Supplementary: Figures

1 Clients data distribution

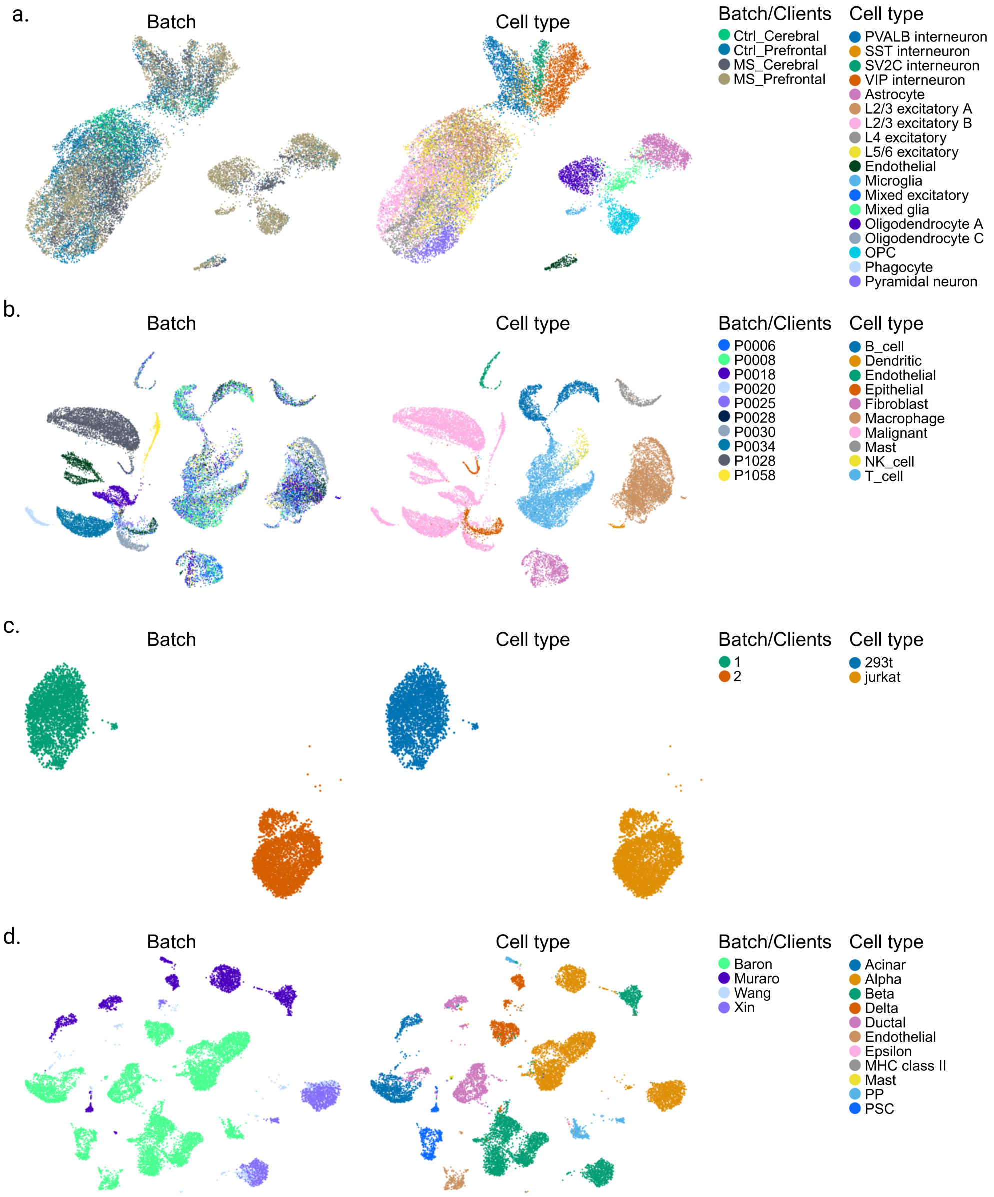


Fig S1: UMAP visualization showing batch and cell type distributions for the reference datasets. In the federated scenarios, each batch is assigned to a specific client: **a.** Multiple Sclerosis (MS) dataset. **b.** Lung-Kim dataset. **c.** CellLine (CL) dataset. **d.** Human Pancreas (HP) dataset.

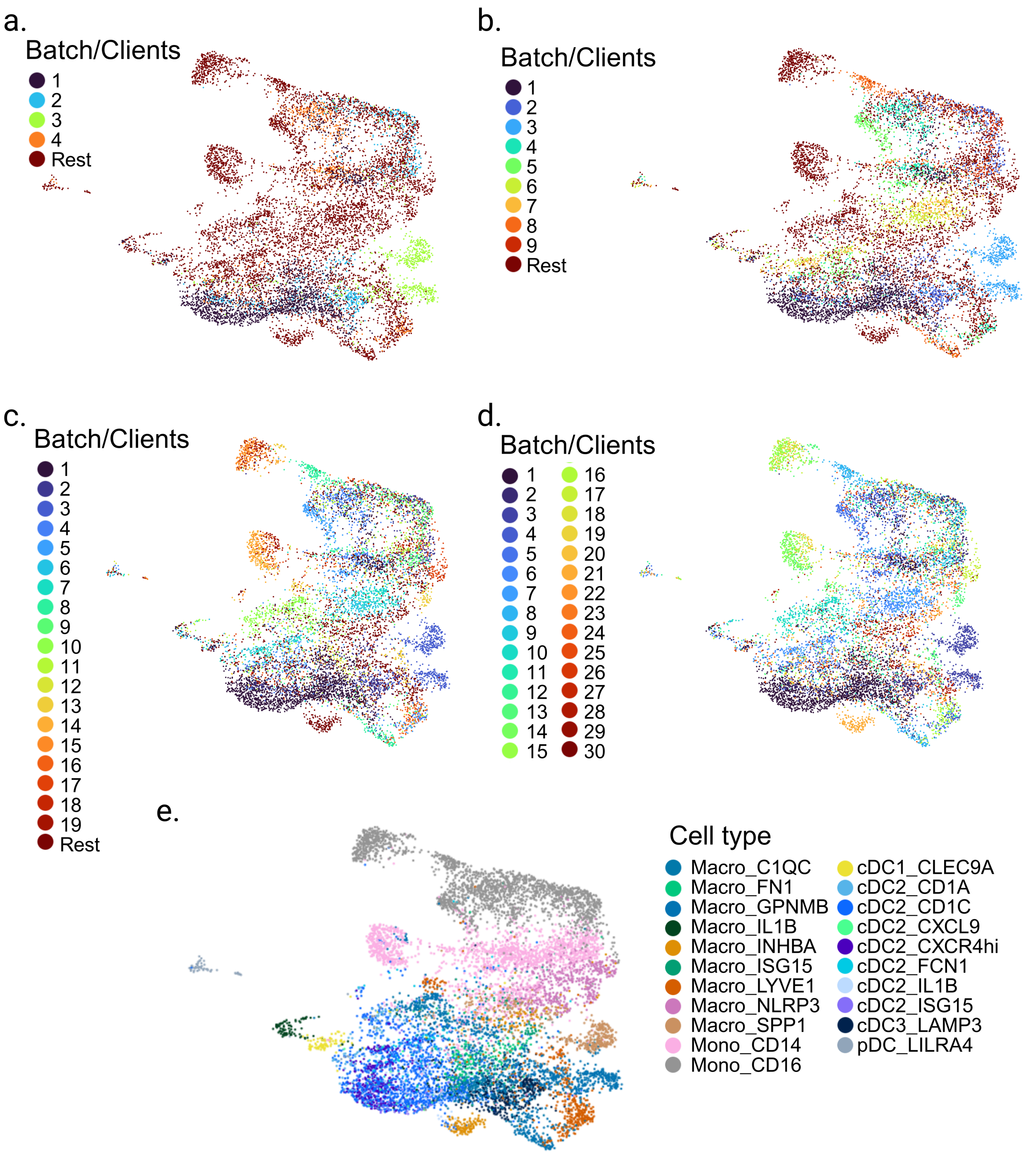


Fig S2: UMAP visualization of the Myeloid reference dataset across 30 batches and multiple cell types. In four federated scenarios, batches are sorted by size and assigned to clients: **a.** **Top5** – The four largest batches are assigned to clients 1–4, with the remaining batches grouped as client "rest." **b.** **Top10** – The nine largest batches are assigned to clients 1–9, with the remaining batches grouped as client "rest." **c.** **Top20** – The nineteen largest batches are assigned to clients 1–19, with the remaining batches grouped as client "rest." **d.** **Top30** – All thirty batches are treated as individual clients. **e.** UMAP of the Myeloid reference dataset annotated by known cell types.

2 Annotation

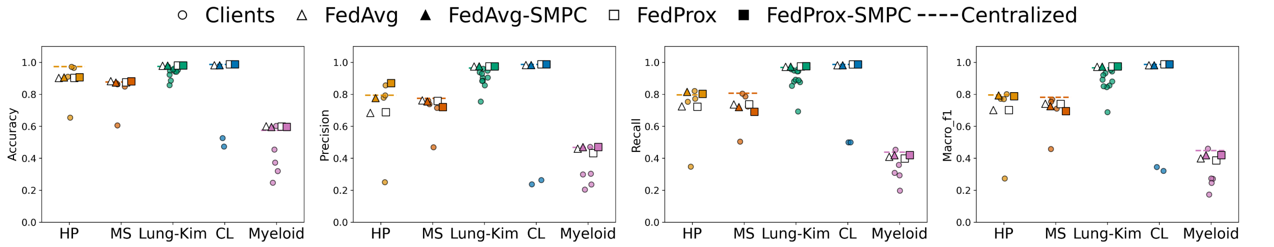


Fig S3: Performance comparison of the centralized model (dashed horizontal line), local client-level model (○), and federated models (FedAvg (△), FedAvg-SMPC (▲), FedProx (□), and FedProx-SMPC (■)) across Accuracy, Precision, Recall, and Macro-F1 (left to right) for cell type annotation on HP, MS, Lung-Kim, CL, and Myeloid (Top5) datasets. The centralized and client-level models were trained for 20 epochs, while the federated models were trained for 20 communication rounds with one local epoch per round.

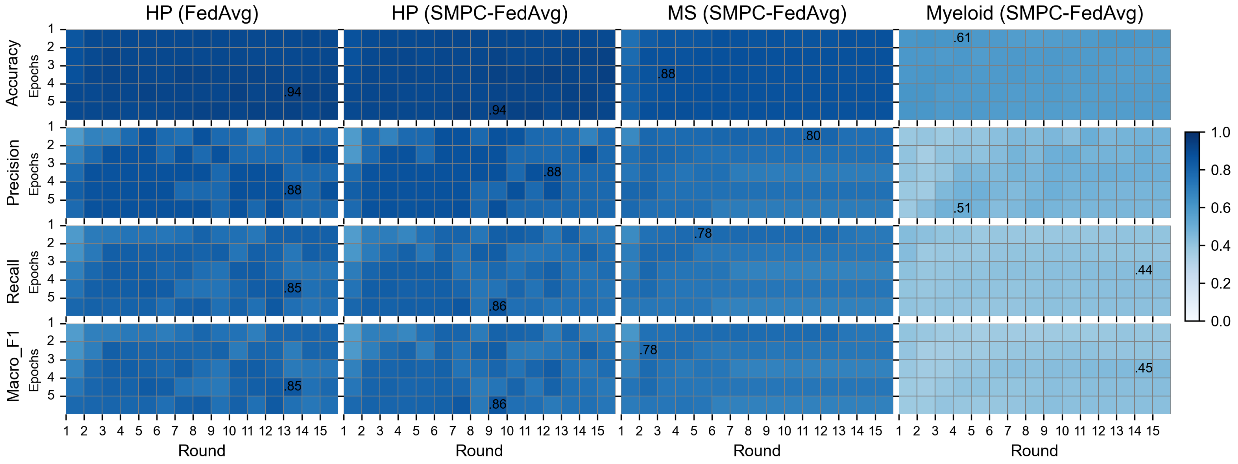


Fig S4: Performance heatmaps over 15 communication rounds and 1-5 local epochs for different federated aggregation methods and datasets. From left to right: HP (FedAvg), HP (SMPC-FedAvg), MS (SMPC-FedAvg), and Myeloid (SMPC-FedAvg). The effect of hyperparameters in each setting is evaluated in terms of Accuracy, Precision, Recall, and Macro-F1 (top to bottom).

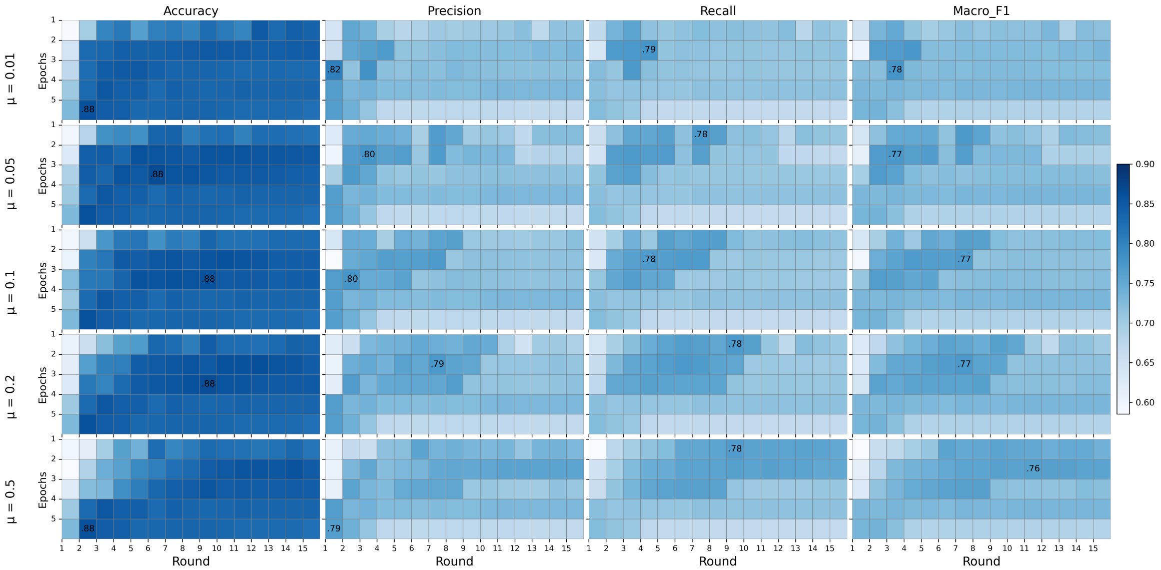


Fig S5: Performance of FedProx-SMPC across 15 communication rounds and 1-5 local epochs for different μ values ($\mu \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$; rows top to bottom) on the MS dataset, evaluated in terms of Accuracy, Precision, Recall, and Macro-F1 (columns left to right).

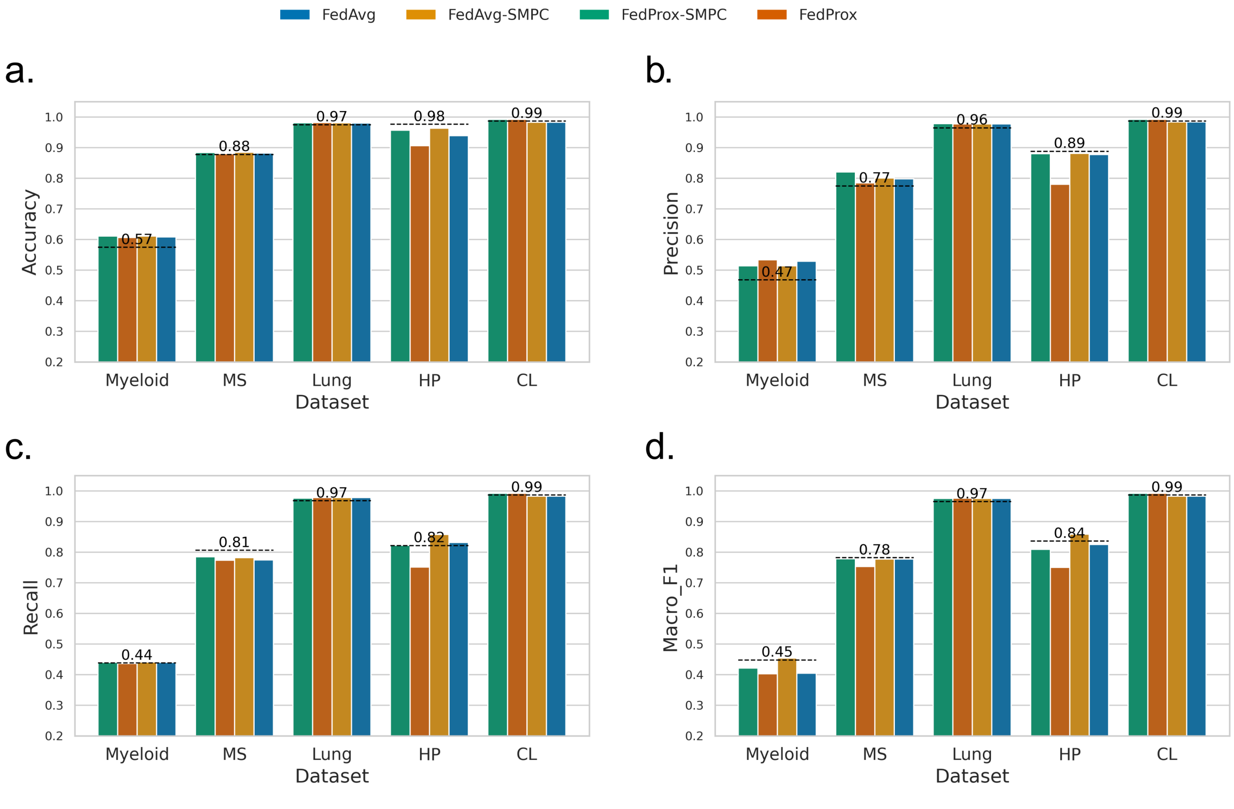
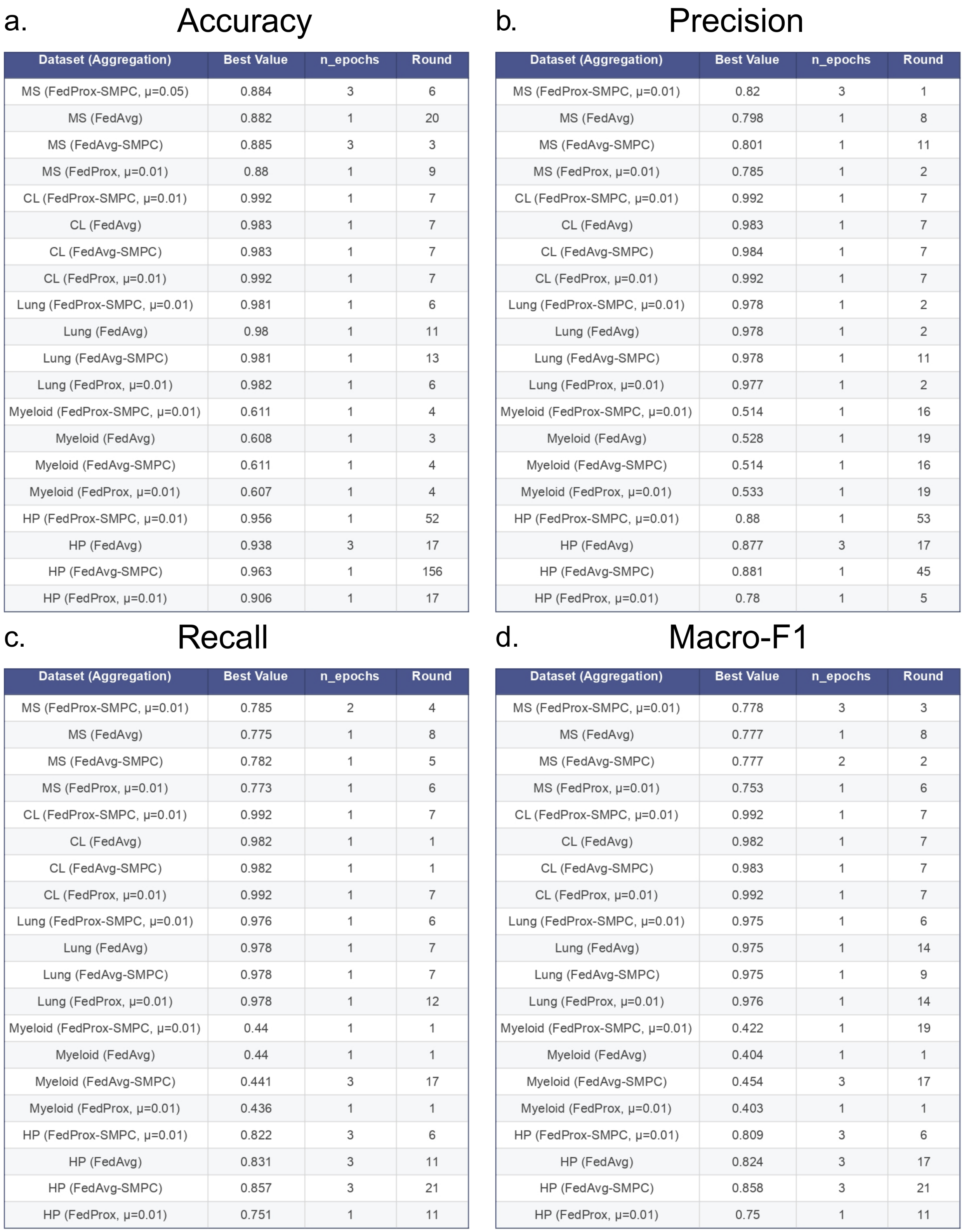


Fig S6: Best performance achieved by FedAvg and FedProx aggregation strategies, with and without SMPC (i.e., FedAvg, FedAvg-SMPC, FedProx, and FedProx-SMPC), across the Myeloid (Top5), MS, Lung, HP, and CL datasets. Dashed lines indicate centralized (non-federated) performance for each dataset and metric. **a.** Accuracy, **b.** Precision, **c.** Recall, and **d.** Macro-F1.



a. Accuracy

Dataset (Aggregation)	Best Value	n_epochs	Round
MS (FedProx-SMPC, μ=0.05)	0.884	3	6
MS (FedAvg)	0.882	1	20
MS (FedAvg-SMPC)	0.885	3	3
MS (FedProx, μ=0.01)	0.88	1	9
CL (FedProx-SMPC, μ=0.01)	0.992	1	7
CL (FedAvg)	0.983	1	7
CL (FedAvg-SMPC)	0.983	1	7
CL (FedProx, μ=0.01)	0.992	1	7
Lung (FedProx-SMPC, μ=0.01)	0.981	1	6
Lung (FedAvg)	0.98	1	11
Lung (FedAvg-SMPC)	0.981	1	13
Lung (FedProx, μ=0.01)	0.982	1	6
Myeloid (FedProx-SMPC, μ=0.01)	0.611	1	4
Myeloid (FedAvg)	0.608	1	3
Myeloid (FedAvg-SMPC)	0.611	1	4
Myeloid (FedProx, μ=0.01)	0.607	1	4
HP (FedProx-SMPC, μ=0.01)	0.956	1	52
HP (FedAvg)	0.938	3	17
HP (FedAvg-SMPC)	0.963	1	156
HP (FedProx, μ=0.01)	0.906	1	17

b. Precision

Dataset (Aggregation)	Best Value	n_epochs	Round
MS (FedProx-SMPC, μ=0.01)	0.82	3	1
MS (FedAvg)	0.798	1	8
MS (FedAvg-SMPC)	0.801	1	11
MS (FedProx, μ=0.01)	0.785	1	2
CL (FedProx-SMPC, μ=0.01)	0.992	1	7
CL (FedAvg)	0.983	1	7
CL (FedAvg-SMPC)	0.984	1	7
CL (FedProx, μ=0.01)	0.992	1	7
Lung (FedProx-SMPC, μ=0.01)	0.978	1	2
Lung (FedAvg)	0.978	1	2
Lung (FedAvg-SMPC)	0.978	1	11
Lung (FedProx, μ=0.01)	0.977	1	2
Myeloid (FedProx-SMPC, μ=0.01)	0.514	1	16
Myeloid (FedAvg)	0.528	1	19
Myeloid (FedAvg-SMPC)	0.514	1	16
Myeloid (FedProx, μ=0.01)	0.533	1	19
HP (FedProx-SMPC, μ=0.01)	0.88	1	53
HP (FedAvg)	0.877	3	17
HP (FedAvg-SMPC)	0.881	1	45
HP (FedProx, μ=0.01)	0.78	1	5

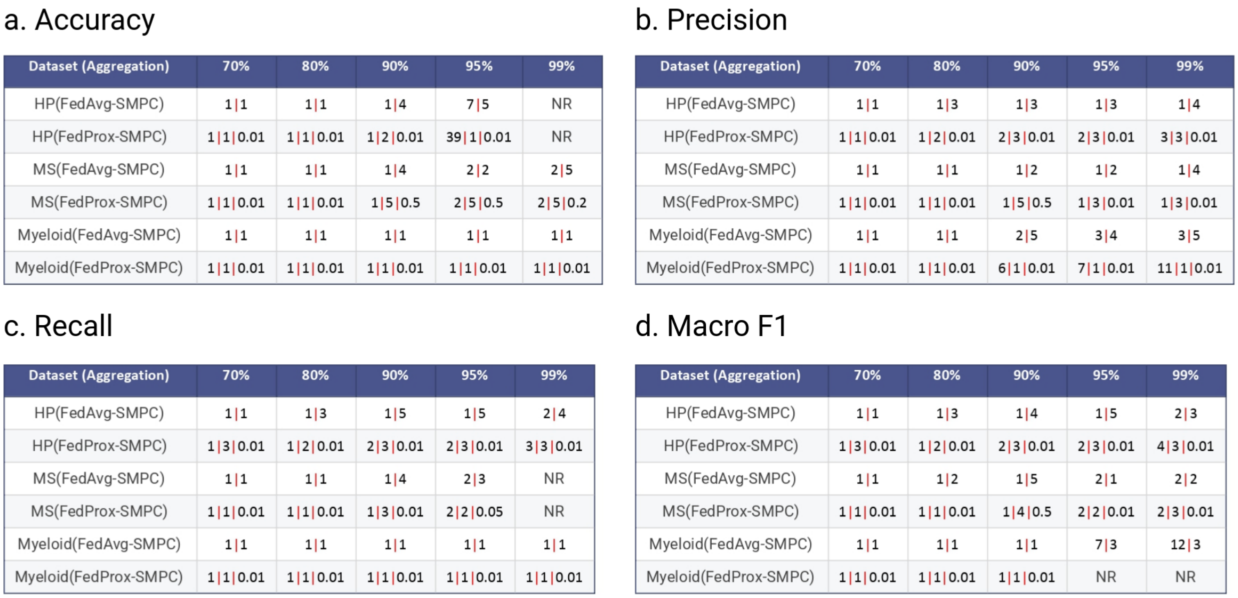
c. Recall

Dataset (Aggregation)	Best Value	n_epochs	Round
MS (FedProx-SMPC, μ=0.01)	0.785	2	4
MS (FedAvg)	0.775	1	8
MS (FedAvg-SMPC)	0.782	1	5
MS (FedProx, μ=0.01)	0.773	1	6
CL (FedProx-SMPC, μ=0.01)	0.992	1	7
CL (FedAvg)	0.982	1	1
CL (FedAvg-SMPC)	0.982	1	1
CL (FedProx, μ=0.01)	0.992	1	7
Lung (FedProx-SMPC, μ=0.01)	0.976	1	6
Lung (FedAvg)	0.978	1	7
Lung (FedAvg-SMPC)	0.978	1	7
Lung (FedProx, μ=0.01)	0.978	1	12
Myeloid (FedProx-SMPC, μ=0.01)	0.44	1	1
Myeloid (FedAvg)	0.44	1	1
Myeloid (FedAvg-SMPC)	0.441	3	17
Myeloid (FedProx, μ=0.01)	0.436	1	1
HP (FedProx-SMPC, μ=0.01)	0.822	3	6
HP (FedAvg)	0.831	3	11
HP (FedAvg-SMPC)	0.857	3	21
HP (FedProx, μ=0.01)	0.751	1	11

d. Macro-F1

Dataset (Aggregation)	Best Value	n_epochs	Round
MS (FedProx-SMPC, μ=0.01)	0.778	3	3
MS (FedAvg)	0.777	1	8
MS (FedAvg-SMPC)	0.777	2	2
MS (FedProx, μ=0.01)	0.753	1	6
CL (FedProx-SMPC, μ=0.01)	0.992	1	7
CL (FedAvg)	0.982	1	7
CL (FedAvg-SMPC)	0.983	1	7
CL (FedProx, μ=0.01)	0.992	1	7
Lung (FedProx-SMPC, μ=0.01)	0.975	1	6
Lung (FedAvg)	0.975	1	14
Lung (FedAvg-SMPC)	0.975	1	9
Lung (FedProx, μ=0.01)	0.976	1	14
Myeloid (FedProx-SMPC, μ=0.01)	0.422	1	19
Myeloid (FedAvg)	0.404	1	1
Myeloid (FedAvg-SMPC)	0.454	3	17
Myeloid (FedProx, μ=0.01)	0.403	1	1
HP (FedProx-SMPC, μ=0.01)	0.809	3	6
HP (FedAvg)	0.824	3	17
HP (FedAvg-SMPC)	0.858	3	21
HP (FedProx, μ=0.01)	0.75	1	11

Fig S7: Configurations of FedAvg, FedAvg-SMPC, FedProx, and FedProx-SMPC that achieved the best performance across the Myeloid (Top5), MS, Lung, HP, and CL datasets. For each aggregation strategy, the number of local epochs and communication rounds is reported alongside the corresponding performance value. For FedProx, the μ value is also indicated. **a.** Accuracy, **b.** Precision, **c.** Recall, **d.** Macro-F1.



a. Accuracy

Dataset (Aggregation)	70%	80%	90%	95%	99%
HP(FedAvg-SMPC)	1|1	1|1	1|4	7|5	NR
HP(FedProx-SMPC)	1|1|0.01	1|1|0.01	1|2|0.01	39|1|0.01	NR
MS(FedAvg-SMPC)	1|1	1|1	1|4	2|2	2|5
MS(FedProx-SMPC)	1|1|0.01	1|1|0.01	1|5|0.5	2|5|0.5	2|5|0.2
Myeloid(FedAvg-SMPC)	1|1	1|1	1|1	1|1	1|1
Myeloid(FedProx-SMPC)	1|1|0.01	1|1|0.01	1|1|0.01	1|1|0.01	1|1|0.01

b. Precision

Dataset (Aggregation)	70%	80%	90%	95%	99%
HP(FedAvg-SMPC)	1|1	1|3	1|3	1|3	1|4
HP(FedProx-SMPC)	1|1|0.01	1|2|0.01	2|3|0.01	2|3|0.01	3|3|0.01
MS(FedAvg-SMPC)	1|1	1|1	1|2	1|2	1|4
MS(FedProx-SMPC)	1|1|0.01	1|1|0.01	1|5|0.5	1|3|0.01	1|3|0.01
Myeloid(FedAvg-SMPC)	1|1	1|1	2|5	3|4	3|5
Myeloid(FedProx-SMPC)	1|1|0.01	1|1|0.01	6|1|0.01	7|1|0.01	11|1|0.01

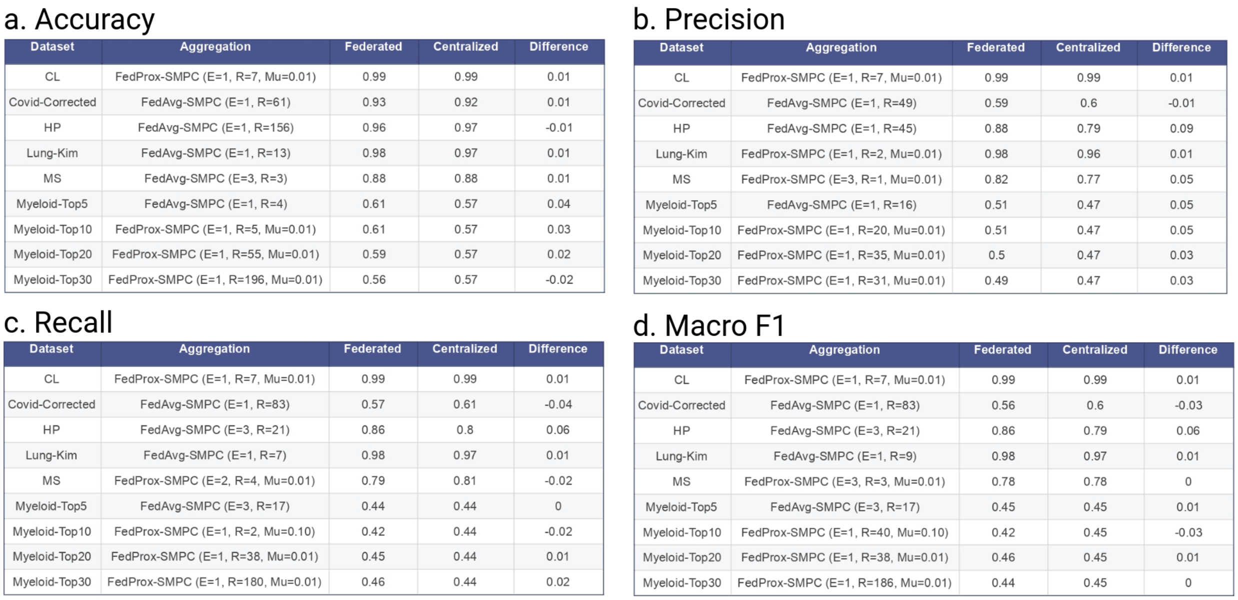
c. Recall

Dataset (Aggregation)	70%	80%	90%	95%	99%
HP(FedAvg-SMPC)	1|1	1|3	1|5	1|5	2|4
HP(FedProx-SMPC)	1|3|0.01	1|2|0.01	2|3|0.01	2|3|0.01	3|3|0.01
MS(FedAvg-SMPC)	1|1	1|1	1|4	2|3	NR
MS(FedProx-SMPC)	1|1|0.01	1|1|0.01	1|3|0.01	2|2|0.05	NR
Myeloid(FedAvg-SMPC)	1|1	1|1	1|1	1|1	1|1
Myeloid(FedProx-SMPC)	1|1|0.01	1|1|0.01	1|1|0.01	1|1|0.01	1|1|0.01

d. Macro F1

Dataset (Aggregation)	70%	80%	90%	95%	99%
HP(FedAvg-SMPC)	1|1	1|3	1|4	1|5	2|3
HP(FedProx-SMPC)	1|3|0.01	1|2|0.01	2|3|0.01	2|3|0.01	4|3|0.01
MS(FedAvg-SMPC)	1|1	1|2	1|5	2|1	2|2
MS(FedProx-SMPC)	1|1|0.01	1|1|0.01	1|4|0.5	2|2|0.01	2|3|0.01
Myeloid(FedAvg-SMPC)	1|1	1|1	1|1	7|3	12|3
Myeloid(FedProx-SMPC)	1|1|0.01	1|1|0.01	1|1|0.01	NR	NR

Fig S8: Communication efficiency analysis of privacy-preserving federated aggregation strategies across the HP, MS, and Myeloid (Top5) datasets. The analysis shows the required communication rounds and local training epochs needed to reach 70%, 80%, 90%, 95%, and 99% of the centralized baseline performance. Each cell displays the number of communication rounds, local epochs, and μ value (only applicable for FedProx), separated by the symbol |. Entries marked as "NR" indicate that the corresponding performance threshold was not reached within the evaluated training budget. **a.** Accuracy, **b.** Precision, **c.** Recall, **d.** Macro-F1.



a. Accuracy

Dataset	Aggregation	Federated	Centralized	Difference
CL	FedProx-SMPC (E=1, R=7, Mu=0.01)	0.99	0.99	0.01
Covid-Corrected	FedAvg-SMPC (E=1, R=61)	0.93	0.92	0.01
HP	FedAvg-SMPC (E=1, R=156)	0.96	0.97	-0.01
Lung-Kim	FedAvg-SMPC (E=1, R=13)	0.98	0.97	0.01
MS	FedAvg-SMPC (E=3, R=3)	0.88	0.88	0.01
Myeloid-Top5	FedAvg-SMPC (E=1, R=4)	0.61	0.57	0.04
Myeloid-Top10	FedProx-SMPC (E=1, R=5, Mu=0.01)	0.61	0.57	0.03
Myeloid-Top20	FedProx-SMPC (E=1, R=55, Mu=0.01)	0.59	0.57	0.02
Myeloid-Top30	FedProx-SMPC (E=1, R=196, Mu=0.01)	0.56	0.57	-0.02

b. Precision

Dataset	Aggregation	Federated	Centralized	Difference
CL	FedProx-SMPC (E=1, R=7, Mu=0.01)	0.99	0.99	0.01
Covid-Corrected	FedAvg-SMPC (E=1, R=49)	0.59	0.6	-0.01
HP	FedAvg-SMPC (E=1, R=45)	0.88	0.79	0.09
Lung-Kim	FedProx-SMPC (E=1, R=2, Mu=0.01)	0.98	0.96	0.01
MS	FedProx-SMPC (E=3, R=1, Mu=0.01)	0.82	0.77	0.05
Myeloid-Top5	FedAvg-SMPC (E=1, R=16)	0.51	0.47	0.05
Myeloid-Top10	FedProx-SMPC (E=1, R=20, Mu=0.01)	0.51	0.47	0.05
Myeloid-Top20	FedProx-SMPC (E=1, R=35, Mu=0.01)	0.5	0.47	0.03
Myeloid-Top30	FedProx-SMPC (E=1, R=31, Mu=0.01)	0.49	0.47	0.03

c. Recall

Dataset	Aggregation	Federated	Centralized	Difference
CL	FedProx-SMPC (E=1, R=7, Mu=0.01)	0.99	0.99	0.01
Covid-Corrected	FedAvg-SMPC (E=1, R=83)	0.57	0.61	-0.04
HP	FedAvg-SMPC (E=3, R=21)	0.86	0.8	0.06
Lung-Kim	FedAvg-SMPC (E=1, R=7)	0.98	0.97	0.01
MS	FedProx-SMPC (E=2, R=4, Mu=0.01)	0.79	0.81	-0.02
Myeloid-Top5	FedAvg-SMPC (E=3, R=17)	0.44	0.44	0
Myeloid-Top10	FedProx-SMPC (E=1, R=2, Mu=0.10)	0.42	0.44	-0.02
Myeloid-Top20	FedProx-SMPC (E=1, R=38, Mu=0.01)	0.45	0.44	0.01
Myeloid-Top30	FedProx-SMPC (E=1, R=180, Mu=0.01)	0.46	0.44	0.02

d. Macro F1

Dataset	Aggregation	Federated	Centralized	Difference
CL	FedProx-SMPC (E=1, R=7, Mu=0.01)	0.99	0.99	0.01
Covid-Corrected	FedAvg-SMPC (E=1, R=83)	0.56	0.6	-0.03
HP	FedAvg-SMPC (E=3, R=21)	0.86	0.79	0.06
Lung-Kim	FedAvg-SMPC (E=1, R=9)	0.98	0.97	0.01
MS	FedProx-SMPC (E=3, R=3, Mu=0.01)	0.78	0.78	0
Myeloid-Top5	FedAvg-SMPC (E=3, R=17)	0.45	0.45	0.01
Myeloid-Top10	FedProx-SMPC (E=1, R=40, Mu=0.10)	0.42	0.45	-0.03
Myeloid-Top20	FedProx-SMPC (E=1, R=38, Mu=0.01)	0.46	0.45	0.01
Myeloid-Top30	FedProx-SMPC (E=1, R=186, Mu=0.01)	0.44	0.45	0

Figure S9: Performance comparison between the best privacy-preserving federated model and the centralized model for cell type annotation on the CL, Covid-corrected (Covid-19), HP, Lung-Kim, MS, and Myeloid datasets (all scenarios: Top5–Top30), evaluated across various metrics. All values are rounded to two decimal places for simplicity. **a.** Accuracy, **b.** Precision, **c.** Recall, **d.** Macro-F1.

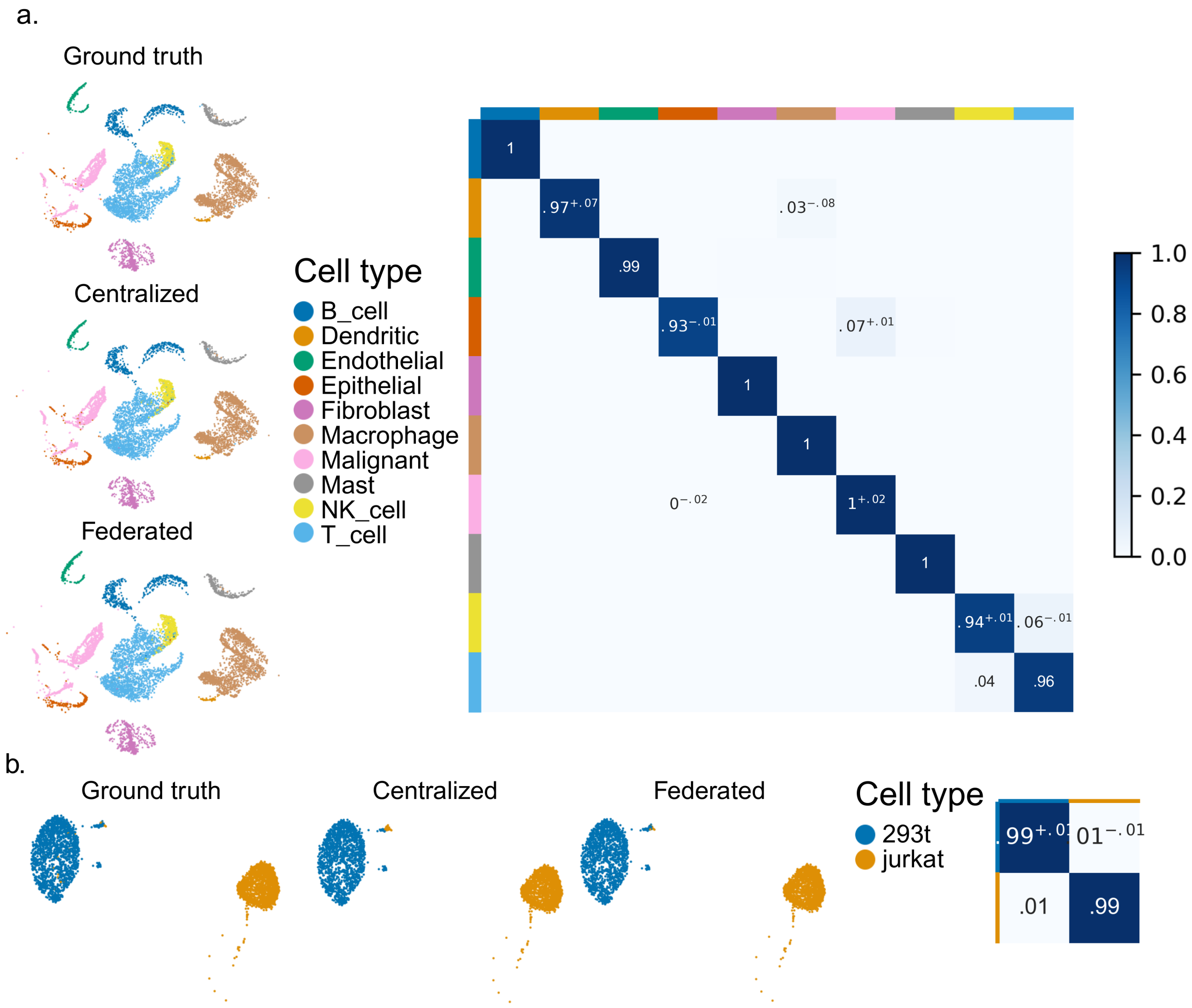
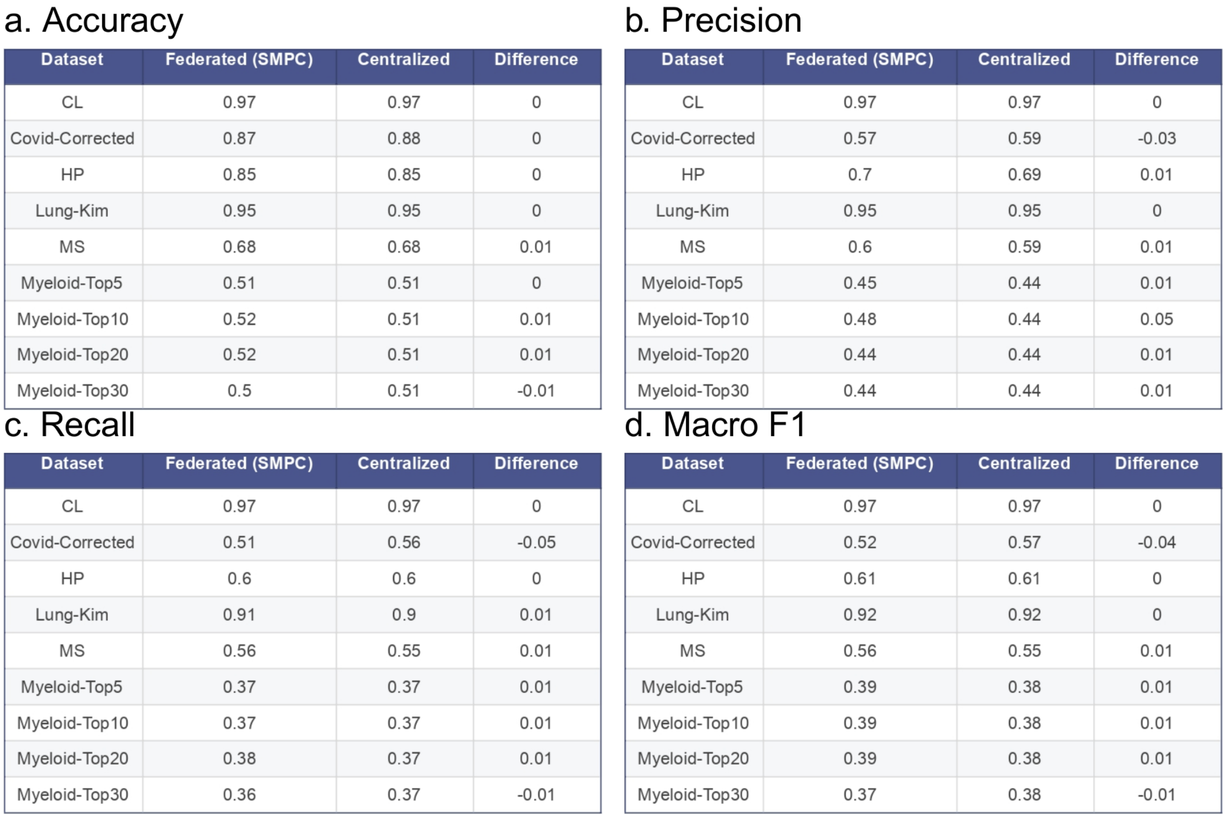


Fig S10: UMAP visualizations of ground truth, centralized predictions, and federated predictions of cell types, alongside confusion matrices showing federated performance in terms of per-class recall for each cell type. **a. Lung-Kim:** Federated results obtained using FedAvg-SMPC, trained for 1 local epoch over 7 communication rounds. **b. Cell Line (CL):** Results obtained using FedProx ($\mu = 0.01$), trained for 1 local epoch over 7 communication rounds.

3 Reference mapping



a. Accuracy

Dataset	Federated (SMPC)	Centralized	Difference
CL	0.97	0.97	0
Covid-Corrected	0.87	0.88	0
HP	0.85	0.85	0
Lung-Kim	0.95	0.95	0
MS	0.68	0.68	0.01
Myeloid-Top5	0.51	0.51	0
Myeloid-Top10	0.52	0.51	0.01
Myeloid-Top20	0.52	0.51	0.01
Myeloid-Top30	0.5	0.51	-0.01

b. Precision

Dataset	Federated (SMPC)	Centralized	Difference
CL	0.97	0.97	0
Covid-Corrected	0.57	0.59	-0.03
HP	0.7	0.69	0.01
Lung-Kim	0.95	0.95	0
MS	0.6	0.59	0.01
Myeloid-Top5	0.45	0.44	0.01
Myeloid-Top10	0.48	0.44	0.05
Myeloid-Top20	0.44	0.44	0.01
Myeloid-Top30	0.44	0.44	0.01

c. Recall

Dataset	Federated (SMPC)	Centralized	Difference
CL	0.97	0.97	0
Covid-Corrected	0.51	0.56	-0.05
HP	0.6	0.6	0
Lung-Kim	0.91	0.9	0.01
MS	0.56	0.55	0.01
Myeloid-Top5	0.37	0.37	0.01
Myeloid-Top10	0.37	0.37	0.01
Myeloid-Top20	0.38	0.37	0.01
Myeloid-Top30	0.36	0.37	-0.01

d. Macro F1

Dataset	Federated (SMPC)	Centralized	Difference
CL	0.97	0.97	0
Covid-Corrected	0.52	0.57	-0.04
HP	0.61	0.61	0
Lung-Kim	0.92	0.92	0
MS	0.56	0.55	0.01
Myeloid-Top5	0.39	0.38	0.01
Myeloid-Top10	0.39	0.38	0.01
Myeloid-Top20	0.39	0.38	0.01
Myeloid-Top30	0.37	0.38	-0.01

Figure S11: Performance comparison between the privacy-preserving federated refernce mapping and the centralized reference mapping on CL, Covid-corrected (Covid-19), HP, Lung-Kim, MS, and Myeloid datasets (all scenarios: Top5–Top30) daatsets, evaluated across various metrics. All values are rounded to two decimal places for simplicity. **a.** Accuracy, **b.** Precision, **c.** Recall, **d.** Macro-F1.

4 Scalability

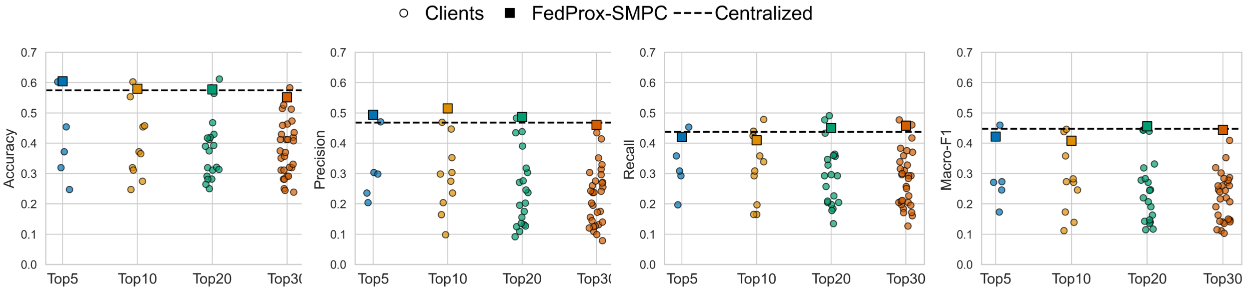


Fig S12: Scalability analysis of federated cell type annotation on the Myeloid dataset using FedProx-SMPC ($\mu = 0.01$) compared to the centralized model (dashed horizontal line) and the local client-level model (○), evaluated in terms of Accuracy, Precision, Recall, and Macro-F1 (left to right). Four scalability scenarios were designed with increasing numbers of clients (Top5, Top10, Top20, Top30). FedProx-SMPC was trained for 1 local epoch with varying numbers of communication rounds: 19 rounds for Top5, 20 for Top10, 38 for Top20, and 186 for Top30.

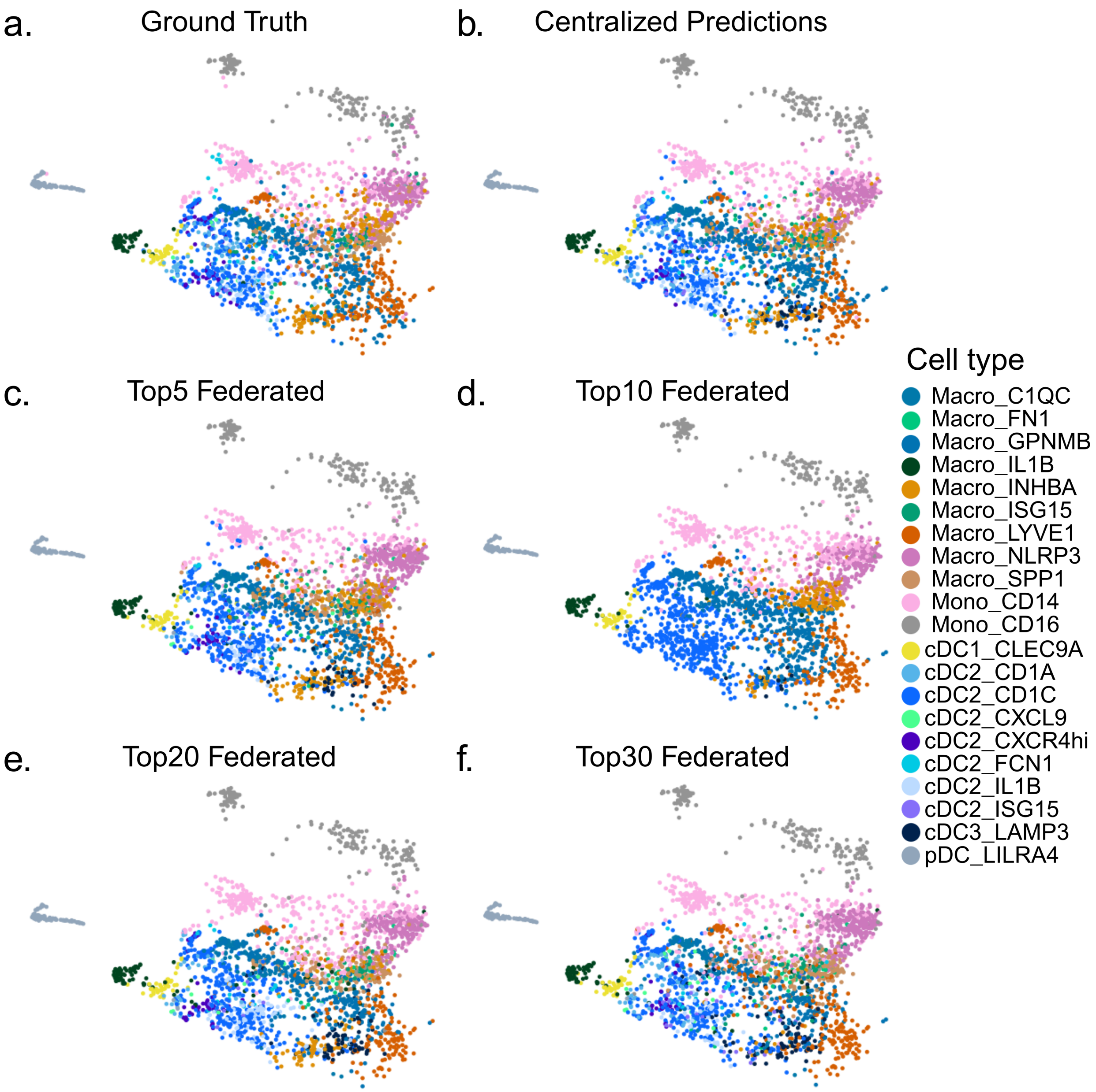


Fig S13: UMAP visualizations of ground truth cell types compared to predictions by the centralized model and FedProx-SMPC ($\mu = 0.01$) across various scalability scenarios of the Myeloid dataset. FedProx-SMPC was trained for one local epoch. **a.** Ground truth. **b.** Centralized: trained for 20 epochs. **c.** Top5: FedProx-SMPC trained for 19 communication rounds. **d.** Top10: FedProx-SMPC trained for 20 communication rounds. **e.** Top20: FedProx-SMPC trained for 38 communication rounds. **f.** Top30: FedProx-SMPC trained for 186 communication rounds.

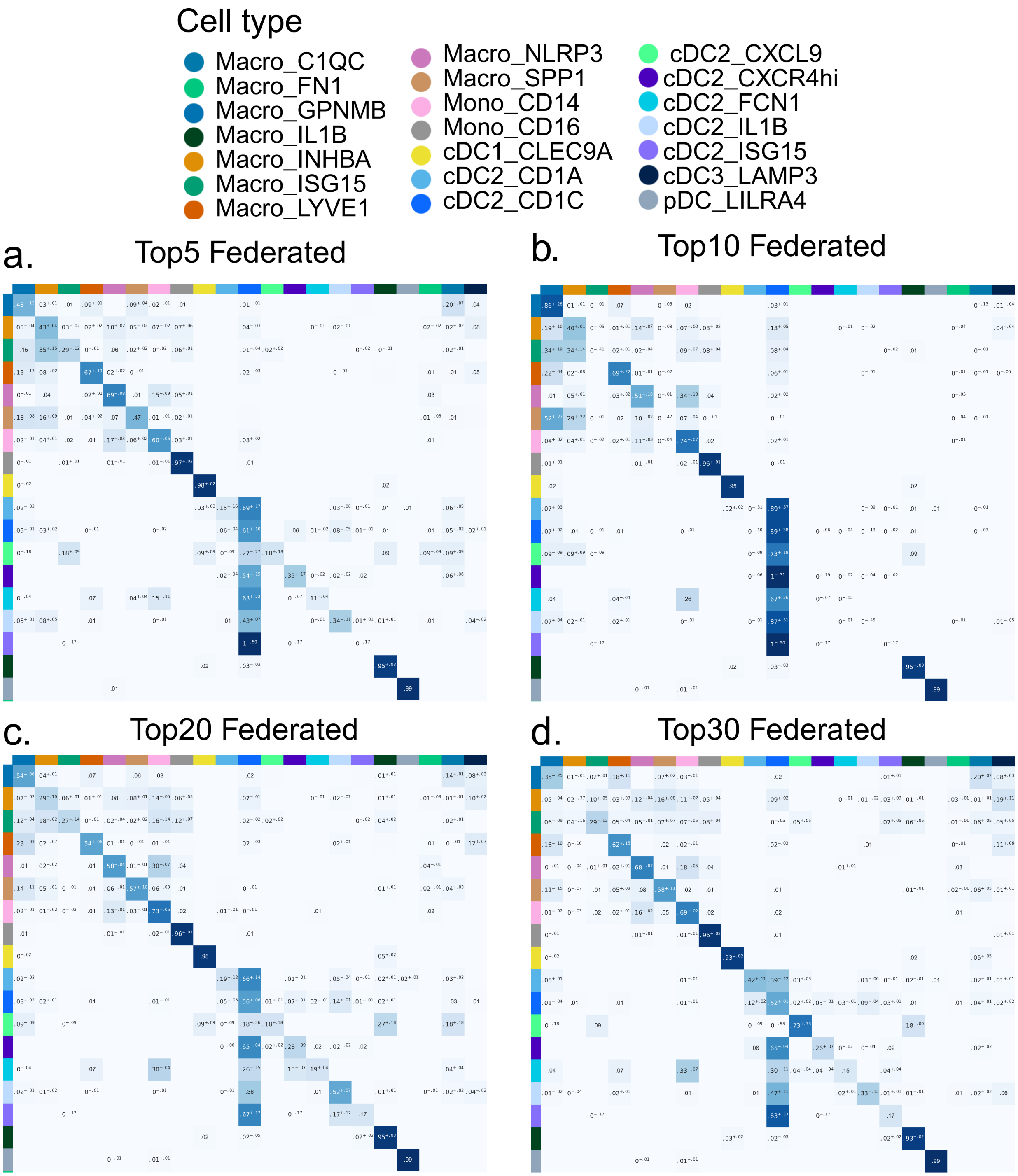


Fig S14: Confusion matrices of FedProx-SMPC ($\mu = 0.01$, trained for one local epoch) across various scalability scenarios of the Myeloid dataset. Each row represents the true cell type, and each column represents the predicted label. Per-class recall values are annotated, with superscripted differences relative to the centralized model. A plus sign indicates recall improvement, and a minus sign indicates a decrease. **a. Top5** (19 communication rounds). **b. Top10** (20 rounds). **c. Top20** (38 rounds). **d. Top30** (186 rounds).

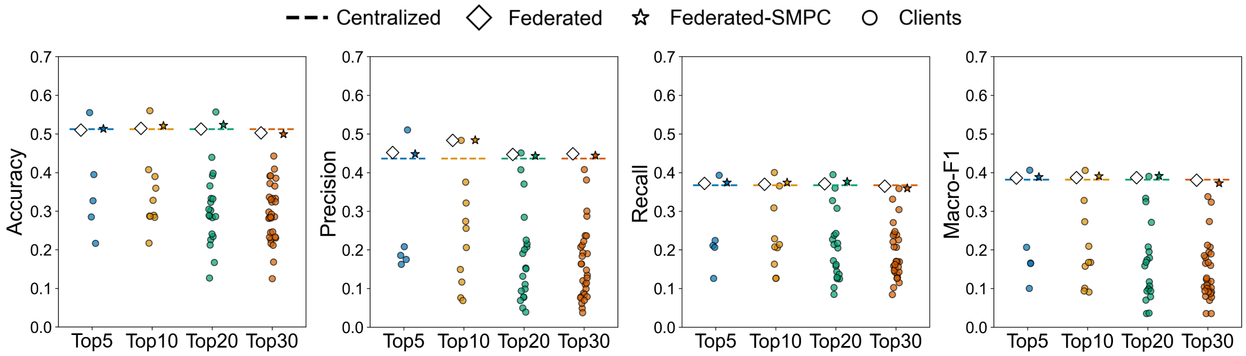


Fig S15: Performance comparison between the privacy-preserving federated models (Federated (⋄) and Federated-SMPC (⋆)), the centralized model (dashed horizontal line), and the local client-level model (∘) for reference mapping across various scalability scenarios of the Myeloid dataset, evaluated in terms of Accuracy, Precision, Recall, and Macro-F1 (left to right).

5 Batch effect

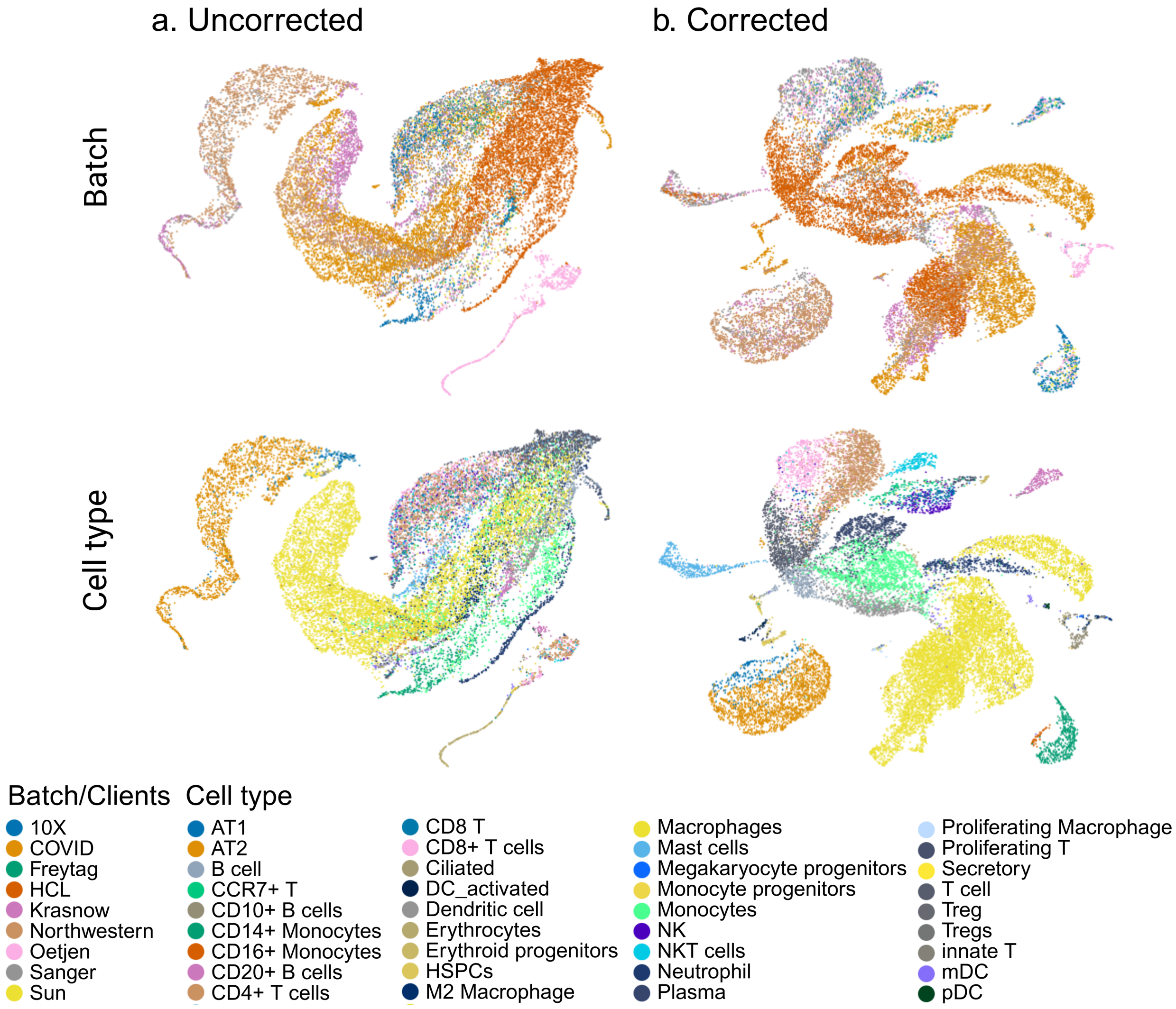


Fig S16: UMAP visualizations of cell types and batches in the COVID-19 dataset. **a.** Uncorrected COVID-19 dataset. **b.** Batch-effect corrected COVID-19 dataset using the scGen model.

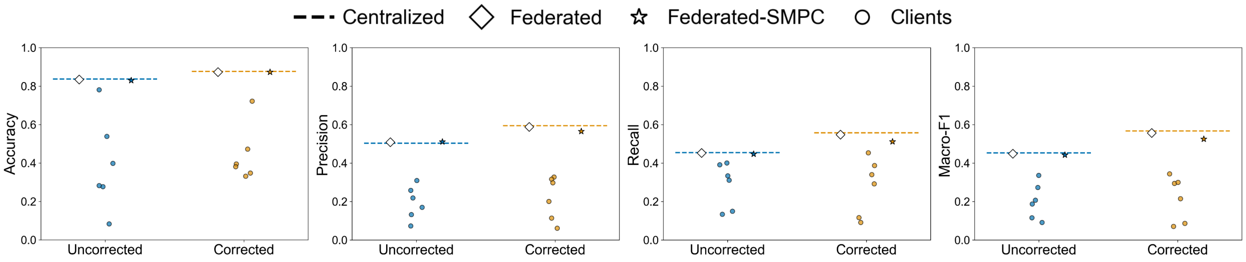


Fig S17: Performance comparison between the privacy-preserving federated models (Federated (◇) and Federated-SMPC (⋆)), the centralized model (dashed horizontal line), and the local client-level model (○) for reference mapping of the COVID-19 dataset, evaluated in terms of Accuracy, Precision, Recall, and Macro-F1 (left to right).