Membership Inference on Synthetic Single-Cell Genomic Data

Steven Golob[1] **Patrick McKeever**[1] **Sikha Pentyala**[1]
Martine De Cock[1,2] **Jonathan Peck**[2,3]
[1]School of Engineering and Technology, University of Washington Tacoma, USA
[2]Department of Mathematics, Computer Science and Statistics, Ghent University, Belgium
[3]VIB Center for Inflammation Research, Ghent, Belgium
golobs@uw.edu

Abstract

Single-cell RNA sequencing (scRNA-seq) data is subject to strict access control due to its sensitive nature, motivating the use of synthetic data generation (SDG) for privacy-preserving data sharing. We present the first adversarial privacy attack that performs meaningfully above random guessing against state-of-the-art scRNA-seq SDG methods. Our attack enables donor-level membership inference, demonstrating that leading SDG techniques fail to adequately mask which individuals were used to train the generator. We show that privacy leakage increases as the number of training donors decreases. The leak persists even when the adversary holds only cells that the generator never saw, or when the adversary sees *only one cell* from a donor. Although the attack is designed to exploit vulnerabilities in scDesign2, we find that it also succeeds against synthetic data generated by every other high-fidelity method we evaluated, including scDesign3, scVI, and ZINB-WaVE. This transferability indicates that an adversary can infer sensitive information from synthetic data without access to the training procedure, model parameters, or even the underlying generation algorithm. Finally, we add noise to the parameters of scDesign2 as a first-line defense. Noise on the copula covariance brings the standard attack near chance at little cost in fidelity. It does not stop a variant of the attack that also uses the per-gene marginals, and noise on the marginals stops that variant only once fidelity is lost.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) measures gene expression at cellular resolution (a granularity so consequential it was named Nature's Method of the Year [31]) enabling discoveries in transcriptome profiling, personalized medicine, and tumor microenvironment characterization that bulk sequencing cannot [5, 20, 25]. Because scRNA-seq data is tied to individual donors and can reveal sensitive genetic and disease information, it is subject to strict access controls [33, 34, 35]. Indeed, even de-identified single-cell expression matrices can be linked back to individuals: one recent study re-identified donors from scRNA-seq counts via eQTL genotypes with 99.9% accuracy [45]. Because genetic variants are heritable, such exposure risks harm not only to individuals but to their biological relatives [39].

These barriers answer valid privacy concerns, but they have a cost. Strict access policies induce a "privacy chill" among data custodians and slow research [10], in contrast to the rapid workflows behind recent advances in AI. Synthetic data generation (SDG) [18], i.e., training a generative model on real data to produce artificial samples that preserve key statistical properties such as per-gene expression distributions and inter-gene correlations [12, 49], has been proposed as a path to privacy-preserving sharing, with applications to benchmarking, rare cell-type augmentation, and data release [27, 28, 43]. The European Genome Data Infrastructure has already begun releasing SDG-derived cancer genomic data to satisfy EU privacy regulations [37].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

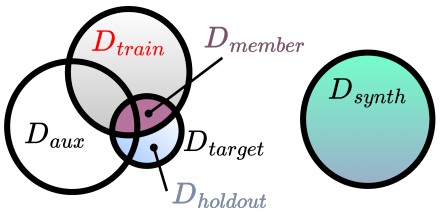

Figure 1: Threat model. In a MIA, the adversary has access to synthetic data D_{synth} but not training data D_{train}, and must infer $D_{train} \cap D_{target}$ from a set of target instances D_{target}. The adversary may or may not have access to auxiliary data D_{aux}.

What is less clear is whether synthetic data actually protects donor privacy. Membership inference attacks (MIAs) that ask whether a specific individual's data was used to train a given model are a standard auditing tool [4, 40]. Yet existing MIAs on synthetic data [11, 13, 42], typically designed for demographic tabular data with dozens of features, fail to perform above random guessing on genomic data with thousands of features [36, 50]. No MIA specifically designed for synthetic genomic data has been shown to succeed. The CAMDA Health Privacy Challenge at ISMB 2025[1] highlighted this gap explicitly, calling for both private scRNA-seq SDG methods and privacy audits, and none of its participants mounted an effective attack. When the challenge ran again at ISMB 2026, scMAMA-MIA was a winner of its single-cell track.

We propose **scMAMA-MIA**, an adaptation of the MAMA-MIA attack [11] to *scDesign2* and *scDesign3* (Gaussian copula variant) [43, 41], the highest-fidelity scRNA-seq SDG methods in the CAMDA2025 evaluation [29]. Both fit a marginal distribution for each gene and cell type and a Gaussian copula over gene correlations, and scMAMA-MIA exploits this structure. It also succeeds against the other high-fidelity generators we evaluated (scDesign3 with a vine copula, scVI [26], and ZINB-WaVE [38]), which suggests that the vulnerability is not specific to one architecture.[2]

On three scRNA-seq datasets, scMAMA-MIA substantially outperforms existing tabular MIAs against scDesign2 (Section 4). We then show something more striking. The adversary does not need the cells that were used for training, because in small cohorts other cells from the same donors also reveal their membership (Section 4.4). We accompany the attack with quality evaluations of these SOTA SDG methods to characterize the privacy/utility trade-off, and a study of how adding noise during generation, as a first-line defense, shifts this trade-off.

2 Problem Description and Threat Models

A membership inference attack (MIA) empirically audits the privacy risk of a system (typically a discriminative or generative model) derived from real data by attempting to infer which real samples were used to train it [40]. We focus on MIAs against generative models: given synthetic dataset D_{synth} and a set of target instances D_{target}, the adversary infers which targets were used to train the generator (see Fig. 1). Membership alone can be sensitive. If a donor's data is found to have trained the generator of a synthetic cancer dataset, this reveals that they have cancer [50].

Synthetic Data Generation An scRNA-seq dataset is a count matrix $D \in \mathbb{Z}_{\geq 0}^{n \times m}$ of n cells by m genes. Each cell (row) belongs to one of $|T|$ cell types and one of P donors (as in "patients"), and D_t and D_p denote the cells of type t and of donor p.

A data holder runs an SDG algorithm on such a dataset D_{train} to create a synthetic dataset D_{synth}. scDesign2 [43] fits a marginal for each gene and cell type by maximum likelihood over Negative Binomial, ZINB, Poisson, and Zero-Inflated Poisson families. Genes with sufficient non-zero expression ("group 2") are jointly modeled via a Gaussian copula to preserve gene-gene correlations; the remainder ("group 1") are sampled independently.[3] A copula models a multivariate distribution by separating the marginal distributions of individual variables from their dependency structure [32]. Formally, given random variables $X_1, \ldots, X_k$ with CDFs $F_1, \ldots, F_k$, the probability integral transform (PIT) maps each $X_i \mapsto U_i = F_i(X_i)$, yielding uniform marginals whose joint distribution is a copula. A Gaussian copula captures dependency via a covariance matrix $\mathcal{R}$ over Gaussian-transformed marginals:

$$C_{\mathcal{R},k}(u_1, \ldots, u_k) = \Phi_{\mathcal{R},k}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_k)) \quad (1)$$

[1] `https://bipress.boku.ac.at/camda2025/the-camda-contest-challenges/`

[2] scMAMA-MIA is one instance of a more general attack that we call MahalaMIA, which measures the Mahalanobis distance of a target to a covariance matrix that the generator preserves. In follow-up work, we and our colleagues apply MahalaMIA to synthetic bulk RNA-seq data [21], where it was a winner of the bulk track of the same CAMDA 2026 challenge.

[3] scDesign2 applies a distributional transform to handle discrete count distributions [2].

Algorithm 1 scMAMA-MIA membership inference attack, (**BB-aux**)

Require: synthetic data D_{synth}, target cells D_{target}, list of cell types T, and scDesign2 algorithm ϱ
Ensure: membership predictions for each donor $p \in D_{target}$

$[C^1, C^2, \ldots, C^{|T|}] \leftarrow \varrho(D_{synth})$
for each cell type $t \in T$ **do**
  Let $D_{target,t}$ be the set of target cells of type t
  Let $\mathcal{R}_t$ be the covariance matrix, defined in C^t
  Let G be the genes included in $\mathcal{R}_t$
  $U \leftarrow [0,1]^{n_t \times |G|}$
  for expression levels X_g^t of gene $g \in G$, for all cells x in $D_{target,t}$ **do**
    Let F_g^t be the fitted marginal CDF for gene g from C^t
    $U[:, g] \leftarrow F_g^t(X_g^t)$
    $z_x \leftarrow \Phi^{-1}(U[x, :])$
    $d_s \leftarrow \sqrt{z_x^\top \mathcal{R}_t^{-1} z_x}$
    $\lambda_t[x] \leftarrow 1/d_s$
  end for
end for
for all donors p in D_{target} **do**
  $\lambda_p \leftarrow \text{avg}(\{\lambda_t[x] \mid x \in D_{target,p},\ t \in T\})$
end for
return $\sigma(\lambda)$

where Φ^{-1} is the inverse standard Gaussian CDF and $\Phi_{\mathcal{R},k}$ is the multivariate normal CDF with covariance $\mathcal{R}$. To sample, one draws $Z \sim \mathcal{N}(0, \mathcal{R})$, applies the PIT to obtain uniform $U_i = \Phi(Z_i)$, then applies the inverse marginal CDFs $F_1^{-1}, \ldots, F_k^{-1}$ to recover samples on the original scale with the correlations of $\mathcal{R}$.

scDesign3 [41] extends scDesign2 to spatial data and additional omics modalities, with a user-selectable Gaussian (default) or vine copula (which we refer to as scDesign3-G and scDesign3-V); the vine copula is substantially slower and not recommended for >1000 genes. scMAMA-MIA (Sec. 3) targets the Gaussian copula variant of both models; we also report performance against the vine copula variant in Sec. 4, with full adaptation left to future work.

Threat Models We consider four threat models varying two axes of adversarial knowledge: whether the adversary has access to the fitted scDesign2 Gaussian copula (**white-box**, WB) or only to D_{synth} (**black-box**, BB), and whether they have access to an auxiliary real dataset from the same distribution (**+aux / -aux**), as in [1, 17]. Although scMAMA-MIA is designed around scDesign2's Gaussian copula, the black-box variants operate on D_{synth} alone and do not assume which SDG produced it. The adversary observes the target cells D_{target} with their donor IDs and cell types, and must infer, for each donor, whether that donor was in D_{train} (a donor's cells are either all in D_{train} or all held out).

3 Our Approach to Privacy Auditing: scMAMA-MIA

MAMA-MIA (Marginals Measurement Aggregation-based MIA) is a class of MIAs that exploits knowledge of an SDG's architecture to infer membership [11]. The heuristic proceeds in three steps: (1) **identify** which statistics of the training data are preserved by the SDG ("focal-points"); (2) **simulate** the SDG to record which focal-points are selected; and (3) **aggregate** focal-point measurements on target instances to score their similarity to the synthetic data. MAMA-MIA has been shown effective against several SDGs on demographic data. Below, we adapt each of the three steps to scDesign2.

3.1 Choice of Focal-Points

In **Step 1** of MAMA-MIA, we take the per-cell-type Gaussian copulas as focal-points, since scDesign2 measures them on the real training data and preserves them in the synthetic data. Each copula holds (i) marginal parameters for every gene and (ii) the covariance matrix $\mathcal{R}$ over the group 2 genes. We use only group 2 genes in our attack. An *enhanced* version also uses the group 1 genes' marginals (Section 3.3).

In the *white-box* (WB) setting, the attacker directly accesses the true copula $\varrho(D_{train})$ fit by scDesign2 on the hidden training data. In the *black-box* (BB) setting, the attacker instead constructs a proxy

copula by running scDesign2 on D_{synth} and extracting the resulting copula $\varrho(D_{synth})$ before the generation phase. We fit the proxy with scDesign2 whichever generator produced D_{synth}, since its fit gives the marginals and covariance that our distance needs. The proxy works because scDesign2 assigns a gene to group 2 by how often the gene is non-zero, and the synthetic data reproduces those rates, so most group 2 genes of the training data are also group 2 genes of the synthetic data. **Step 2** of MAMA-MIA (simulating the SDG repeatedly to record focal-point variation) is unnecessary here, because scDesign2 selects its focal-points deterministically, without noise.

3.2 Measuring and Aggregating

Step 3 is the most involved. For each cell type t, we measure how well a target cell's expression levels fit the focal-point statistics extracted from the copulas. Specifically, for each group 2 gene g with fitted marginal CDF F_g^t (parametrized by zero-inflation $\hat{\pi}$, dispersion $\hat{\theta}$, and mean $\hat{\mu}$), we apply the PIT to map observed counts into the uniform space: $U[:, g] = F_g^t(X_g^t)$, where $X_g^t \in \mathbb{Z}_{\geq 0}^{n_t}$ are the observed counts for gene g across all n_t target cells of type t. This mapping is necessary because the covariance matrix $\mathcal{R}_t$ in the Gaussian copula is defined over PIT-transformed marginals, not raw counts. We then compute the Mahalanobis distance [8] between each cell's transformed expression vector z_x (which is a distance from zero in the PIT-transformed space):

$$d_s = \sqrt{z_x^T \mathcal{R}^{-1} z_x}. \tag{2}$$

A greater d_s indicates that a target cell's expression is out-of-distribution relative to the copula's learned statistics, accounting for gene-gene covariances. We therefore take membership score $\lambda = d_s^{-1}$: target cells close to the copula's distribution are more likely to have contributed to it. $\mathcal{R}_t$ is singular when a cell type has fewer cells than copula genes, so $\mathcal{R}_t^{-1}$ denotes the pseudo-inverse (Appendix D compares alternatives). While a high Mahalanobis distance for any individual cell may simply reflect atypical expression, *aggregating* λ over all cells from the same donor yields a stable signal. Cell types differ in their number of copula genes, and d_s grows with that number, so scores from different cell types are not directly comparable. We therefore standardize $\log \lambda$ within each cell type (a z-score over that type's target cells, followed by a sigmoid) before pooling. Finally, we average cell-level scores per donor and pass through a sigmoid to produce donor-level membership scores $p = \sigma(\lambda) \in [0, 1]$; these are ranking statistics, not calibrated probabilities.

3.3 Threat Model Variants

Algorithm 1 gives the **BB-aux** variant, which assumes the least adversarial knowledge. With the trained SDG model, auxiliary data, or both, the attack changes as follows:

- **WB vs. BB**: The WB variant uses the trained model's copula $\varrho(D_{train})$ in place of the proxy $\varrho(D_{synth})$. The other SDG methods do not fit a Gaussian copula, so the WB variant applies only to scDesign2 and scDesign3-G.
- $\pm$**aux**: With auxiliary data D_{aux}, the adversary also computes the distance d_a under $\varrho(D_{aux})$ and sets $\lambda = d_a/(d_s + d_a)$ in place of d_s^{-1}. This compares how well a cell fits the synthetic copula with how well it fits a copula of other real data, and improves accuracy in practice (Figure 3). The gene set G then becomes the intersection of the group 2 genes of both copulas (on average, the intersection holds 86.6% of them).

Enhanced attack. The distance d_s ignores the group 1 genes, which are the majority (Table 1) and for which scDesign2 fits a marginal but no covariance. When auxiliary data is available, the enhanced attack also asks, for each such gene, whether the cell's count x_g is more likely under the marginal fit to D_{synth} than under the one fit to D_{aux}. The standard score is already the sigmoid of a log-ratio, since $d_a/(d_s + d_a) = \sigma(\log d_a - \log d_s)$, so we add the log-likelihood ratios to that logit, which is the first term of ℓ:

$$\ell = \log \frac{d_a}{d_s} + \gamma \sum_{g \in G_1} \log \frac{p_g^{synth}(x_g)}{p_g^{aux}(x_g)}, \qquad \gamma = \frac{1}{\sqrt{|G_1|}}, \tag{3}$$

where G_1 holds the genes that are fit in both copulas and are outside G. The spread of a sum of $|G_1|$ terms grows like $\sqrt{|G_1|}$, and this γ cancels that growth, so the second term does not swamp the first in cell types with many group 1 genes. γ is fixed by the gene count and has no free parameter. We

Dataset	Donors	Cells	Cell types	HVGs	scDesign2 genes (mean ±std) *group 1*	*group 2*
OneK1K	981	1,267,733	14	1,118	864 ±284	145 ±73
AIDA	508	1,058,909	33	2,021	1,214 ±465	446 ±98
HFRA	22	108,838	9	3,664	3,044 ±816	363 ±110

Table 1: Dataset statistics and scDesign2 gene group assignments. Group 1 genes are modeled via independent marginals; group 2 genes are jointly modeled via the Gaussian copulas. HVGs are the highly variable genes retained for all experiments.

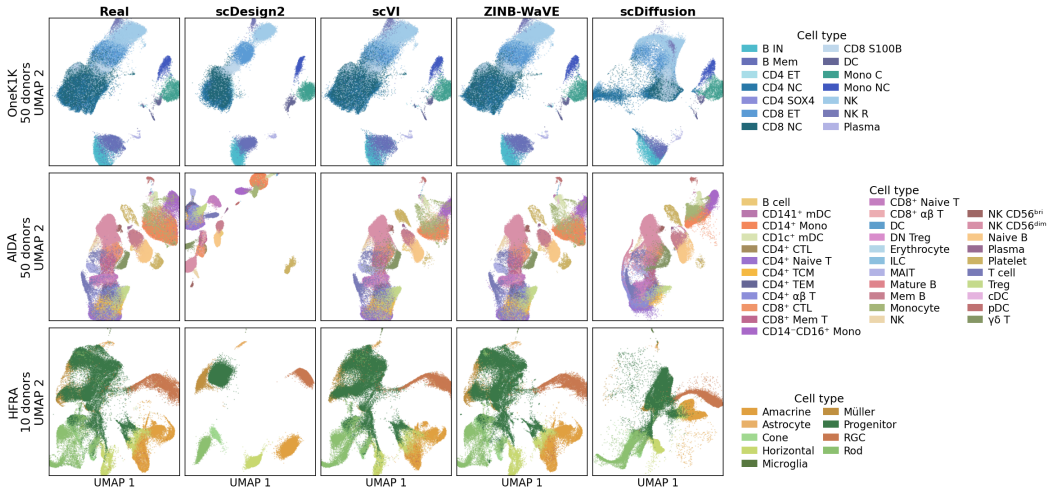

Figure 2: UMAPs for real data samples and the corresponding synthetic data (computed jointly) to illustrate how much information each SDG method is able to preserve.

standardize ℓ within each cell type and average per donor, as for the standard attack. The enhanced attack needs D_{aux}. Without p_g^{aux} to divide by, it fails (Appendix B).

4 Experiments

We evaluate scMAMA-MIA on synthetic data from three real scRNA-seq datasets across six SDG methods, averaging all results over five trials per configuration (each with a different donor sample for training, holdout, and auxiliary sets[4]). Code is available at `https://github.com/steveng9/scRNA-seq_privacy_audits`. [5]

4.1 Datasets, SDG Methods, and Experimental Setup

We evaluate on three scRNA-seq datasets of varying scale, demographic composition, and donor count (Table 1). **OneK1K** [48] contains peripheral blood mononuclear cells from 981 donors across 14 cell types, and served as the primary benchmark in the CAMDA2025 and CAMDA2026 Health Privacy Challenges. **AIDA** [22] comprises healthy immune cells from 508 donors across 33 cell types and seven Asian population groups from five countries; its demographic breadth heightens its sensitivity. **HFRA** [24] is a Human Fetal Retina Atlas from just 22 donors across 9 cell types, yielding substantially more cells per donor than the other two. Both AIDA and HFRA are retrieved from the CZ Cell x Gene database.[6]

We follow [43] and select highly variable genes (HVGs) `sc.pp.highly_variable_genes` in Scanpy with `min_mean`=0.0125, `max_mean`=3, and `min_disp`=0.5. UMAPs [15] of representative training and synthetic samples are provided in Figure 2.

[4] but importantly, all generators for a trial are trained on the same D_{train}.

[5] LLMs aided in the organization of code, installing of external libraries, and running experiment suites. The core methods are owned and implemented by the authors.

[6] `https://cellxgene.cziscience.com/`

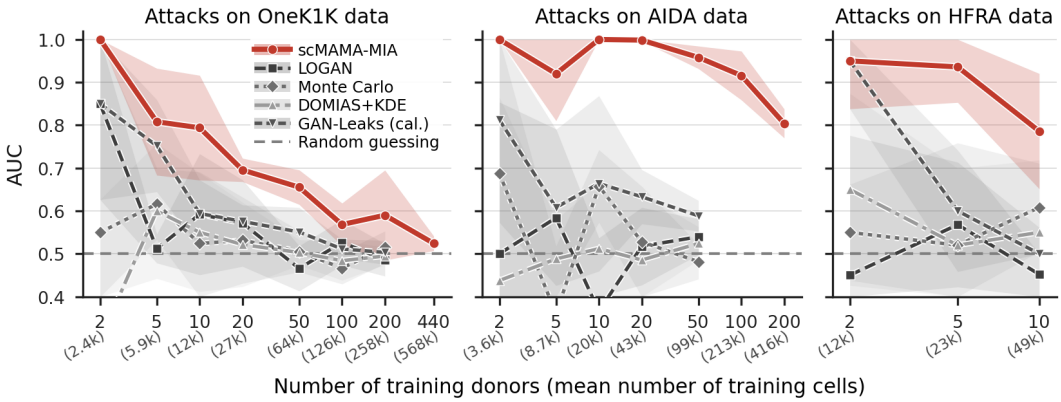

Figure 4: **scMAMA-MIA matches or outperforms every baseline attack on all three datasets.** AUC on **scDesign2** data by number of training donors (mean number of training cells in parentheses), with all attacks in the BB+aux threat model. Bands show ± 1 standard deviation. D_{aux} is the 108 remaining donors for AIDA at 200 donors. HFRA's 22 donors cannot supply a disjoint D_{aux} at 10 donors or more, so there it includes 8 of the target donors.

In addition to **scDesign2** and **scDesign3** (Gaussian and vine copula variants; Sec. 2), we evaluate against four further SOTA SDG methods: **scVI** [26], a variational autoencoder with a ZINB observation model; **scDiffusion** [27], a diffusion model conditioned on cell-type embeddings; **ZINB-WaVE** [38], a zero-inflated factor model; and **NMF** [46], which decomposes expression data via non-negative matrix factorization and samples synthetic cells from per-cluster distributions.

We experiment with training sets of {2, 5, 10, 20, 50, 100, 200} donors (and 440 on OneK1K), retaining all cells per sampled donor. The target set D_{target} holds the training donors and an equal number of other donors (the holdout set), so exactly half of target donors are members, as in the standard MIA setup [1, 17, 30, 47]. A member donor's target cells are therefore the cells the generator was trained on; Section 4.4 removes this assumption. Auxiliary data D_{aux} is sampled similarly (minimum 10 donors), drawing from donors outside D_{target} first, so that D_{aux} is disjoint from both the training and holdout donors. HFRA experiments use $\leq$11 donors, given its small donor pool.

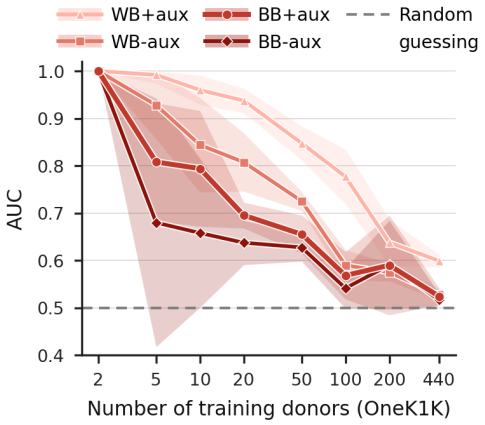

Figure 3: AUC of scMAMA-MIA on **scDesign2** data for OneK1K under the four threat models: white-box (WB) or black-box (BB), each with or without auxiliary data.

4.2 Success Rate of scMAMA-MIA against scDesign2

Figure 4 compares the donor-level AUC of scMAMA-MIA on scDesign2 data with that of four attacks from the CAMDA2025 benchmark (LOGAN [14], GAN-Leaks (calibrated) [6], DOMIAS with KDE [44], and Monte Carlo [16]), all in the BB+aux threat model.

scMAMA-MIA matches or exceeds every baseline at every donor count on every dataset. At 50 donors on OneK1K, for example, it reaches 0.66 and the best baseline 0.55. Among the baselines, GAN-Leaks (calibrated, i.e., also leverages D_{aux}) tends to be the strongest competitor, yet still falls well short of scMAMA-MIA, particularly at larger donor sizes, where most baselines regress toward random-chance performance. To run the baselines on larger cohorts at all, we implemented batched versions of GAN-Leaks, DOMIAS+KDE, and Monte Carlo; even so, these attacks could not complete[7] on AIDA above 50 donors or OneK1K above 200 donors, which accounts for the gaps in

[7]Experiments are conducted on an AMD Ryzen Threadripper PRO 5965WX with 24 cores.

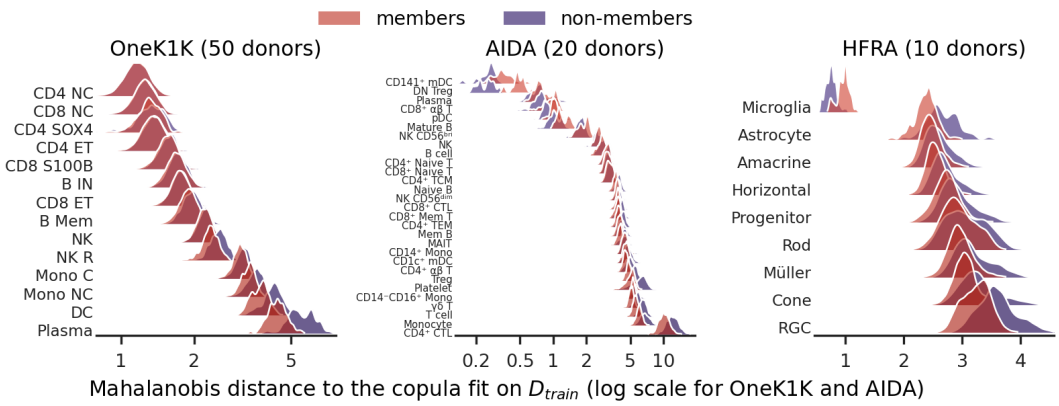

Figure 5: **In most cell types, member cells lie closer than non-member cells to the copula fit on the training data.** The distribution of Mahalanobis distances taken during an attack on **scDesign2**, between member and non-member cells and the copula's covariance matrix $\mathcal{R}$ learned from the training data. Rows are sorted by median distance, and their shading darkens slightly with it. The distance axis is logarithmic for OneK1K and AIDA. AIDA omits four cell types with fewer than 10 member or non-member cells. The few cell types in which members lie farther away are fit on about 200 cells or fewer (Appendix G).

Figure 4. At 50 donors, scMAMA-MIA runs 2.6 to 12.3 times faster than these three; only LOGAN, whose inference is much lighter, is faster than scMAMA-MIA (by 2 to 3 times).

In small cohorts, scMAMA-MIA identifies the members almost without error, with a perfect or near-perfect AUC (0.95 to 1.00) at the smallest donor counts across all three datasets. On AIDA, this extends to 10 and 20 donors (roughly 43,000 cells). scMAMA-MIA succeeds because it exploits the specific way scDesign2 fits its model, which the existing attacks, treating the generator as a generic black box, do not use.

AUC averages over every false-positive rate (FPR), but an adversary who acts on a prediction can afford few false positives [4]. So we also measure the true-positive rate (TPR) at low FPRs (Appendix F). At a 1% FPR, the attack identifies half of the members of a 20-donor OneK1K cohort (enhanced attack) and nearly all members of a 20-donor AIDA cohort (standard attack), so its success in small cohorts does not come from averaging over FPRs. With 440 OneK1K donors in the training cohort, it is barely better than guessing.

Figure 3 compares the four threat models against scDesign2 on OneK1K. As expected, WB+aux is the strongest overall, and auxiliary data generally improves attack success. The more interesting comparison is between WB-aux and BB+aux. WB-aux outperforms BB+aux at most donor counts, despite having no auxiliary data at all. So white-box access to the generator's learned parameters tells the adversary more than an auxiliary dataset does.

Attack AUCs on HFRA are lower than on AIDA at every donor count (at most 0.95), likely because HFRA's cell types are larger (Section 4.3 and Appendix G).

We validate scMAMA-MIA with the Mahalanobis distances behind the attack (Step 3). Figure 5 shows them for member and non-member cells, per cell type and dataset. In most cell types, member cells have lower mean distances than non-members, reflecting their contribution to the fitted copula, though the distributions overlap at the single-cell level. The exceptions are rare cell types fit on about 200 cells or fewer, the regime in which the attack reverses (Appendix G). This overlap is why the attack pools the evidence over many cells of a donor. Distances are drawn from one random trial (50 donors for OneK1K, 20 for AIDA, 10 for HFRA). Distance magnitudes differ across cell types. The attack is insensitive to this, because it standardizes scores within each cell type (Section 3).

Appendix A breaks the results down by sex and ethnicity, and shows that privacy is unequally distributed across population groups. Appendix G scores donors on one cell type at a time. The cell types fit on a few hundred to a few thousand cells reveal the most, and the AUC falls as a cell type grows.

SDG Method	Fidelity			Privacy (scMAMA-MIA AUC)		
	LISI↑	ARI↑	MMD$\times 10^3$ ↓	BB-aux	BB+aux	*enhanced*
scDesign2	0.87 ± .00	0.48 ± .02	0.45 ± .03	0.63 ± .03	0.66 ± .04	0.79 ± .06
scDesign3-G	0.88 ± .00	0.46 ± .04	0.28 ± .03	0.61 ± .03	0.65 ± .03	0.78 ± .07
scDesign3-V	0.88 ± .00	0.46 ± .03	0.31 ± .01	0.61 ± .02	0.64 ± .04	0.79 ± .06
scVI	0.90 ± .00	0.40 ± .05	0.28 ± .03	0.57 ± .03	0.59 ± .05	0.69 ± .09
ZINB-WaVE	0.89 ± .01	0.45 ± .03	1.02 ± 1.66	0.59 ± .04	0.66 ± .08	0.78 ± .12
NMF	0.18 ± .04	0.29 ± .02	27.56 ± 2.99	0.55 ± .05	0.56 ± .05	0.51 ± .08
scDiffusion	0.49 ± .01	0.39 ± .05	45.69 ± 3.99	0.48 ± .05	0.47 ± .03	0.47 ± .02

Table 2: **Fidelity and privacy for all SDG methods, without added noise (OneK1K, 50 donors).** MMD is between real and synthetic cells in PCA space, reported $\times 10^3$ (45.69 = 0.0457), and is not comparable to MMD values reported elsewhere. BB±aux: scMAMA-MIA black-box attack without/with auxiliary data; *enhanced*: the BB+aux attack with the log-likelihood-ratio term of Eq. 3.

4.3 Success Rate of scMAMA-MIA against other SDG Methods

Having established scMAMA-MIA's effectiveness against scDesign2, we evaluate whether it generalizes to other SDG methods. Table 2 reports the attack against every SDG method, with three fidelity metrics: LISI (neighborhood mixing between real and synthetic cells), ARI (cluster agreement), and MMD (distributional distance) [23, 19]. Five SDGs have high fidelity (scDesign2, scDesign3-G, scDesign3-V, scVI, and ZINB-WaVE, all with LISI at or above 0.87), and against all five the BB+aux AUC is between 0.59 and 0.66, well above chance. scVI has the lowest, which suggests modest robustness from its latent variational structure, but no immunity. The enhanced attack (Eq. 3) raises these to between 0.69 and 0.79. The attack transfers in the same way to smaller cohorts and to AIDA. At 10 donors, for example, the enhanced attack scores above 0.8 against all five generators on both datasets (Appendix Table 9). The enhanced attack is not always the stronger of the two, though. Against scDesign2 it falls below the standard attack on HFRA, and on AIDA from 10 to 100 donors (Appendix Table 10).

Against NMF and scDiffusion, every variant scores near 0.5. Their fidelity is far lower (LISI of 0.18 and 0.49), so the attack likely fails because their synthetic data carries little information about the training data, not because they protect privacy. A low attack AUC is therefore a meaningful privacy signal only when the synthetic data also has high fidelity.

Fidelity explains only part of the differences between datasets. On HFRA, scDesign2 has lower fidelity (Table 7), its UMAPs resolve the fine retinal subtypes poorly (Figure 2), and the attack is weaker. Interestingly, on AIDA, real and synthetic cells overlap little in the UMAPs, yet the attack reaches a high AUC. The size of the cell types predicts leakage better (Appendix G). scDesign2 fits one copula per cell type, and down to a few hundred cells, the fewer cells a copula is fit on, the more it reveals about their donors. At 10 donors, the median AIDA cell type has about 300 training cells and the median HFRA cell type about 3,500, and at equal cell-type size the standard attack does about as well on HFRA as on AIDA.

4.4 Does the Attack Need the Training Cells?

So far, the target cells of a member donor have been the cells the generator was trained on, and all of them (about 1,300 per donor in OneK1K and 2,000 in AIDA). An adversary may instead hold only a few cells of a donor, or a different sample from the one used for training. We test both cases against scDesign2.

Dataset, donors	Cells per target donor			
	1	10	100	all
OneK1K, 20	0.53 ± .01	0.61 ± .02	0.69 ± .03	0.70 ± .03
OneK1K, 50	0.54 ± .03	0.60 ± .03	0.65 ± .03	0.66 ± .04
OneK1K, 100	0.52 ± .02	0.55 ± .04	0.56 ± .05	0.57 ± .05
OneK1K, 200	0.56 ± .10	0.58 ± .11	0.59 ± .10	0.59 ± .11
AIDA, 10	0.67 ± .03	0.88 ± .02	0.99 ± .01	1.00 ± .00
AIDA, 100	0.57 ± .01	0.68 ± .04	0.85 ± .06	0.92 ± .06

Table 3: **The attack needs few cells.** Standard BB+aux AUC against scDesign2 with k random cells per target donor (20 draws per trial; mean ± std over 5 trials). "All" uses every cell: a median of 1,270 per donor on OneK1K and 2,050 on AIDA.

Few cells. We re-score the standard BB+aux attack on k cells drawn at random from each target donor (Table 3). The signal saturates quickly. With 100 cells, under a tenth of a typical donor's cells, the attack is within 0.01 of its all-cell AUC on OneK1K and within 0.07 on AIDA. This is because a donor's score is an average over their cells, and 100 cells already pin that average down. For the same reason, donors with more cells are not easier to identify (there is no meaningful correlation between a donor's cell count and their score,

Dataset	Target cells	Training donors						
		2	5	10	20	50	100	200
OneK1K	training	$1.00 \pm .00$	$0.98 \pm .05$	$0.96 \pm .07$	$0.99 \pm .01$	$0.98 \pm .03$	$0.91 \pm .06$	$0.75 \pm .02$
	unseen	$0.90 \pm .20$	$0.89 \pm .11$	$0.71 \pm .13$	$0.62 \pm .07$	$0.65 \pm .04$	$0.56 \pm .05$	$0.54 \pm .02$
AIDA	training	$1.00 \pm .00$	$0.94 \pm .08$	$0.99 \pm .02$	$0.99 \pm .01$	$1.00 \pm .01$	$1.00 \pm .01$	–
	unseen	$0.85 \pm .30$	$0.84 \pm .18$	$0.82 \pm .13$	$0.73 \pm .11$	$0.63 \pm .05$	$0.59 \pm .03$	–
HFRA	training	$1.00 \pm .00$	$0.92 \pm .08$	–	–	–	–	–
	unseen	$0.70 \pm .24$	$0.61 \pm .16$	–	–	–	–	–

Table 4: **The attack does not need the training cells.** Donor-level AUC of the enhanced BB+aux attack against scDesign2 trained on 200 cells per donor, when the targets are those same cells (training) or 200 other cells of the same donors (unseen); mean $\pm$ std over 5 trials. Unseen cells still reveal membership, mostly in cohorts of 50 donors or fewer.

with a Pearson r between -0.14 and 0.05). At the other extreme, a single cell per donor gives an AUC of 0.67 on AIDA at 10 donors, 0.57 at 100 donors, and 0.53 to 0.56 on OneK1K. So one cell, out of the roughly 2,000 that an AIDA donor contributes, already tells better than chance whether that donor's data trained the generator.

Unseen cells. For each training donor we draw two disjoint sets of 200 cells (donors with fewer than 400 cells are left out). The generator is trained on the first set only. The adversary's targets are either the second set, which the generator never saw, or the first set as a control. Each non-member donor also contributes 200 cells. Table 4 reports the enhanced BB+aux attack. What's striking is that cells the generator never saw still reveal membership! On OneK1K at 5 donors, the training cells give an AUC of 0.98 and the unseen cells nearly as much, 0.89. The gap widens as the cohort grows (0.98 vs. 0.65 at 50 donors), and by 200 donors unseen cells are near chance. The standard attack shows the same gap (at 50 donors, 0.81 with training cells and 0.59 with unseen cells on OneK1K; 1.00 and 0.56 on AIDA). So does scVI when it is trained and attacked the same way (on OneK1K, 0.87 with training cells and 0.73 with unseen cells at 10 donors, and 0.79 and 0.57 at 100).

This tells us where the leak comes from. Unseen cells can reveal membership only if the generator retains something that all of a donor's cells share, so scDesign2 leaks the donor's expression profile, not only the cells it was trained on. It also makes the results of the previous sections an upper bound for an adversary who holds a different sample from a donor. That adversary still succeeds in cohorts of 50 donors or fewer (Table 4).

4.5 Effect of Noise on Data Quality and Defending Privacy

As a first-line defense, we add noise to scDesign2's covariance estimates, empirically characterizing the resulting privacy–utility trade-off. The noise scale is a heuristic and carries no formal privacy guarantee. Each entry $\mathcal{R}_{jk}$ of each cell type's copula covariance receives additive noise drawn from $\mathcal{N}(0, \sigma^2)$, where

$$\sigma = \frac{\eta \cdot |G| \cdot k_{\max}}{N}, \tag{4}$$

with $|G|$ the number of copula genes, $k_{\max}$ the maximum number of cells contributed by any single donor to this cell type, and N the total training cells for this cell type. The factor $|G|$ makes the calibration conservative. Removing one donor changes each covariance entry by about $k_{\max}/N$, the donor's share of the cell type's cells, so the whole matrix moves by up to about $|G|\, k_{\max}/N$ in Frobenius norm. Scaling σ to this whole-matrix change gives each entry $|G|$ times the noise that a per-entry bound would. The noised matrix E is symmetrized (i.e. $(E + E^T)/2$), projected to the nearest positive semidefinite matrix, and re-normalized to a correlation matrix. We sweep $\eta \in \{10^{-4}, \ldots, 10^{1}\}$, so that the noise ranges from negligible to well beyond typical gene–gene correlations of 0.05–0.30.

Table 5 shows the results on OneK1K at 50 donors. Fidelity is nearly unchanged through $\eta = 10^{-2}$. At $\eta = 10^{-1}$ the standard BB+aux attack falls to 0.54, near chance, while only LISI drops noticeably (from 0.87 to 0.72). Covariance noise therefore stops the standard attack at a modest cost in fidelity.

But it does not stop the enhanced attack, because the enhanced attack's additional evidence comes from the marginals, which this noise leaves untouched. In a second sweep on OneK1K and AIDA at 50 donors (Appendix E), the enhanced attack keeps an AUC of about 0.7 even when the noise on each covariance entry has a standard deviation of 2, larger than any correlation can be. Noise on the marginal parameters does bring it near chance, but only once real and synthetic cells no longer mix at

Privacy Param	LISI↑	ARI↑	MMD$\times 10^3$ ↓	BB-aux	BB+aux
$\eta = 0$ (least privacy)	$0.87 \pm .00$	$0.48 \pm .02$	$0.45 \pm .03$	$0.63 \pm .03$	$0.66 \pm .04$
$\eta = 10^{-4}$	$0.87 \pm .00$	$0.49 \pm .01$	$0.44 \pm .02$	$0.63 \pm .03$	$0.64 \pm .04$
$\eta = 10^{-3}$	$0.87 \pm .00$	$0.47 \pm .04$	$0.44 \pm .04$	$0.62 \pm .03$	$0.63 \pm .04$
$\eta = 10^{-2}$	$0.85 \pm .00$	$0.46 \pm .04$	$0.42 \pm .02$	$0.59 \pm .02$	$0.60 \pm .03$
$\eta = 10^{-1}$	$0.72 \pm .01$	$0.51 \pm .03$	$0.47 \pm .02$	$0.55 \pm .03$	$0.54 \pm .01$
$\eta = 1$	$0.62 \pm .01$	$0.47 \pm .03$	$0.53 \pm .03$	$0.52 \pm .02$	$0.49 \pm .03$
$\eta = 10^{1}$	$0.61 \pm .01$	$0.54 \pm .02$	$0.55 \pm .01$	$0.52 \pm .02$	$0.49 \pm .02$

Table 5: **Privacy–utility tradeoff** for scDesign2 (OneK1K data, 50 donors). η Gaussian noise-to-signal parameter added to the copula covariance (larger = more noise). LISI ↑, ARI ↑, MMD ↓: fidelity metrics (see Section 4.3).

all (LISI of 0). We found no noise level that defends against the enhanced attack while preserving fidelity.

Formal donor-level differential privacy [9] is harder still, because it has to protect the $|G| \times |G|$ covariance entries and hundreds of marginal parameters in every cell type, against a donor who may contribute thousands of cells to one cell type. Calibrating that noise to a reasonable privacy budget without destroying utility is an open problem, which we leave to future work.

5 Discussion

We introduced scMAMA-MIA, the first membership inference attack effective against synthetic scRNA-seq data, exploiting the donor-specific statistical fingerprints embedded in scDesign2's Gaussian copula. Against all three datasets, scMAMA-MIA substantially outperforms existing MIA baselines and reaches perfect AUC at small donor counts even in the black-box setting. It generalizes to other SDG methods despite being designed around the copula structure, suggesting that the underlying vulnerability is broader than any single generative model. Membership also shows in cells of a member donor that the generator never saw, mainly in cohorts of 50 donors or fewer (Section 4.4). So the generator leaks something about the donor, not only about the cells it was trained on, and an adversary does not need those cells.

Our quality analysis shows that a low attack AUC is a meaningful privacy signal only when the synthetic data is also high-fidelity. NMF and scDiffusion resist the attack because their synthetic data is too unlike the real data to carry a membership signal, not because they protect privacy. Among the high-fidelity SDGs, scVI is the least vulnerable, which warrants further study.

Noise on the copula covariance brings the standard attack close to chance at a modest cost in fidelity. But the enhanced attack, which also uses the per-gene marginals, survives every level of covariance noise we tried, and noise on the marginals stops it only after fidelity is lost. A defense for scDesign2 therefore has to protect the marginals as well as the covariance.

Practically, we recommend that model parameters (e.g. copula covariances and marginal fits) never be released alongside synthetic data, as white-box access consistently amplifies attack success. Increasing the donor cohort size also reduces leakage, as individual expression profiles are increasingly able to "hide among the crowd." The open problem is a generator that gives a formal donor-level differential privacy guarantee for all of these parameters and still keeps enough fidelity to be useful. scMAMA-MIA offers a way to audit such a generator before its data is released.

Limitations Despite these findings, scMAMA-MIA is primarily designed for covariance-based SDGs (e.g. scDesign2 / scDesign3-G), and while it transfers empirically to other models, its behavior is not theoretically characterized across diffusion-based or likelihood-free generators. We also focus on a fixed donor-level threat model and do not consider more adaptive adversaries or richer external genomic reference integration. The noise defenses are heuristic and carry no formal guarantee.

Acknowledgments and Disclosure of Funding

This material is based upon work supported by the U.S. National Science Foundation under Grant Nos. 2451163 and 2523406, and by NSF National AI Research Resource (NAIRR) Pilot allocations 240485 (CloudBank AWS) and 240091 (TACC Frontera). This research was, in part, funded by the National Institutes of Health (NIH) Agreement No. 1OT2OD032581. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH. Steven Golob is supported by an NSF CSGrad4US Fellowship and was supported by an IEEE Graduate Student Travel Grant. Sikha Pentyala and Martine De Cock are affiliated with the University of Washington eScience Institute.

References

[1] Tristan Allard, Louis Béziaud, and Sébastien Gambs. SNAKE challenge: Sanitization algorithms under attack. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5010–5014, 2023.

[2] Athanassios N Avramidis, Nabil Channouf, and Pierre L'Ecuyer. Efficient correlation matching for fitting discrete multivariate distributions with arbitrary marginals and normal-copula dependence. *INFORMS Journal on Computing*, 21(1):88–106, 2009.

[3] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914, 2022.

[5] Xu Chang, Yunxi Zheng, and Kai Xu. Single-Cell RNA Sequencing: Technological progress and biomedical application in cancer research. *Molecular Biotechnology*, 66(7):1497–1519, 2024.

[6] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. GAN-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 343–362, 2020.

[7] Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero. Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010.

[8] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L. Massart. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000.

[9] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[10] Khaled El Emam, Anita Fineberg, Elizabeth Jonker, and Lisa Pilgram. Perspectives of Canadian privacy regulators on anonymization practices and anonymized information: a qualitative study. *International Data Privacy Law*, 14(4):391–403, 2024.

[11] Steven Golob, Sikha Pentyala, Anuar Maratkhan, and Martine De Cock. Privacy vulnerabilities in marginals-based synthetic data. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 977–995, 2025.

[12] Christopher Heje Grønbech, Maximillian Fornitz Vording, Pascal N Timshel, Casper Kaae Sønderby, Tune H Pers, and Ole Winther. scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics*, 36(16):4415–4422, 2020.

[13] Florent Guépin, Matthieu Meeus, Ana-Maria Creţu, and Yves-Alexandre de Montjoye. Synthetic is all you need: removing the auxiliary data assumption for membership inference attacks against synthetic data. In *Computer Security. ESORICS 2023 International Workshops*, volume 14398 of *Lecture Notes in Computer Science*, pages 182–198. Springer, 2023.

[14] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019(1):133–152, 2019.

[15] John Healy and Leland McInnes. Uniform manifold approximation and projection. *Nature Reviews Methods Primers*, 4(1):82, 2024.

[16] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Monte Carlo and reconstruction membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019(4):232–249, 2019.

[17] Florimond Houssiau, James Jordon, Samuel N Cohen, Owen Daniel, Andrew Elliott, James Geddes, Callum Mole, Camila Rangel-Smith, and Lukasz Szpruch. TAPAS: a toolbox for adversarial privacy auditing of synthetic data. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research (SyntheticData4ML)*, 2022. arXiv:2211.06550.

[18] Yuzheng Hu, Fan Wu, Qinbin Li, Yunhui Long, Gonzalo Munilla Garrido, Chang Ge, Bolin Ding, David Forsyth, Bo Li, and Dawn Song. SoK: Privacy-preserving data synthesis. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 4696–4713, 2024.

[19] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, December 1985.

[20] Alexandru Isaic, Nadica Motofelea, Teodora Hoinoiu, Alexandru Catalin Motofelea, Ioan Cristian Leancu, Emanuela Stan, Simona R. Gheorghe, Alina Gabriela Dutu, and Andreea Crintea. Next-Generation Sequencing: A review of its transformative impact on cancer diagnosis, treatment, and resistance management. *Diagnostics*, 15(19):2425, 2025.

[21] Charlene Jarrell, Daniil Filienko, Jonathan D. H. Kim, Emma Szebenyi, Sikha Pentyala, Steven Golob, and Martine De Cock. Does synthetic bulk RNA-seq data protect donors? Privacy auditing through membership inference attacks. In *Annual International Conference on Critical Assessment of Massive Data Analysis (CAMDA), collocated with ISMB*, 2026.

[22] Kian Hong Kock, Le Min Tan, Kyung Yeon Han, Yoshinari Ando, Damita Jevapatarakul, Ankita Chatterjee, Quy Xiao Xuan Lin, Eliora Violain Buyamin, Radhika Sonthalia, Deepa Rajagopalan, Yoshihiko Tomofuji, et al. Asian diversity in human immune cells. *Cell*, 188(8):2288–2306.e24, 2025.

[23] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, December 2019.

[24] Emil Kriukov, Jonathan R Soucy, Everett Labrecque, and Petr Baranov. Unraveling the developmental heterogeneity of human retinal ganglion cells within the developing retina to study to the continuity of maturation. *Developmental Biology*, 528:91–104, 2025.

[25] Jidong Lang, William C. Cho, Tao Huang, Taoyang Wu, and Junlin Xu. Editorial: Applications of RNA-seq in cancer and tumor research. *Frontiers in Genetics*, 14, 2023.

[26] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

[27] Erpai Luo, Minsheng Hao, Lei Wei, and Xuegong Zhang. scDiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics*, 40(9):btae518, 2024.

[28] Matteo Marouf, Pierre Machart, Vikas Bansal, Christoph Kilian, Daniel S. Magruder, Christian F. Krebs, and Stefan Bonn. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nature Communications*, 11(1):166, 2020.

[29] Patrick McKeever, Daniil Filienko, Steven Golob, Shane Menzies, Sikha Pentyala, Jineta Banerjee, Luca Foschini, and Martine De Cock. Comparison of single cell RNA synthetic data generators: a CAMDA health challenge analysis. In *Annual International Conference on Critical Assessment of Massive Data Analysis (CAMDA), collocated with ISMB/ECCB*, 2025.

[30] Matthieu Meeus, Florent Guepin, Ana-Maria Creţu, and Yves-Alexandre de Montjoye. Achilles' heels: vulnerable record identification in synthetic data publishing. In *Computer Security – ESORICS 2023*, Lecture Notes in Computer Science, pages 380–399. Springer, 2023.

[31] Nature Methods Editorial. Method of the Year 2013. *Nature Methods*, 11(1):1, 2014.

[32] Roger B Nelsen. *An Introduction to Copulas*. Springer, 2006.

[33] NIH. Implementation update for data management and access practices under the genomic data sharing policy, 2024. Number: NOT-OD-24-157, Available at `https://grants.nih.gov/grants/guide/notice-files/NOT-OD-24-157.html`.

[34] NIH. Request for information on responsibly developing and sharing generative artificial intelligence tools using NIH controlled access data, 2025. Number: NOT-OD-25-118, Available at: `grants.nih.gov/grants/guide/notice-files/NOT-OD-25-118.html`.

[35] Marie Oestreich, Dingfan Chen, Joachim L. Schultze, Mario Fritz, and Matthias Becker. Privacy considerations for sharing genomics data. *EXCLI Journal*, 20:1243–1260, 2021.

[36] Bristena Oprisanu, Georgi Ganev, and Emiliano De Cristofaro. On utility and privacy in synthetic genomic data. In *Proceedings of the 29th Network and Distributed System Security Symposium*, San Diego, CA, USA, 2022.

[37] E Pascucci, F A Causio, G Pesole, M Chiara, M Morelli, G Tonon, F Nicassio, R Pastorino, G E Calabrò, and S Boccia. Progressing towards personalised medicine: the Genomic Data Infrastructure (GDI) project. *European Journal of Public Health*, 34(Supplement_3):ckae144.1956, 2024.

[38] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(1):284, 2018.

[39] Mahsa Shabani and Pascal Borry. Challenges of web-based personal genomic data sharing. *Life Sciences, Society and Policy*, 11(1):3, 2015.

[40] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.

[41] Dongyuan Song, Qingyang Wang, Guanao Yan, Tianyang Liu, Tianyi Sun, and Jingyi Jessica Li. scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nature Biotechnology*, 42(2):247–252, 2024.

[42] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data–anonymisation groundhog day. In *31st USENIX Security Symposium*, pages 1451–1468, 2022.

[43] Tianyi Sun, Dongyuan Song, Wei Vivian Li, and Jingyi Jessica Li. scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome Biology*, 22(1):163, 2021.

[44] Boris van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. Membership inference attacks against synthetic data through overfitting detection. In *International Conference on Artificial Intelligence and Statistics*, pages 3493–3514. PMLR, 2023.

[45] Conor R. Walker, Xiaoting Li, Manav Chakravarthy, William Lounsbery-Scaife, Yoolim A. Choi, Ritambhara Singh, and Gamze Gürsoy. Private information leakage from single-cell count matrices. *Cell*, 187(23):6537–6549.e10, 2024.

[46] Andrew Wicks. Differentially-private non-negative matrix factorization. In *Annual International Conference on Critical Assessment of Massive Data Analysis (CAMDA), collocated with ISMB/ECCB*, 2025.

[47] Xiaoyu Wu, Yifei Pang, Terrance Liu, and Zhiwei Steven Wu. Winning the MIDST challenge: New membership inference attacks on diffusion models for tabular data synthesis. *CoRR*, abs/2503.12008, March 2025.

[48] Seyhan Yazar, Jose Alquicira-Hernandez, Kristof Wing, Anne Senabouth, M. Grace Gordon, Stacey Andersen, Qinyi Lu, Antonia Rowson, Thomas R. P. Taylor, Linda Clarke, Katia Maccora, Christine Chen, Anthony L. Cook, Chun Jimmie Ye, Kirsten A. Fairfax, Alex W. Hewitt, and Joseph E. Powell. Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. *Science*, 376(6589):eabf3041, 2022.

[49] Tianjiao Zhang, Zhongqian Zhao, Jixiang Ren, Ziheng Zhang, Hongfei Zhang, and Guohua Wang. cfDiffusion: diffusion-based efficient generation of high quality scRNA-seq data with classifier-free guidance. *Briefings in Bioinformatics*, 26(1):bbaf071, 2025.

[50] Ziqi Zhang, Chao Yan, and Bradley A. Malin. Membership inference attacks against synthetic health data. *Journal of Biomedical Informatics*, 125:103977, 2022.

A By Demographic Subgroup

Privacy leakage may be more severe for demographic subgroups underrepresented in training data [3]. Table 6 shows donor-level AUC broken down by sex and ethnicity on the AIDA dataset (BB+aux, 20 donors). For scDesign2 and scDesign3-G, scMAMA-MIA achieves perfect AUC uniformly across all subgroups. Neural network- and matrix-factorization-based SDGs exhibit substantially higher variance in attack success across subgroups, likely because these models are more sensitive to distributional shifts between groups, amplifying differences in data quality and representation into unequal privacy exposure. For sex, privacy leakage tends to be higher for females (55.1% of donors) than for males (most clearly in scVI, 0.74 vs. 0.59, and ZINB-WaVE, 0.84 vs. 0.68), consistent with the hypothesis that overrepresented groups are more vulnerable due to better model fit. For ethnicity, AUCs fluctuate considerably across groups and SDGs (e.g. scDiffusion ranges from 0.17 for Singaporean Indian to 0.80 for Singaporean Chinese), with no consistent relationship to group size; cell-type composition and batch effects are likely confounders.

Subgroup	scDesign2	scDesign3-G	scDesign3-V	scVI	scDiffusion	NMF	ZINB-WaVE
Sex							
female (55.1%)	1.00±0.00	1.00±0.00	0.66±0.32	0.74±0.13	0.55±0.24	0.68±0.20	0.84±0.10
male (44.9%)	1.00±0.00	1.00±0.00	0.63±0.29	0.59±0.12	0.48±0.14	0.54±0.12	0.68±0.07
Ethnicity							
Japanese (27.1%)	1.00±0.00	0.99±0.01	0.57±0.34	0.65±0.18	0.63±0.20	0.64±0.19	0.73±0.16
Korean (34.6%)	1.00±0.00	1.00±0.00	0.70±0.32	0.69±0.15	0.35±0.15	0.62±0.19	0.84±0.09
Singaporean Chinese (11.7%)	1.00±0.00	1.00±0.00	0.85±0.20	0.69±0.11	0.80±0.19	0.60±0.24	0.85±0.18
Singaporean Indian (10.1%)	1.00±0.00	1.00±0.00	0.50±0.41	0.67±0.24	0.17±0.24	0.71±0.16	0.71±0.26
Singaporean Malay (9.5%)	1.00±0.00	1.00±0.00	0.58±0.43	0.80±0.40	0.32±0.19	0.53±0.45	1.00±0.00

Table 6: scMAMA-MIA (BB+aux) disparate-impact breakdown by sex and ethnicity on the AIDA dataset, averaged over 5 samples with 20 donors. Subgroup cell-count prevalence in parentheses.

Algorithm 2 scMAMA-MIA (**BB-aux**), ***enhanced*** with copula's one-way marginal parameters

Require: synthetic data D_{synth}, target cells D_{target}, list of cell types T, and scDesign2 algorithm ϱ
Ensure: membership predictions for each donor $p \in D_{target}$

$[C^1, C^2, \ldots, C^{|T|}] \leftarrow \varrho(D_{synth})$
for each cell type $t \in T$ **do**
  Let $D_{target,t}$ be the set of target cells of type t
  Let $\mathcal{R}_t$ be the covariance matrix, defined in C^t
  Let G_1 be the "group 1" genes in C^t
  Let G_2 be the "group 2" genes included in covariance matrix $\mathcal{R}_t$
  $U \leftarrow [0,1]^{n_t \times |G_2|}$
  for expression levels X_g^t of gene $g \in G_2$, for all cells x in $D_{target,t}$ **do**
    Let F_g^t be the fitted marginal CDF for gene g from C^t
    $U[:, g] \leftarrow F_g^t(X_g^t)$
    $z_x \leftarrow \Phi^{-1}(U[x, :])$
    $\mathcal{B}_x \leftarrow \sum_{g \in G_1} \log p_g^{\text{synth}}(x_g)$
    $\lambda_t[x] \leftarrow 1/d_s + \gamma_{\text{eff}} \cdot \mathcal{B}_x$
  end for
end for
for all donors p in D_{target} **do**
  $\lambda_p \leftarrow \text{avg}(\{\lambda_t[x] \mid x \in D_{target,p},\ t \in T\})$
end for
return $\sigma(\lambda)$

B Why the Enhanced Attack Needs Auxiliary Data

The rows marked "-aux (enhanced)" in Table 10 are at or below chance, and below the standard -aux rows above them. Algorithm 2 is the attack behind those rows. Without D_{aux} it cannot form the ratio in Eq. 3, so it keeps the numerator alone:

$$\sum_{g \in G_1} \log p_g^{synth}(x_g) \qquad \text{in place of} \qquad \sum_{g \in G_1} \log \frac{p_g^{synth}(x_g)}{p_g^{aux}(x_g)}.$$

The sum on the left measures how ordinary a cell is. Take a gene that is zero in 95% of the cells of a type. A cell with a zero count adds $\log 0.95$ to the sum and a cell with a count of 7 adds a large negative number, whichever donor the cell came from. Over all of G_1, the sum mostly records how many unusual counts the cell has.

The ratio on the right removes that part. The marginals fit to D_{synth} and to D_{aux} agree closely on what is ordinary (both give the zero count a probability near 0.95), so the ordinary part cancels. What remains is the small amount by which the fit to D_{synth} leans toward the counts of its own training donors, and that remainder is the membership signal.

The saved scores show both effects. In the black-box trials with 10 or more donors, the numerator alone varies about 14 times as much from cell to cell as the ratio does (the median, within a cell type, of the two standard deviations' quotient). Used as the whole score, the ratio gives a mean donor-level AUC of 0.79 and the numerator alone gives 0.53. After the weight γ, the numerator alone still varies about 20 times as much as the log-distance it is added to, so it replaces the distance signal, which is why the enhanced -aux rows fall below the standard ones. A smaller γ only moves the score back toward the standard attack, which is the case $\gamma = 0$. The enhanced attack is therefore defined only with auxiliary data.

C Extended Results

This section gives the full results behind the main text. Table 7 reports the fidelity of scDesign2 at every cohort size. LISI falls slowly as the cohort grows on OneK1K (0.91 to 0.86) and AIDA (0.87 to 0.71), and HFRA has the lowest LISI at every size. Table 10 reports every attack on scDesign2 at every cohort size; Figures 4 and 3 plot a subset of it. Table 8 reports attack success against each generator for the largest OneK1K cohort. Table 11 reports the fidelity of each generator at 10 and 20 donors on OneK1K, and of scDesign2 at each level of covariance noise. Table 9 reports the enhanced attack against each generator at 10 to 50 donors on OneK1K and AIDA.

	OneK1K			AIDA			HFRA		
Donors	**LISI ↑**	**ARI ↑**	**MMD $\times 10^3$ ↓**	**LISI ↑**	**ARI ↑**	**MMD $\times 10^3$ ↓**	**LISI ↑**	**ARI ↑**	**MMD $\times 10^3$ ↓**
2	0.91±.01	0.54±.16	0.03±.00	0.87±.01	0.76±.08	0.41±.05	0.56±.02	0.41±.14	0.51±.11
5	0.90±.01	0.53±.04	0.03±.01	0.83±.02	0.71±.11	0.35±.06	0.37±.07	0.44±.08	0.59±.11
10	0.88±.01	0.50±.06	0.04±.01	0.81±.01	0.74±.06	0.38±.04	0.27±.04	0.48±.06	0.65±.05
20	0.88±.00	0.50±.04	0.04±.00	0.78±.02	0.74±.04	0.35±.04	0.31±.06	0.45±.07	0.63±.12
50	0.87±.00	0.48±.02	0.04±.00	0.75±.01	0.77±.02	0.35±.02	–	–	–
100	0.86±.00	0.48±.02	0.04±.00	0.73±.00	0.74±.05	0.34±.02	–	–	–
200	0.86±.00	0.46±.04	0.04±.00	0.71±.00	0.73±.07	0.34±.01	–	–	–

Table 7: Quality metrics (LISI, ARI, and MMD) for **scDesign2** across datasets and donor counts. Values are mean ± std over 5 trials. ↑ = higher is better; ↓ = lower is better. MMD values scaled by 10^3. HFRA has only 22 donors, so its 20-donor row uses 11 training donors.

D Inverting the Copula Correlation Matrix

The Mahalanobis distance needs the inverse of each cell type's copula correlation matrix R, and R often has none. It is singular whenever a cell type has fewer cells than copula genes. In one 10-donor OneK1K trial, for example, a rare cell type had 33 synthetic cells and 238 copula genes, so its R could describe at most 33 independent directions of variation among the 238 genes. All results in the main text use the Moore–Penrose pseudo-inverse, which gives zero weight to the directions in which R has no variance. That is, a target cell can stray from the synthetic data in those directions as far as it likes without changing its distance.

SDG Method	Noise param.	BB-aux	BB+aux	*enhanced* BB+aux
scDesign2	(no noise)	$0.52 \pm .01$	$0.52 \pm .02$	$0.56 \pm .01$
	$\eta = 10^{-4}$	$0.52 \pm .01$	$0.53 \pm .02$	$0.56 \pm .01$
	$\eta = 10^{1}$	$0.51 \pm .01$	$0.51 \pm .01$	$0.54 \pm .01$
scVI		$0.51 \pm .01$	$0.51 \pm .02$	$0.55 \pm .02$
ZINB-WaVE		$0.51 \pm .01$	$0.53 \pm .03$	$0.57 \pm .02$

Table 8: **Donor-level AUC of scMAMA-MIA with 440 training donors** (OneK1K; mean $\pm$ std over 5 trials). The auxiliary set is disjoint from the target donors. The scDesign2 row repeats the 440-donor column of Table 10. At this cohort size, no attack exceeds an AUC of 0.57 against any generator.

	OneK1K			AIDA		
Generator	10	20	50	10	20	50
scDesign2	$0.89 \pm .08$	$0.89 \pm .07$	$0.79 \pm .05$	$0.95 \pm .05$	$0.90 \pm .08$	$0.90 \pm .05$
scDesign3-G	$0.89 \pm .08$	$0.88 \pm .07$	$0.78 \pm .06$	$0.98 \pm .02$	$0.91 \pm .09$	$0.89 \pm .05$
scDesign3-V	$0.90 \pm .08$	$0.89 \pm .07$	$0.79 \pm .06$	$0.87 \pm .15$	$0.74 \pm .22$	–
scVI	$0.81 \pm .11$	$0.83 \pm .04$	$0.69 \pm .08$	$0.85 \pm .11$	$0.77 \pm .11$	$0.72 \pm .04$
ZINB-WaVE	$0.85 \pm .08$	$0.88 \pm .05$	$0.78 \pm .11$	$0.89 \pm .09$	$0.81 \pm .09$	$0.75 \pm .02$
NMF	$0.77 \pm .13$	$0.67 \pm .08$	$0.51 \pm .07$	$0.75 \pm .18$	$0.72 \pm .12$	$0.60 \pm .05$

Table 9: Enhanced BB+aux AUC of scMAMA-MIA against six generators, by number of training donors (mean $\pm$ one standard deviation over 5 trials). The attack fits the same scDesign2 proxy copula to each generator's synthetic data. NMF, the one low-fidelity generator in the table, is the only one whose AUC falls to chance by 50 donors on OneK1K. We have no scDesign3-V result at 50 donors on AIDA.

We asked whether another way of inverting R gives a stronger attack, and compared the pseudo-inverse with three alternatives on every scDesign2 trial of Section 4 up to 100 donors. Ridge adds a small multiple of the identity before inverting, $(R + \alpha I)^{-1}$, so it keeps every direction and gives the empty ones a large, finite weight of $1/\alpha$. OAS shrinkage [7] also mixes R with the identity, with a weight set by the number of cells the copula was fit on. PCA goes the other way. It keeps the top 50% or 90% of the eigenvectors of the synthetic-data R and measures the distance in that subspace, so it ignores even more directions than the pseudo-inverse.

Table 12 gives the results. Ridge raises the AUC of the standard attacks on OneK1K and AIDA by 0.01 to 0.02, and its strength α hardly matters between 10^{-8} and 10^{-2}. It barely moves the enhanced attack (by at most 0.006), and on HFRA it lowers the standard BB+aux attack slightly. OAS stays within 0.013 of the pseudo-inverse, usually below it. PCA is lower almost everywhere, and keeping half of the components pushes the BB-aux attack below chance on OneK1K and AIDA. So the directions that the pseudo-inverse ignores carry some membership signal, and ignoring more of them costs more. Figure 6 (Appendix G) shows where ridge gains. In cell types with fewer than about 100 training cells, which mostly have fewer training cells than copula genes, the pseudo-inverse attacks reverse, with median AUCs between 0.00 and 0.14. The figure's second column repeats the -aux attacks with ridge, and there the median AUC of these cell types is 1.00. We suspect that a target cell that strays in a direction the synthetic data never varies in is evidence against membership, which ridge counts and the pseudo-inverse discards. We have not tested this explanation.

Before looking at these results, we fixed a rule for when to replace the pseudo-inverse. A candidate was chosen on two datasets and scored on the third, using the enhanced BB+aux attack, and it had to gain at least 0.01 AUC on average over the three held-out datasets. Ridge with $\alpha = 10^{-4}$ was chosen in all three folds, but its average gain was 0.002, so the main text keeps the pseudo-inverse. An attacker who uses ridge would do slightly better than the standard-attack AUCs we report for OneK1K and AIDA.

E Noise on the Covariance and on the Marginals

Section 4.5 perturbs only the copula covariance. Here we compare that with perturbing the per-gene marginals, using one scale s for both. For the covariance, we add $\mathcal{N}(0, s^2)$ noise to each entry of $\mathcal{R}$, then symmetrize, project to the nearest positive semidefinite matrix and re-normalize to a correlation matrix, as in Section 4.5. For the marginals, we multiply each gene's mean and dispersion by e^{ϵ} with $\epsilon \sim \mathcal{N}(0, s^2)$, and add $\mathcal{N}(0, s^2)$ noise to the logit of its zero-inflation probability.

Table 10: Donor-level AUC against scDesign2 (no added noise), by number of training donors: scMAMA-MIA under four threat models, standard and enhanced, and four baseline attacks (BB+aux). The submitted version reported OneK1K at 490 donors and AIDA at 200 donors with auxiliary donors that overlapped the targets. The 440-donor column and the AIDA 200-donor +aux entries use a disjoint auxiliary set. HFRA cannot supply one at 10 or 20 donors, and its 20-donor column uses 11 training donors.

Method	2	5	10	20	50	100	200	440
OneK1K								
WB+aux	1.00	0.99	0.96	0.94	0.85	0.78	0.64	0.60
WB+aux (enhanced)	1.00	0.95	0.89	0.95	0.85	0.76	0.73	0.61
WB-aux	1.00	0.93	0.84	0.81	0.73	0.59	0.58	0.53
WB-aux (enhanced)	0.30	0.42	0.47	0.46	0.55	0.51	0.52	0.51
BB+aux	1.00	0.81	0.79	0.70	0.66	0.57	0.59	0.52
BB+aux (enhanced)	1.00	0.97	0.89	0.89	0.79	0.72	0.65	0.56
BB-aux	1.00	0.68	0.66	0.64	0.63	0.54	0.59	0.52
BB-aux (enhanced)	0.25	0.39	0.46	0.46	0.55	0.51	0.57	0.51
LOGAN	0.85	0.51	0.59	0.57	0.47	0.53	0.49	—
GAN-Leaks	0.85	0.75	0.59	0.58	0.55	0.51	0.50	—
DOMIAS+KDE	0.30	0.60	0.55	0.52	0.50	0.48	0.50	—
MC	0.55	0.62	0.52	0.53	0.51	0.47	0.52	—
AIDA								
WB+aux	1.00	0.92	1.00	1.00	1.00	0.98	0.92	—
WB+aux (enhanced)	1.00	0.97	0.98	0.94	0.95	0.88	0.85	—
WB-aux	1.00	0.92	1.00	0.96	0.81	0.68	0.63	—
WB-aux (enhanced)	0.55	0.67	0.70	0.58	0.54	0.51	0.54	—
BB+aux	1.00	0.92	1.00	1.00	0.96	0.92	0.80	—
BB+aux (enhanced)	1.00	0.96	0.95	0.90	0.90	0.82	0.79	—
BB-aux	1.00	0.92	0.97	0.87	0.69	0.60	0.60	—
BB-aux (enhanced)	0.55	0.66	0.69	0.56	0.53	0.51	0.53	—
LOGAN	0.50	0.58	0.37	0.52	0.54	—	—	—
GAN-Leaks	0.81	0.61	0.66	0.63	0.59	—	—	—
DOMIAS+KDE	0.44	0.49	0.51	0.49	0.52	—	—	—
MC	0.69	0.36	0.66	0.53	0.48	—	—	—
HFRA								
				(11)				
WB+aux	0.95	0.94	—	—	—	—	—	—
WB+aux (enhanced)	0.70	0.59	—	—	—	—	—	—
WB-aux	0.85	0.74	0.65	0.65	—	—	—	—
WB-aux (enhanced)	0.60	0.52	0.56	0.62	—	—	—	—
BB+aux	0.95	0.94	—	—	—	—	—	—
BB+aux (enhanced)	0.70	0.59	—	—	—	—	—	—
BB-aux	0.85	0.75	0.64	0.63	—	—	—	—
BB-aux (enhanced)	0.60	0.52	0.55	0.61	—	—	—	—
LOGAN	0.45	0.57	0.45	0.46	—	—	—	—
GAN-Leaks	0.95	0.60	0.50	0.50	—	—	—	—
DOMIAS+KDE	0.65	0.52	0.55	0.50	—	—	—	—
MC	0.55	0.52	0.61	0.59	—	—	—	—

Method	Setting	OneK1K, 10 donors LISI↑	ARI↑	MMD$\times 10^3$ ↓	OneK1K, 20 donors LISI↑	ARI↑	MMD$\times 10^3$ ↓
scDesign2	no privacy	0.882	0.496	0.427	0.879	0.500	0.407
scDesign2	$\eta = 10^{-4}$	0.885	0.484	0.419	0.881	0.456	0.399
	$\eta = 10^{-3}$	0.882	0.474	0.404	0.880	0.504	0.396
	$\eta = 10^{-2}$	0.835	0.524	0.422	0.842	0.510	0.366
	$\eta = 10^{-1}$	0.738	0.570	0.543	0.726	0.525	0.454
	$\eta = 10^{0}$	0.719	0.518	0.562	0.692	0.555	0.488
	$\eta = 10^{1}$	0.714	0.539	0.555	0.682	0.532	0.496
scDesign3-G		0.901	0.487	0.277	0.891	0.464	0.252
scDesign3-V		0.896	0.483	0.313	0.892	0.477	0.290
scVI		0.913	0.439	0.347	0.907	0.434	0.318
scDiffusion		0.332	0.223	95.788	0.527	0.206	15.416
NMF	no DP	0.404	0.327	22.356	0.343	0.327	21.848
NMF	$\varepsilon = 10^{3}$	0.366	0.304	20.324	0.253	0.311	22.359
	$\varepsilon = 10^{2}$	0.370	0.295	20.498	0.245	0.311	22.622
	$\varepsilon = 10^{1}$	0.283	0.275	23.744	0.165	0.294	31.869
	$\varepsilon = 2.8$	0.019	0.293	66.479	0.035	0.290	56.870
	$\varepsilon = 10^{0}$	0.068	0.280	57.839	0.009	0.274	74.295

Table 11: **Fidelity metrics on the OneK1K dataset for 10 and 20 donor settings**. We report LISI (higher is better), ARI (higher is better), and MMD $\times 10^3$ (lower is better) across synthetic data generation (SDG) methods and levels of covariance noise η (Section 4.5). LISI falls as the noise grows; ARI does not.

	OneK1K			**AIDA**			**HFRA**		
Inverse of R	**BB+aux**	**BB-aux**	**Enh.**	**BB+aux**	**BB-aux**	**Enh.**	**BB+aux**	**BB-aux**	**Enh.**
Pseudo-inverse (main text)	0.754	0.691	0.876	0.965	0.842	0.922	**0.862**	0.717	**0.685**
Ridge, $\alpha = 10^{-8}$	0.768	**0.710**	**0.877**	**0.980**	0.862	**0.928**	0.858	**0.726**	**0.685**
Ridge, $\alpha = 10^{-6}$	0.768	**0.710**	**0.877**	**0.980**	0.862	**0.928**	0.858	**0.726**	**0.685**
Ridge, $\alpha = 10^{-4}$	0.768	**0.710**	**0.877**	**0.980**	**0.863**	**0.928**	0.858	**0.726**	**0.685**
Ridge, $\alpha = 10^{-2}$	**0.769**	0.708	**0.877**	**0.980**	**0.863**	**0.928**	0.860	0.725	**0.685**
OAS shrinkage	0.746	0.688	0.874	0.972	0.829	0.910	0.852	0.718	**0.685**
PCA, top 90% of components	0.714	0.608	0.863	0.971	0.810	0.916	0.843	0.673	0.684
PCA, top 50% of components	0.687	0.361	0.850	0.906	0.415	0.882	0.790	0.506	0.680

Table 12: Donor-level AUC of scMAMA-MIA against scDesign2 under different ways of inverting the copula correlation matrix R. Each entry is the mean over 5 trials, then over the donor counts of the dataset (2 to 100 donors for OneK1K and AIDA, 2 to 20 for HFRA). BB$\pm$aux are the standard black-box attacks and Enh. is the enhanced BB+aux attack. The best value in each column is in bold. The four ridge strengths agree to within 0.002. They match or beat the pseudo-inverse in every column except HFRA BB+aux, where they are 0.003 to 0.004 lower.

Table 13 shows scDesign2 at 50 donors. Covariance noise brings the standard attack to chance but leaves the enhanced attack near 0.7 even at $s = 2$, because the marginals that the enhanced attack also reads are untouched. Marginal noise lowers both attacks, but real and synthetic cells stop mixing (LISI of 0) before the enhanced attack reaches chance. That happens at $s = 1$ on OneK1K and at $s = 0.5$ on AIDA, where the enhanced attack is still at 0.59 and 0.75.

F True-Positive Rate at Low False-Positive Rates

Table 14 gives the true-positive rate of the standard and enhanced BB+aux attacks at false-positive rates of 1%, 5%, 10% and 20%, with a 95% interval on each AUC. Table 15 covers the other three threat models. Two attacks can be closer in AUC than at low FPR. On OneK1K at 20 donors, the standard and enhanced attacks have AUCs of 0.70 and 0.89, and they identify 4% and 50% of the members at a 1% FPR.

	OneK1K					AIDA				
Noise	LISI↑	ARI↑	MMD$\times 10^3$ ↓	BB+aux	Enh.	LISI↑	ARI↑	MMD$\times 10^3$ ↓	BB+aux	Enh.
No noise	0.87	0.46	0.4	$0.66 \pm .04$	$0.79 \pm .05$	0.76	0.74	3.5	$0.96 \pm .02$	$0.90 \pm .05$
Covariance, $s = 0.5$	0.76	0.51	0.4	$0.55 \pm .02$	$0.76 \pm .04$	0.55	0.75	3.7	$0.62 \pm .03$	$0.78 \pm .04$
Covariance, $s = 1$	0.70	0.48	0.5	$0.53 \pm .01$	$0.73 \pm .02$	0.49	0.76	3.7	$0.57 \pm .04$	$0.76 \pm .02$
Covariance, $s = 2$	0.65	0.49	0.5	$0.50 \pm .02$	$0.69 \pm .02$	0.44	0.75	3.7	$0.55 \pm .04$	$0.73 \pm .01$
Marginals, $s = 0.2$	0.85	0.48	1.3	$0.68 \pm .04$	$0.77 \pm .04$	0.46	0.77	5.0	$0.89 \pm .03$	$0.86 \pm .05$
Marginals, $s = 0.5$	0.48	0.48	8.0	$0.61 \pm .03$	$0.67 \pm .03$	0.00	0.55	14.3	$0.68 \pm .05$	$0.75 \pm .04$
Marginals, $s = 1$	0.00	0.47	33.9	$0.56 \pm .03$	$0.59 \pm .03$	0.00	0.54	44.0	$0.58 \pm .04$	$0.58 \pm .07$
Marginals, $s = 2$	0.00	0.47	101.4	$0.55 \pm .01$	$0.54 \pm .02$	0.00	0.55	116.0	$0.54 \pm .04$	$0.52 \pm .06$

Table 13: Noise on the scDesign2 copula covariance or on its per-gene marginals, at 50 donors (mean over 5 trials; $\pm$ is one standard deviation). BB+aux is the standard attack and Enh. the enhanced BB+aux attack. Covariance noise brings the standard attack to chance on OneK1K but leaves the enhanced attack at 0.69 or above. Marginal noise brings the enhanced attack near chance only once LISI is 0.

		Standard BB+aux						Enhanced BB+aux					
			TPR at FPR						TPR at FPR				
Dataset	**Donors**	AUC [95% CI]	1%	5%	10%	20%	Prec.	AUC [95% CI]	1%	5%	10%	20%	Prec.
OneK1K	2	1.00 [1.00, 1.00]	0.80	0.80	1.00	1.00	0.53	1.00 [1.00, 1.00]	1.00	1.00	1.00	1.00	0.53
	5	0.81 [0.67, 0.93]	0.40	0.48	0.64	0.72	0.42	0.97 [0.90, 1.00]	0.12	0.96	0.96	0.96	0.52
	10	0.79 [0.71, 0.88]	0.02	0.10	0.38	0.56	0.30	0.89 [0.82, 0.96]	0.02	0.46	0.64	0.80	0.42
	20	0.70 [0.62, 0.77]	0.04	0.13	0.21	0.38	0.19	0.89 [0.84, 0.94]	0.50	0.63	0.80	0.86	0.47
	50	0.66 [0.61, 0.70]	0.03	0.13	0.24	0.36	0.21	0.79 [0.75, 0.82]	0.18	0.32	0.48	0.60	0.35
	100	0.57 [0.54, 0.60]	0.03	0.07	0.12	0.24	0.12	0.72 [0.69, 0.75]	0.05	0.14	0.25	0.41	0.21
	200	0.59 [0.57, 0.61]	0.11	0.16	0.22	0.33	0.20	0.65 [0.63, 0.68]	0.12	0.17	0.25	0.37	0.22
	440	0.52 [0.51, 0.54]	0.01	0.06	0.11	0.22	0.11	0.56 [0.54, 0.57]	0.01	0.07	0.12	0.24	0.12
AIDA	2	1.00 [1.00, 1.00]	1.00	1.00	1.00	1.00	0.53	1.00 [1.00, 1.00]	1.00	1.00	1.00	1.00	0.53
	5	0.92 [0.80, 1.00]	0.00	0.96	0.96	0.96	0.52	0.96 [0.89, 1.00]	0.88	0.88	0.88	0.92	0.49
	10	1.00 [1.00, 1.00]	1.00	1.00	1.00	1.00	0.53	0.95 [0.90, 0.99]	0.76	0.80	0.90	0.90	0.50
	20	1.00 [0.99, 1.00]	0.95	0.98	0.99	1.00	0.52	0.90 [0.86, 0.94]	0.47	0.61	0.70	0.81	0.44
	50	0.96 [0.94, 0.97]	0.54	0.77	0.90	0.96	0.50	0.90 [0.87, 0.93]	0.16	0.44	0.61	0.83	0.40
	100	0.92 [0.90, 0.93]	0.25	0.61	0.75	0.85	0.45	0.82 [0.80, 0.85]	0.20	0.42	0.52	0.68	0.37
HFRA	2	0.95 [0.80, 1.00]	0.70	0.70	1.00	1.00	0.53	0.70 [0.50, 0.90]	0.20	0.20	0.40	0.60	0.31
	5	0.94 [0.86, 0.99]	0.36	0.64	0.88	0.92	0.49	0.59 [0.44, 0.75]	0.12	0.12	0.20	0.28	0.18

Table 14: **True-positive rate at low false-positive rates, BB+aux against scDesign2.** For each setting the donors of the 5 trials are pooled into one ROC curve, and TPR is the largest true-positive rate whose false-positive rate does not exceed the target. Prec. is the precision at a 10% FPR when one candidate donor in ten is a member; guessing gives 0.10. The 95% interval on the AUC resamples donors within each trial (2,000 resamples). With 2 or 5 donors per trial the pooled curve has 10 or 25 non-members, so a 1% FPR allows no false positive.

G Cell-Type Size and Attack Success

scDesign2 fits one copula per cell type, so a rare cell type is a copula fit on few cells. Figure 6 scores donors on one cell type at a time and plots the AUC against the number of training cells of that type. Between 300 and 1,000 training cells the median AUC over the three datasets is 0.89 to 1.00 in every panel. It falls as the cell type grows, to between 0.50 and 0.63 above 30,000 cells. On AIDA at 50 donors, for example, platelets (about 1,200 training cells) give an enhanced BB+aux AUC of 0.97 and CD14-positive monocytes (about 16,700) give 0.73. The same decline accounts for the fall in AUC with donor count in Figure 4, since each added donor adds cells to every copula.

Below about 300 cells the attacks that use the pseudo-inverse collapse. Under 100 cells the median AUC of BB-aux, WB-aux and the standard BB+aux attack is between 0.00 and 0.14. An AUC near 0 still separates members from non-members, with the sign reversed. The collapse belongs to our attack and not to the data. It occurs in the cell types with fewer training cells than copula genes (98% of the points under 100 cells), whose correlation matrix is singular. Between 100 and 300 cells, the median WB-aux AUC is 0.59 for such cell types and 1.00 for the others. The second column of the figure repeats the -aux attack with the ridge inverse of Appendix D ($\alpha = 10^{-4}$), and the median under 100 cells becomes 1.00 for both BB-aux and WB-aux. So the smallest cell types reveal the

		WB+aux			WB-aux			BB-aux		
Dataset	**Donors**	AUC	TPR@1%	TPR@10%	AUC	TPR@1%	TPR@10%	AUC	TPR@1%	TPR@10%
OneK1K	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90	1.00
	5	0.99	0.92	1.00	0.93	0.60	0.84	0.68	0.12	0.24
	10	0.96	0.38	0.92	0.84	0.04	0.34	0.66	0.00	0.04
	20	0.94	0.58	0.86	0.81	0.02	0.28	0.64	0.03	0.08
	50	0.85	0.14	0.56	0.73	0.08	0.21	0.63	0.06	0.16
	100	0.78	0.08	0.40	0.59	0.02	0.10	0.54	0.01	0.08
	200	0.64	0.02	0.22	0.58	0.02	0.12	0.59	0.11	0.21
	440	0.60	0.02	0.17	0.53	0.01	0.11	0.52	0.01	0.11
AIDA	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	5	0.92	0.00	0.96	0.92	0.00	0.96	0.92	0.00	0.96
	10	1.00	1.00	1.00	1.00	0.98	1.00	0.97	0.80	0.90
	20	1.00	1.00	1.00	0.96	0.55	0.91	0.87	0.29	0.61
	50	1.00	0.94	1.00	0.81	0.19	0.45	0.69	0.08	0.27
	100	0.98	0.78	0.97	0.68	0.07	0.29	0.60	0.04	0.18
	200	–	–	–	0.63	0.05	0.24	0.60	0.04	0.19
HFRA	2	0.95	0.80	1.00	0.85	0.50	0.70	0.85	0.50	0.70
	5	0.94	0.48	0.84	0.74	0.08	0.52	0.75	0.16	0.52
	10	–	–	–	0.65	0.02	0.08	0.64	0.04	0.04
	20	–	–	–	0.63	0.02	0.11	0.45	0.00	0.00

Table 15: Standard scMAMA-MIA against scDesign2 under the other three threat models: AUC (mean over trials) and TPR at a 1% and a 10% FPR, computed as in Table 14. A dash marks a setting whose auxiliary donors overlap the targets.

most about their donors. They hold few of a donor's cells, which is why the ridge inverse changes the donor-level AUC by only 0.01 to 0.02.

The datasets follow similar curves, which suggests that the differences between datasets in Figure 4 come largely from the size of their cell types, and not from anything else about their cells. HFRA and AIDA show this most clearly. HFRA has more cells per donor than AIDA (about 4,900 against 2,100) and divides them among 9 cell types instead of 33. So at 10 donors its median cell type is fit on about 3,500 training cells and AIDA's on about 300, and HFRA's points sit to the right of AIDA's in Figure 6, where every dataset's AUC is lower. At equal size, the two datasets agree. Above 300 training cells, the median AUCs of HFRA and AIDA are within 0.08 of each other in every size range for the standard BB-aux attack, which uses no auxiliary data and so is unaffected by HFRA's overlapping auxiliary donors. HFRA is therefore likely harder to attack than AIDA because its cell types are larger. The enhanced attack is the exception. At equal size it is weaker on HFRA than on AIDA, just as it is weaker than the standard attack on HFRA overall (Table 10). OneK1K fits this account less well. With 14 cell types, its median cell type at 50 donors is fit on about 2,300 cells, against 1,400 for AIDA. Above 300 training cells, the two datasets' enhanced BB+aux medians are within 0.06 of each other in every size range, but for the standard BB+aux attack AIDA is up to 0.21 above OneK1K at equal size. We have also not tested the explanation directly, for example by subsampling a cell type before fitting.

H Which Genes Carry the Membership Signal?

If a few genes carried most of the membership signal, a data holder could protect those genes specially, and an adversary could focus on them. scMAMA-MIA gives no explicit weight to any gene, so we measured each gene's contribution in two ways. First, we scored every donor with one gene alone, using that gene's log-likelihood ratio (its term in the enhanced attack, Eq. 3) averaged over the donor's cells of one cell type, and took the AUC of those scores. We call this the gene's stand-alone AUC. Second, we removed the gene and recorded how far the AUC of the enhanced BB+aux attack fell. We did this for all 1,118 highly variable genes in each of the 14 cell types of OneK1K, against scDesign2, at 50, 100, 200, and 440 donors with 5 trials each. That gives 15,319 pairs of a cell type and a gene, each averaged over up to 20 runs.

The signal is spread thinly over almost all of the genes. 90% of the pairs have a stand-alone AUC above 0.5, but the largest is only 0.541, and removing any one gene lowers the attack's AUC by at most 0.003. The ranking of the genes is also unstable. Between two disjoint sets of trials, the stand-alone AUCs have a Spearman correlation of only 0.11. So there is no small set of genes that the attack depends on, and none that a defense could single out.

We then asked whether the signal concentrates in genes under genetic control, which seemed the most likely place for a donor's fingerprint. The expression of some genes depends on a nearby genetic

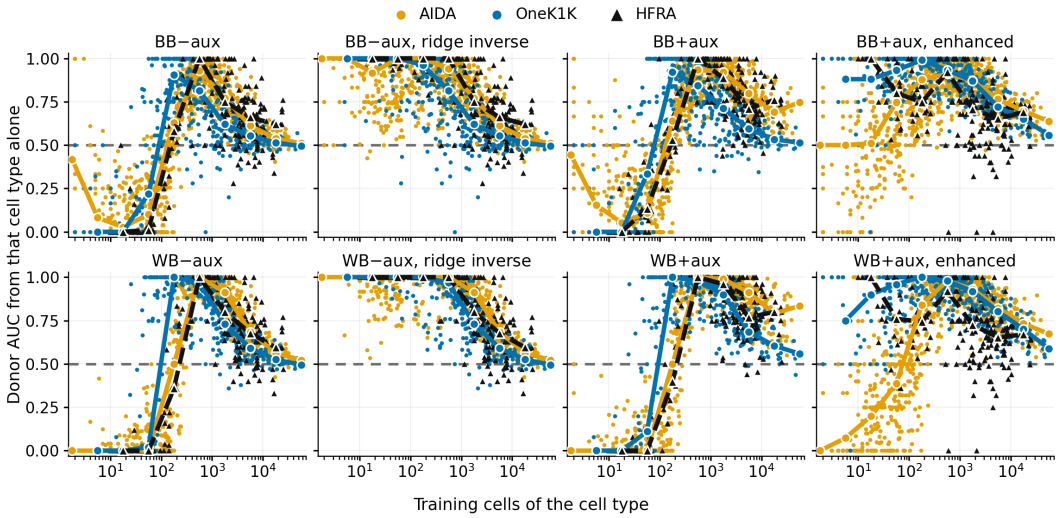

Figure 6: **The ridge inverse removes the collapse of the attack on cell types with fewer than about 100 training cells.** Donor-level AUC against scDesign2 when donors are scored on one cell type alone, by that cell type's number of training cells. Columns run from the weakest attack to the strongest. The second column uses the ridge inverse (Appendix D) and the others the pseudo-inverse. Each point is one cell type in one trial (5 trials; 2 to 100 donors, 2 to 20 for HFRA). Each line joins the median AUC of a dataset's points within buckets of training cells (1 to 3, 3 to 10, 10 to 30, and so on); buckets with fewer than 5 points are skipped. The dashed line is chance. HFRA's +aux points at 10 and 20 donors use auxiliary donors that overlap the targets.

variant, called an expression quantitative trait locus (eQTL), and we call such a gene an eGene. For each cell type, the OneK1K authors list its eGenes [48]. A donor's genotype is the same in every one of their cells and differs from other donors', so an eGene's expression is partly a property of the donor, and we expected eGenes to help identify donors. At first they seem to. Of our pairs, 2,047 are eGenes in that same cell type. Their mean stand-alone AUC is 0.5124, against 0.5096 for the other genes ($p = 3 \times 10^{-49}$, Mann–Whitney test), and they are 1.7 times as likely to be in the top tenth of a cell type's genes by signal (15.8% of eGenes, against 9.1% of the other genes).

But expression level accounts for this difference. An eQTL is easier to detect in a gene with many counts, so 28% of the genes in a cell type's top expression decile are eGenes, against 1% of those in the bottom decile. And highly expressed genes carry more signal, whether or not they are eGenes (a mean stand-alone AUC of 0.512 in the top decile and 0.504 in the bottom). When we compare eGenes only with the other genes in the same expression decile of the same cell type, the difference in mean AUC vanishes (0.0000, $p = 0.82$, permutation test). The eGenes are more donor-specific, as expected. The share of a gene's expression variance that lies between donors averages 0.023 for eGenes and 0.013 for the other genes, and the gap survives the matching ($p = 3 \times 10^{-4}$). But in both groups nearly all of the variance is between cells of the same donor, and once expression is matched, this share does not predict a gene's signal (Spearman correlation -0.004).

Removing the eGenes outright gives a similar picture. We reran the attack without each cell type's own eGenes (14% of its genes) and, as a control, without the same number of other genes drawn from the same expression deciles. At 50 donors the enhanced BB+aux AUC falls from 0.786 to 0.745 without the eGenes and to 0.778 without the control genes. At 440 donors it falls from 0.557 to 0.550 and 0.549. Over all 20 trials, removing the eGenes costs 0.012 more AUC than removing the control genes (standard error 0.008). The gap is largest at 50 donors, where it is 0.034 for both the enhanced attack (standard error 0.024) and the standard attack (standard error 0.009), and it is within 0.006 at 200 and 440 donors. So eGenes may matter somewhat more than equally expressed genes when donors are few, which five trials per donor count cannot settle. Either way, the membership signal is not a fingerprint held in a few genes under genetic control. It is a small contribution from nearly every gene, which the attack adds up over genes and then over a donor's cells. The whole analysis covers one dataset, one generator, and the black-box setting, and genes absent from the OneK1K list count as non-eGenes whether or not they were tested.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 1

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations at the end of section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: No theoretical results.

   Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Section 4 and 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide a public GitHub repository with clear instructions to provide open access to code. The datasets are sensitive and restricted and can be accessed via data agreements and/or the *CellxGene* website mentioned in Section 4.1.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report mean with standard deviations in main results in Tables 2 and 3 and report standard deviations as colored bands in figures.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section 4, in footnote 7.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research include sensitive datasets which was obtained through approved data agreements. No individuals were identified and we followed safe data practices and abide by Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Section 1 and 5.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We work with data with removed personally identifiable information. Access to these datasets is restricted to research use under a data usage agreement. We use models that are already publicly available and our research assesses the safety risks of releasing such models and datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We detail, acknowledge, and cite all related works. We arduously verify that all of our versions of software we are comparing against are implemented / installed correctly.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

Justification: Our codebase, given in Section 4, is well-documented with external validation in mind.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with direct human subjects. The human genomics data was obtained using data agreements for research use (openly accessible datasets)

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with direct human subjects. The human genomics data was obtained using data agreements for research use (openly accessible datasets)

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: See Section 4, footnote 5.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.