Fishing for Privacy: Machine Unlearning for Single-Cell VAEs

David Benson
Columbia University
dmb2262@columbia.edu

Abstract

Single-cell RNA sequencing models can memorize individual training samples, which is a problem when the data contains sensitive biological information. This paper tests whether machine unlearning can remove specific samples from a variational autoencoder (VAE) so that membership inference attacks (MIAs) can no longer detect them. Eight unlearning methods were evaluated against four attack families on two datasets (PBMC-33k and Tabula Muris). All eight methods fail on structured (biologically coherent) forget sets. Methods that treat unlearning as a small parameter perturbation (retain-only fine-tuning, gradient ascent, SSD, SCRUB) preserve utility perfectly but produce no measurable privacy improvement. Fisher scrubbing and contrastive latent unlearning make the model detectably worse rather than detectably better. Extra-gradient co-training shows high variance across seeds (mean advantage $= 0.300$, 95% CI $[0.226, 0.374]$). DP-SGD trained from scratch on the retain set comes closest to the retrain baseline (advantage $= 0.072$ vs. 0.046), but at a real utility cost and by construction, not by unlearning. The core finding is that memorization concentrates in biologically coherent subpopulations. Structured clusters show baseline MIA AUC of 0.78–0.89, while scattered random cells show 0.41–0.53. A Fisher information analysis reveals the structural cause: the VAE's shared decoder produces $17\times$ higher Fisher overlap between forget and retain sets than a classifier on the same data, so selective parameter perturbation cannot cleanly separate the two. Full retraining remains the only dependable option for structured forget sets.

1 Introduction

Deep learning models memorize training data, and for rare subpopulations this memorization may be necessary for generalization [Feldman, 2020]. When those models are trained on sensitive biological information, an attacker can determine whether a specific sample was part of the training set through a membership inference attack (MIA). This paper tests machine unlearning for VAEs trained on single-cell RNA sequencing (scRNA-seq) data, where gene expression profiles can reveal disease status, genetic predispositions, and other personal health information. Rare cell types like cancer cells or immune subtypes may be individually identifiable even within large, aggregated datasets.

The work here addresses subpopulation unlearning, meaning the removal of entire biological groups such as rare cell types or tissue-specific populations, not individual samples. This corresponds to scenarios where all cells from a specific disease subtype or tissue must be deleted. The experiments use cell type clusters as forget sets. Donor-level or patient-level deletion would require donor annotations, which neither dataset provides.

Given a trained VAE with parameters θ, a forget set $\mathcal{F}$, and a retain set $\mathcal{R}$, produce updated parameters θ' such that a post-hoc MIA cannot tell forget-set samples apart from samples that were never in training.

Preprint. Under review.

Privacy leakage is measured by membership inference advantage (adapted from Yeom et al. [2018]): advantage $= 2|\text{AUC} - 0.5|$. This is direction-agnostic, so AUC $= 0.38$ (over-unlearning) and AUC $= 0.62$ (under-unlearning) both give advantage $= 0.24$, since an adversary can always flip predictions. Advantage is 0 at chance and 1 for a perfect attack. Unlearning succeeds when the post-hoc advantage falls within the 95% CI of the retrain model's advantage.

Contributions.

1. An evaluation protocol for MIAs in single-cell VAEs that uses biologically matched negatives (nearest neighbors in latent space from the unseen set) to separate membership signal from cell-type signal, validated with within-cluster holdout controls.
2. Evidence that memorization concentrates in structured subpopulations. Coherent biological clusters have baseline MIA AUC of 0.78–0.89; scattered random cells have AUC of 0.41–0.53. The unlearning problem only matters for structured sets.
3. A systematic comparison of eight unlearning methods (frozen critics, extra-gradient, Fisher, retain-only fine-tuning, gradient ascent, SSD, SCRUB, contrastive latent) plus a DP-SGD baseline, against four attack families, with utility evaluation across four metrics (ELBO, KL divergence, ARI, marker gene correlation).
4. A catalog of failure modes: posterior collapse (Fisher on structured sets), critic exploitation (frozen adversarial training), over-unlearning/Streisand effect [Golatkar et al., 2020, Hayes et al., 2024] (contrastive latent, aggressive extra-gradient λ), parameter-space ineffectiveness (SSD, SCRUB), and dataset dependence (extra-gradient fails on Tabula Muris and multi-cluster compositions).
5. A Fisher information analysis showing that the VAE's shared decoder creates cosine similarity of 0.306 between forget-set and retain-set Fisher diagonals, $17\times$ higher than a classifier on the same data (0.018). A proposition formalizes this gap for linear decoders, showing that the Fisher cosine factorizes into residual-profile and latent-moment components, with the residual-profile factor scaling as $1 - O(M/D)$ for generative models and $O(1/\sqrt{C})$ for classifiers. Parameter-space unlearning methods designed for classifiers rely on a low-overlap regime that does not hold in the VAE.

2 Related work

Cao and Yang [2015] introduced formal definitions of machine unlearning. Bourtoule et al. [2021] proposed SISA training, which partitions data so that forgetting requires retraining only the affected shards. Approximate unlearning has theoretical guarantees for restricted model classes: Ginart et al. [2019] gave deletion-efficient algorithms for k-means, Guo et al. [2020] proposed certified removal via Newton-step updates for convex models, Sekhari et al. [2021] derived sample complexity bounds under strong convexity, Neel et al. [2021] showed that noisy gradient descent admits deletion guarantees in smooth settings, and Izzo et al. [2021] used projective residual updates for linear models. These results all depend on convexity or smoothness assumptions that deep generative models violate.

Golatkar et al. [2020] introduced Fisher information scrubbing ("Eternal Sunshine"), which identifies parameters most influenced by the samples to forget and adds noise proportional to that influence. Foster et al. [2024] proposed Selective Synaptic Dampening (SSD), which dampens parameters proportional to their forget-set Fisher importance. Kurmanji et al. [2023] introduced SCRUB, a teacher-student framework where the student matches the teacher on retain data while diverging on forget data. All three methods assume that forget-set influence concentrates in identifiable parameter subsets; Section 6 shows that assumption breaks down in shared-decoder architectures. Basu et al. [2021] found that influence functions produce unreliable estimates in non-convex models, weakening the case for influence-based parameter selection in unlearning.

Nasr et al. [2018] formulated privacy-preserving training as a min-max game where the model minimizes both task loss and attacker success while the attacker maximizes membership detection. Chavdarova et al. [2019] showed that extragradient updates stabilize GAN training by damping oscillations in simultaneous gradient descent-ascent. The present work applies extragradient to the VAE-attacker min-max problem.

Table 1: Dataset and split summary. HVGs = highly variable genes. Matched negatives are the k-nearest unseen cells to the structured forget set in baseline latent space ($k = 10$). Tabula Muris scattered uses 285 matched negatives.

	PBMC-33k	Tabula Muris
Cells	33,088	41,647
HVGs	2,000	2,000
Clusters	14	35
Tissues	1 (blood)	12
Train / Unseen	28,124 / 4,964	35,399 / 6,248
Structured forget set	Cluster 13 (30 megakaryocytes)	Cluster 33 (82 cardiac muscle)
Scattered forget set	35 random from train	30 random from train
Matched negatives	194	137

Abadi et al. [2016] introduced DP-SGD, which provides formal differential privacy guarantees through per-sample gradient clipping and Gaussian noise injection. DP-SGD trains from scratch on the retain set and is used here as a formal privacy baseline rather than an unlearning method.

Shokri et al. [2017] introduced shadow model attacks for membership inference. Carlini et al. [2022] proposed the likelihood ratio attack (LiRA), which trains multiple shadow models and compares per-example difficulty scores against calibrated null distributions. Hayes et al. [2024] showed that inexact unlearning can make forgotten samples *more* detectable, a Streisand effect [Golatkar et al., 2020]. Thudi et al. [2022] argued that without auditable definitions, unlearning claims are unverifiable in practice, a problem confirmed by the results here (Section 7).

Moon et al. [2024] used latent-space interventions to remove class-level features from pre-trained GANs and VAEs. Their goal is preventing generation of a class, not preventing detection that specific samples were in training, which is the distinction here.

Lopez et al. [2018] developed scVI, a VAE for scRNA-seq data using negative binomial likelihood, which is the architectural basis for the models tested here.

3 Datasets and setup

The PBMC-33k dataset consists of 33,088 peripheral blood mononuclear cells from 10x Genomics, preprocessed with Scanpy [Wolf et al., 2018], each represented by 2,000 highly variable gene counts. The structured forget set is cluster 13, which contains 30 rare megakaryocytes. The Tabula Muris dataset [The Tabula Muris Consortium, 2018] has 41,647 cells from 12 mouse tissues with 35 Leiden clusters; the structured forget set is cluster 33 (82 cardiac muscle cells). Table 1 gives a full comparison.

To control for biological confounds in MIA evaluation, matched negatives were selected as the unseen cells closest to the forget set in the baseline model's latent space (k-NN with $k = 10$). Without this control, an attacker could distinguish forget from unseen cells by cell type alone rather than by memorization. As a robustness check, a within-cluster holdout variant (using only unseen cells from cluster 13, $k = 5$) was also tested; it reduced baseline AUC from 0.769 to 0.527, confirming the biology confound but preserving relative method rankings (Appendix C).

4 Methods

4.1 VAE architecture

The VAE follows scVI design principles [Lopez et al., 2018] with a negative binomial likelihood for count data overdispersion. The encoder maps 2,000 input genes through layers of size [1024, 512, 128] to produce the latent mean μ and log-variance $\log \sigma^2$ (latent dimension 32). The decoder reverses this, mapping z through [128, 512, 1024] to negative binomial parameters. Layer normalization and dropout (0.1) are applied after each hidden layer. The training objective is the negative ELBO:

$$\mathcal{L}(\theta, \phi; x) = -\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + \mathrm{KL}(q_\phi(z|x) \| p(z)) \tag{1}$$

4.2 Unlearning methods

Eight unlearning approaches were tested, grouped below from simplest to most complex. A DP-SGD model trained from scratch on the retain set is included as a formal privacy baseline.

Retain-only fine-tuning. The simplest approach is to fine-tune the trained VAE on only the retain set ($\mathcal{R} = \mathcal{D} \backslash \mathcal{F}$), which is 99.9% of the original training data. This preserves utility (ELBO $= 363.2$, marker $r = 0.832$) but produces no measurable unlearning.

Gradient ascent. Maximize the loss on the forget set for 10–100 steps (learning rates in $[10^{-5}, 10^{-4}]$), then fine-tune on the retain set. No hyperparameter combination produced measurable unlearning.

Frozen critics. Freeze pre-trained MIA attackers and update only the VAE to minimize their success. This fails completely (AUC > 0.98) because the VAE exploits critic-specific blind spots without removing the underlying membership signal.

Extra-gradient co-training. Frame unlearning as a min-max game:

$$\min_{\theta} \max_{\psi} \mathcal{L}_{\text{VAE}}(\theta; \mathcal{R}) - \lambda \cdot \mathcal{L}_{\text{att}}(\theta, \psi; \mathcal{F}) \tag{2}$$

Use a two-step extragradient update [Chavdarova et al., 2019] to damp oscillations:

$$\theta^{k+1/2} = \theta^k - \eta_\theta \nabla_\theta \mathcal{L}(\theta^k, \psi^k) \tag{3}$$

$$\theta^{k+1} = \theta^k - \eta_\theta \nabla_\theta \mathcal{L}(\theta^{k+1/2}, \psi^{k+1/2}) \tag{4}$$

Training uses two-timescale update rules (TTUR) with attacker LR $5\times$ higher than VAE LR, three co-trained critics, and 50 epochs (stopping before equilibrium). A λ sweep over $\{5, 10, 15, 20\}$ found $\lambda = 10$ optimal on PBMC; $\lambda = 5$ over-unlearns and $\lambda = 15$+ under-unlearns.

Fisher information scrubbing. Fisher scrubbing [Golatkar et al., 2020] computes the diagonal Fisher information matrix on the forget set and updates parameters inversely proportional to curvature:

$$\theta \leftarrow \theta + \alpha \cdot (F + \lambda I)^{-1} \cdot \nabla_\theta \mathcal{L}_{\mathcal{F}}(\theta) \tag{5}$$

Hyperparameters: $\alpha = 10^{-4}$, $\lambda = 0.1$, 100 scrubbing steps, 10 fine-tuning epochs on the retain set.

Selective Synaptic Dampening (SSD). SSD [Foster et al., 2024] computes the diagonal Fisher information on the forget set, then dampens parameters proportional to their importance: $\theta \leftarrow \theta \cdot \max(1 - \alpha \cdot F_{\mathcal{F}}, 0)$, where α controls dampening strength. Unlike Fisher scrubbing, SSD multiplicatively shrinks parameters rather than adding gradient-scaled noise. Tested with $\alpha \in \{0.5, 1.0, 5.0\}$, 3 seeds each.

Contrastive latent unlearning. A VAE-specific approach that pushes forget-set latent representations toward the prior $p(z) = \mathcal{N}(0, I)$ while preserving retain-set representations near their original locations. The loss combines a KL-to-prior term on forget samples with a representation-preservation term on retain samples: $\mathcal{L} = \gamma \cdot \text{KL}(q_\phi(z|x_f) \| p(z)) + \lambda \cdot \|\mu_\phi(x_r) - \mu_{\text{orig}}(x_r)\|^2$, followed by retain-only fine-tuning. If forget samples map to the prior, they become indistinguishable from random latent samples. Tested with $\gamma \in \{0.1, 1.0, 10.0\}$, 3 seeds each.

SCRUB. SCRUB [Kurmanji et al., 2023] uses teacher-student distillation. The original (trained) model is the teacher. The student is trained to match the teacher's output on retain data while maximizing divergence on forget data: $\mathcal{L} = \mathcal{L}_{\text{match}}(\theta_s, \theta_t; \mathcal{R}) - \alpha_f \cdot \mathcal{L}_{\text{match}}(\theta_s, \theta_t; \mathcal{F})$, with alternating optimization (forget steps then retain steps). Adapted from classification to VAE outputs (reconstruction and KL). Tested with $\alpha_f \in \{0.1, 1.0, 10.0\}$, 3 seeds each.

DP-SGD baseline. DP-SGD [Abadi et al., 2016] trains a fresh VAE from scratch on only the retain set with per-sample gradient clipping (max norm $= 1.0$) and calibrated Gaussian noise to achieve a target privacy budget (ε, δ). Since the forget set is excluded from training entirely, this is not an unlearning method but a formal privacy baseline: membership inference should be at chance by construction. Tested with $\varepsilon \in \{1, 10, 50\}$, $\delta = 10^{-5}$, 50 epochs, 3 seeds for $\varepsilon = 10$.

Table 2: All methods on PBMC-33k structured forget set (cluster 13, $n = 30$). Multi-seed results evaluated with a fresh MLP attacker trained on baseline forget vs. matched negatives (baseline AUC $= 0.783$, retrain $= 0.523$). †Single-seed entries use a separate attacker (baseline $= 0.769$, retrain $= 0.481$). Advantage is computed per seed as $2|\text{AUC}_s - 0.5|$ then averaged; this differs from $2|\overline{\text{AUC}} - 0.5|$ for high-variance methods (e.g., extra-gradient: mean advantage $= 0.300$ vs. $2|0.429 - 0.5| = 0.142$). No approximate method achieves mean advantage ≤ 0.266 (retrain CI upper bound).

Method	Seeds	AUC	Advantage	Marker r	Status
Baseline (no unlearning)	–	0.783	0.565	0.831	–
Retain-only fine-tune	5	0.665 ± 0.007	0.331	0.832	FAIL
Gradient ascent	5	0.702 ± 0.004	0.404	0.832	FAIL
SSD ($\alpha = 1.0$)	3	0.725 ± 0.001	0.450	0.831	FAIL
SCRUB ($\alpha_f = 1.0$)	3	0.737 ± 0.002	0.474	0.832	FAIL
Contrastive latent ($\gamma = 1.0$)	3	0.153 ± 0.032	0.695	0.832	FAIL (Streisand)
Fisher scrubbing	3	0.814 ± 0.003	0.628	–	FAIL (worse)
Frozen critics† ($\lambda = 10$)	1	0.992	0.984	–	FAIL
Extra-grad ($\lambda = 10$)	10	0.429 ± 0.142	0.300	0.789	FAIL
DP-SGD ($\varepsilon = 10$)	3	0.464 ± 0.024	0.072	0.787	Near target
Full retrain	–	0.523	0.046	0.829	TARGET

4.3 Attack suite

Beyond the trained MLP attacker (69-dim features, spectral normalization, [256, 256] hidden layers, dropout 0.3), three additional attack families were tested to assess robustness:

- **Threshold attacks** (low-assumption): threshold on reconstruction loss, KL divergence, or ELBO individually. These require only black-box access to the model's loss components.
- **Likelihood-ratio attack** (high-assumption): compares per-sample loss under the target model vs. a reference model (retrain). Requires access to a second model.
- **k-NN latent attack** (high-assumption): classifies membership based on distance to known training samples in latent space ($k = 10$). Requires access to training set embeddings.

The retrain model is the evaluation target, with advantage $= 0.046$ (AUC $= 0.523$, bootstrap 95% CI for advantage: $[0.004, 0.266]$). The wide CI reflects the small matched evaluation set ($\sim$200 samples). Unlearning succeeds when a method's mean advantage falls within this interval.

5 Experiments

Reference values throughout are baseline MIA AUC $= 0.783$ (advantage $= 0.565$), retrain AUC $= 0.523$ (advantage $= 0.046$, CI upper $= 0.266$). Single-seed entries marked † in Table 2 use a separate evaluation attacker (baseline $= 0.769$, retrain $= 0.481$).

5.1 Unlearning methods on structured forget set

Table 2 shows all eight methods plus the DP-SGD baseline.

Utility-preserving methods that fail on privacy. Retain-only fine-tuning, gradient ascent, SSD, and SCRUB all preserve utility almost perfectly (marker $r \geq 0.831$, matching baseline) but produce no useful privacy reduction. Thirty forget-set cells leave too small a gradient signal relative to the 28,094 retain cells. SSD [Foster et al., 2024] dampens parameters proportional to their Fisher importance on the forget set, but the structured subpopulation's influence is distributed across the parameter space. At $\alpha = 1.0$ (3 seeds), advantage $= 0.450$. At $\alpha = 5.0$ (single seed), advantage drops to 0.268, near the retrain threshold (0.266), but utility is preserved identically (marker $r = 0.831$), suggesting the dampening is not removing membership information so much as adding noise at high α. SCRUB [Kurmanji et al., 2023] alternates forget-divergence and retain-matching steps,

Table 3: Fisher unlearning on PBMC-33k by forget set type (3 seeds). Fisher reduces AUC on scattered sets, where baseline memorization is already weak. On structured sets it increases AUC through posterior collapse.

Forget type	Mean AUC	Std	Baseline AUC	Status
Structured	0.814	0.003	0.769	FAIL (worse)
Scattered	0.499	0.004	0.525	At chance

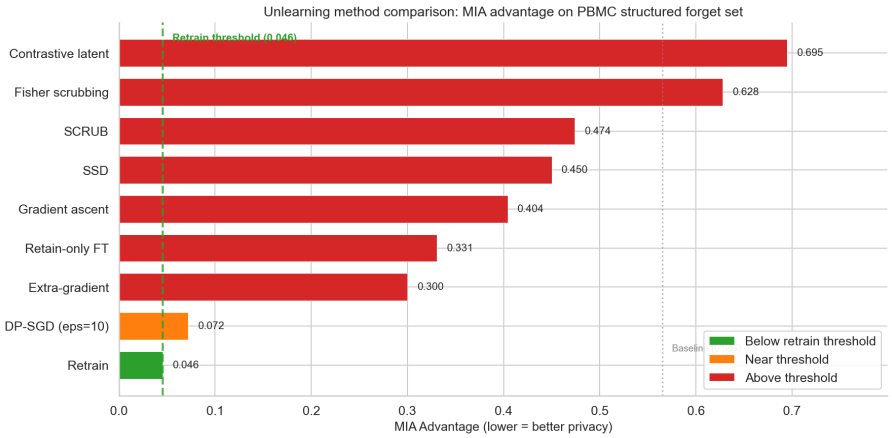

Figure 1: MIA advantage by method on PBMC-33k structured forget set ($n = 30$). The dashed line marks the retrain advantage (0.046). DP-SGD comes closest (0.072) but trains from scratch. No post-hoc method falls below the threshold.

but the retain objective dominates regardless of α_f, with advantage $= 0.474$ at $\alpha_f = 1.0$, and a sweep from 0.1 to 10.0 moves advantage by less than 0.02.

Methods that create detectable artifacts. Contrastive latent unlearning produces a Streisand effect [Hayes et al., 2024], with AUC dropping to 0.153 (advantage $= 0.695$), worse than the baseline's 0.565. Pushing forget-set representations away from their natural latent location does not erase membership information; it creates a new artifact the attacker detects instead. Fisher scrubbing similarly makes the model worse rather than better (advantage $= 0.628$), with KL collapsing from 10.55 to 0.007 on the forget set. Frozen critics (AUC $= 0.992^{\dagger}$) exploit fixed decision boundaries without removing the underlying signal.

Adversarial and formal methods. Extra-gradient with $\lambda = 10$ (10 seeds) gives mean advantage $= 0.300$ with 95% CI $[0.226, 0.374]$, exceeding the retrain CI upper bound (0.266). Per-seed variance is high ($\sigma_{\text{AUC}} = 0.142$), and the method fails entirely on Tabula Muris (Section 5.4).

DP-SGD at $\varepsilon = 10$ (3 seeds) achieves advantage $= 0.072 \pm 0.048$, the closest to the retrain target (0.046). But DP-SGD trains from scratch on the retain set; the forget set was never in training, so this is privacy by exclusion, not by unlearning. The utility cost is real. ELBO rises from 364 to 403 and marker r drops from 0.831 to 0.787. An ε sweep shows a non-monotonic tradeoff. At $\varepsilon = 1$, the model degrades so severely (marker $r = 0.495$) that the damage itself becomes detectable (advantage $= 0.292$); $\varepsilon = 50$ over-trains slightly (advantage $= 0.123$).

5.2 Fisher unlearning by forget set type

Table 3 separates the two cases. On scattered forget sets, Fisher achieves AUC $= 0.499$ (near chance), but the baseline is already 0.525, so the actual privacy gain is $\Delta = 0.026$. These cells were barely memorized to begin with. On structured sets, Fisher makes things worse. The memorization problem is concentrated in structured subpopulations, and the one method that works on scattered sets fails where it is most needed.

Table 4: Attack suite results on PBMC-33k structured forget set. Advantage $= 2|\text{AUC} - 0.5|$. Low-assumption attacks use only loss components; high-assumption attacks require a reference model or training set embeddings. The trained MLP (canonical attacker) is shown for comparison.

	Attack	Baseline	Retrain	EG λ=10	Fisher
Low	Recon. threshold	0.438	0.438	0.438	0.438
	KL threshold	0.841	0.733	0.192	0.272
	ELBO threshold	0.838	0.725	0.023	0.454
High	Likelihood ratio	0.839	0.839	0.734	0.744
	k-NN latent	0.986	0.971	0.897	0.927
	Trained MLP	0.538†	0.038†	0.036*†	0.628†

*Single seed; multi-seed mean advantage $= 0.300$. †Separate evaluation attacker.

Table 5: Cross-dataset comparison (PBMC-33k vs. Tabula Muris). PBMC values use the single-seed attacker (baseline $= 0.769$, retrain $= 0.481$; cf. Table 2 multi-seed attacker where baseline $= 0.783$, retrain $= 0.523$). The Tabula Muris retrain AUC (0.944) exceeds the baseline (0.891), confirming that the attacker detects biological structure rather than membership on this dataset.

Method	PBMC		Tabula Muris		Gen.?
	AUC	Baseline	AUC	Baseline	
Retrain (structured)	0.481	0.769	0.944	0.891	–
EG $\lambda = 10$ (structured)	0.482	0.769	0.874	0.891	No
Fisher (structured)	0.814	0.769	0.946	0.891	Yes (fails both)
Fisher (scattered)	0.499	0.525	0.568	0.411	–

5.3 Attack diversity

Table 4 shows how four attack families perform across methods. Three patterns emerge.

First, reconstruction-loss thresholding is identical across all models (advantage $= 0.438$), because all models reconstruct the forget set similarly. The membership signal lives in the KL and latent structure, not in reconstruction quality.

Second, ELBO-based thresholding is the most informative low-assumption attack for extra-gradient, which reduces ELBO advantage from 0.838 (baseline) to 0.023 (near chance). KL thresholding tells a complementary story. Extra-gradient reduces KL advantage from 0.841 to 0.192, and Fisher reduces it to 0.272, near the retrain threshold. However, Fisher's low KL advantage reflects posterior collapse (KL itself is destroyed rather than normalized), not genuine unlearning.

Third, high-assumption attacks (k-NN latent, likelihood ratio) detect all models, including retrain (advantage ≥ 0.73). These attacks pick up on biological structure because megakaryocytes occupy a distinct region of latent space regardless of whether the model was trained on them. This means high-assumption attacks in this setting measure cell-type identity, not membership. The trained MLP attacker, which combines all signal types with learned weights, provides the most balanced evaluation.

5.4 Cross-dataset validation

Table 5 shows results on Tabula Muris. The retrain model itself has AUC $= 0.944$, exceeding the baseline (0.891). Since retrain never saw cluster 33, this confirms that the attacker detects cardiac muscle cell biology rather than membership. Against this confounded baseline, extra-gradient does not transfer: AUC $= 0.874$. A λ sweep over $\{5, 7, 10, 15, 20\}$ gave identical results; every run early-stopped at epoch 6 with AUC ≈ 0.87. This is not a hyperparameter problem. The multi-tissue structure in Tabula Muris creates biological signals that the VAE cannot remove while still reconstructing the data.

Fisher fails on structured sets in both datasets (PBMC 0.814, TM 0.946), confirming that posterior collapse on correlated forget sets is a general property of this approach. On scattered sets the picture

Table 6: Utility metrics on PBMC-33k held-out set (4,964 cells). Marker r is the average Pearson correlation across 8 cell-type marker genes. Full per-gene results in Appendix A.

	ELBO ↓	KL	Marker r ↑	ARI ↑
Baseline	364.2	10.55	0.831	0.452
Retrain	365.6	9.69	0.829	0.488
Retain-FT	363.2	10.59	0.832	0.461
Grad-ascent	363.3	10.53	0.832	0.441
SSD ($\alpha = 1.0$)	363.9	10.40	0.831	0.449
SCRUB ($\alpha_f = 1.0$)	363.8	10.41	0.832	0.451
Contrastive ($\gamma = 1.0$)	364.0	10.38	0.832	0.458
Extra-grad ($\lambda = 10$, 10-seed)	403.7	4.70	0.789	0.508
Fisher	489.8	0.007	0.628	0.481
DP-SGD ($\varepsilon = 10$, 3-seed)	403.3	8.08	0.787	0.482

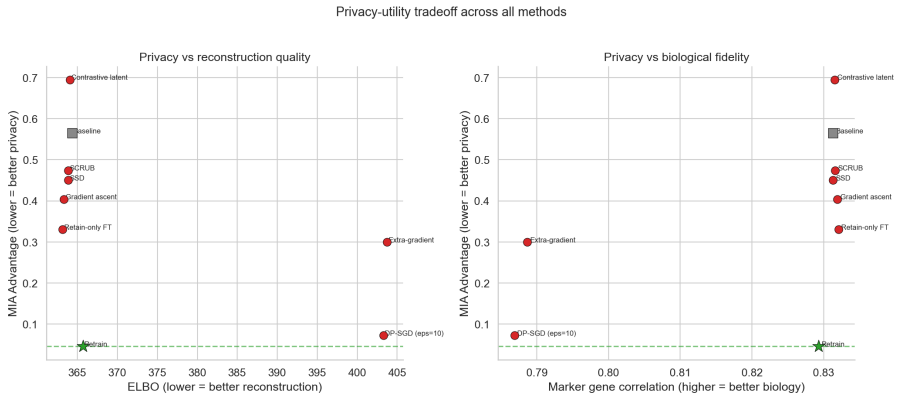

Figure 2: Privacy-utility tradeoff across all methods. Left: advantage vs. ELBO. Right: advantage vs. marker gene correlation. Methods that preserve utility (upper cluster) fail on privacy; methods that reduce advantage (DP-SGD, extra-gradient) pay a utility cost. Only retrain achieves both.

is muddier. Fisher AUC $= 0.568$ vs. a Tabula Muris baseline of 0.411 (AUC below 0.5 means the labels are flipped; advantage $= 0.178$ baseline, $= 0.136$ Fisher).

5.5 Utility evaluation

Utility was measured on the held-out set (4,964 cells) across four metrics (Table 6). Five methods (retain-FT, gradient ascent, SSD, SCRUB, contrastive) preserve utility almost identically to baseline (ELBO ≈ 364, marker $r \geq 0.831$). These methods barely change the model, which is why they fail on privacy. Extra-gradient and DP-SGD both degrade ELBO by $\sim$40 and marker r by $\sim 5\%$, the cost of reducing membership signal. Fisher is worst, with ELBO $= 490$ and marker $r = 0.628$ due to posterior collapse (KL $= 0.007$). ARI is preserved across all methods, indicating that global cluster structure survives even when gene-level reconstruction degrades.

Figure 2 shows the privacy-utility tradeoff across methods. The five utility-preserving methods (retain-FT, gradient ascent, SSD, SCRUB, contrastive) cluster in the upper-left with excellent ELBO but high advantage. Extra-gradient and DP-SGD trade utility for privacy but neither reaches the retrain target without substantial reconstruction degradation. No method occupies the lower-left region (low advantage, low ELBO) except retrain.

6 Why parameter-space methods fail

The previous section shows that all eight methods fail to match retrain-level privacy on the structured forget set. Four methods (retain-FT, gradient ascent, SSD, SCRUB) produce no privacy improve-

Table 7: Fisher information overlap between forget and retain sets. Cosine similarity measures how entangled forget-set influence is with retain-set influence in parameter space. Higher cosine means less room for selective parameter perturbation. The classifier's low overlap shows that class-specific output weights allow targeted forgetting; the VAE's shared decoder does not.

Model	Layer category	Parameters	Cosine sim.	Eff. rank ratio (F)
VAE	Encoder	2,642,816	0.273	0.030
VAE	Bottleneck	8,256	0.291	0.033
VAE	Decoder hidden	598,912	0.232	0.041
VAE	Decoder output	4,100,000	0.362	0.006
VAE	**Global**	**7,349,984**	**0.306**	0.010
Classifier	**Output layer**	**462**	**0.018**	0.037

ment; three (contrastive, Fisher, frozen critics) create detectable artifacts. This section provides a structural explanation.

Fisher information overlap. Parameter-space unlearning methods work by identifying which parameters are "important" for the forget set and modifying those parameters. The standard tool for this is the diagonal Fisher information matrix, $F_i = \mathbb{E}[(\partial\mathcal{L}/\partial\theta_i)^2]$, which measures how much each parameter contributes to the loss on a given dataset. Fisher scrubbing, SSD, and SCRUB all use variants of this idea.

For this approach to work, the forget-set Fisher F^f and retain-set Fisher F^r must be separable, meaning there must exist parameters that are important for the forget set but not the retain set. The cosine similarity $\cos(F^f, F^r) = \langle F^f, F^r\rangle/(\|F^f\|\|F^r\|)$ measures this separability. If it is close to 1, no parameter-space perturbation can reduce forget-set loss without proportionally affecting retain-set loss.

The diagonal Fisher was computed on the forget set (30 megakaryocytes) and retain set (28,094 cells) using the baseline PBMC model. Table 7 reports the results per layer category, alongside a linear classifier (logistic regression on frozen VAE latent codes, 14 Leiden clusters) trained on the same data.

The VAE's global Fisher cosine similarity is 0.306, driven by two effects. First, the dispersion output layer (2.05M parameters, 28% of the model) has cosine $= 0.998$, meaning these parameters respond identically to forget and retain data. Second, the mean output layer ($\cos = 0.360$, 2.05M parameters) shares the same 1,024-dimensional hidden representation across all 2,000 output genes, so perturbations to any column affect reconstruction of all cell types. At the individual gene level, 73% of genes (1,462/2,000) have per-gene Fisher cosine > 0.5, and the median is 0.877.

The classifier's cosine similarity is 0.018, $17\times$ lower. The forget set (cluster 13, megakaryocytes) contributes Fisher information primarily to the 32 weights connecting the latent space to the cluster-13 logit. Retain-set Fisher spreads across all 14 class logits. These weight sets are nearly orthogonal, making selective perturbation possible.

A structural bound on Fisher overlap. The pattern above admits a closed-form explanation for linear decoders.

Proposition 1 *Let $f(z) = Wz + b$ with $W \in \mathbb{R}^{D\times H}$ and squared-error loss. If residuals $e_d = x_d - f(z)_d$ are independent of latent activations z_h, the diagonal Fisher factorizes as $F_{dh} = 4\,\mathbb{E}[e_d^2]\,\mathbb{E}[z_h^2]$, and*

$$\cos(F^{\mathcal{F}}, F^{\mathcal{R}}) = \cos(\sigma^{\mathcal{F}}, \sigma^{\mathcal{R}}) \cdot \cos(\nu^{\mathcal{F}}, \nu^{\mathcal{R}}) \tag{6}$$

where $\sigma^S \in \mathbb{R}^D$ has entries $\sigma_d^S = \mathbb{E}_{x\sim S}[e_d^2]$ (residual variance per output dimension) and $\nu^S \in \mathbb{R}^H$ has entries $\nu_h^S = \mathbb{E}_{x\sim S}[z_h^2]$ (latent second moment).

Proof sketch. $\partial\mathcal{L}/\partial W_{dh} = -2e_d z_h$, so $F_{dh} = 4\,\mathbb{E}[e_d^2 z_h^2] = 4\,\sigma_d\nu_h$ under independence. Then $\langle F^{\mathcal{F}}, F^{\mathcal{R}}\rangle = 16\,\langle\sigma^{\mathcal{F}}, \sigma^{\mathcal{R}}\rangle\,\langle\nu^{\mathcal{F}}, \nu^{\mathcal{R}}\rangle$ and $\|F^S\| = 4\,\|\sigma^S\|\,\|\nu^S\|$. Dividing gives (6). □

Remark. The factorization quantifies a privacy-utility tradeoff for Fisher-based methods. For a perturbation that weights each parameter by its forget-set Fisher importance ($\delta\theta_i^2 \propto F_i^{\mathcal{F}}$), the ratio

of retain-set to forget-set curvature disturbance is $\cos(F^{\mathcal{F}}, F^{\mathcal{R}}) \cdot \|F^{\mathcal{R}}\|/\|F^{\mathcal{F}}\|$. When the cosine is near zero (classifier), perturbation directions exist that primarily affect the forget set. At cosine $= 0.306$ (VAE), every unit of forget-set curvature change imposes a proportional cost on the retain set.

Corollary 2 (Dimensional scaling) *Under the conditions of Proposition 1:*

1. *If $D - M$ of the D output dimensions satisfy $\sigma_d^{\mathcal{F}} = \sigma_d^{\mathcal{R}}$ and the remaining M dimensions have $\sigma_d^S \leq V\bar{\sigma}$ for all S, then*

$$\cos(\sigma^{\mathcal{F}}, \sigma^{\mathcal{R}}) \;\geq\; \frac{D - M}{D - M + MV^2}.$$

2. *For a single-class forget set where $\sigma^{\mathcal{F}}$ is supported on a single coordinate k, $\cos(\sigma^{\mathcal{F}}, \sigma^{\mathcal{R}}) = \sigma_k^{\mathcal{R}}/\|\sigma^{\mathcal{R}}\|$, which equals $1/\sqrt{C}$ for balanced C-class residuals.*

Proof. (i) The inner product $\langle \sigma^{\mathcal{F}}, \sigma^{\mathcal{R}} \rangle \geq \sum_{\text{shared}} \bar{\sigma}_d^2 = A$, while $\|\sigma^S\|^2 \leq A + MV^2\bar{\sigma}^2$ for each set S. Then $\cos \geq A/(A + MV^2\bar{\sigma}^2)$. Writing $A = (D - M)\bar{\sigma}^2$ gives the bound, which tends to 1 as $D/M \to \infty$. (ii) When $\sigma^{\mathcal{F}} = \sigma_k^{\mathcal{F}} \mathbf{e}_k$, the cosine reduces to $\sigma_k^{\mathcal{R}}/\|\sigma^{\mathcal{R}}\|$. For balanced classes ($\sigma_j^{\mathcal{R}} = c$ for all j), this is $c/(c\sqrt{C}) = 1/\sqrt{C}$. □

This factorization separates Fisher cosine into two interpretable components. The input to fc_mean is the 1,024-dimensional decoder hidden output h, not the 32-dimensional latent z. KL regularization constrains z toward $\mathcal{N}(0, I)$ but does not act on h. Because forget and retain data pass through the same shared decoder network, the hidden activation profiles are similar ($\cos(\nu^{\mathcal{F}}, \nu^{\mathcal{R}}) = 0.80$). Corollary 2 formalizes the gap between these settings. In the VAE, roughly 50–100 marker genes out of $D = 2{,}000$ have substantially different residual variance for the forget subpopulation. Under the linear-decoder conditions of Proposition 1, the relevant σ is the data-direct residual variance per gene, for which $\cos(\sigma^{\mathcal{F}}, \sigma^{\mathcal{R}}) = 0.83$, above the lower bound of 0.68 from part (i) with $M = 100$, $V = 3$. The Fisher-marginal σ gives a lower cosine of 0.51 because the negative binomial likelihood couples output dimensions, departing from the linear-MSE assumptions. For the classifier ($C = 14$), part (ii) gives an upper bound of $1/\sqrt{14} \approx 0.27$; the measured value is lower still because the forget-class gradient is more concentrated than a point mass at 95% accuracy.

Empirical verification. The fc_mean Fisher matrix is approximately rank-1: the leading singular value explains 94% of the Frobenius norm for the forget set and 96% for the retain set, consistent with the outer-product form $F_{dh} \approx 4\sigma_d\nu_h$. The factorized prediction $\cos(\sigma^{\mathcal{F}}, \sigma^{\mathcal{R}}) \cdot \cos(\nu^{\mathcal{F}}, \nu^{\mathcal{R}}) = 0.41$ overestimates the measured cosine of 0.37 by 11%. The dominant source of error is the softmax nonlinearity in the decoder, which couples all output dimensions through its Jacobian. Fisher-marginal ν matches data-direct hidden second moments (cosine > 0.99), but Fisher-marginal σ diverges from data-direct residual variance (cosine $= 0.50$–0.73), and the product computed directly from data statistics (0.65) overestimates by 77%. The Kronecker structure of the actual Fisher (94–96% rank-1) is tighter than the data-level interpretation of σ as residual variance per gene would suggest. The $17\times$ gap over the classifier (0.018) is consistent with the structural prediction.

Controlling for model capacity. The linear classifier has 462 parameters against the VAE's 7.35M. A deep MLP classifier (2,000 → [512, 128] → 14 with layer normalization and dropout, 1.09M parameters) trained on the same raw gene expression (95.2% accuracy) separates the effect of model capacity from architecture. The shared hidden layers have cosine $= 0.262$, in the same range as the VAE encoder (0.273) and decoder hidden layers (0.232). The class-specific output layer has cosine $= 0.010$, comparable to the linear probe (0.018). Fisher overlap therefore depends on whether parameters are shared across classes or specific to individual classes, not on model size.

Architecture generalization. Repeating the analysis with a VAE using latent dimension $z = 8$ instead of $z = 32$ (same hidden layers) yields a global cosine of 0.846, higher than the baseline. The bottleneck layers drive this increase (cosine $= 0.858$ vs. 0.291 for $z = 32$), because a smaller latent space concentrates the representation. The encoder (0.314) and decoder hidden layers (0.216) remain in the same range as $z = 32$. Reducing the latent dimension does not reduce Fisher overlap.

Cluster-conditional decoder. If shared output parameters cause the overlap, conditioning the decoder on cell type should reduce it. A conditional VAE was trained in which the mean output layer receives the final hidden representation concatenated with a 14-dimensional cluster one-hot vector (adding 28K parameters). The fc_mean weight matrix decomposes into shared columns (corresponding to the 1,024 hidden features) and cluster-specific columns (corresponding to the 14 one-hot inputs). The shared columns have cosine $= 0.222$, reflecting irreducible overlap from the shared hidden representation. The cluster-specific columns have cosine ≈ 0 (1.2×10^{-8}), achieving the same near-orthogonality as the classifier output weights. To test whether this structural separation improves unlearning, Fisher scrubbing was applied to the conditional VAE with the same hyperparameters as the standard experiments. The conditional VAE baseline shows higher memorization (advantage $= 0.73$) than the standard VAE (0.54), likely because the cluster-specific columns give the model additional capacity to distinguish cluster 13. After scrubbing, the advantage drops to 0.72, a negligible reduction, while the standard VAE achieves advantage 0.63 after scrubbing. Fisher overlap persists throughout the shared network in the conditional model, with encoder cosine $= 0.433$, bottleneck $= 0.508$, and decoder hidden layers $= 0.346$. Since the MIA attacker draws 64 of its 69 feature dimensions from encoder outputs (latent mean and log-variance), the high encoder overlap is the dominant factor. Routing cluster-specific information through dedicated output parameters does not reduce the memorization already encoded in the shared encoder and hidden layers.

Implication. This difference explains the empirical results. Fisher scrubbing, SSD, and SCRUB all modify parameters proportional to forget-set Fisher magnitude. When the forget-set Fisher overlaps heavily with the retain-set Fisher (cosine = 0.306 for the VAE), these modifications necessarily damage retain performance or leave the forget set's influence intact. The same structural problem applies to any generative model with shared output parameters. Classifier unlearning methods can succeed because class-specific output weights create a low-overlap regime that does not exist in the standard VAE architecture.

Diagonal Fisher approximation. The analysis above uses the diagonal Fisher, which ignores off-diagonal curvature. Kunstner et al. [2019] showed that the diagonal approximation can misrepresent the true Fisher in deep networks, particularly when parameters interact through shared activations. Two points limit this concern. First, Fisher scrubbing [Golatkar et al., 2020], SSD [Foster et al., 2024], and SCRUB [Kurmanji et al., 2023] all use the diagonal Fisher internally, so the overlap measured above characterizes these methods as implemented, not an idealized version of them. If the diagonal Fisher is the wrong tool, then these methods are already using the wrong tool, and the overlap analysis correctly identifies why they fail. Second, the log-Fisher correlation ($r = 0.73$) and the per-gene analysis (73% of genes with cosine > 0.5) provide converging evidence at different granularities, reducing the chance that diagonal artifacts drive the conclusion.

7 Discussion

The results in Sections 5–6 show three failure modes, a structural explanation for one of them, and an irreducible tradeoff between privacy and utility. The numbers are reported above. Here the question is what these failures imply for practice.

Unlearning as a detectable operation. Contrastive latent unlearning, Fisher scrubbing, and frozen critics all share a problem. The operation of unlearning itself leaves a trace. Contrastive pushes latent representations to an unnatural location; Fisher collapses the posterior on forget-set-specific dimensions; frozen critics drive the model toward critic-specific blind spots. In each case the model's behavior on forget-set samples becomes *more* distinguishable from its behavior on never-seen samples, not less. This is a variant of the Streisand effect [Golatkar et al., 2020, Hayes et al., 2024], and it poses a design constraint for future methods. Any unlearning procedure that modifies the model's response to forget-set inputs in a consistent direction risks creating exactly the signal it aims to remove.

Verification without a retrain oracle. A practical gap across all methods is the absence of a verification mechanism. None of the eight methods can confirm, at the end of the procedure, that membership information has been removed. The only ground truth is comparison against the retrain model. In a real deployment, the data controller would not have the retrain model available (retrain-

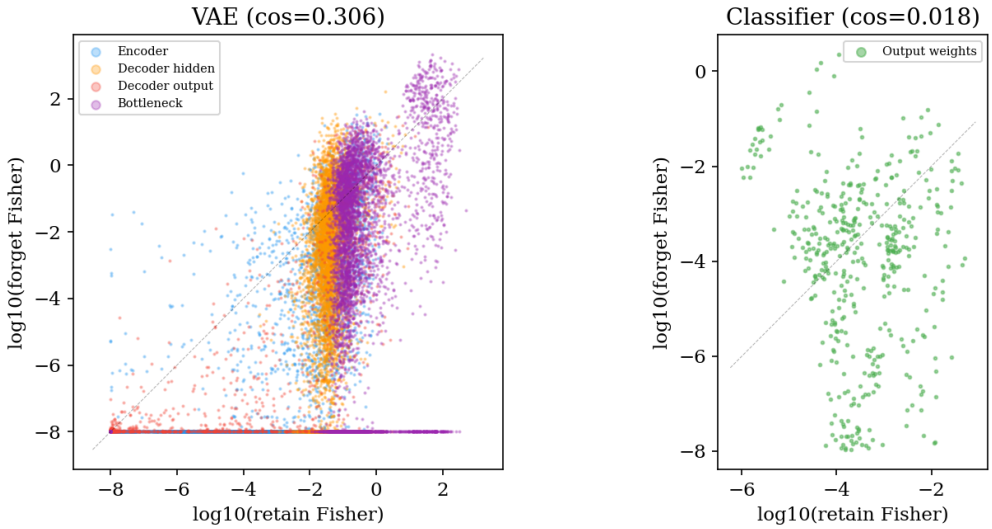

Figure 3: Per-parameter Fisher magnitude (log scale) for forget vs. retain sets. Left: VAE parameters are correlated (log-Fisher $r = 0.73$), confirming that the same parameters matter for both sets. Right: classifier parameters show no correlation ($\cos = 0.018$), with forget-set Fisher concentrated in a few class-specific weights.

ing is the alternative they want to avoid). This means approximate unlearning in this setting is both technically difficult and unverifiable. DP-SGD sidesteps verification by offering a formal guarantee, but it is not an unlearning method because it requires training from scratch and accepting a utility cost (marker r drops by 5%, ELBO rises by $\sim$40).

Biology vs. membership in MIA evaluation. The attack diversity results (Table 4) expose a confound that affects all MIA-based unlearning evaluation on biological data. High-assumption attacks (k-NN latent, likelihood ratio) show advantage ≥ 0.83 on the retrain model, because megakaryocytes occupy a distinct latent region regardless of training history. These attacks measure cell-type identity, not membership. The trained MLP attacker, which combines multiple feature types with learned weights, provides the most informative evaluation by separating the two signals. Future work on MIA evaluation for structured forget sets should be explicit about which signal an attack detects. This confound is likely present in any domain where forget sets correspond to coherent data subpopulations (disease subtypes, demographic groups, geographic clusters).

Limitations. All privacy guarantees here are empirical, not formal (except DP-SGD, which provides formal (ε, δ) guarantees but at substantial utility cost). Tabula Muris evaluation is confounded by tissue-of-origin signals. Even the retrain model (which never saw the forget set) has AUC $= 0.944$, exceeding the baseline. Only two VAE latent dimensions were tested ($z = 32$ and $z = 8$). The size ablation (Appendix B) conflates forget-set size with biological heterogeneity (larger sets span more clusters). Forget sets are cell-type clusters, not donors or patients, so the clinical relevance is limited.

Broader impact. This work studies methods for removing sensitive biological data from trained models. The results are largely negative. All approximate methods fail on the structured forget sets where memorization is strongest, and the paper documents those failures. The matched-negative evaluation protocol and attack diversity analysis may be useful to others working on privacy in single-cell genomics.

8 Conclusion

Memorization in single-cell VAEs is concentrated in structured subpopulations. Scattered cells have low baseline MIA AUC (0.41–0.53) and barely need unlearning. Structured cells, which are the ones that pose a privacy risk, resist all eight approximate methods tested here.

The methods fall into three failure modes. Four methods (retain-FT, gradient ascent, SSD, SCRUB) preserve utility perfectly but produce no privacy improvement. Three methods (contrastive, Fisher, frozen critics) create detectable artifacts that make things worse. Extra-gradient co-training reduces advantage to 0.300 but with high per-seed variance and no generalization to Tabula Muris. DP-SGD at $\varepsilon = 10$ achieves advantage $= 0.072$ by excluding the forget set from training entirely, establishing the cost of formal privacy guarantees (marker r drops by 5%, ELBO rises by ~ 40).

Full retraining remains the only reliable option. The Fisher overlap analysis (Section 6) identifies the structural cause. The VAE's shared decoder creates cosine similarity of 0.306 between forget-set and retain-set Fisher information, $17\times$ higher than the equivalent classifier (0.018). This gap holds across model capacities (deep MLP classifier with 1.09M shared-hidden parameters still shows class-specific output cosine of 0.010) and latent dimensions ($z = 8$ yields even higher overlap than $z = 32$). A cluster-conditional decoder reduces overlap in the class-specific output columns to near zero, but irreducible overlap in the shared hidden representation remains.

The contributions of this paper are the evaluation protocol (matched negatives, advantage metric, attack diversity, multi-seed validation), the systematic comparison of eight methods with utility characterization, the Fisher overlap analysis explaining why parameter-space methods fail for generative models, and the empirical documentation of where memorization concentrates.

Future work. Adaptive λ scheduling during extra-gradient training could extend the method to larger or multi-cluster forget sets. Methods that combine adversarial verification with parameter-space updates (rather than treating them separately) are worth testing. The conditional decoder experiment shows that cluster-specific output weights achieve near-zero Fisher overlap, but the shared encoder (cosine $= 0.433$), bottleneck (0.508), and hidden layers (0.346) retain high overlap, and the MIA attacker draws 64 of 69 features from encoder outputs. Reducing this overlap would require modular pathways through the entire network, not just the output layer, at the cost of losing shared representation learning across cell types. Donor-level forget sets, which require donor annotations not available in the current datasets, would strengthen the clinical relevance.

Acknowledgments

Code: `https://github.com/db-d2/Machine_Unlearning`.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM Conference on Computer and Communications Security (CCS)*, pages 308–318, 2016.

Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *International Conference on Learning Representations (ICLR)*, 2021.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, pages 141–159, 2021.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, pages 463–480, 2015.

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

Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Vitaly Feldman. Does learning require memorization? A short tale about a long tail. In *ACM Symposium on Theory of Computing (STOC)*, pages 954–959, 2020.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *AAAI Conference on Artificial Intelligence*, pages 12043–12051, 2024.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Zou. Making AI forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 3513–3526, 2019.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9301–9309, 2020.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *International Conference on Machine Learning (ICML)*, pages 3832–3842, 2020.

Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2024.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2008–2016, 2021.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

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.

Saemi Moon, Seunghyuk Cho, and Dongwoo Kim. Feature unlearning for pre-trained GANs and VAEs. In *AAAI Conference on Artificial Intelligence*, volume 38, pages 21420–21428, 2024.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *ACM Conference on Computer and Communications Security (CCS)*, pages 634–646, 2018.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 931–962, 2021.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 18075–18086, 2021.

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

The Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562(7727):367–372, 2018.

Anvith Thudi, Hengrui Jia, Ilia Shumailov, and Nicolas Papernot. On the necessity of auditable algorithmic definitions for machine unlearning. In *USENIX Security Symposium*, pages 4007–4022, 2022.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium (CSF)*, pages 268–282, 2018.

A Full utility metrics

Table 8: Per-gene marker correlation (r) on PBMC-33k held-out set for 8 cell-type marker genes. Fisher shows severe degradation on NK markers (GNLY, NKG7), consistent with posterior collapse disrupting specific decoder pathways.

Gene	Baseline	Retrain	Retain-FT	Grad-asc.	Extra-grad	Fisher
CD3D	0.767	0.761	0.770	0.769	0.602	0.601
CD3E	0.645	0.642	0.644	0.648	0.532	0.552
MS4A1	0.809	0.812	0.810	0.808	0.693	0.761
CD79A	0.900	0.900	0.901	0.901	0.727	0.815
CD14	0.798	0.797	0.800	0.797	0.767	0.644
LYZ	0.901	0.901	0.900	0.899	0.875	0.835
NKG7	0.930	0.928	0.928	0.929	0.698	0.515
GNLY	0.900	0.894	0.903	0.903	0.636	0.304
Mean	0.831	0.829	0.832	0.832	0.691	0.628

Extra-gradient preserves reconstruction quality for most genes but degrades T-cell (CD3D/E) and NK-cell (NKG7/GNLY) markers; the per-gene values shown are from a single seed and the 10-seed mean marker r (0.789) is higher than this seed. Fisher is worst on GNLY ($r = 0.304$ vs. baseline 0.900) and NKG7 ($r = 0.515$), consistent with posterior collapse disrupting cell-type-specific decoder mappings. Retain-only fine-tuning, gradient ascent, SSD, SCRUB, and contrastive latent unlearning are all indistinguishable from baseline on all genes (marker $r \geq 0.831$), consistent with these methods making only minimal parameter changes. DP-SGD at $\varepsilon = 10$ shows a similar per-gene pattern to extra-gradient, with moderate degradation on T-cell and NK markers.

B Forget set size ablation

Table 9: PBMC-33k structured size ablation for extra-gradient $\lambda = 10$ (3 seeds except $n = 30$ which uses the 10-seed evaluation from Table 2). Larger forget sets span multiple clusters and the method fails to unlearn them.

Size	Clusters	Mean AUC	95% CI	Advantage	Status
10	13 (subset)	0.263 ± 0.055	[0.200, 0.326]	0.474	Over-unlearn
30	13 (all)	0.429 ± 0.142	[0.327, 0.530]	0.300	FAIL
50	13 + 12	0.503 ± 0.109	[0.379, 0.626]	0.203	Borderline
100	13 + 12 + 11	0.611 ± 0.028	[0.580, 0.642]	0.222	Insufficient

At $n = 10$, extra-gradient over-unlearns (advantage $= 0.474$). At $n = 30$ (10 seeds), mean AUC is 0.429 with high variance. At $n = 50$, the mean AUC of 0.503 is near the retrain floor, and mean advantage $= 0.203$ falls within the retrain CI ($[0.004, 0.266]$). However, per-seed AUCs span 0.41 to 0.66, so individual seeds range from over-unlearning to under-unlearning. At $n = 100$, the method under-unlearns. A separate scattered-set ablation ($n = 10$ to 500, 10 seeds) confirmed that scattered cells have low baseline memorization (AUC 0.52–0.63) and Fisher provides at most marginal improvement (Δ of -0.01 to 0.04).

C Matched negative validation

The default matched negatives (194 cells, k-NN from unseen set with $k = 10$) include cells from multiple clusters (98.5% non-megakaryocytes), which means the attacker could partially rely on cell-type differences rather than pure membership. A within-cluster control using only the 5 unseen cluster-13 cells produced: baseline AUC $= 0.527$, retrain AUC $= 0.609$, extra-gradient AUC $= 0.463$. The biology confound inflates the default baseline (0.769 vs. 0.527). Retrain AUC exceeds 0.5 in the within-cluster setting because both groups (30 forget, 5 unseen) were excluded from

retraining, and with only 5 negatives the attacker overfits to within-cluster heterogeneity. Within-cluster matching would be preferable but is limited by the small number of held-out cells from rare clusters.

NeurIPS Paper Checklist

1. **Claims.** Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Yes. The abstract states that all eight approximate methods fail on structured sets and that memorization concentrates in structured subpopulations. Both claims are supported by Tables 2–5.

2. **Limitations.** Does the paper discuss the limitations of the work performed by the authors?

   Yes. Section 7 includes a dedicated limitations paragraph covering empirical-only guarantees, computational cost, architectural scope, and the use of cell-type (rather than donor) forget sets.

3. **Theory.** For each theoretical result, does the paper provide the full set of assumptions and a complete proof?

   Yes. Section 6 presents Proposition 1, which derives the Fisher cosine factorization for linear decoders under a residual-latent independence assumption. The proof sketch is provided. The assumption and the gap between the linear prediction and the measured nonlinear values are discussed. The diagonal Fisher approximation is explicitly addressed.

4. **Reproducibility.** 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?

   Yes. All hyperparameters, architectures, seeds, data splits, and evaluation procedures are specified in Sections 3–5. Code is publicly available.

5. **Open access to data and code.** Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results?

   Yes. Code is at `https://github.com/db-d2/Machine_Unlearning`. Both datasets (PBMC-33k from 10x Genomics, Tabula Muris) are publicly available.

6. **Experimental setting/details.** Does the paper specify all the training and test details necessary to understand the results?

   Yes. Data splits, optimizer settings, learning rates, batch sizes, number of epochs, architecture dimensions, attacker features, and evaluation thresholds are given in Sections 3–4.

7. **Experiment statistical significance.** Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Yes. Multi-seed experiments (3–10 seeds) with 95% confidence intervals are reported for all main results.

8. **Experiments compute resources.** For each experiment, does the paper provide sufficient information on the computer resources needed to reproduce the experiments?

   Yes. Experiments were run on an Apple M-series laptop (CPU only). Typical training times: baseline VAE $\sim$20 min, extra-gradient $\sim$80 min, DP-SGD $\sim$250 min (50 epochs), retain-FT/gradient ascent/SSD/SCRUB/contrastive $\sim$1–3 min each, Fisher $\sim$2 min.

9. **Code of ethics.** Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Yes. This is privacy research conducted on publicly available datasets. No human subjects were involved.

10. **Broader impacts.** Does the paper discuss both potential positive societal impacts and negative societal impacts of the work?

    Yes. Section 7 includes a broader impact statement. The work studies methods for removing sensitive data from models; the results are largely negative (most methods fail), and the paper documents those failures.

11. **Safeguards.** Does the paper describe safeguards that have been put in place for responsible release of data or models?

    N/A. No new models or datasets are released. All experiments use existing public data.

12. **Licenses.** 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?

    Yes. PBMC-33k (10x Genomics, CC-BY), Tabula Muris (Nature, CC-BY), and all referenced methods are properly cited.

13. **New assets.** Are new assets introduced in the paper well documented and is the URL or other publicly available method for accessing these assets provided?

    Yes. The code repository is documented and linked in the acknowledgments.

14. **Crowdsourcing and research with human subjects.** For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots?

    N/A. No crowdsourcing or human subjects research was conducted.

15. **IRB approvals or equivalent for research with human subjects.** Does the paper describe potential risks incurred by study participants, whether compensation was adequate, and whether informed consent was obtained?

    N/A. Only publicly available datasets were used.

16. **Declaration of LLM usage.** Does the paper declare the use of LLMs?

    No. LLMs were not used in the research methodology, experimental design, or analysis.