

<https://doi.org/10.1038/s41540-026-00667-6>

SwarmMAP: swarm learning for decentralized cell type annotation in single cell sequencing data



Oliver Lester Saldanha^{1,2,8}, Vivien Goepp^{3,8}, Kevin Pfeiffer¹, Hyojin Kim³, Jie Fu Zhu^{1,2}, Rafael Kramann³, Sikander Hayat^{3,4,5}✉ & Jakob Nikolas Kather^{1,6,7}✉

Rapid technological progress now enables large-scale generation of single-cell data. Many laboratories can produce single-cell transcriptomic profiles from diverse tissues. A key step in single-cell analysis is unsupervised clustering followed by cell-type annotation, yet there is no agreement on marker genes, and annotation is typically done manually, making it irreproducible and poorly scalable. Privacy constraints in human datasets further complicate data sharing. There is a need for standardized, automated, and privacy-preserving cell-type annotation across datasets. We developed SwarmMAP, which applies Swarm Learning to train machine-learning models for cell-type classification in a decentralized setting without exchanging raw data between centers. SwarmMAP achieves F1-scores of 0.93, 0.98, and 0.88 in heart, lung, and breast datasets, respectively. Swarm Learning models reach an average performance of 0.907, comparable to models trained on centralized data (p-val = 0.937, Mann-Whitney U Test). Increasing the number of datasets improves prediction accuracy and supports classification across broader cell-type diversity. These results show that Swarm Learning provides an effective approach for automated cell-type annotation. SwarmMAP is available at <https://github.com/hayatlab/SwarmMAP>.

Recent technological advances in single-cell sequencing have led to a plethora of scientific discoveries improving our understanding of human tissue and diseases¹, including COVID^{2,3}, lung⁴, cardiovascular^{5–7}, renal diseases⁸, 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 genes¹². 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 are critical issues while working with human patient data¹³.

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 using transfer learning^{18–20}. However, their applicability remains limited, as transfer learning methods strongly depend on large, well-annotated reference datasets and often fail to generalize when the query data differ in tissue, condition, or cell-type composition from the reference data. Hence, generalizable machine learning models need to be trained on large, multi-

¹Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Saxony, Germany. ²Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany. ³Department of Medicine 2, RWTH Aachen University, Medical Faculty, Aachen, North Rhine-Westphalia, Germany. ⁴Cardiovascular Research Institute, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ⁵Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ⁶Department of Medicine I, Faculty of Medicine and University Hospital Carl Gustav Carus, Technical University Dresden, Dresden, Saxony, Germany. ⁷Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Baden-Wuerttemberg, Germany. ⁸These authors contributed equally: Oliver Lester Saldanha, Vivien Goepp. ✉e-mail: sikander.hayat@mssm.edu; jakob_nikolas.kather@tu-dresden.de

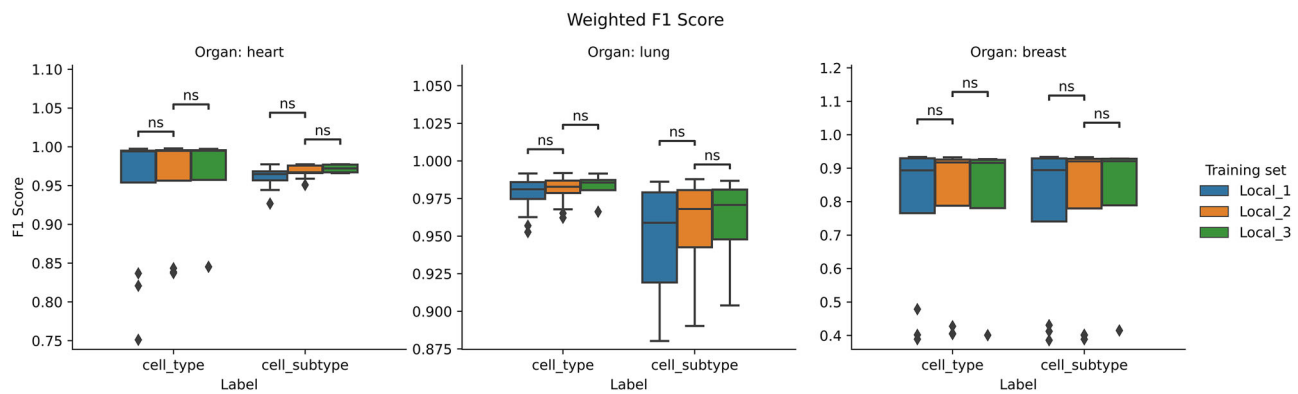


Fig. 1 | Classification performance of the local learning framework. Weighted F1 scores are averaged over all simulation runs. F1 score distributions for classifying main cell types and subtypes using Local_1 (training on 1 dataset), Local_2 (training on 2 datasets), and Local_3 (training on 3 datasets)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

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 medicine²¹.

Swarm learning is a computational technique to co-train machine learning models at multiple institutions in a decentralized way, without exchanging underlying data²². Swarm learning does not require a central coordinator of the network and thus avoids monopolization of resources and machine learning models²³. 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. Importantly, SwarmMAP achieves robust integration without explicit batch correction, leveraging the diversity and breadth of datasets distributed across peers.

Results

Machine learning-based cell type prediction in multiple datasets

The average weighted F1 score for main cell-type classification in heart datasets is 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 (Fig. 1). Similar values are obtained from the lung and breast datasets (Fig. 1). 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 Fig. 4.

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

“Cell type annotation improves with cell count”), making the distribution of F1 scores more dispersed. This is especially true for breast data, in which classification is harder (see section “Annotation is challenging for rare cell subtypes”). 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).

Classification performances vary greatly between cell-types

Model performance (F1-score) per cell type varies between different cell types (Fig. 2). The corresponding results for cell subtypes are represented in Supplementary Fig. 11. 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 Supplementary Fig. 8), and the difficulty to define ischemic cells biologically (see UMAP representation in Supplementary Fig. 1). In addition, ischemic cells are 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 datasets, all cell types are well classified, except “respiratory basal cells”, which are classified as “epithelial cells”. For the breast datasets, classification is more challenging, with four cell types with an 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 “Annotation is challenging for rare subtypes”).

Cell type annotation improves with cell count

To investigate whether 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 3 represents the F1 score as a function of the Gini impurity index of cell types. The results show that the classification performance decreases as the rarity of the cell types increases.

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

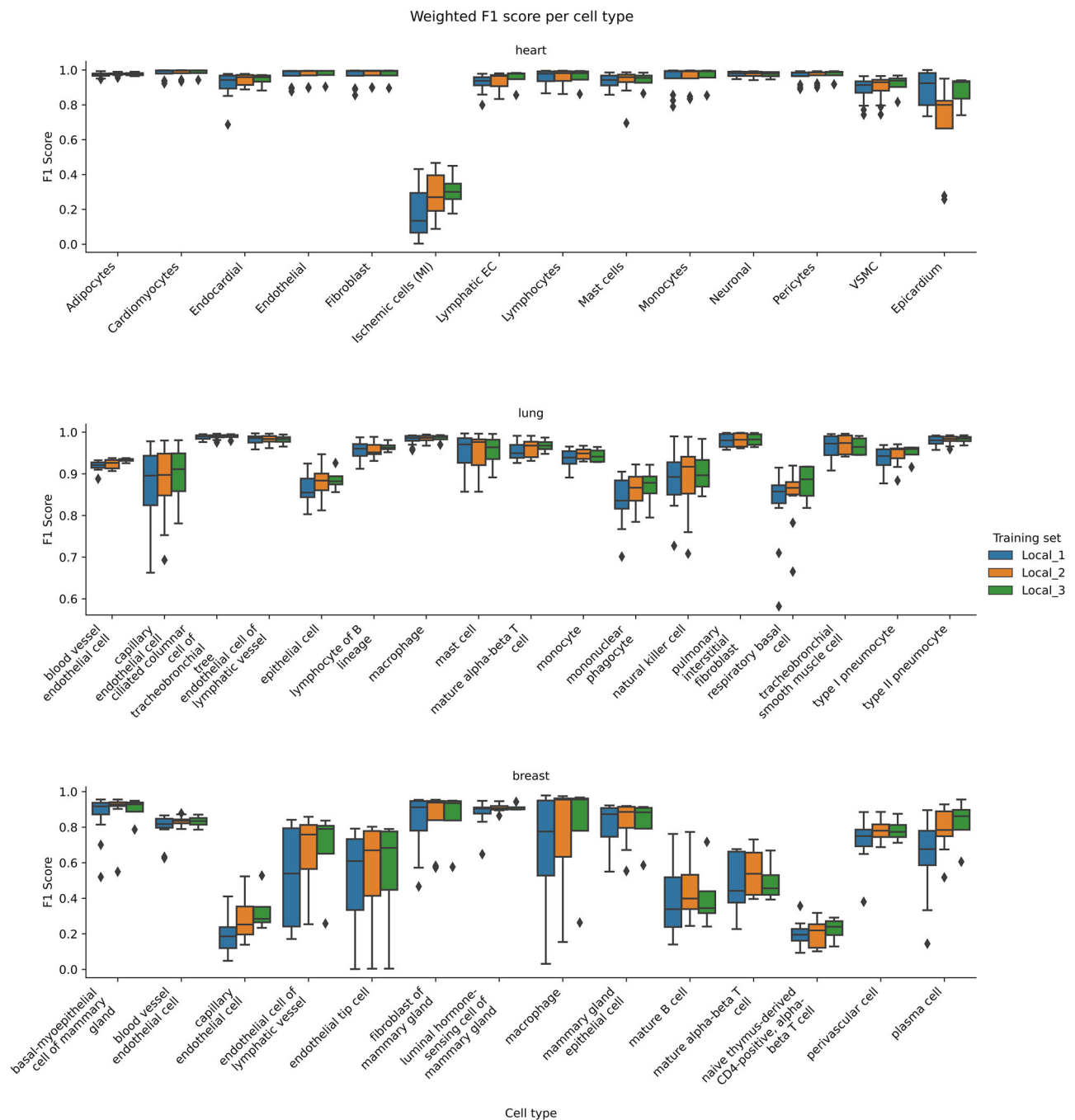


Fig. 2 | Classification performance of the local learning framework. F1 scores are averaged over all simulation runs for each cell type. Overall, there is a notable disparity in classification accuracy between cell types.

statistically confirms that the classification improves for cell types that are more common.

Swarm learning performance across cell types

The SL classifier is compared to the LL classifier trained on 3 studies (“Local_3”). Figure 4 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.

Swarm learning performance per cell type

The classification accuracy shows largely high accuracy for most cell types across the three organs, as shown in the normalized confusion matrices (Fig. 5 and Supplementary Fig. 6). 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) hinder efficient classification, which is reflected in lower SL performance, while “Ischemic cells” and “Lymphatic ECs” are difficult to differentiate using

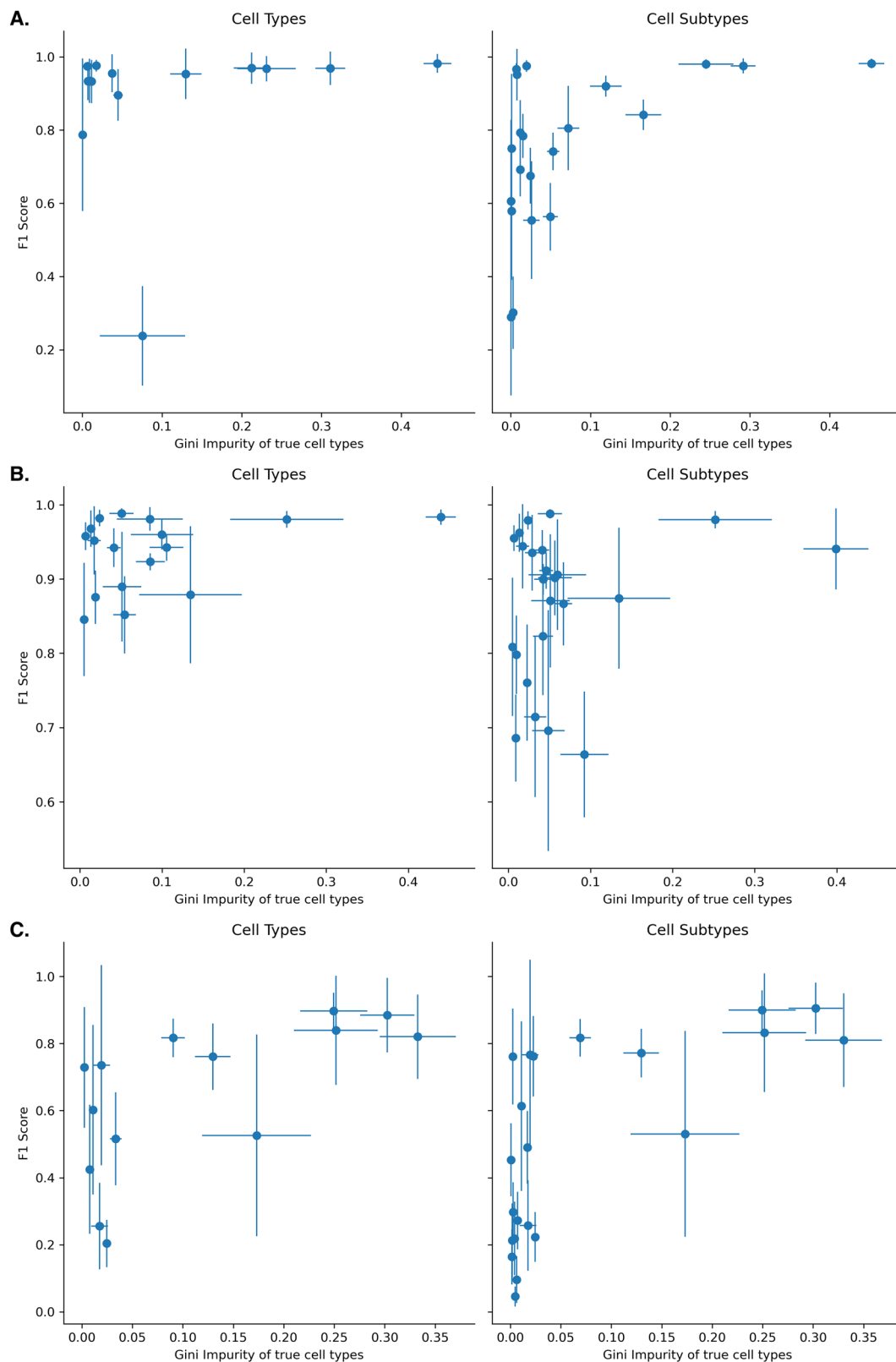


Fig. 3 | Classification performance as a function of the rarity of the cell types. **A** Heart, **B** lung, and **C** breast. The x-axis represents the Gini impurity index of true cell types (higher values are rarer classes). The y-axis represents the F1 score. Each

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.

expression profiles. For the 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-myoeptithelial 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 cells: mature B cells, and two T cell types (mature alpha-beta T cells and naive thymus-derived CD4-positive, alpha-beta T cells). The very specific cell type, naive thymus-derived CD4-positive, alpha-beta T cell, has predictions scattered over 5 cell types. In the LL setting, endothelial tip cells are predicted as mammary gland epithelial cells. 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 Fig. 12.

Annotation is challenging for rare subtypes

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.

Table 1 | Parameters of the linear regression models fitting the F1 score as a function of cell type rarity

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

Significant p-values are in bold.

For the heart, in the LL setting, the only cell subtype that is misclassified in the relatively 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 (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 killer cells, and myofibroblast cells as fibroblasts. In the 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 the 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-myoeptithelial 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 that 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

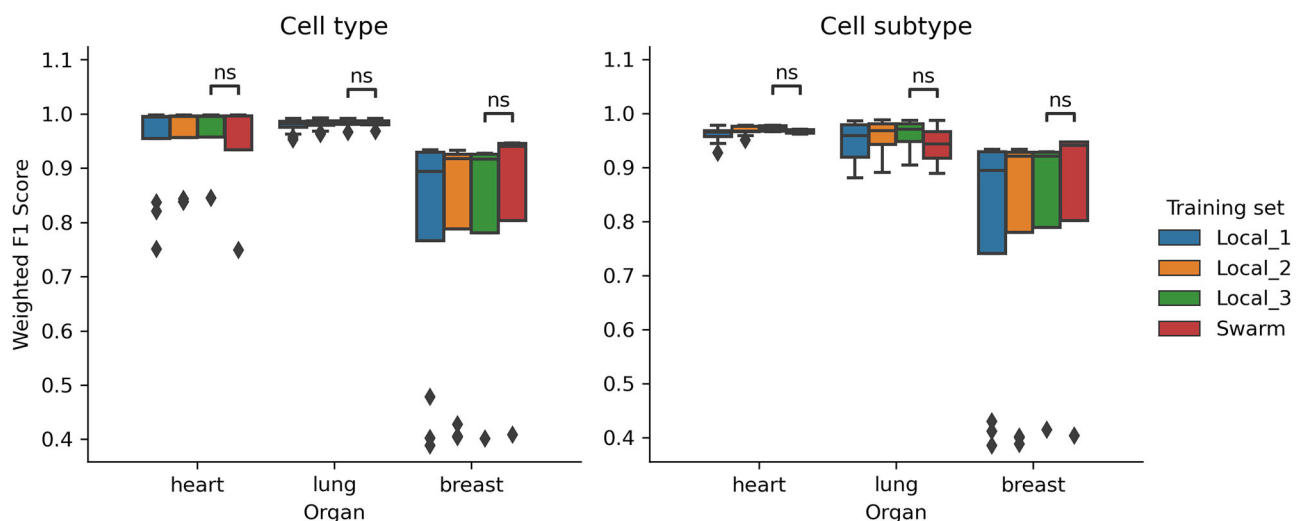


Fig. 4 | Weighted F1 score across all cell types and subtypes for LL and SL settings. Significance levels are provided by Mann–Whitney tests. Overall, the performance of local and Swarm learning approaches is comparable, showing that there is no

performance loss even when training is done without sharing datasets and models in the Swarm learning setting.

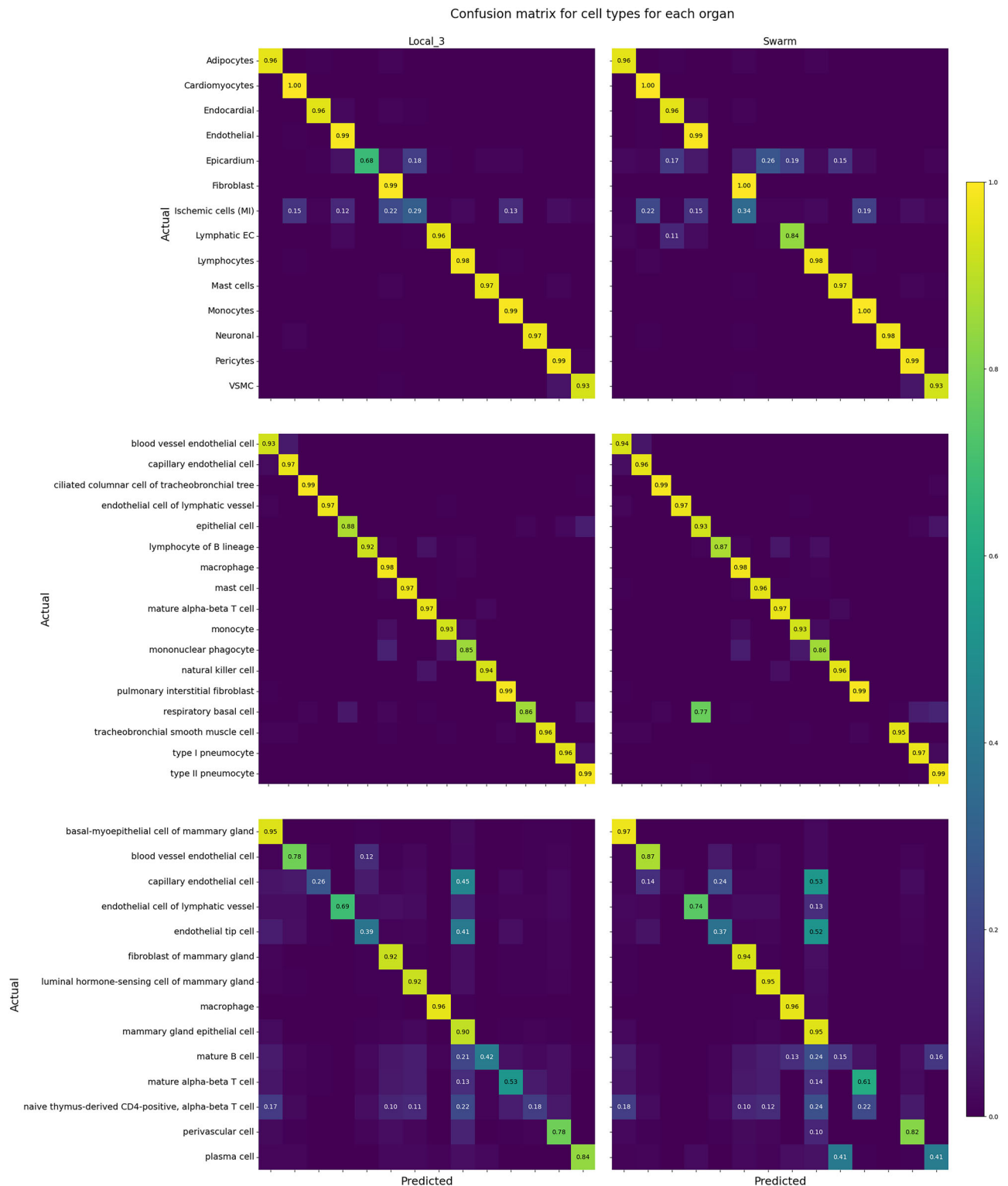


Fig. 5 | Confusion matrices for classifying cell types using Local_3 (left) versus Swarm_3 (right) classifier on heart (top), lung (center), and breast (bottom) datasets. The accuracies are averaged over all simulation runs and normalized by row.

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 that 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.

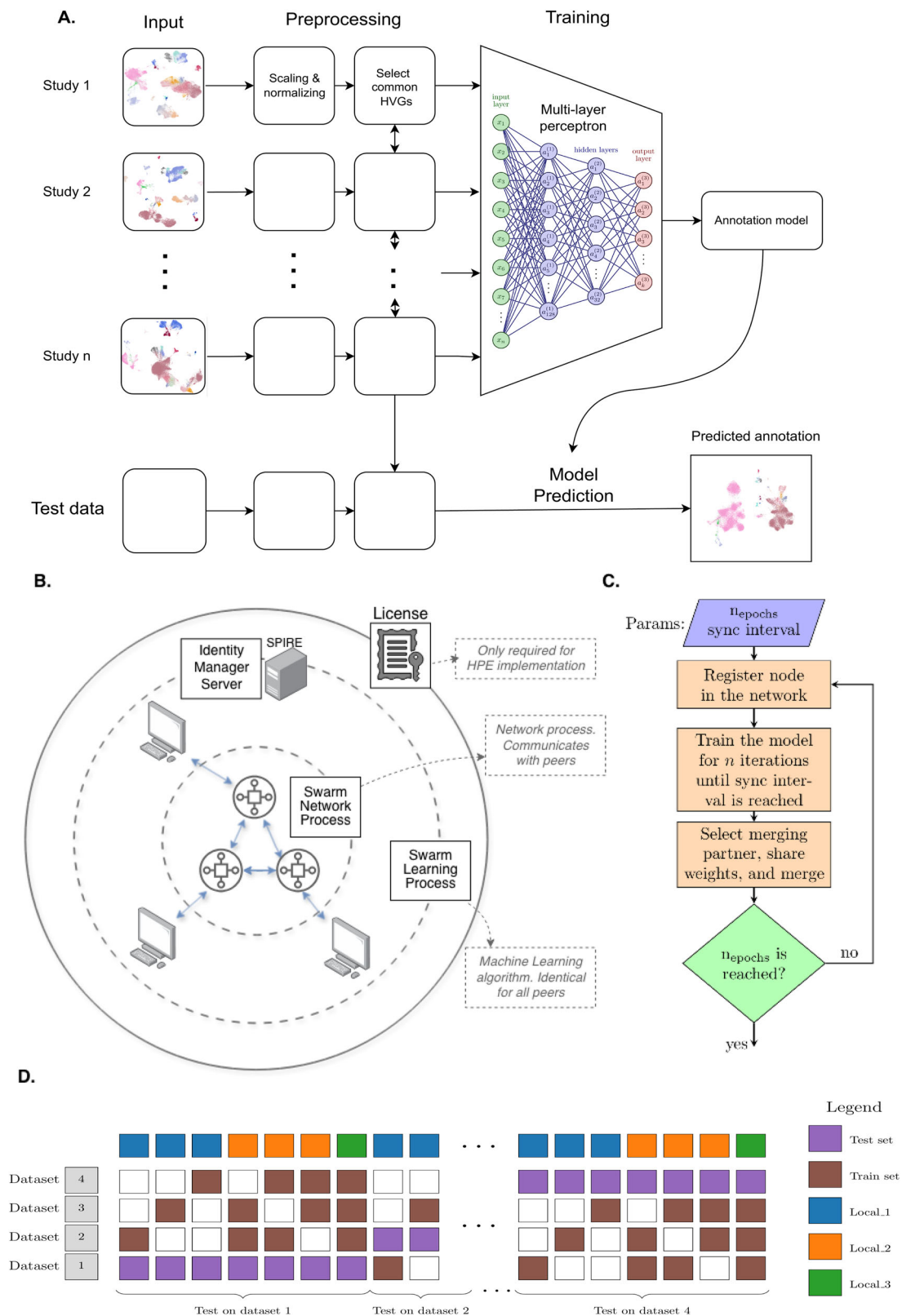


Fig. 6 | Overview of the SwarmMAP workflow. **A** Training Pipeline for our annotation classifier: scaling and normalization is done independently for each study, except for selecting the highly variable genes (HVGs), where the union of the top 2000 HVGs across studies is used; **B** Swarm learning (SL) framework, where no data is shared and decentralized sharing of models' parameters is enabled by a Swarm Network on which each peer trains its own MLP classifier; **C** Procedure of the Swarm

Network, which synchronizes training between peers; **D** Experimental design for local learning (LL): for each organ, all combinations of train and test sets are used. In the settings where 1, 2, or 3 training sets are used (Local_1, Local_2, and Local_3 resp.), a single MLP classifier is trained on the combined training sets, whereas with Swarm learning (SL), all 3 training sets are used in the SL network.

Runtime comparison

The average training times of LL and SL are presented in Supplementary Table 5. For cell types, the runtimes of Local_3 versus SL are 631 s vs 2780 s for heart, 272 s vs 1120 s for breast, and 138 s vs 610 s for lung.

The increased runtime in SL primarily reflects synchronization latency and parameter exchange between distributed sites, rather than the computational cost of the model itself, which remains low (simple MLP with tabular input). SL offers privacy-preserving, multi-institutional training without data sharing in exchange for additional communication overhead at a runtime trade-off we consider acceptable for secure and collaborative medical AI development.

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 5 and five highlight (i) the challenge in annotating cell types that are present in small numbers and (ii) the higher accuracy of annotating cell types that are well-defined and present in higher numbers.

In the heart datasets tested here, clusters annotated as ischemic cells were difficult to classify in both LL and SL settings. This is due to ischemic cells being an aggregated cluster comprising 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 genes²⁶. 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 id “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 Fig. 6). 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 of training hierarchical models for all annotation levels. Using the cell ontology as a priori information, hierarchical classifiers²⁷ 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.

Methods

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 (Fig. 6A). In LL, a single model is trained using a common training dataset while in SL, each peer keeps his data and model private (Fig. 6B), as a blockchain-enabled Swarm learning network allows each peer's model to learn from the other models (6C). In SL, neither preprocessing nor model training performs any batch correction, as each peer operates independently without access to other peers' data.

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.

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 atlas⁴ and breast atlas⁹, 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 (Supplementary Fig. 13). Training and testing are 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 1 and 2 using UMAP²⁸.

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 for 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 terms²⁹. 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 the 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 is provided in dendrograms (Supplementary Fig. 7), and the correspondence between cell types and cell subtypes is represented in a Sankey diagram (Supplementary Fig. 5).

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 (see Supplementary Fig. 10 for the difference between using HVGs and a low-dimensional embedding). 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 Fig. 9. 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). In SL, each peer shares their top 2000 HVGs, the union of these lists is computed, and all peers train their model using the union HVGs list. When using SwarmMAP, private union can be used to enable all peers to know the intersection of HVGs without sharing other peers' HVG lists. Since the genes are part of a list of known human reference gene list, private set intersection methods and De Morgan's laws can be used to implement a simple private union protocol. Consequently, sharing HVGs does not constitute a privacy breach.

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 Fig. 9 displays the F1 scores for each cell type for several numbers 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 throughout³⁰.

The final dataset consists of 1,196,647, 243,031, and 516,565 cells in heart, lung, and breast collections, respectively (with 1,117,502 cells for heart subtype, see Supplementary Fig. 13) 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.

Experimental design

The experiment design for local learning (LL) is detailed in Fig. 6D. 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 experiment designs with three training sets are used, using all combinations of the test set.

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 activation functions ReLU and LeakyReLU;
- 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.

Swarm learning

Swarm learning (SL) allows decentralized collaborative training of machine learning (ML) models on multiple physically distinct computing systems (peers)²². 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 the 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^{23,24}. 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.

Swarm learning communications protocol

In our implementation, all participating peers were connected via the HPE Swarm Learning community edition, which employs a private, permissioned blockchain network to manage node authentication, parameter exchange, and synchronization. Each node operates inside a Docker-based container that communicates only encrypted model parameters, not raw data, through the smart contracts established by the blockchain layer.

During training, every node independently updates its model on local data (Fig. 6C) until the predefined sync interval ($\text{sync frequency} = 1.00$) is reached. At each synchronization event, model weights are securely shared among peers through the blockchain's consensus protocol. The designated leader node (the first to complete its batch) collects encrypted

parameter updates from at least two peers, performs weighted averaging of the model parameters, and redistributes the aggregated model to all nodes. This ensures that only statistical representations (model weights) are exchanged, with no access to intermediate activations, gradients linked to specific data samples, or raw images.

All nodes were connected via a 1 Gbit/s private network with passwordless SSH and synchronized clocks, operating under Ubuntu 20.04 within isolated containers. This setup minimizes any risk of privacy leakage by design, as the blockchain layer enforces node identity management (via SPIFFE/SPIRE certificates) and prevents unauthorized inspection or tampering with transmitted model parameters.

Our SL implementation requires the “SL community edition” by Hewlett Packard Enterprise (Spring, Texas, United States), which is publicly available under an Apache 2.0 license along with detailed instructions and troubleshooting at <https://github.com/HewlettPackard/swarm-learning>.

Comparison to federated learning

Employing Swarm learning for cell type classification in single-cell RNA sequencing data is a novel approach. Wang et al.³¹ introduced scFed, a federated learning framework for cell type annotation. Federated learning (FL) also aims to preserve data privacy but has structural limitations that SL addresses. Both FL and SL keep training data local to each client. The difference lies in how model parameters are handled and who controls them.

In FL, a central server collects parameters from each client, aggregates them, and redistributes an updated global model. This central actor becomes a point of concentration for model parameters, enabling single-point access to model updates and creating a target for attacks, misuse, or inference of private data. Centralization also introduces trust asymmetry and reduces fault tolerance.

In SL, no central coordinator exists. Clients interact through a decentralized, trustless blockchain network. Model updates are merged through consensus directly on the distributed ledger. No participant unilaterally receives or controls all clients’ parameters. This eliminates a privileged access point and distributes trust, responsibility, and control across all participants.

Privacy is therefore not assumed but enforced through security. The architecture prevents unilateral access to model parameters, distributes authority, and ensures tamper resistance. Security mechanisms—ledger immutability, consensus-based updates, permissioned onboarding—guarantee privacy by design.

As summarized in ref. 22, SL improves over FL by: (a) keeping data at local sites; (b) avoiding any exchange of raw data; (c) providing high-level data security; (c) enabling transparent and secure onboarding of decentralized members without a central authority; (c) merging parameters with equal rights across participants; (c) protecting machine learning models from attacks.

In FL, privacy depends on trusted central custodians handling model parameters. In SL, privacy emerges from enforced security—no central access, no privileged node, only verifiable, decentralized coordination.

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 is 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. Garnett³² uses marker genes already curated to annotate cells using a regularized multinomial linear classifier. ACTINN³³ uses an MLP classifier with 3 hidden layers (100, 50, and 25 neurons, respectively). Celltypist¹⁴ uses L2 regularized logistic regression. Supervised Contrastive Learning for Single Cell (SCLSC)³⁴ employs contrastive learning to learn an embedding representation for cell types and a KNN classifier to annotate cells. dev-CellPy is a machine learning-enabled pipeline for automated annotation of complex multilayered single-cell transcriptomic data, based on the XGBoost classifier. scTab³⁵ introduces a feature-attention-based classifier model for single-cell transcriptomic data based on TabNet, a deep learning classifier

for tabular data³⁶. The classification performance of TabNet was compared with several models, including XGBoost³⁷ and MLP.³⁵ 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 considers 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 TabNet³⁸ and is considered to be more flexible for classification tasks, 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 Methods section “Data Preprocessing”) and with the same experiment design (Local_1, Local_2, and Local_3, see Methods section “Experimental design”). MLP compares slightly favorably to XGBoost while being faster to train by a factor of 2 to 4 (see Supplementary Fig. 3). Thus, SwarmMAP is built using an MLP classifier.

Data availability

The 12 datasets are publicly available from CellxGene at <https://cellxgene.cziscience.com/datasets> and the Broad Institute Single Cell portal at https://singlecell.broadinstitute.org/single_cell. The download links are provided in Supplementary Table 3. The processed datasets are available on Zenodo at <https://doi.org/10.5281/zenodo.17747500>.

Code availability

The SwarmMAP method and the code to reproduce the results in this study are available at <https://github.com/hayatlab/SwarmMAP>. SwarmMAP relies on the “SL community edition” by Hewlett Packard Enterprise (Spring, Texas, United States), which is publicly available under an Apache 2.0 license along with detailed instructions and troubleshooting at <https://github.com/HewlettPackard/swarm-learning>.

Received: 31 July 2025; Accepted: 4 February 2026;

Published online: 18 February 2026

References

1. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The human cell atlas: from vision to reality. *Nature* **550**, 451–453 (2017).
2. Schreiber, F. et al. Dissecting CD8+ T cell pathology of severe SARS-CoV-2 infection by single-cell immunoprofiling. *Front. Immunol.* **13**, 1066176 (2022).
3. Stephenson, E. et al. Single-cell multi-omics analysis of the immune response in COVID-19. *Nat. Med.* **27**, 904–916 (2021).
4. Sikkema, L. et al. An integrated cell atlas of the lung in health and disease. *Nat. Med.* **29**, 1563–1577 (2023).
5. Bleckwehl, T. et al. Encompassing view of spatial and single-cell rna sequencing renews the role of the microvasculature in human atherosclerosis. *Nat. Cardiovasc. Res.* **4**, 26–44 (2024).
6. Kuppe, C. et al. Spatial multi-omic map of human myocardial infarction. *Nature* **608**, 766–777 (2022).
7. Kanemaru, K. et al. Spatially resolved multiomics of human cardiac niches. *Nature* **619**, 801–810 (2023).
8. Kuppe, C. et al. Decoding myofibroblast origins in human kidney fibrosis. *Nature* **589**, 281–286 (2021).
9. Reed, A. D. et al. A single-cell atlas enables mapping of homeostatic cellular shifts in the adult human breast. *Nat. Genet.* **56**, 652–662 (2024).
10. Khaliq, A. M. et al. Refining colorectal cancer classification and clinical stratification through a single-cell atlas. *Genome Biol.* **23**, 113 (2022).

11. Kang, J. et al. Systematic dissection of tumor-normal single-cell ecosystems across a thousand tumors of 30 cancer types. *Nat. Commun.* **15**, 4067 (2024).
12. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
13. Lohmöller, J. et al. sce (match): privacy-preserving cluster matching of single-cell data. In *Proc. IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)* (IEEE, 2024).
14. Domínguez Conde, C. et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
15. Hou, R., Denisenko, E. & Forrest, A. R. R. scMatch: a single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics* **35**, 4688–4695 (2019).
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**, 1756–1760 (2022).
17. Xu, Y., Kramann, R., McCord, R. P. & Hayat, S. MASI enables fast model-free standardization and integration of single-cell transcriptomics data. *Commun. Biol.* **6**, 1–14 (2023).
18. Kang, J. B. et al. Efficient and precise single-cell reference atlas mapping with Symphony. *Nat. Commun.* **12**, 5890 (2021).
19. Xu, C. et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
20. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
21. Walker, C. R. et al. Private information leakage from single-cell count matrices. *Cell* **187**, 6537–6549 (2024).
22. Warnat-Herresthal, S. et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature* **594**, 265–270 (2021).
23. Saldanha, O. L. et al. Swarm learning for decentralized artificial intelligence in cancer histopathology. *Nat. Med.* **28**, 1232–1239 (2022).
24. Saldanha, O. L. et al. Direct prediction of genetic aberrations from pathology images in gastric cancer with swarm learning. *Gastric Cancer* **26**, 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? *Annu. Rev. Cancer Biol.* **7**, 57–71 (2023).
26. Ishii, Y., Langberg, J., Rosborough, K. & Mikawa, T. Endothelial cell lineages of the heart. *Cell Tissue Res.* **335**, 67–73 (2009).
27. Silla, C. N. & Freitas, A. A. A survey of hierarchical classification across different application domains. *Data Min. Knowl. Discov.* **22**, 31–72 (2011).
28. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2019).
29. Diehl, A. D. et al. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semant.* **7**, 44 (2016).
30. Zhao, R., Lu, J., Zhou, W., Zhao, N. & Ji, H. A systematic evaluation of highly variable gene selection methods for single-cell RNA-seq. *Genome Biol.* **26**, 424 (2025).
31. Wang, S. et al. scFed: federated learning for cell type classification with scRNA-seq. *Brief. Bioinformatics* **25**, bbad507 (2024).
32. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* **16**, 983–986 (2019).
33. Ma, F. & Pellegrini, M. ACTINN: automated identification of cell types in single cell RNA sequencing. *Bioinformatics* **36**, 533–538 (2020).
34. Heryanto, Y. D., Zhang, Y. -z & Imoto, S. Predicting cell types with supervised contrastive learning on cells and their types. *Sci. Rep.* **14**, 430 (2024).
35. Fischer, F., Fischer, D. S. & Mukhin, R. et al. scTab: Scaling cross-tissue single-cell annotation models. *Nat. Commun.* **15**, 6611 (2024).
36. Arik, S. Ö. & Pfister, T. TabNet: Attentive Interpretable Tabular Learning, In *Proc. AAAI Conference on Artificial Intelligence*, **35**, 6679–6687 <https://doi.org/10.1609/aaai.v35i8.16826>.
37. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794 (Association for Computing Machinery, 2016).
38. Shwartz-Ziv, R. & Armon, A. Tabular data: deep learning is not all you need. *Inf. Fusion* **81**, 84–90 (2022).

Acknowledgements

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

Author contributions

Conceptualization: O.L.S., J.N.K., and S.H.; Methodology: V.G., K.P., O.L.S., H.K., J.F.Z., J.N.K., and S.H.; Software: V.G., and K.P.; Formal analysis: V.G., and K.P.; Resources: J.N.K., R.K., and S.H.; Data curation: V.G., H.K., and K.P.; Writing—original draft: V.G., K.P., and S.H.; Writing—review and editing: V.G., O.L.S., K.P., R.K., J.N.K., and S.H.; Visualization: V.G. and K.P.; Supervision: J.N.K. and S.H.; Project administration: O.L.S. and S.H.; Funding acquisition: J.N.K. and S.H.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

S.H. is a consultant for Turbine AI, co-founder and shareholder of Sequantrix GmbH, and has received research funding from AskBio and Novo Nordisk. JNK declares consulting services for Bioprimus, 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 G.S.K.; and has received honoraria by AstraZeneca, Bayer, Daiichi Sankyo, Eisai, Janssen, Merck, M.S.D., B.M.S., Roche, Pfizer, and Fresenius. No other competing financial interests are declared by any of the remaining authors.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41540-026-00667-6>.

Correspondence and requests for materials should be addressed to Sikander Hayat or Jakob Nikolas Kather.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026