

Privacy-preserving federated tensor decomposition of single-cell immune data: recovering multicellular programs across institutions

Axel Faes*, Maryam Amir Haeri

*Learning, Data Analytics and Technology, University of Twente, Drienerlolaan 5, 7522
NB, Enschede, The Netherlands*

Abstract

Background and Objectives: Tensor decomposition of single-cell data identifies multicellular programs but needs pooled data, which consent and data-protection rules often prohibit. We develop a federated estimator and evaluate its approximation to the pooled analysis and the leakage of released objects.

Methods: Sites centre their data on a federated global mean and send a truncated SVD summary of the local program subspace; a coordinator takes the top singular vectors of the stacked summaries. The merge is approximate, with error set by the omitted local singular components. When sites profile different cell types, donor Gram matrices are summed. A membership-inference attack measures leakage. Seven public atlases were analysed.

Results: In a lupus cohort, federated recovery closely approximates the pooled result (subspace correlation 0.967; AUC difference 0.000, 95% CI $[-0.004, +0.012]$), also across multi-ancestry splits and the three original COVID-19 collection sites (correlation 0.989). When no site observes all cell types, donor Gram matrices recover the full program (correlation 1.000, exact by construction). Membership-inference AUC is 0.906 for site-specific subspaces, 0.609 for the merged global subspace and 1.000 for the donor Gram. The program out-predicts the best single cell type in lung (AUC 0.959 vs 0.910) and liver and is comparable elsewhere.

Conclusions: Cell-type-incomplete sites can be federated where fixed-feature federated PCA cannot. The stacked merge is not an additive

*Corresponding author.

Email address: `axel.faes@utwente.nl` (Axel Faes)

aggregation, and the released program carries no formal privacy guarantee. The predictive advantage of multicellularity is tissue-dependent. The code and the public data are released.

Keywords: federated learning, tensor decomposition, single-cell transcriptomics, membership inference, privacy, immune variation

1. Introduction

Population-scale single-cell RNA sequencing has made it possible to study how the *coordinated* state of the immune system varies between individuals. An established framing represents a patient cohort as a third-order tensor of donor $\times$ cell-type $\times$ gene pseudobulk expression and decomposes it to recover *multicellular programs*: factors whose loadings span several cell types and capture inter-individual axes of variation that no single cell type expresses in isolation. Interpretable tensor decomposition (scITD) introduced this view and used it to stratify lupus and COVID-19 patients [1]; related tensor methods model cell-cell communication across contexts [2] and condition-resolved variation [3]. These methods are *centralized*: they assume every cell is available in one place.

Such programs place each patient on continuous axes of immune state and link coordinated transcriptional shifts to disease severity, treatment response, and genetics. Their value grows with cohort size and diversity, so the most informative analyses span many institutions and ancestries, the setting hardest to centralize. Immune single-cell atlases are assembled across hospitals, biobanks, and national cohorts [4], and consent and data-protection constraints frequently preclude pooling patient-derived single-cell profiles [5]. Federated analysis (computing over data that remain at their source) is the

natural response [6], and federated methods now exist for single-cell batch correction [7], cell-type classification under homomorphic encryption [8], and quantitative-trait-locus mapping [9]. None recovers multicellular *programs*: the tensor-decomposition task that scITD addresses centrally has no federated counterpart.

We address this gap (Fig. 1). The principal contributions and their limitations are as follows:

1. **Vertical-FL recovery of cross-panel multicellular programs**: recovery of the full multicellular program when *no site observes all cell types*, by coupling through the shared donor mode. In this vertical federated setting (shared donors, partitioned cell-type panels), fixed-feature federated PCA cannot be applied at all.
2. A **confounding-robust federated estimator**: sites share a low-dimensional program subspace, and the coordinator merges the local summaries by stacked SVD after federated global-mean centering, approximating the centralized decomposition with error set by the omitted local singular components. The stacked calculation requires the individual site summaries and is therefore not equivalent to standard additive secure aggregation [10]. This merge instantiates distributed PCA [11, 12]; our additions are the global-mean centering, which makes it robust to site-label confounding (Table 1), and, with contribution 1, the cross-panel donor coupling.
3. A **membership-inference privacy evaluation** [13] that quantifies the leakage of each evaluated release object. These experiments describe performance under specified attacks and do not establish a gen-

eral privacy guarantee; formal differential privacy [14] is out of reach at present single-cell cohort sizes (Section 5).

4. **A cross-tissue map of when multicellular recovery matters** (§4.5): across seven cohorts the multicellular program out-predicts the best single cell type only where cross-cell-type coordination is strong: interstitial lung disease is the primary result, a smaller fibrosing-cholangiopathy liver cohort corroborates it, and the program is at parity elsewhere.

2. Related Work

Centralized multicellular decomposition. scITD [1] factorizes a donor $\times$ cell-type $\times$ gene tensor by Tucker decomposition to recover multicellular programs, MOFA [15] integrates modalities by matrix factorization, and DIALOGUE [16] recovers programs by cross-cell-type association. All operate on pooled data.

Federated single-cell analysis. Existing federated single-cell methods target batch-effect correction [7], cell-type classification under homomorphic encryption [8], and foundation-model pretraining [17].

Federated and private factorization. Distributed PCA over partitioned data is well characterized [11, 12], and the donor-mode SVD we federate is an instance. Differentially private low-rank factorization has been studied for PCA and tensor decompositions [18, 19], but meaningful private subspace recovery needs $n = \tilde{O}(d)$ samples [19], beyond current single-cell atlases.

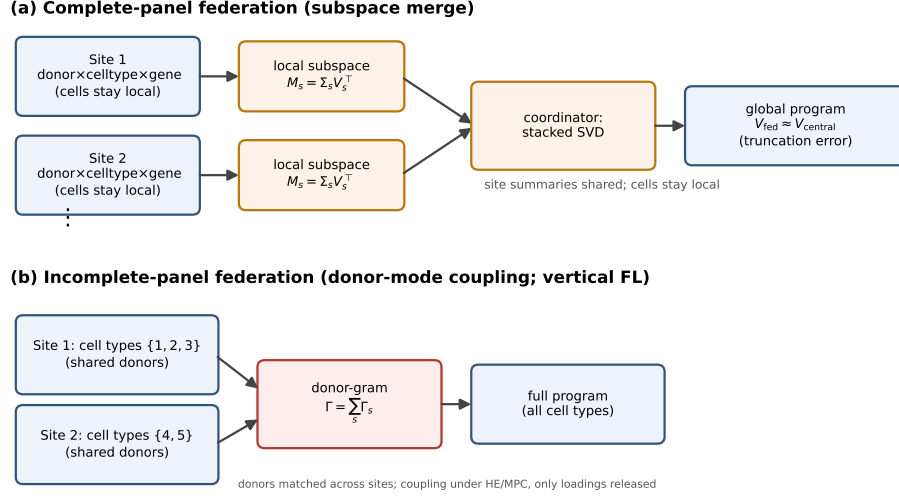


Figure 1: Federated multicellular program estimation. (a) Complete-panel: sites transmit rank- L local summaries, and a coordinator applies stacked SVD to approximate the rank- K centralized program. (b) Incomplete-panel (vertical FL): sites holding different cell-type panels for shared donors contribute donor Gram matrices, and secure computation of the coupling supports a loadings-only release.

Federated genomics privacy. privateQTL [9] performs secure federated quantitative-trait-locus mapping via multiparty computation. The re-identification risk is concrete: it has been demonstrated from aggregate allele-frequency statistics [20] and, more recently, from single-cell count matrices [21]. Our target is the *unsupervised multicellular program*.

3. Methods

3.1. Multicellular tensor and centralized decomposition

Let D , C , G and S denote the numbers of donors, cell types, genes and sites. The final number of global components is K , and $L \geq K$ denotes the number of local singular components each site communicates. Let r_{deg}

denote the summed UMI count of gene g over all cells of type c in donor d (pseudobulk) [22]. We normalize each (donor, cell type) profile to log counts-per-million, $\tilde{r}_{dcg} = \log(1 + 10^6 r_{dcg} / \sum_{g'} r_{dcg'})$, retain the G most variable genes, and *standardize each gene within each cell type across donors*:

$$\mathcal{T}_{dcg} = \frac{\tilde{r}_{dcg} - \mu_{cg}}{\sigma_{cg}}, \quad \mu_{cg} = \frac{1}{|\mathcal{D}_c|} \sum_{d \in \mathcal{D}_c} \tilde{r}_{dcg}, \quad (1)$$

with σ_{cg} the corresponding standard deviation and $\mathcal{D}_c$ the donors in which type c is observed (≥ 20 cells). For horizontally partitioned donors, the counts, sums and sums of squares needed for μ_{cg} and σ_{cg} are aggregated across sites, as is variable-gene selection. In the present implementation a missing cell-type slab is set to zero after standardization, which corresponds to mean imputation and is not treated as a masked value by ordinary SVD. This centering removes cell-type baseline so $\mathcal{T} \in \mathbb{R}^{D \times C \times G}$ carries *inter-individual* variation, the scITD convention [1]. The donor-mode unfolding $X \in \mathbb{R}^{D \times CG}$ flattens the cell-type and gene axes [23, 24]. A rank- K decomposition takes the top- K right singular vectors $V \in \mathbb{R}^{CG \times K}$ of the column-centred $X - \bar{x}$; each column, reshaped to $C \times G$, is a *multicellular program* whose per-cell-type loading norms indicate which cell types participate, and the projected donor coordinates $Z = (X - \bar{x})V$ place each donor on every program. We use $K = 10$ and $G = 1500$ throughout; donors are retained if at least four cell types are present.

3.2. Federated estimator

In the complete-panel setting, donors are partitioned across S sites over the same CG features; site s holds rows X_s and never transmits them. The coordinator obtains the federated global mean $\bar{x} = \frac{1}{D} \sum_s n_s \bar{x}_s$ from the per-site means $\bar{x}_s$, each averaged over site s 's donors across all CG features. Each site centres locally, computes a truncated SVD $X_s - \bar{x} = U_s \Sigma_s V_s^\top$, and transmits only the scaled loadings $M_s = \Sigma_s^{(L)} V_s^{(L)\top} \in \mathbb{R}^{L \times CG}$, where the local communication rank L is distinct from the final rank K . The coordinator stacks $M = [M_1; \dots; M_S]$ and takes the top- K right singular vectors of M as the global program basis V_{fed} ; donor coordinates are computed locally by projection. Here

$$M^\top M = \sum_s V_s^{(L)} (\Sigma_s^{(L)})^2 V_s^{(L)\top}, \quad (2)$$

which in general differs from the centralized covariance $X_c^\top X_c$ with $X_c = X - \bar{x}$. Writing $X_{s,c} = X_s - \bar{x}$,

$$\left\| X_c^\top X_c - M^\top M \right\|_2 \leq \sum_{s=1}^S \sigma_{L+1}^2(X_{s,c}), \quad (3)$$

The two matrices coincide, $M^\top M = X_c^\top X_c$, only when every site retains all nonzero local singular components. The stacked-SVD basis therefore approximates the centralized subspace, with error controlled by the omitted local spectra [11]. The merge also resolves the per-site rotation gauge, so no explicit factor alignment is required (Algorithm 1).

Algorithm 1 Federated complete-panel multicellular decomposition

- 1: **Input:** site unfoldings $\{X_s\}_{s=1}^S$, output rank K , local rank $L \geq K$
 - 2: each site sends $(n_s, \bar{x}_s)$; coordinator sets $\bar{x} = \frac{1}{D} \sum_s n_s \bar{x}_s$, broadcasts
 - 3: **for** each site s in parallel **do**
 - 4: $U_s \Sigma_s V_s^\top \leftarrow \text{SVD}(X_s - \bar{x})$; send $M_s = \Sigma_s^{(L)} V_s^{(L)\top}$
 - 5: **end for**
 - 6: coordinator: $V_{\text{fed}} \leftarrow$ top- K right vectors of $[M_1; \dots; M_S]$
 - 7: each site computes donor coordinates $Z_s = (X_s - \bar{x}) V_{\text{fed}}$ *locally*
 - 8: **Output:** program basis V_{fed}
-

3.3. Complexity and communication

Each site computes a rank- L truncated SVD at cost $O(n_s C G L)$ and transmits $L \times C G$ values per round *independent of its cell count*; the coordinator’s merge is $O(S L C G K)$. A raw expression matrix has approximately $\text{cells} \times G$ entries, in addition to cell and donor metadata. The procedure uses one communication round, and its result is approximate whenever local components are truncated.

3.4. Cell-type-incomplete federation

When sites profile *different* cell-type panels, the (cell-type $\times$ gene) feature space differs per site and fixed-feature federated PCA does not apply. This is a *vertical* federated setting (shared donors, partitioned features) [25]. We exploit the shared donor mode: with donors common across sites and disjoint cell-type panels, each site contributes a donor-Gram block $\Gamma_s = (X_s - \bar{x}_s)(X_s - \bar{x}_s)^\top$ over *its* cell-type columns, where here $\bar{x}_s$ is the mean over all donors of site s ’s columns (distinct from the complete-panel per-site mean of Section 3.2). The coordinator sums $\Gamma = \sum_s \Gamma_s = (X - \bar{x})(X - \bar{x})^\top$ and eigendecomposes it to the left singular vectors U and donor coordinates

Algorithm 2 Incomplete-panel recovery (vertical FL, donor-mode coupling)

- 1: **Input:** shared donors; site cell-type panels $\{C_s\}$ (disjoint, $\cup_s C_s = \text{all types}$)
 - 2: **for** each site s in parallel **do**
 - 3: $\Gamma_s \leftarrow (X_s - \bar{x}_s)(X_s - \bar{x}_s)^\top$ over its columns C_s $\triangleright D \times D$
 - 4: **end for**
 - 5: coordinator (HE/MPC): $\Gamma \leftarrow \sum_s \Gamma_s$; $U\Sigma^2U^\top \leftarrow \text{rank-}K \text{ eigendecomposition of } \Gamma$
 - 6: **for** each site s in parallel **do**
 - 7: $V_s \leftarrow (X_s - \bar{x}_s)^\top U\Sigma^{-1}$ $\triangleright$ orthonormal right singular vectors; only loadings revealed
 - 8: **end for**
 - 9: coordinator concatenates $\{V_s\}$ into the full $C \times G$ program V
 - 10: **Output:** full multicellular program V (no site observed all C)
-

$Z = U\Sigma$, after which each site returns the loadings for its own cell types and the coordinator concatenates them into the full $C \times G$ program, spanning cell types no single site observed. Because the sites partition the columns over a shared donor set, $\bar{x}_s$ is the global column mean restricted to site s 's columns, so the summed Gram *equals* the centralized Gram and recovery is exact *by construction*. Recovering cross-panel programs carries a privacy cost, since the donor Gram exposes donor-level structure and requires DP or homomorphic encryption [26, 27]. It also presumes donors are matched and consistently ordered across sites (a private-set-intersection step) under harmonized cell-type annotations (Algorithm 2).

3.5. Privacy model and membership inference

Threat model. We assume honest-but-curious parties: all follow the protocol but may inspect any message they receive. The only quantities leaving a site are per-site means and the low-dimensional summaries M_s

(or Gram blocks Γ_s); raw cells and donor coordinates Z never leave a site. The complete-panel coordinator must obtain the individual M_s to build the stacked matrix. Additive secure aggregation [10] reveals only $\sum_s M_s$, and $(\sum_s M_s)^\top (\sum_s M_s)$ contains cross-site terms, so it does not implement the stack. Algorithm 1 therefore has low communication but exposes site-level summaries unless a separate MPC or homomorphic-encryption protocol performs the stacked SVD. Additive secure aggregation does form $\sum_s \Gamma_s$ in the incomplete-panel algorithm, though each Γ_s is donor-indexed, and that protocol must further specify which parties learn U , Σ and V_s .

Membership inference. Hiding the intermediate messages does not by itself bound what the *released* program reveals about a donor. We quantify this with a reconstruction-residual membership-inference attack [13, 28]. For a released $CG \times K$ subspace V fit on a member set $\mathcal{M}$, define a donor’s projection residual $\rho(x) = \|(x - \bar{x}) - (x - \bar{x})VV^\top\|_2$. The attacker already holds the candidate donor vector x and tests whether it contributed to fitting V . Donors in $\mathcal{M}$ tend to have smaller ρ than held-out non-members; the membership score is $-\rho$, and the attack AUC over 30 balanced member/non-member splits measures leakage (0.5 = none). We evaluate four release objects: (i) per-donor coordinates Z ; (ii) site-specific subspaces $\{V_s\}$; (iii) the merged global subspace V_{fed} ; and (iv) the incomplete-panel donor Gram Γ , whose eigenvectors and eigenvalues reconstruct $Z = U\Sigma$. These objects carry different information, so their attack AUCs are not measurements under one threat model.

3.6. Data and evaluation

We use seven public atlases. **SLE**: Perez et al. [29], 1.26 M PBMCs from 261 donors (162 SLE, 99 control), 11 cell types, Asian and European ancestry. **COVID-19**: Stephenson et al. [30], PBMCs from 130 donors across three UK centres (Newcastle, Cambridge, Sanger); we use the 117 donors with initial-timepoint samples, 18 compartments, and a severity contrast over the 73 severity-graded donors (22 severe/critical/death vs. 51 asymptomatic/mild/moderate). **IBD**: the IBDverse Crohn’s-disease gut atlas [31], 9 compartments, used for the cross-disease meta-analysis and the complementarity map. **ILD**: the interstitial-lung-disease atlas of Natri et al. [32] (109 donors and 33 cell types after the study-specific quality-control and inclusion filters, 63 ILD vs. 46 normal), the most strongly cross-cell-type-coordinated cohort. **Liver**: the PSC/PBC fibrosing-cholangiopathy atlas of Andrews et al. [33] (16 donors, 10 PSC/PBC vs. 6 normal, 27 cell types; caudate-lobe scRNA). **Kidney**: the KPMP adult-kidney atlas of Lake et al. [34] (66 donors, 24 cell types; 18 chronic kidney disease and 17 acute kidney injury vs. 31 normal; CZ CELLxGENE collection bcb61471). **Cardiomyopathy**: the DCM/ACM heart atlas of Reichart et al. [35] (79 donors, 10 cell types; 61 cardiomyopathy vs. 18 normal; CZ CELLxGENE collection e75342a8). All are processed identically (Eq. 1) [36]; cell types follow each study’s published broad annotation, and the gut and lung atlases use their respective tissues only.

Recovery is assessed by three metrics. (i) *Subspace canonical correlation* between the centralized and federated program bases $V_c, V_f \in \mathbb{R}^{CG \times K}$: the mean of the cosines of the principal angles, i.e. the mean singular value

of $Q_c^\top Q_f$ for orthonormal bases Q (1.0 = identical subspace). *(ii) Program face validity:* for a joint loading $V_k \in \mathbb{R}^{C \times G}$, the signed per-gene loading is $\ell_{gk} = C^{-1} \sum_{c=1}^C V_{k,cg}$. We then apply a Mann-Whitney test of whether $|\ell_{gk}|$ is higher for a curated interferon-stimulated-gene (ISG) set than for the remaining genes, reported as the corresponding AUC. *(iii) Phenotype separation:* the ROC AUC of the disease program’s donor scores against the clinical label (SLE case/control; COVID severe/critical/death vs. asymptomatic/mild/moderate), with donor-bootstrap 95% confidence intervals (300 resamples) on all AUCs, Δ AUC, and subspace correlations. We additionally report a *label-free* five-fold cross-validation in which the disease program is chosen on each training fold by ISG enrichment alone (never by the label) and scored on the held-out fold (per-fold AUC, sign-resolved). The component sign is determined from training data only, and preprocessing, component selection, single-cell-type selection and hyperparameter selection are all repeated within each training fold.

4. Results

4.1. Centralized recovery and biological validity

On the SLE atlas the decomposition yields a dominant multicellular program: a single donor-mode factor whose $C \times G$ loading spans all eleven cell types, peaking on classical monocytes but with coherent positive loadings across B, T, NK, and dendritic compartments, a genuinely *multicellular* signature of coordinated inter-individual variation. Its donor scores separate SLE from control at AUC 0.958 (donor-bootstrap 95% CI [0.79, 0.98]; the absolute value carries real uncertainty, whereas the federated-minus-centralized

difference below is far more tightly constrained), and its per-gene loading ℓ_{gk} is strongly enriched for ISGs (enrichment AUC 0.998, $p = 7.4 \times 10^{-13}$). This monocyte peak is the established hallmark of the lupus type-I interferon response, matching centralized scITD [1] and giving a ground-truth target.

4.2. Federated recovery closely approximates the centralized analysis

The federated estimator closely approximates the centralized program in every evaluated regime (Table 3, Fig. 2). Donor-bootstrap analysis gives $\Delta\text{AUC} = 0.000$ (95% CI $[-0.004, +0.012]$; the interval contains zero, so no difference is detectable under this analysis and equivalence is unproven) and subspace correlation 0.967 (CI $[0.92, 0.99]$). Selecting the program *without labels* (by ISG enrichment on training folds) and evaluating on held-out donors gives AUC 0.960. Recovery also holds under partitions that stress the estimator. A realistic *multi-ancestry* split places Asian-ancestry donors (78% SLE) at one site and the rest (51% SLE) at another, the label skew real consortia face, yet recovery is unchanged (AUC 0.958, correlation 0.981, ISG enrichment 0.998). An adversarial *processing-cohort* split, one site 0% cases and another 90%, is likewise recovered (AUC 0.956, correlation 0.980): because every site centres on the federated global mean, the merge is insensitive to such site-label confounding.

On the COVID-19 atlas we federate multicellular-program recovery across the cohort’s three *original* institutional collection sites with no simulated partitions: Newcastle (61 donors), Cambridge (45), and the Wellcome Sanger Institute (11). Federation recovers the centralized severity-program subspace at correlation 0.989 (severity program cross-validated AUC 0.678

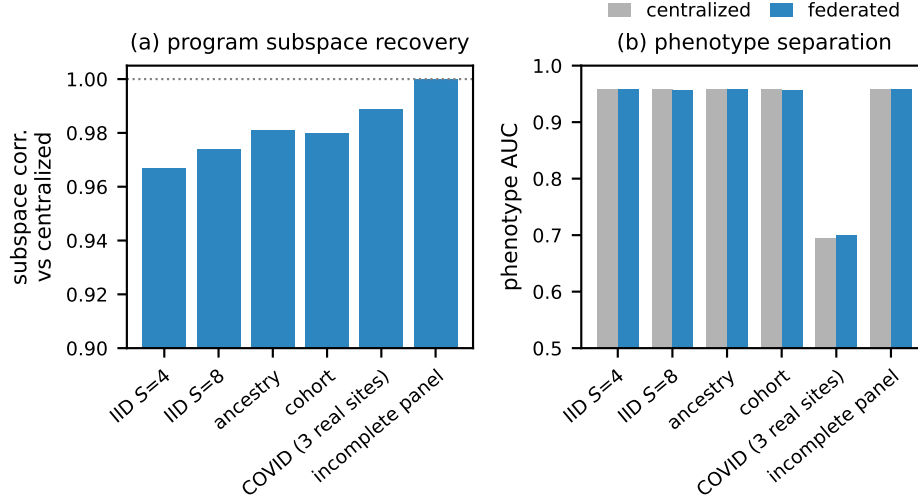


Figure 2: Federated recovery across regimes. (a) Subspace correlation to the centralized program (≥ 0.96 , 1.000 for the exact incomplete-panel scheme). (b) Phenotype-separation AUC, federated vs. centralized (COVID is the weaker severity program).

over the 73 severity-graded donors). The smallest site ($n=11$) contributes a rank-deficient local subspace yet does not degrade the merge, since the stacked SVD pools rank across sites.

4.3. Recovery from cell-type-incomplete sites

Partitioning the 11 SLE cell types into disjoint panels of ≤ 3 types per site, the donor-gram estimator recovers the full program (donor-score AUC 0.958, matching the centralized 0.958 and the best single site 0.954). The subspace correlation of 1.000 is exact *by construction* under complete donor alignment, consistent centering and complete numerical panels, and should be read as an algebraic identity. Two baselines place it. Standard fixed-feature federated PCA cannot run at all: with disjoint panels the cell types common to every site are empty, so there is no shared feature space to

align. Per-site PCA followed by fusion over the shared donors does run but only *approximately* recovers the centralized program (subspace correlation 0.977, against 1.000 for the donor-gram coupling), because each site ranks components within its own panel and loses a program that is weak per panel but strong once cell types are combined.

4.4. Privacy: membership inference

Table 4 reports membership-inference AUC for the reconstruction-residual attack over 30 splits (0.5 = no leakage). A coordinator given the individual complete-panel site subspaces obtains AUC 0.906, whereas the attack on the merged global subspace obtains 0.609: the global release leaks less than the site-specific releases, though that does not establish low membership risk in general. For the *incomplete-panel* scheme, revealing the donor Gram gives attack AUC 1.000, since its eigendecomposition recovers the donor coordinates exactly (correlation 1.000). A loadings-only release gives 0.609 under the same attack, but producing it without revealing Γ , U or Z needs a concrete MPC or homomorphic-encryption protocol.

Two further analyses sharpen the picture (Fig. 3). First, a complementary *attribute-inference* attack: training a classifier from donor scores on the released program to predict a sensitive attribute, the program leaks self-reported ancestry at AUC 0.81 and sex at 0.65. The interferon program covaries with ancestry in this cohort and no secure-computation protocol removes that, so consortia should treat both the messages and the released program as attribute-revealing. Second, the site-specific-subspace membership leakage *grows as sites shrink*: MIA-AUC rises from 0.61 at 130 donors/site

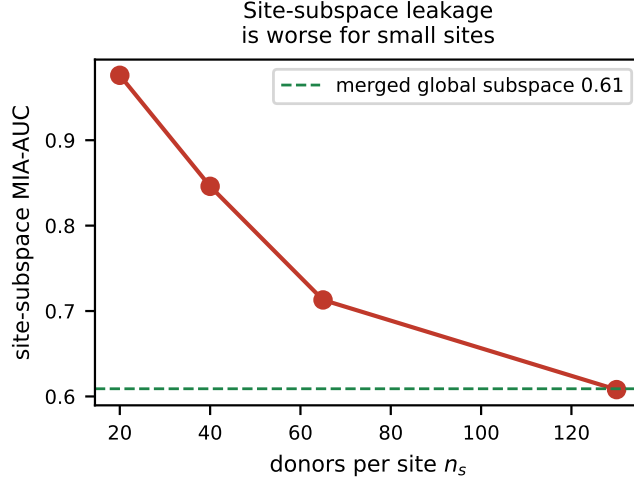


Figure 3: Membership-inference AUC for site-specific subspaces increases as site size decreases (mean over 30 member/non-member splits). The dashed line is the attack AUC for the merged global subspace (0.61).

to 0.98 at 20 donors/site, because small local subspaces overfit their members. A secure stacked-SVD protocol hiding the individual site summaries is therefore most valuable in the many-small-site federations that motivate this work.

4.5. When multicellularity matters: a cross-tissue map

The contribution is faithful *recovery*; whether the recovered program also *out-predicts* the best single cell type is a separate, tissue-dependent question. We map it across the cohorts by nested donor-stratified five-fold cross-validation (the predictive factor is selected on training donors only and scored on held-out donors), benchmarking the multicellular program against the best single-compartment classifier selected the same way, with donor-bootstrap 95% confidence intervals on the gap.

Multicellularity wins where cross-cell-type coordination is the biology. On the interstitial-lung-disease atlas [32] the multicellular program attains cross-validated AUC 0.959 against 0.910 for the best single compartment, a gap of +0.049 whose 95% CI [+0.006, +0.101] excludes zero and that is positive in all 10 cross-validation seeds; the winning factor spans ≈ 24 effective cell types, an alveolar-epithelial, macrophage and endothelial fibrotic axis that no single compartment carries. The federated estimator recovers this program faithfully (subspace correlation 0.981 to centralized, AUC 0.969), so the predictive advantage survives federation.

A second fibrotic tissue corroborates this pattern. On the fibrosing-cholangiopathy liver atlas [33] the multicellular program attains nested donor-stratified cross-validated AUC 0.975 (10-seed mean), with a positive multi-minus-single gap in all 10 seeds; a leave-one-donor-out reanalysis gives AUC 1.000, and a 200-permutation label-shuffle gives empirical $p = 0.005$ (the matched five-fold permutation gives $p = 0.010$), showing performance above that obtained with shuffled labels under the implemented pipeline. The leading factor spans ≈ 23 effective cell types, an immune-stromal fibrotic niche consistent with the lung result. We read this as corroboration short of a co-equal anchor: the cohort is small (16 donors, only 6 controls), the single-cell-type baseline is unstable at this n (0.56 ± 0.20 over seeds), so we lead on the multicellular separation and the permutation p in preference to the raw gap, and no independent open replication was obtainable. A near-perfect AUC at $n = 16$ cannot exclude a confounder perfectly correlated with PSC-vs-normal, since the permutation test addresses overfitting and leaves confounding open, though the broad program and the separation

holding for pure PSC-vs-normal (8 vs 6 donors, leave-one-donor-out AUC 1.000) argue against a narrow technical artifact.

Elsewhere, where one compartment dominates, the program is at parity. The gap to the best single cell type is +0.013 in SLE blood (multicellular 0.958 vs. classical monocyte 0.945), -0.064 in COVID severity (a single monocyte compartment is better), +0.008 in Crohn’s gut, +0.024 in kidney injury (CI includes zero), and -0.015 in cardiomyopathy (cardiac fibroblast alone is better). Of seven cohorts spanning blood, gut, kidney, heart, lung and liver, the multicellular program exceeds the best single compartment only in the two fibrotic, strongly coordinated tissues.

4.6. Robustness and communication cost

The federated recovery is insensitive to the two main hyperparameters (Table 2). Across rank $K \in \{2, \dots, 20\}$ the disease-program AUC stays 0.955–0.958 and ISG enrichment ≥ 0.997 ; the subspace correlation fluctuates with K because lower-variance components are less stably ordered. Across site counts $S \in \{2, \dots, 32\}$ (down to eight donors per site), the federated AUC remains 0.957–0.958 and subspace correlation ≥ 0.96 , so the stacked-SVD merge tolerates many small, rank-deficient sites. All reported experiments set the local communication rank equal to the final rank, $L = K = 10$. Each site transmits LCG values per round (165,000 for SLE, 270,000 for COVID, 135,000 for IBD at $L=10$). The SLE atlas holds 1.26 M cells over $G=1500$ genes, about 1.9×10^9 raw expression values in total and 4.7×10^8 per site at $S=4$, three orders of magnitude above the transmitted summary. The numerical SVD runtime is low in the evaluated implementation (Fig. 4):

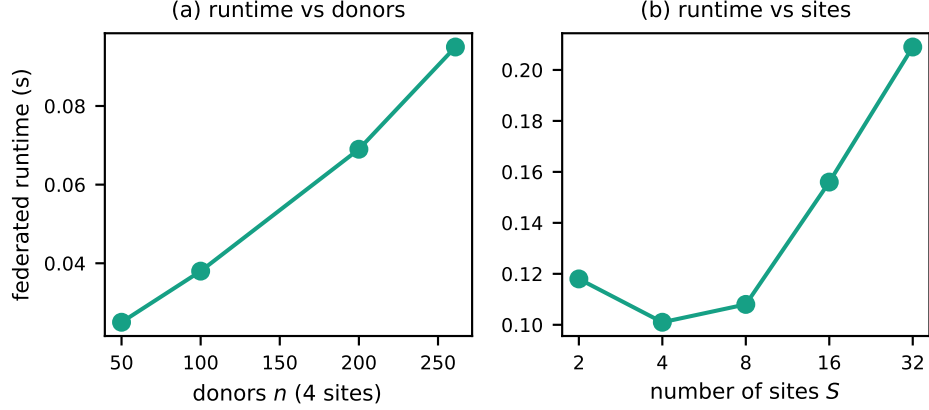


Figure 4: Scalability of the numerical SVD calculations: (a) measured runtime as donor count increases and (b) as site count increases. These values exclude preprocessing, network transfer and cryptographic computation.

on the SLE atlas the per-site SVDs plus merge complete in 0.10 s versus 0.16 s for the centralized SVD, growing linearly in donors (0.025 s at $n=50$ to 0.095 s at $n=261$) and sub-linearly in sites (0.10–0.21 s for $S = 2$ to 32). Global-mean centering keeps the merge robust to site-label confounding, the realistic failure mode of cross-institution federation. Sweeping the case-fraction skew between two sites from balance to extreme (10 seeds; Table 1), it holds the program at AUC between 0.957 and 0.958 throughout, whereas naive per-site centering degrades monotonically from 0.957 at balance to 0.861 when one site is 90% cases. The merged *subspace* stays similar either way (canonical correlation ≥ 0.98); naive centering corrupts the disease program’s donor scores, which each site shifts by a label-correlated mean.

Finally, the donor-mode unfolded PCA matters: replacing it with a trilinear CP-ALS of the same rank collapses recovery to SLE-AUC 0.534 (best-of- K 0.592) versus 0.958 for the SVD, because CP forces a single rank-one gene

Table 1: Confounding robustness (SLE, 2 sites; each entry is the mean over 10 seeds): federated phenotype-AUC vs. case-fraction skew between sites, under federated global-mean vs. naive per-site centering.

site-0 case fraction	0.5	0.6	0.7	0.8	0.9	0.97
global-mean (ours)	0.958	0.957	0.957	0.957	0.957	0.957
naive per-site	0.957	0.956	0.949	0.929	0.861	0.861

Table 2: Ablations (SLE). Centralized AUC = 0.958. The two sweeps draw independent random donor partitions, so a configuration appearing in both panels (e.g. $K=10$ with 4 sites) need not give identical values.

rank K (4 sites)				sites S ($K=10$)		
K	subcorr	fed AUC	ISG	S	subcorr	fed AUC
2	0.989	0.955	0.997	2	0.979	0.958
5	0.935	0.956	0.998	4	0.963	0.957
8	0.979	0.958	0.998	8	0.970	0.957
10	0.967	0.958	0.999	16	0.989	0.957
20	0.968	0.958	0.998	32	1.000	0.958

pattern shared across cell types and underfits the donor axis. Letting each program carry a full $C \times G$ loading is essential. With one unrestricted joint cell-type-by-gene loading per component, the fitted model is donor-mode unfolded PCA rather than a Tucker decomposition with separate donor, cell-type and gene factors and a core tensor; the contribution here is the federated estimation procedure. The per-gene loading used for enrichment and cross-disease comparisons is the cell-type average ℓ_{gk} .

4.7. Cross-disease federated meta-analysis

The framework can also identify programs *shared across diseases* held at different institutions. Treating the three atlases as three disease-sites, we harmonize them to four immune compartments (B, T, monocyte, plasma; present in all annotations) and 800 top-variable genes from the 22,644

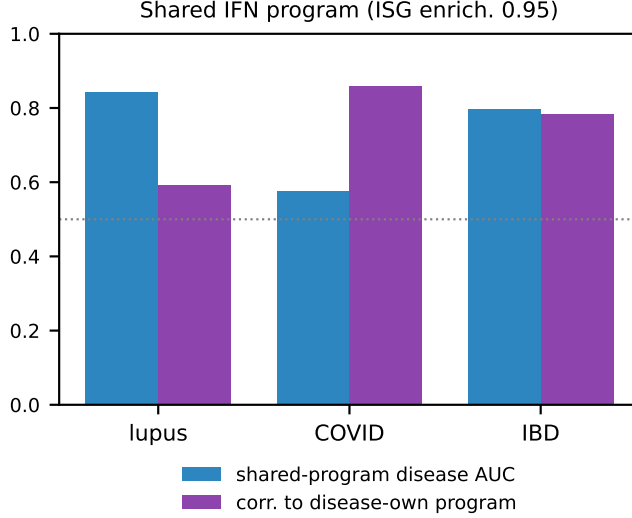


Figure 5: Cross-disease federated meta-analysis. A shared interferon-type multicellular program (ISG enrichment 0.95), recovered federated across lupus, COVID and IBD as three disease-sites, associates with disease (blue) and aligns with each disease’s own program (purple).

shared by all three, and federate (donors are fully disjoint across diseases). The merge recovers a single *shared multicellular program*, strongly interferon-typed (ISG enrichment AUC 0.948), that associates with disease in lupus (AUC 0.842) and Crohn’s inflammation (AUC 0.796), more weakly in COVID severity (AUC 0.576), and whose per-gene loading ℓ_{gk} aligns with each disease’s own program (correlation 0.59/0.86/0.78 for lupus/COVID/IBD; Fig. 5). A shared interferon-type immune axis is thus recoverable across the three disease cohorts without pooling cells. We report this as a feasibility demonstration. The weak COVID association indicates the shared axis captures interferon biology common to the three and carries little COVID-specific severity signal.

Table 3: Federated vs. centralized recovery (SLE case/control AUC; COVID severity program). Centralized SLE-AUC = 0.958. The COVID figure is the in-sample best-program AUC; its cross-validated severity AUC is 0.678.

Partition	Sites	Fed. AUC	Subspace corr.	ISG enr.
SLE, IID	4	0.958	0.967	0.999
SLE, IID	8	0.957	0.974	0.998
SLE, multi-ancestry	2	0.958	0.981	0.998
SLE, processing cohort	4	0.956	0.980	0.998
COVID, real sites	3	0.700	0.989	n/a

Table 4: Membership-inference attack AUC by release object (0.5 = no leakage; reconstruction-residual attack; mean over 30 member/non-member splits, SLE).

Released object	MIA-AUC (95% CI)
Complete-panel: site-specific subspaces	0.906 [0.84, 0.96]
Complete-panel: merged global subspace	0.609 [0.55, 0.68]
Incomplete-panel: loadings only (Γ in HE/MPC)	0.609
Incomplete-panel: donor Gram revealed	1.000
(<i>reference</i>) per-donor coordinates Z	1.000

5. Discussion

Local rank truncation prevents a general equality claim, and standard additive secure aggregation does not implement the stacked-SVD merge.

Practical implications. Multicellular-program analyses such as scITD can now run across multi-institution and multi-ancestry cohorts without moving cells: a consortium agrees a gene panel and rank, exchanges site-level program summaries, and obtains a close approximation of the pooled result.

Three limitations bound the contribution. *First, the predictive value of multicellularity is tissue-dependent.* On compartment-dominated blood, gut, kidney and heart cohorts the dominant disease program is also legible from a single cell type, so there the contribution is faithful federated *recovery* and

the incomplete-panel capability, with no added prediction. Where cross-cell-type coordination is strong, multicellularity *does* add predictive power (interstitial lung disease, §4.5); whether that regime is common across diseases remains open, demonstrated here in two fibrotic tissues, of which only the lung cohort ($n = 109$) is well powered. *Second, formal differential privacy is out of reach at current cohort sizes.* Donor-level DP [14] on the released program requires $n = \tilde{O}(d)$ samples in the ambient program dimension [18, 19, 37], whereas single-cell atlases have $n \approx 10^2$, and on these cohorts the noisy power method fails to recover the program at any meaningful budget. Formal DP, accounted under zero-concentrated DP composition [38], becomes realistic only at consortium scale ($n \gtrsim 10^3$). *Third, the incomplete-panel capability requires stronger cryptography:* revealing the donor Gram exposes donor coordinates (MIA-AUC 1.000). No secure stacked-SVD protocol and no end-to-end cryptographic systems experiment are implemented, so the reported complete-panel communication of $L \times CG$ values per site assumes summaries are transmitted directly and excludes the cost of MPC or homomorphic encryption.

Future work. Two settings are natural: (i) *multi-omic coupling*, where paired modalities (RNA and chromatin accessibility, or surface protein) couple through shared regulatory structure, so the vertical-FL form of our incomplete-panel scheme applies directly; and (ii) *spatial niches*, where a coordinated multi-cell-type signature is the object of interest by construction. Methodologically, a differentially private or homomorphic-encryption backend for the donor-Gram coupling [39] and confounding-robust aggregation for severe site-label skew are the main open directions. The multicellular-

tensor view is complementary to MOFA [15] and cell-cell communication tensors [2].

6. Conclusion

On a systemic lupus erythematosus atlas the federated estimator recovers the canonical interferon multicellular program and closely approximates the centralized decomposition, with the residual difference set by the local components each site omits. The approximation holds across institution-scale and multi-ancestry partitions and the three original collection sites of a COVID-19 atlas. When no site observes all cell types, donor-mode coupling still recovers the full program (subspace correlation 1.000, exact given aligned donors and consistent centering), which fixed-feature federated PCA cannot. Membership inference gives attack AUC 0.906 for site-specific subspaces and 0.609 for the merged global subspace, whose release on its own requires a secure stacked-SVD protocol, since the merge is not an additive aggregation. Where the biology is intrinsically multicellular (interstitial lung disease, corroborated in a fibrosing-cholangiopathy liver cohort) the recovered program also out-predicts the best single cell type; elsewhere it stays at parity while still being recovered faithfully. A cryptographic protocol matched to the stacked-SVD merge, and formal privacy for the released program, are the open steps toward multi-cohort single-cell immunology.

CRediT authorship contribution statement

Axel Faes: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original

draft, Writing – review & editing. **Maryam Amir Haeri:** Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics

No new human data were collected. All data are public, de-identified single-cell atlases accessed through CZ CELLxGENE.

Funding

This work was conducted within the project “Predicting the Development of Primary Liver Cancer Using the AI-HCC Model”, which is funded by ZonMw, The Netherlands Organisation for Health Research and Development.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used a generative AI assistant to support drafting and editing of the manuscript text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data availability

The code and the scripts that regenerate every table and figure are at <https://github.com/TheAxeC/federated-multicellular-tensor> (MIT licence). All data are public atlases on CZ CELLxGENE with the collection identifiers SLE 436154da, COVID-19 ddfad306, IBDverse 7f7fdf50, PSC/PBC liver 0c8a364b, ILD 07e12576, kidney bcb61471 and cardiomyopathy e75342a8.

References

- [1] J. Mitchel, M. G. Gordon, R. K. Perez, E. Biederstedt, R. Bueno, C. J. Ye, P. V. Kharchenko, Coordinated, multicellular patterns of transcriptional variation that stratify patient cohorts are revealed by tensor decomposition, *Nature Biotechnology* 43 (2025) 1192–1201. doi:10.1038/s41587-024-02411-z.
- [2] E. Armingol, H. M. Baghdassarian, C. Martino, A. Perez-Lopez, C. Aamodt, R. Knight, N. E. Lewis, Context-aware deconvolution of cell–cell communication with Tensor-cell2cell, *Nature Communications* 13 (2022) 3665. doi:10.1038/s41467-022-31369-2.
- [3] A. Ramirez, B. T. Orcutt-Jahns, S. Pascoe, A. Abraham, B. Remigio, N. Thomas, A. S. Meyer, Integrative, high-resolution analysis of single-cell gene expression across experimental conditions with PARAFAC2-RISE, *Cell Systems* 16 (6) (2025) 101294. doi:10.1016/j.cels.2025.101294.
- [4] A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, A. Farmer, L. Fugger, B. Göttgens, N. Hacohen, M. Haniffa, M. Hemberg, S. Kim, P. Klenerman, A. Kriegstein, E. Lein, S. Linnarsson, E. Lundberg, J. Lundeberg, P. Majumder, J. C. Marioni, M. Merad, M. Mhlanga, M. Nawijn, M. Netea, G. Nolan, D. Pe’er, A. Phillipakis, C. P. Ponting, S. Quake, W. Reik, O. Rozenblatt-Rosen, J. Sanes, R. Satija, T. N. Schumacher, A. Shalek, E. Shapiro, P. Sharma, J. W. Shin, O. Stegle, M. Stratton, M. J. T. Stubbington, F. J. Theis, M. Uhlen, A. van

- Oudenaarden, A. Wagner, F. Watt, J. Weissman, B. Wold, R. Xavier, N. Yosef, Human Cell Atlas Meeting Participants, The Human Cell Atlas, *eLife* 6 (2017) e27041. doi:10.7554/eLife.27041.
- [5] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R. M. Summers, A. Trask, D. Xu, M. Baust, M. J. Cardoso, The future of digital health with federated learning, *npj Digital Medicine* 3 (2020) 119. doi:10.1038/s41746-020-00323-1.
- [6] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, B. Agüera y Arcas, Communication-efficient learning of deep networks from decentralized data, in: *Proc. 20th Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, PMLR 54, 2017, pp. 1273–1282. doi:10.48550/arXiv.1602.05629.
URL <https://proceedings.mlr.press/v54/mcmahan17a.html>
- [7] M. Bakhtiari, S. Bonn, F. Theis, O. Zolotareva, J. Baumbach, FedscGen: privacy-preserving federated batch effect correction of single-cell RNA sequencing data, *Genome Biology* 26 (1) (2025) 216. doi:10.1186/s13059-025-03684-6.
- [8] S. Sav, J.-P. Bossuat, J. R. Troncoso-Pastoriza, M. Claassen, J.-P. Hubaux, Privacy-preserving federated neural network learning for disease-associated cell classification, *Patterns* 3 (5) (2022) 100487. doi:10.1016/j.patter.2022.100487.
- [9] Y. A. Choi, Y. Kim, P. Miao, T. Lappalainen, G. Gürsoy, Secure and federated quantitative trait loci mapping with privateQTL, *Cell Genomics* 5 (2) (2025) 100769. doi:10.1016/j.xgen.2025.100769.
- [10] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, K. Seth, Practical secure aggregation for privacy-preserving machine learning, in: *Proc. 2017 ACM SIGSAC Conf. on Computer and Communications Security (CCS)*, 2017, pp. 1175–1191. doi:10.1145/3133956.3133982.
- [11] J. Fan, D. Wang, K. Wang, Z. Zhu, Distributed estimation of principal eigenspaces, *Annals of Statistics* 47 (6) (2019) 3009–3031. doi:10.1214/18-AOS1713.
- [12] A. Grammenos, R. Mendoza-Smith, J. Crowcroft, C. Mascolo, Federated principal component analysis, in: *Advances*

- in Neural Information Processing Systems (NeurIPS) 33, 2020. doi:10.48550/arXiv.1907.08059.
URL <https://proceedings.neurips.cc/paper/2020/hash/47a658229eb2368a99f1d032c8848542-Abstract.html>
- [13] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: Proc. 2017 IEEE Symposium on Security and Privacy (S&P), 2017, pp. 3–18. doi:10.1109/SP.2017.41.
 - [14] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, Foundations and Trends in Theoretical Computer Science 9 (3–4) (2014) 211–487. doi:10.1561/04000000042.
 - [15] R. Argelaguet, D. Arnol, D. Bredikhin, Y. Deloro, B. Velten, J. C. Marioni, O. Stegle, MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data, Genome Biology 21 (2020) 111. doi:10.1186/s13059-020-02015-1.
 - [16] L. Jerby-Arnon, A. Regev, DIALOGUE maps multicellular programs in tissue from single-cell or spatial transcriptomics data, Nature Biotechnology 40 (2022) 1467–1477. doi:10.1038/s41587-022-01288-0.
 - [17] J. Ding, J. Lin, S. Jiang, Y. Wang, Z. Miao, Z. Fang, J. Tang, M. Li, X. Qiu, Toward a privacy-preserving predictive foundation model of single-cell transcriptomics with federated learning and tabular modeling, bioRxiv preprint 2025.01.06.631427 (2025). doi:10.1101/2025.01.06.631427.
 - [18] M. Hardt, E. Price, The noisy power method: A meta algorithm with applications, in: Advances in Neural Information Processing Systems (NeurIPS), Vol. 27, 2014, pp. 2861–2869. doi:10.48550/arXiv.1311.2495.
 - [19] X. Liu, W. Kong, P. Jain, S. Oh, DP-PCA: Statistically optimal and differentially private PCA, in: Advances in Neural Information Processing Systems (NeurIPS), 2022. doi:10.48550/arXiv.2205.13709.
 - [20] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, D. W. Craig, Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays, PLoS Genetics 4 (8) (2008) e1000167. doi:10.1371/journal.pgen.1000167.

- [21] C. R. Walker, X. Li, M. Chakravarthy, W. Lounsbery-Scaife, Y. A. Choi, R. Singh, G. Gürsoy, Private information leakage from single-cell count matrices, *Cell* 187 (23) (2024) 6537–6549.e10. doi:10.1016/j.cell.2024.09.012.
- [22] H. L. Crowell, C. Soneson, P.-L. Germain, D. Calini, L. Collin, C. Raposo, D. Malhotra, M. D. Robinson, muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data, *Nature Communications* 11 (2020) 6077. doi:10.1038/s41467-020-19894-4.
- [23] T. G. Kolda, B. W. Bader, Tensor decompositions and applications, *SIAM Review* 51 (3) (2009) 455–500. doi:10.1137/07070111X.
- [24] L. De Lathauwer, B. De Moor, J. Vandewalle, A multilinear singular value decomposition, *SIAM Journal on Matrix Analysis and Applications* 21 (4) (2000) 1253–1278. doi:10.1137/S0895479896305696.
- [25] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, *ACM Transactions on Intelligent Systems and Technology* 10 (2) (2019) 12:1–12:19. doi:10.1145/3298981.
- [26] C. Gentry, Fully homomorphic encryption using ideal lattices, in: *Proc. 41st Annual ACM Symposium on Theory of Computing (STOC)*, 2009, pp. 169–178. doi:10.1145/1536414.1536440.
- [27] J. H. Cheon, A. Kim, M. Kim, Y. Song, Homomorphic encryption for arithmetic of approximate numbers, in: *Advances in Cryptology – ASIACRYPT 2017, Part I*, LNCS 10624, 2017, pp. 409–437. doi:10.1007/978-3-319-70694-8_15.
- [28] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, F. Tramèr, Membership inference attacks from first principles, in: *Proc. 2022 IEEE Symposium on Security and Privacy (S&P)*, 2022, pp. 1897–1914. doi:10.1109/SP46214.2022.9833649.
- [29] R. K. Perez, M. G. Gordon, M. Subramaniam, M. C. Kim, G. C. Hartoularos, S. Targ, Y. Sun, A. Ogorodnikov, R. Bueno, A. Lu, M. Thompson, N. Rappoport, A. Dahl, C. M. Lanata, M. Matloubian, L. Maliskova, S. S. Kwek, T. Li, M. Slyper, J. Waldman, D. Dionne, O. Rozenblatt-Rosen, L. Fong, M. Dall’Era, B. Balliu, A. Regev, J. Yazdany, L. A. Criswell, N. Zaitlen, C. J. Ye, Single-cell RNA-seq reveals

cell type-specific molecular and genetic associations to lupus, *Science* 376 (6589) (2022) eabf1970. doi:10.1126/science.abf1970.

- [30] E. Stephenson, G. Reynolds, R. A. Botting, F. J. Calero-Nieto, M. D. Morgan, Z. K. Tuong, K. Bach, W. Sungnak, K. B. Worlock, M. Yoshida, N. Kumasaka, K. Kania, J. Engelbert, B. Olabi, J. S. Spengarová, N. K. Wilson, N. Mende, L. Jardine, L. C. S. Gardner, I. Goh, D. Horsfall, J. McGrath, S. Webb, M. W. Mather, R. G. H. Lindeboom, E. Dann, N. Huang, K. Polanski, E. Prigmore, F. Gothe, J. Scott, R. P. Payne, K. F. Baker, A. T. Hanrath, I. C. D. Schim van der Loeff, A. S. Barr, A. Sanchez-Gonzalez, L. Bergamaschi, F. Mescia, J. L. Barnes, E. Kilich, A. de Wilton, A. Saigal, A. Saleh, S. M. Janes, C. M. Smith, N. Gopee, C. Wilson, P. Coupland, J. M. Coxhead, V. Y. Kiselev, S. van Dongen, J. Bacardit, H. W. King, Cambridge Institute of Therapeutic Immunology and Infectious Disease-National Institute of Health Research (CITIID-NIHR) COVID-19 BioResource Collaboration, A. J. Rostron, A. J. Simpson, S. Hambleton, E. Laurenti, P. A. Lyons, K. B. Meyer, M. Z. Nikolić, C. J. A. Duncan, K. G. C. Smith, S. A. Teichmann, M. R. Clatworthy, J. C. Marioni, B. Göttgens, M. Haniffa, Single-cell multi-omics analysis of the immune response in COVID-19, *Nature Medicine* 27 (5) (2021) 904–916. doi:10.1038/s41591-021-01329-2.
- [31] T. Alegbe, B. T. Harris, L. Fachal, L. Ramirez-Navarro, M. Tutert, M. Krzak, M. Ghouraba, M. Strickland, M. Ozols, C. E. Cohen, S. Khullar, E. Khabirova, N. I. Panousis, D. Ochoa, N. Wana, M. X. Hu, J. Skelton, J. Ostermayer, K. A. X. Cheam, D. Leland Taylor, Y. Gu, C. Dawson, T. Thompson, K. Arestang, N. Nishad, B. Brezina, C. Q. Caballes, W. Garri, S. Leonard, V. Iyer, M. Parkes, C. Wallace, R. E. McIntyre, C. C. Martin, G.-R. Jones, T. Raine, C. A. Anderson, Cell-type-resolved genetic variation shapes inflammatory bowel disease risk, *Nature* 656 (8126) (2026) 129–139. doi:10.1038/s41586-026-10627-z.
- [32] H. M. Natri, C. B. Del Azodi, L. Peter, C. J. Taylor, S. Chugh, R. Kendle, M.-i. Chung, D. K. Flaherty, B. K. Matlock, C. L. Calvi, T. S. Blackwell, L. B. Ware, M. Bacchetta, R. Walia, C. M. Shaver, J. A. Kropski, D. J. McCarthy, N. E. Banovich, Cell-type-specific and disease-associated expression quantitative trait loci in the human lung, *Nature Genetics* 56 (4) (2024) 595–604. doi:10.1038/s41588-024-01702-0.
- [33] T. S. Andrews, D. Nakib, C. T. Perciani, X. Z. Ma, L. Liu, E. Win-

- ter, D. Camat, S. W. Chung, P. Lumanto, J. Manuel, S. Mangroo, B. Hansen, B. Arpinder, C. Thoeni, B. Sayed, J. Feld, A. Gehring, A. Gulamhusein, G. M. Hirschfield, A. Ricciuto, G. D. Bader, I. D. McGilvray, S. MacParland, Single-cell, single-nucleus, and spatial transcriptomics characterization of the immunological landscape in the healthy and PSC human liver, *Journal of Hepatology* 80 (5) (2024) 730–743. doi:10.1016/j.jhep.2023.12.023.
- [34] B. B. Lake, R. Menon, S. Winfree, Q. Hu, R. Melo Ferreira, K. Kalthor, D. Barwinska, E. A. Otto, M. Ferkowicz, D. Diep, N. Plongthongkum, A. Knoten, S. Urata, L. H. Mariani, A. S. Naik, S. Eddy, B. Zhang, Y. Wu, D. Salamon, J. C. Williams, X. Wang, K. S. Balderrama, P. J. Hoover, E. Murray, J. L. Marshall, T. Noel, A. Vijayan, A. Hartman, F. Chen, S. S. Waikar, S. E. Rosas, F. P. Wilson, P. M. Palevsky, K. Kiryluk, J. R. Sedor, R. D. Toto, C. R. Parikh, E. H. Kim, R. Satija, A. Greka, E. Z. Macosko, P. V. Kharchenko, J. P. Gaut, J. B. Hodggin, KPMP Consortium, M. T. Eadon, P. C. Dagher, T. M. El-Achkar, K. Zhang, M. Kretzler, S. Jain, An atlas of healthy and injured cell states and niches in the human kidney, *Nature* 619 (7970) (2023) 585–594. doi:10.1038/s41586-023-05769-3.
- [35] D. Reichart, E. L. Lindberg, H. Maatz, A. M. A. Miranda, A. Viveiros, N. Shvetsov, A. Gärtner, E. R. Nadelmann, M. Lee, K. Kanamaru, J. Ruiz-Orera, V. Strohmenger, D. M. DeLaughter, G. Patone, H. Zhang, A. Woehler, C. Lippert, Y. Kim, E. Adami, J. M. Gorham, S. N. Barnett, K. Brown, R. J. Buchan, R. A. Chowdhury, C. Constantinou, J. Cranley, L. E. Felkin, H. Fox, A. Ghauri, J. Gummert, M. Kanda, R. Li, L. Mach, B. McDonough, S. Samari, F. Shahriaran, C. Yapp, C. Stanasiuk, P. I. Theotokis, F. J. Theis, A. van den Bogardt, H. Wakimoto, J. S. Ware, C. L. Worth, P. J. R. Barton, Y.-A. Lee, S. A. Teichmann, H. Milting, M. Nosedá, G. Y. Oudit, M. Heinig, J. G. Seidman, N. Hubner, C. E. Seidman, Pathogenic variants damage cell composition and single cell transcription in cardiomyopathies, *Science* 377 (6606) (2022) eabo1984. doi:10.1126/science.abo1984.
- [36] F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis, *Genome Biology* 19 (2018) 15. doi:10.1186/s13059-017-1382-0.
- [37] V. Singhal, T. Steinke, Privately learning subspaces, in: *Advances in Neural Information Processing Systems* (NeurIPS) 34, 2021.

doi:10.48550/arXiv.2106.00001.

URL <https://proceedings.neurips.cc/paper/2021/hash/09b69adcd7cbae914c6204984097d2da-Abstract.html>

- [38] M. Bun, T. Steinke, Concentrated differential privacy: Simplifications, extensions, and lower bounds, in: Theory of Cryptography (TCC-B), LNCS 9985, 2016, pp. 635–658. doi:10.1007/978-3-662-53641-4_24.
- [39] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep learning with differential privacy, in: Proc. 2016 ACM SIGSAC Conf. on Computer and Communications Security (CCS), 2016, pp. 308–318. doi:10.1145/2976749.2978318.