# Machine Unlearning for Single-Cell VAEs: Fishing for Privacy

David Benson
Columbia University
dmb2262@columbia.edu

*Abstract*—**Single-cell RNA sequencing models trained on patient data may memorize individual samples, creating privacy risks. This study examines machine unlearning for variational autoencoders (VAEs), asking whether specific training samples can be removed so that membership inference attacks cannot distinguish "forgotten" cells from truly unseen cells. Two approaches are compared: adversarial unlearning, where the VAE is trained to fool an attacker, and Fisher information scrubbing, which perturbs parameters based on their influence on specific samples. On PBMC-33k data, frozen-critic adversarial methods fail completely (AUC $>$ 0.98). Extragradient co-training with $\lambda = 10$ achieves the target band (AUC = 0.48 vs floor = 0.48) for structured forget sets, but requires approximately 42 minutes compared to 10 minutes for full retraining. Fisher unlearning succeeds on scattered forget sets (AUC = 0.50) but fails on structured sets (AUC = 0.81). For this dataset, full retraining remains the most practical approach since it is faster and guarantees data removal.**

## I. Introduction

Deep learning models can memorize their training data and for models trained on sensitive patient information, this creates a privacy problem. An attacker might be able to determine whether a specific individual's data was used to train the model. This is called a membership inference attack (MIA).

Privacy regulations like GDPR give individuals the "right to be forgotten," the ability to request that their data be deleted. For machine learning models, this is challenging. Simply deleting the original data file does not remove what the model learned from that data. The obvious solution is to retrain the model from scratch without the deleted samples, but this can be computationally expensive for large models.

Machine unlearning aims to modify a trained model so that it behaves as if certain samples were never in the training set, without full retraining. This study applies machine unlearning to VAEs trained on single-cell RNA sequencing (scRNA-seq) data. Single-cell data is particularly sensitive because gene expression profiles can reveal disease status, genetic predispositions, and other personal health information. Furthermore, rare cell types (e.g., cancer cells or immune subtypes) may be uniquely identifiable even within aggregated datasets.

**Input and Output.** The input to the unlearning algorithm consists of (1) a trained VAE with parameters $\theta$, (2) a "forget set" $\mathcal{F}$ of samples to remove, and (3) a "retain set" $\mathcal{R}$ of samples to keep. The output is updated parameters $\theta'$ such that a membership inference attacker cannot distinguish forget-set samples from samples that were never in training.

**Evaluation Criterion.** Unlearning succeeds if a post-hoc membership inference attack achieves AUC within $\pm 0.03$ of the "retrain floor," defined as the AUC measured on a model retrained from scratch without the forget set. An AUC near 0.5 means the attacker cannot distinguish members from non-members, which is the goal.

## II. Related Work

**Machine Unlearning.** Cao and Yang [2] introduced formal definitions of machine unlearning. Bourtoule et al. [1] proposed SISA training, which partitions data so that forgetting only requires retraining affected shards.

**Fisher Information Scrubbing.** Golatkar et al. [4] introduced "Eternal Sunshine," which uses Fisher information to identify parameters most influenced by samples to forget, then adds noise proportional to that influence.

**Adversarial Training for Privacy.** Nasr et al. [7] formulated privacy-preserving training as a min-max game for classifiers in which the model minimizes both task loss and attacker success while the attacker maximizes membership detection. Their approach uses alternating gradient descent.

**Extragradient Methods.** Chavdarova et al. [3] showed that extragradient updates stabilize GAN training by damping the rotational dynamics of simultaneous gradient descent-ascent. This study applies extragradient to the VAE-attacker min-max game for unlearning.

**Membership Inference Attacks.** Shokri et al. [8] introduced shadow model attacks for membership inference. Hayes et al. [5] showed that inexact unlearning can create a "Streisand effect" where forgotten samples become more detectable, not less.

**Single-Cell VAEs.** Lopez et al. [6] developed scVI, a VAE for scRNA-seq data using negative binomial likelihood. This study implements a similar architecture.

## III. Dataset

### A. PBMC-33k

The dataset consists of 33,088 peripheral blood mononuclear cells (PBMCs) from 10x Genomics, each represented by expression counts for 2,000 highly variable genes. Cell types include T cells, B cells, NK cells, monocytes, and rare megakaryocytes (cluster 13, containing 30 cells). The rare cluster is of particular interest because rare cell types are potentially easier to memorize.

### B. Preprocessing

Standard scRNA-seq preprocessing was applied using scanpy [9], including quality control filtering, library size normalization (CPM), log transformation, selection of top 2,000 highly variable genes, and Leiden clustering for cell type annotation.

### C. Data Splits

Two forget set configurations were tested. The structured set contains all 30 cells from cluster 13, forming a coherent biological group. The scattered set contains 30 randomly selected cells with no shared structure.

**Matched Negatives.** A naive MIA evaluation would compare forget cells to random unseen cells. However, this is confounded

| Split | Size | Description |
|---|---:|---|
| Train | 28,124 | 85% of data |
| Unseen | 4,964 | 15% held out |
| Forget (structured) | 30 | Cluster 13 (rare cells) |
| Forget (scattered) | 30 | Random from train |
| Retain | 28,094 | Train minus forget |
| Matched negatives | 194 | k-NN from unseen |

TABLE I
DATA SPLITS USED IN EXPERIMENTS.

by biological structure. If forget cells are rare megakaryocytes and unseen cells are common T cells, the attacker might distinguish them based on cell type rather than membership. To control for this, "matched negatives" were selected as the 194 unseen cells closest to the forget set in latent space (k-NN with k=7). These are biologically similar to forget cells but were never in training, providing a fair comparison.

## IV. Methods

### A. VAE Architecture

The VAE follows scVI design principles. Single-cell RNA-seq data consists of discrete read counts with high variance and many zeros. The negative binomial distribution handles this overdispersion better than Gaussian or Poisson alternatives, which is why it has become standard for scRNA-seq modeling.

**Encoder.** Input (2,000 genes) → [1024, 512, 128] → ($\mu$, $\log\sigma^2$), latent dim 32. Layer normalization and dropout (0.1) are applied after each hidden layer.

**Decoder.** Latent $z$ → [128, 512, 1024] → negative binomial parameters (mean and dispersion).

The model minimizes the negative ELBO

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

Training uses Adam optimizer with learning rate $10^{-3}$, batch size 128, and KL annealing over the first 10 epochs. The model is trained for 100 epochs with early stopping based on validation ELBO.

### B. Membership Inference Attack

The MIA is a multilayer perceptron (MLP) with spectral normalization, designed to detect whether a sample was in the training set based on model behavior. The input is a 69-dimensional feature vector comprising the latent mean $\mu$ (32 dims), latent log-variance $\log\sigma^2$ (32 dims), reconstruction loss (1 dim), KL divergence (1 dim), ELBO (1 dim), and L2 norms of latent statistics (2 dims).

The attacker architecture uses two hidden layers of 256 units each with dropout (0.3) and spectral normalization to prevent overfitting. Spectral normalization constrains the Lipschitz constant of each layer, which stabilizes training and prevents the attacker from memorizing the small training set.

The attacker is trained on a balanced dataset of forget-set samples (positive class) and matched negatives from the unseen set (negative class). Importantly, a fresh post-hoc attacker is trained after each unlearning method completes. This is critical because an attacker trained during the unlearning process might be fooled by adversarial artifacts, while a post-hoc attacker reveals whether membership signal truly remains.

The success criterion is post-hoc AUC $\in [0.451, 0.511]$, i.e., within $\pm 0.03$ of the retrain floor (0.481). The retrain floor represents the best achievable privacy since a model retrained without the forget set has no memory of those samples.

### C. Adversarial Unlearning

The adversarial approach frames unlearning as a min-max game

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

**Frozen Critics.** This approach freezes a pre-trained attacker and only updates the VAE. Two variants were tested. Single-critic uses one frozen attacker. Multi-critic uses an ensemble of three frozen attackers with different initializations, hoping diversity would prevent the VAE from finding universal blind spots. Both fail completely (AUC $>$ 0.98) because the VAE learns to fool even diverse critics without truly forgetting.

**Extra-gradient Co-training.** Chavdarova et al. [3] introduced a two-step update that dampens oscillations

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

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

With large $\lambda$ (e.g., 10), the game is privacy-dominant. Training uses two-timescale update rule (TTUR) with attacker learning rate $5\times$ higher than VAE learning rate. The faster attacker updates help the critic track the changing VAE representations, preventing the VAE from easily escaping. Training runs for 50 epochs and stops before the game reaches equilibrium. Extended training allows the VAE to eventually find ways to encode membership information that the current critics miss, so early stopping is essential.

### D. Fisher Information Unlearning

Fisher scrubbing perturbs parameters to reduce "memory" of the forget set

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

where $F_{ii}$ is the diagonal Fisher information for parameter $i$.

**Hyperparameters.** The scrubbing step size is $\alpha = 10^{-4}$, regularization $\lambda = 0.1$, with 100 scrubbing steps followed by 10 epochs of fine-tuning on the retain set to restore utility.

**Limitation for VAEs.** Unlike classifiers where each class has relatively independent parameters, VAE decoders store shared generative mappings across all data. When scrubbing parameters important for a rare cluster, those same parameters often contribute to reconstructing other cell types. This creates a dilemma since aggressive scrubbing damages utility for retained cells while conservative scrubbing leaves membership signal intact. For structured forget sets, scrubbing can cause "collapse to prior" where the decoder learns to ignore latent codes for forgotten cells, producing generic outputs. This collapse itself becomes a detectable signature, as the model behaves anomalously on these inputs.

## V. Experiments and Results

### A. Experimental Setup

**Baseline.** VAE trained on full training set (28,124 cells), 100 epochs. **Retrain.** VAE retrained from scratch on retain set only. **Hyperparameters.** Learning rate $10^{-3}$, batch size 128, Adam optimizer.

Reference values are Baseline MIA AUC = 0.769, Retrain floor AUC = 0.481, and Target band = [0.451, 0.511].

**Baseline Memorization.** The baseline AUC of 0.769 confirms that the VAE memorizes training data to a detectable degree.

This is well above random chance (0.5) but not perfect (1.0), suggesting partial memorization. The model learns general patterns shared across cells while also encoding sample-specific information that an attacker can exploit. Rare cell types like cluster 13 are particularly vulnerable because the model sees fewer similar examples during training.

### *B. Adversarial Methods on Structured Forget Set*

| Method | AUC | Gap | Status |
|---|---|---|---|
| Baseline | 0.769 | +0.288 | LEAK |
| Frozen single $\lambda$=5 | 0.997 | +0.516 | FAIL |
| Frozen multi $\lambda$=10 | 0.992 | +0.511 | FAIL |
| Extra-grad $\lambda$=5 | 0.382 | $-0.099$ | FAIL |
| **Extra-grad $\lambda$=10** | **0.482** | **+0.001** | **SUCCESS** |
| Retrain | 0.481 | 0.000 | TARGET |

TABLE II
Adversarial unlearning on structured forget set. Only extra-gradient $\lambda$=10 succeeds.

All frozen-critic methods produce AUC $> 0.98$, which is worse than the baseline (0.769). This counterintuitive result occurs because frozen critics have fixed decision boundaries that the VAE can learn to circumvent without actually removing membership signal.

Extra-gradient with $\lambda$=10 achieves post-hoc AUC of 0.482, within 0.001 of the retrain floor. Multi-seed validation shows 75% success rate (3 of 4 seeds land in the target band). The variance across seeds reflects the non-convex optimization landscape of the min-max game, where different initializations traverse different trajectories through parameter space.

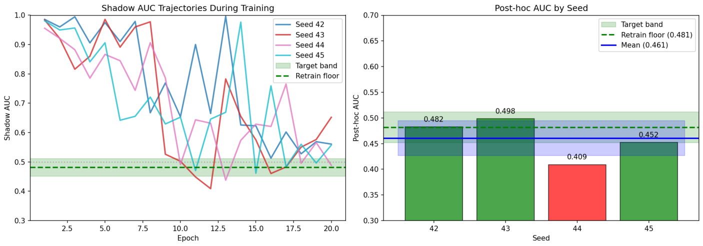

Fig. 1. Shadow AUC trajectories and post-hoc AUC comparison by seed

**The $\lambda$ Tradeoff.** The hyperparameter $\lambda$ controls the privacy-utility balance. With $\lambda$=5, privacy pressure dominates and the model over-unlearns (AUC = 0.38). With $\lambda$=10, the balance shifts toward preserving reconstruction quality while still achieving target privacy. Values below 5 provide insufficient privacy pressure.

### *C. Fisher Unlearning by Forget Set Type*

| Forget Type | Mean AUC | Std | Status |
|---|---|---|---|
| Structured (cluster 13) | 0.814 | 0.003 | FAIL |
| Scattered (random 30) | 0.499 | 0.004 | SUCCESS |

TABLE III
Fisher unlearning results by forget set type.

Fisher achieves AUC 0.499 on scattered forget sets (perfect unlearning) but only 0.814 on structured sets (substantial leakage). The contrast is stark. When forget cells are randomly distributed throughout the training set, they share no common structure that the decoder must preserve. Scrubbing affects each cell independently. When forget cells form a coherent cluster, the decoder has learned cluster-specific mappings that are entangled with general reconstruction capabilities. Scrubbing these parameters disrupts the model's behavior on forgotten cells in ways that an attacker can detect.

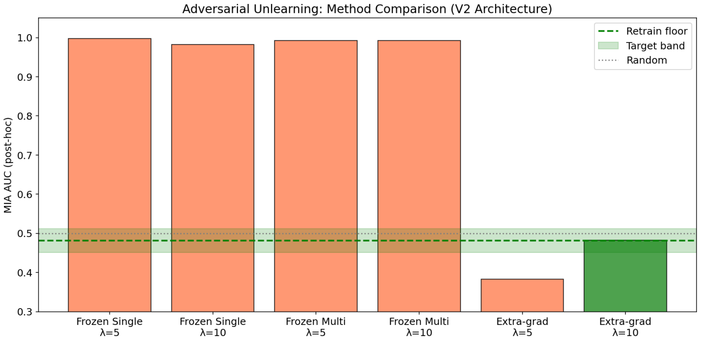

Fig. 2. Adversarial method comparison on structured forget set. Frozen critics (orange) fail completely with AUC near 1.0. Extra-gradient $\lambda$=5 over-unlearns (AUC below target). Only extra-gradient $\lambda$=10 (green) achieves the target band.

### *D. Utility and Computational Cost*

| | Base | Extra-grad | Fisher | Retrain |
|---|---|---|---|---|
| Recon. MSE | 0.572 | 0.572 | 0.572 | 0.572 |
| Marker MSE | 2.54 | 2.54 | 2.55 | 2.54 |
| Class. acc | 0.954 | 0.942 | 0.818 | 0.954 |
| Time | – | 42 min | 2 min | 10 min |

TABLE IV
Utility and computational cost.

Reconstruction quality (MSE) is fully preserved across all methods, indicating that the decoder's ability to reconstruct gene expression is not degraded. Marker gene MSE specifically measures reconstruction of biologically important genes that define cell types.

Classification accuracy measures how well a simple classifier can predict cell type from latent representations. Extra-gradient preserves this structure (0.942 vs 0.954 baseline) while Fisher damages it severely (0.818). This suggests Fisher scrubbing disrupts the latent space geometry, moving cell representations in ways that hurt downstream tasks even though raw reconstruction appears intact.

Extra-gradient takes 4× longer than retraining (42 vs 10 minutes) because it requires joint optimization of the VAE and three attacker networks. Fisher is fast (2 minutes) but only works for scattered forget sets.

## VI. Discussion

### *A. Why Frozen Critics Fail*

Frozen-critic methods achieve AUC $>$ 0.98, which is worse than the baseline (0.769). This counterintuitive result occurs because the VAE learns to exploit critic-specific blind spots. The frozen attacker has a fixed decision boundary, and the VAE can move forget-set representations to regions that fool this specific boundary without actually removing membership signal. A fresh post-hoc attacker, trained on the updated VAE's representations, easily detects these samples because the underlying information was never removed.

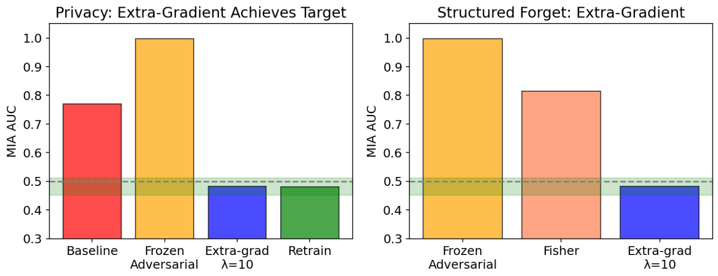

Fig. 3. Summary of main results. Left: Extra-gradient with $\lambda$=10 achieves the target band while frozen adversarial methods fail completely. Right: For structured forget sets, only extra-gradient succeeds; Fisher fails. The green band indicates the target AUC range (±0.03 of retrain floor).

### B. Why Extra-Gradient Succeeds

With $\lambda$=10, the min-max objective is privacy-dominant: the VAE prioritizes fooling the attacker over reconstruction quality. The extra-gradient method dampens the rotational dynamics that cause naive simultaneous gradient descent-ascent to oscillate. This allows optimization to settle into a neighborhood where the VAE has genuinely reduced membership signals, rather than just finding adversarial blind spots. The 75% success rate across seeds suggests multiple local minima exist, with most being privacy-good solutions.

### C. Why Fisher Fails on Structured Sets

VAE decoders learn shared generative mappings across all cell types. When the forget set is a coherent cluster (e.g., all megakaryocytes), the parameters encoding that cluster's characteristics are entangled with parameters needed for other cells. Scrubbing these parameters damages the decoder's ability to reconstruct the forgotten cluster, but this damage itself becomes a detectable signal since the model behaves differently on these cells. This behavior reveals the cells that were targeted for removal. For scattered forget sets with no shared structure, no such entanglement exists.

### D. The Over-Unlearning Problem

Extra-gradient with $\lambda$=5 achieves AUC of 0.38, which is below the target band. This "over-unlearning" makes forget cells look *less* like training data than truly unseen cells. An attacker could exploit this since samples that appear suspiciously "non-member-like" may have been targeted for removal. This connects to the Streisand effect noted by Hayes et al. [5]. The $\lambda$=10 configuration avoids this by balancing privacy pressure against reconstruction fidelity.

### E. Limitations

This study has several limitations. (1) Single dataset (PBMC-33k); results may differ for other datasets or modalities. (2) Small forget set (n=30); larger forget sets may behave differently. (3) Empirical privacy guarantees only, with no formal differential privacy bounds. (4) Extra-gradient is slower than retraining for this model size.

## VII. Conclusion

This study compared adversarial and Fisher-based unlearning for scRNA-seq VAEs. The key findings are that (1) frozen-critic adversarial methods fail completely for VAEs, (2) extra-gradient co-training with $\lambda$=10 achieves retrain-equivalent privacy for structured forget sets, and (3) Fisher unlearning works for scattered forget sets but fails on structured sets due to decoder parameter entanglement.

For small-to-medium models like this VAE, full retraining remains the most practical solution because it is faster (10 minutes vs 42 minutes) and guarantees complete data removal. However, for larger models where retraining is prohibitive, extra-gradient co-training offers a viable alternative.

### A. Future Work

Several directions could extend this work. First, relative Fisher scrubbing could weight parameter importance by the ratio of forget-set to retain-set influence, preserving parameters that matter more for retained cells. Second, latent space projection could directly remove the "cluster direction" in latent space rather than modifying decoder parameters. Third, integrating formal differential privacy bounds would provide theoretical guarantees beyond empirical MIA evaluation. Finally, validation on larger models and datasets (e.g., multi-million cell atlases) would clarify when unlearning becomes preferable to retraining.

## Acknowledgments

Claude Code (Anthropic) was used for debugging assistance, PEP8 standardization, linting, and comment cleanup. Code and data are available at https://github.com/db-d2/Machine_Unlearning.

## References

[1] L. Bourtoule et al., "Machine unlearning," in *IEEE S&P*, 2021.
[2] Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in *IEEE S&P*, 2015.
[3] T. Chavdarova et al., "Reducing noise in GAN training with variance reduced extragradient," in *NeurIPS*, 2019.
[4] A. Golatkar et al., "Eternal sunshine of the spotless net," in *CVPR*, 2020.
[5] J. Hayes et al., "Inexact unlearning needs more careful evaluations to avoid a false sense of privacy," in *ICLR*, 2024.
[6] R. Lopez et al., "Deep generative modeling for single-cell transcriptomics," *Nature Methods*, 2018.
[7] M. Nasr et al., "Machine learning with membership privacy using adversarial regularization," in *ACM CCS*, 2018.
[8] R. Shokri et al., "Membership inference attacks against machine learning models," in *IEEE S&P*, 2017.
[9] F.A. Wolf et al., "Scanpy: large-scale single-cell gene expression data analysis," *Genome Biology*, 2018.