Kropski 2020 [43]	115249	17	24
		Krasnow 2020 [44]	39901	17	24
		Misharin 2021 [45]	47301	17	24
		Misharin Budinger 2018 [46]	40580	17	24
3	Breast [9]	Murrow 2022 [47]	80726	14	22
		Nee 2023 [48]	219239	14	22
		Pal 2021 [49]	116432	14	22
		Twigger 2022 [50]	100168	14	22

Table 1: Datasets used for training and testing SwarmMAP. Each data collection concerns one organ and is composed of 4 datasets, called "studies". For heart, some cells had unknown subtype labels and were filtered out when using subtypes. The corresponding sample sizes are provided in parentheses.

Supplementary Table S12: Detailed composition of datasets

Cell Type	Chaffin 2022	Kuppe 2022	Litvinukova 2020	Reichart 2022
Adipocytes	3914	457	150	599
Cardiomyocytes	149933	58279	42695	132252
Endocardial	5688	1686	46	1390
Endothelial	93673	24064	5994	45095
Epicardium	217	0	9	61
Fibroblast	142441	37012	13030	63816
Ischemic cells (MI)	531	33620	113	839
Lymphatic EC	4584	794	62	321
Lymphocytes	15361	3980	847	6567
Mast cells	3925	157	582	1113
Monocytes	53995	15244	2316	28295
Neuronal	3656	2061	688	3585
Pericytes	66530	9961	16334	67622
VSMC	16248	2034	2087	10094
Total	560696	189349	84953	361649

(a) Heart Data Cell Types

Cell Subtype	Chaffin 2022	Kuppe 2022	Litvinukova 2020	Reichart 2022
B cell	184	132	9	210
CD8-positive, alpha-beta regulatory T cell	3978	1269	277	1485
T cell	6138	1481	235	2250
adipocyte	3903	457	150	599
cardiac muscle cell	146191	55383	36159	119486
circulating angiogenic cell	20365	2840	795	12474
dendritic cell	231	287	218	345
endothelial cell	69080	17640	4156	24899
fibroblast	122773	25517	11814	57790
lymphocyte	19	36	2	42
macrophage	42596	12874	1699	24324
mast cell	3683	147	436	907
megakaryocyte	5	36	10	20
monocyte	11165	2083	399	3624
myofibroblast cell	5876	4952	120	2399
natural killer cell	4761	827	248	2162
neural cell	3656	2061	688	3585
pericyte	66530	9961	16334	67622
plasma cell	94	96	19	300
smooth muscle cell	16512	2203	4997	15108
vein endothelial cell	14476	6043	1151	9414
Total	542216	146325	79916	349045

(b) Heart Data Cell Subtypes

Cell Type	Banovich Kropski 2020	Krasnow 2020	Misharin 2021	Misharin Budinger 2018
blood vessel endothelial cell	7729	2546	1862	458
capillary endothelial cell	7368	10041	903	83
ciliated columnar cell of tracheobronchial tree	6674	821	646	533
endothelial cell of lymphatic vessel	2395	323	721	161
epithelial cell	1335	529	145	391
lymphocyte of B lineage	618	135	204	29
macrophage	30004	11829	24704	14342
mast cell	913	986	46	70
mature alpha-beta T cell	12347	4132	345	157
monocyte	11277	1335	3084	1224
mononuclear phagocyte	2951	577	2732	619
natural killer cell	4179	2845	29	27
pulmonary interstitial fibroblast	2343	979	6787	144
respiratory basal cell	262	245	11	50
tracheobronchial smooth muscle cell	534	670	240	29
type I pneumocyte	4435	480	831	675
type II pneumocyte	19885	1428	4011	21588
Total	115249	39901	47301	40580

(c) Lung Data Cell Types

Cell Subtype	Banovich Kropski 2020	Krasnow 2020	Misharin 2021	Misharin Budinger 2018
B cell	618	135	204	29
CD1c-positive myeloid dendritic cell	1809	302	591	410
CD4-positive, alpha-beta T cell	6455	2184	273	120
CD8-positive, alpha-beta T cell	5892	1948	72	37
alveolar macrophage	16841	11279	22322	13403
alveolar type 1 fibroblast cell	826	766	1424	121
alveolar type 2 fibroblast cell	1517	213	5363	23
capillary endothelial cell	7368	10041	903	83
ciliated columnar cell of tracheobronchial tree	6674	821	646	533
classical monocyte	6026	788	2038	992
elicited macrophage	13163	550	2382	939
endothelial cell of lymphatic vessel	2395	323	721	161
epithelial cell of alveolus of lung	554	219	54	265
epithelial cell of lower respiratory tract	781	310	91	126
lung macrophage	1142	275	2141	209
mast cell	913	986	46	70
natural killer cell	4179	2845	29	27
non-classical monocyte	5251	547	1046	232
pulmonary artery endothelial cell	4103	1328	622	198
respiratory basal cell	262	245	11	50
tracheobronchial smooth muscle cell	534	670	240	29
type I pneumocyte	4435	480	831	675
type II pneumocyte	19885	1428	4011	21588
vein endothelial cell	3626	1218	1240	260
Total	115249	39901	47301	40580

(d) Lung Data Cell Subtypes

Cell Type	Murrow 2022	Nee 2023	Pal 2021	Twigger 2022
basal-myoepithelial cell of mammary gland	16967	58449	16793	13993
blood vessel endothelial cell	3961	13218	6816	2284
capillary endothelial cell	167	350	702	2603
endothelial cell of lymphatic vessel	620	867	822	360
endothelial tip cell	3617	18699	4039	25944
fibroblast of mammary gland	13592	53260	17882	5212
luminal hormone-sensing cell of mammary gland	19397	22624	19187	9551
macrophage	376	23	930	2696
mammary gland epithelial cell	16013	24790	38307	24660
mature B cell	464	475	379	438
mature alpha-beta T cell	1563	1736	1641	2717
naive thymus-derived CD4-positive, alpha-beta T cell	1015	3145	1364	1142
perivascular cell	2877	21482	7326	8485
plasma cell	97	121	244	83
Total	80726	219239	116432	100168

(e) Breast Data Cell Types

Cell Subtype	Murrow 2022	Nee 2023	Pal 2021	Twigger 2022
CD4-positive helper T cell	298	255	287	720
CD8-positive, alpha-beta memory T cell	848	1097	965	1063
Tc1 cell	201	247	285	641
basal-myoepithelial cell of mammary gland	16967	58449	16793	13993
capillary endothelial cell	167	350	702	2603
class switched memory B cell	211	228	257	285
effector memory CD8-positive, alpha-beta T cell	177	90	80	242
endothelial cell of artery	1341	2294	1466	672
endothelial cell of lymphatic vessel	620	867	822	360
endothelial tip cell	3617	18699	4039	25944
fibroblast of mammary gland	13592	53260	17882	5212
luminal adaptive secretory precursor cell of mammary gland	15692	24460	38109	24440
luminal hormone-sensing cell of mammary gland	19397	22624	19187	9551
macrophage	376	23	930	2696
mammary gland epithelial cell	321	330	198	220
mature NK T cell	39	47	24	51
naive B cell	149	159	50	98
naive thymus-derived CD4-positive, alpha-beta T cell	1015	3145	1364	1142
perivascular cell	2877	21482	7326	8485
plasma cell	97	121	244	83
unswitched memory B cell	104	88	72	55
vein endothelial cell	2620	10924	5350	1612
Total	80726	219239	116432	100168

(f) Breast Data Cell Subtypes

Fig. S12: Number of cell for each cell type and subtypes, for each organ.

Supplementary Table 2: Number of donors in each dataset

Organ	Study	Number of donors
Heart	Chaffin 2022	42
	Kuppe 2022	20
	Litvinukova 2020	14
	Reichart 2022	68
Lung	Banovich Kropski 2020	38
	Krasnow 2020	3
	Misharin 2021	2
	Misharin Budinger 2018	8
Breast	Murrow 2022	28
	Nee 2023	22
	Pal 2021	21
	Twigger 2022	18
Total		284

Table 2: Number of donors for each organ and study.

Supplementary Table 3: Data download links

Collection	Organ	Dataset	Download link
1	Heart	Chaffin 2022	https://singlecell.broadinstitute.org/single_cell/study/SCP1303/
		Kuppe 2022	https://cellxgene.cziscience.com/collections/8191c283-0816-424b-9b61-c3e1d6258a77
		Litvinukova 2020	https://cellxgene.cziscience.com/collections/b52eb423-5d0d-4645-b217-e1c6d38b2e72v
		Reichart 2022	https://cellxgene.cziscience.com/collections/b52eb423-5d0d-4645-b217-e1c6d38b2e72e75342a8-0f3b-4ec5-8ee1-245a23e0f7cb
2	Lung	Banovich Kropski 2020 Krasnow 2020 Misharin 2021 Misharin and Budinger 2018	https://cellxgene.cziscience.com/collections/b52eb423-5d0d-4645-b217-e1c6d38b2e726f6d381a-7701-4781-935c-db10d30de293
3	Breast	Murrow 2022 Nee 2023 Pal 2021 Twigger 2022	https://cellxgene.cziscience.com/collections/b52eb423-5d0d-4645-b217-e1c6d38b2e7248259aa8-f168-4bf5-b797-af8e88da6637

Table 3: List of human heart, lung, and breast datasets used in this study. Download links of all datasets for CellxGene platform are provided.

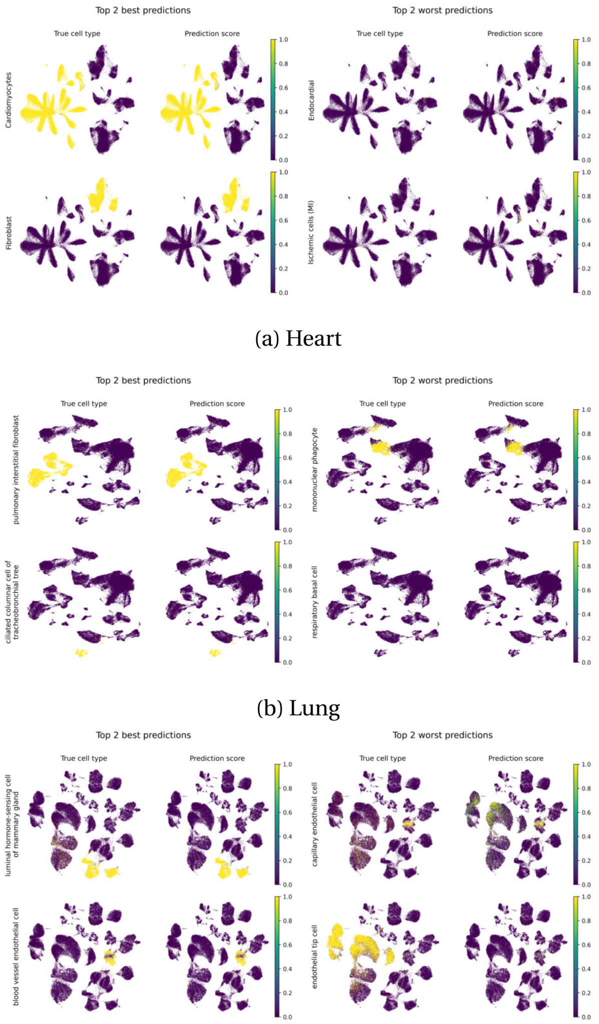
Supplementary Table 4: Averages of weighted F1 scores for LL and SL.

Organ	Label	Training set	Weighted F1 score (Confidence Interval)
Heart	Cell type	Local_1	0.947 (0.896, 0.997)
		Local_2	0.957 (0.917, 0.996)
		Local_3	**0.958** (0.884, 1.032)
		Swarm	0.934 (0.813, 1.056)
Heart	Cell subtype	Local_1	0.961 (0.953, 0.968)
		Local_2	0.968 (0.964, 0.973)
		Local_3	**0.972** (0.966, 0.978)
		Swarm	0.966 (0.962, 0.971)
Lung	Cell type	Local_1	0.978 (0.97, 0.985)
		Local_2	0.981 (0.975, 0.987)
		Local_3	**0.982** (0.971, 0.993)
		Swarm	**0.982** (0.972, 0.992)
Lung	Cell subtype	Local_1	0.945 (0.922, 0.969)
		Local_2	0.954 (0.933, 0.975)
		Local_3	**0.958** (0.921, 0.995)
		Swarm	0.941 (0.899, 0.982)
Breast	Cell type	Local_1	0.786 (0.661, 0.91)
		Local_2	0.794 (0.664, 0.924)
		Local_3	0.79 (0.536, 1.044)
		Swarm	**0.809** (0.547, 1.07)
Breast	Cell subtype	Local_1	0.78 (0.653, 0.908)
		Local_2	0.791 (0.657, 0.926)
		Local_3	0.796 (0.547, 1.046)
		Swarm	**0.808** (0.544, 1.072)

Table 4: Averages of weighted F1 scores for LL and SL across all settings. Confidence intervals are computed using the t-distribution. For each setting, the highest value is highlighted in bold.

Supplementary Figure S13: Visualizing classification prediction score

Both local and Swarm classifiers display heterogeneous classification performance across cell types. To understand the reasons thereof, Figure S13 compares the true label with its prediction score. The results are represented for the two best classified and two worst classified cell types, for every organ. The values are represented in the UMAP of the test study. On one experiment experiment from Local_3 is considered for each organ and the test studies used are “Litvinukova 2020”, “Misharin 2021”, and “Twigger 2022” for heart, lung, and breast collections, respectively.



(a) Heart

(b) Lung

(c) Breast

Fig. S13: Top 2 best (left) and worst (right) classified cell types for lung. For each row, the left pane shows the true cell types and the right pane shows the prediction score.