MaskClass: A Masked Denoising Autoencoder for Scalable and Accurate Single-Cell RNA-seq Analysis

Anonymous Authors[1]

Abstract

The emergence of numerous single-cell ribonucleic acid (RNA) sequencing (scRNA-seq) technologies and protocols over the past decade has enabled the study of RNA expression at single-cell resolution with improved accuracy and scalability, now handling gene expression data from hundreds of thousands to millions of cells. However, scRNA-seq data present analytical challenges, including data sparsity, technical noise, and batch effects, which complicate downstream analyses, such as cell clustering and differential expression analysis. To address these challenges, we introduce MaskClass, a masked denoising autoencoder that significantly improves data quality by selectively hiding parts of gene expression data, learning to reconstruct them while minimizing the impact of noisy dropout values, and using a separate process to recover biological zero-expression measurements. While we focus our primary evaluation on scRNA-seq data, the proposed approach is applicable to other high-dimensional, sparse, and noisy data modalities, including other single-cell omics proteomics, spatial transcriptomics, and related domains characterized by structured missing data and measurement noise. Benchmarking against state-of-the-art denoising methods demonstrates MaskClass's effectiveness in noise reduction on simulated data and significant improvement in cell identity clustering across diverse real-world datasets. MaskClass is available as an open source implementation at `[link omitted for double-blind review]`.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Single-cell RNA sequencing (scRNA-seq) technologies and protocols enable the measurement of RNA expression across individual cells, allowing the study of biological systems at a level of detail and scale previously unattainable, providing more insight into cellular function and response across a variety of contexts, such as disease vs. non-disease states or therapeutic drug response. As this technology continues to grow in scale, the issue of significant technical variation and noise, primarily resulting in sparse count matrices with numerous zeros, is compounded. This issue arises from the low capture efficiency and shallow sequencing depth of current droplet protocols (Zheng et al., 2017), leading to many expressed genes recorded as zero – a phenomenon known as "dropout" event. Notably, not all zeros in the data indicate missing values. True biological zeros coexist with false zeros. This phenomenon complicates downstream analyses such as trajectory inference, differential expression (DE) analyses, and clustering. Using naive imputation methods to impute dropout events can distort genuine differences between cell types or states (Svensson, 2020; Jiang et al., 2022). Consequently, effective methods for denoising and imputation must balance three objectives: (i) recovering likely missing counts, (ii) preserving true biological variability, and (iii) computational scaling to accommodate datasets comprising hundreds of thousands or millions of cells per sample. In this study, we introduce MaskClass, a masked denoising autoencoder that randomly conceals subsets of gene–cell entries during training. This model learns to reconstruct hidden entries while intentionally disregarding noisy data to recover likely missing values without. We benchmark MaskClass against state-of-the-art denoising and imputation methods, including MAGIC (Markov Affinity-based Graph Imputation of Cells) (van Dijk et al., 2018), Single-cell Analysis Via Expression Recovery (SAVER) (Huang et al., 2018), scImpute (Li & Li, 2018), Deep Count Autoencoder (DCA) (Eraslan et al., 2019), ccImpute (Malec et al., 2022), and AutoClass (Li et al., 2022). These methods represent a range of complementary modeling approaches and are influential in advancing the field. However, we acknowledge that many additional techniques exist beyond the scope of this evaluation. MAGIC constructs a k-nearest-neighbor affinity graph in a reduced space (principal component anal-

ysis, PCA), normalizes it into a Markov transition matrix (M), and then raises M to a power t to diffuse information across neighborhoods; the final imputed matrix is $M^t X$, where X is the original expression matrix. In benchmark datasets, MAGIC recovered gene–gene relationships and phenotypic continua that were obscured in the raw data and increased concordance with independent protein measurements (Eraslan et al., 2019). However, diffusion necessarily modifies all entries, including values not affected by dropout, which can over-smooth if t is not well-tuned (Li & Li, 2018; Malec et al., 2022). scImpute models the probability that a zero (or low count) is a dropout versus a true zero for each gene–cell pair by fitting a mixture model; it then imputes only entries with high dropout probability by borrowing information from similar cells identified using genes unlikely to be affected by dropout. This design explicitly aims to avoid altering unaffected measurements while improving the DE and clustering performance on simulated and real datasets. Related critiques note that methods that modify all values (e.g., global smoothing) may introduce bias or erase meaningful heterogeneity if they fail to distinguish between true and false zeroes (Malec et al., 2022; Jiang et al., 2022). Gene-aware Bayesian recovery, in tandem with cell-smoothing strategies, utilizes cross-gene information and uncertainty quantification. SAVER suggests that unique molecular identifier (UMI) counts for each gene in each cell follow a Poisson–Gamma (negative binomial) model, with prior parameters estimated via Poisson Lasso regression on other genes. This approach generates a posterior distribution (and mean) for the latent true expression, providing both point estimates and credible uncertainty intervals (UIs). SAVER has shown proficiency in recovering distributional properties (e.g., Gini coefficients) consistent with RNA FISH, restoring attenuated gene–gene correlations, and avoiding spurious correlations seen with certain graph-smoothing methods, while explicitly quantifying uncertainty for downstream analysis. The rise of deep learning has enabled the development of count-aware autoencoders that merge nonlinear representation learning with likelihood-based reconstructions. DCA replaces the mean squared error with a negative binomial (NB) or zero-inflated NB (ZINB) likelihood, learning gene-specific means and dispersions (and dropout probabilities) while compressing data through a bottleneck layer. This count-modeling loss is crucial because, unlike a standard autoencoder, DCA effectively recovers the cell type structure in simulations with significant dropout and scales linearly to millions of cells, making it attractive for modern atlas-scale datasets. DCA also highlights the distinction between true and false zeros and offers diagnostics to choose between the NB and ZINB models depending on the technology; notably, UMI-based droplet data often do not require zero-inflated models (Eraslan et al., 2019; Svensson, 2020).

Building on autoencoders, AutoClass enhances the bottleneck with a classifier branch trained on pseudo-labels derived from pre-clustering, fostering latent representations that maintain biologically relevant structures, whereas the autoencoder reduces the noise. Unlike distribution-specific methods, AutoClass is explicitly distribution-agnostic and addresses a wide range of noise sources (not just dropout), including amplification bias, library size variation, batch effects, and other non-signal variations. Across simulated and real datasets, the two-branch architecture improved data recovery, clustering, differential expression (DE), and batch correction, demonstrating robustness to the key hyperparameters. Finally, ccImpute represents a consensus clustering-based strategy designed to correct dropout data. Inspired by Single-Cell Consensus Clustering (SC3) (Kiselev et al., 2017), ccImpute computes a cell–cell consensus similarity matrix and imputes likely dropouts by averaging the expression across reliably comparable cells. The authors stressed that imputation methods should minimize the introduction of new noise and avoid altering non-dropout values. They reported enhanced downstream clustering performance compared to several baselines on datasets with known cell identity. This study highlights the limitations of manifold diffusion methods, such as altering unaffected genes and failing to preserve true zeros, and advocates for consensus-based similarity as a stable foundation for imputation.

Collectively, these studies establish core principles for scRNA-seq recovery: leveraging shared structures among cells and genes, respecting the count nature of the data, avoiding erasing genuine heterogeneity, and, when feasible, quantifying uncertainty for downstream tasks (Zheng et al., 2017; Huang et al., 2018; Jiang et al., 2022). Building on this foundation, the present study introduces MaskClass, a masked, self-supervised denoising autoencoder that selectively hides parts of gene expression data and learns to reconstruct them. MaskClass aims to retain these advantages while further mitigating over-smoothing and zero-inflation artifacts, challenges that become increasingly significant as datasets expand in size and complexity.

2. Methods

2.1. Experimental scRNA-seq Datasets

We benchmarked the models using five publicly available scRNA-seq datasets, referenced by the first author of their originating publication: Blakeley (Blakeley et al., 2015), Pollen (Pollen et al., 2014), Darmanis (Darmanis et al., 2015), Usoskin (Usoskin et al., 2015), and Li (Li et al., 2017). These datasets encompass a diverse range of biological origins and technical characteristics, as summarized in Table 1.

The confidence in the ground-truth labels varies across these datasets. The Blakeley, Li, and Pollen datasets provide high-fidelity labels derived from controlled experimental conditions and distinct cell lines. In contrast, Usoskin and Darmanis labels were assigned computationally via clustering and subsequent expert annotation. To ensure that the analysis focused on distinct cellular populations, cells annotated as "hybrids" in the Darmanis dataset were excluded.

We selected these datasets because they exhibit substantial sparsity and technical noise, making denoising and imputation challenging. In our experiments, we selected the top 1000 highly variable genes (HVGs) using `scran` (Lun et al., 2016). Raw counts were library-size normalized by scaling each cell to the median library size and applying a $\log_2(1+\cdot)$ transform, yielding $\mathbf{X}_{\log}$. We retained genes expressed in at least three cells.

2.2. Synthetic Data Generation

To assess the imputation accuracy against a known ground truth, we generated four synthetic datasets using the *splatter* R package (v1.32.0) (Zappia et al., 2017). Each simulation generated $M = 1{,}100$ genes, and the probability of differential expression (likelihood of gene expression differing between cell groups) was set to 0.1 (`de.prob = 0.1`).

To simulate technical noise, dropout midpoints and shapes were randomized per group or batch. We sampled `dropout.mid` from [1,5] and `dropout.shape` from [-1.5,-0.5]. The datasets were designed with varying population structures:

- **Sim-Equal:** Two cell types, equal proportions (50%, 50%); dropout simulated per group.
- **Sim-Unequal:** Three cell types, unequal proportions (60%, 30%, 10%); dropout simulated per group.
- **Sim-Rare:** Four cell types with a rare subpopulation (50%, 25%, 20%, 5%); dropout simulated per group.
- **Sim-Batch:** Three cell types (40%, 30%, 30%) across three batches; dropout simulated per batch (`dropout.type="batch"`). TrueCounts were batch-effect-free (`batch.rmEffect=TRUE`); observed counts included batch effects and dropout.

For each scenario we simulated multiple cell counts $N \in \{1\text{k}, 5\text{k}, 10\text{k}, 15\text{k}, 20\text{k}, 25\text{k}, 50\text{k}, 100\text{k}\}$ ($\text{k} = 10^3$). Each simulation provides a noisy observed count matrix `counts` and a ground-truth matrix `TrueCounts`. Genes expressed in fewer than three cells were removed. We library-size normalized both matrices by scaling each cell to the median library size and applying $\log_2(1+\cdot)$, yielding `logcounts` ($\mathbf{X}_{\log}$) and `logTrueCounts` ($\mathbf{X}^{\star}_{\log}$).

Notation. Throughout, $\mathbf{X}_{\log} \in \mathbb{R}^{N\times G}_{\geq 0}$ denotes the observed log-normalized matrix, $\mathbf{X}^{\star}_{\log}$ the corresponding synthetic ground truth (when available), and $\mathbf{X}'_{\log}$ the reconstructed matrix produced by MaskClass. We write x_{ij}, $x^{\star}_{ij}$, and x'_{ij} for their respective entries. We denote the (unnormalized) count matrix by $\mathbf{C} = \{c_{ij}\}$.

2.3. Model Architecture

MaskClass is a symmetric fully connected autoencoder with encoder $f_\theta : \mathbb{R}^G \to \mathbb{R}^d$ and decoder $g_\phi : \mathbb{R}^d \to \mathbb{R}^G$,

$$\mathbf{h}_i = f_\theta(\tilde{\mathbf{s}}_i), \qquad \hat{\mathbf{s}}_i = g_\phi(\mathbf{h}_i),$$

where $\tilde{\mathbf{s}}_i$ is the corrupted, scaled input (Sections 2.4–2.5). The network uses one hidden layer of width 64 and a bottleneck d=32 with a symmetric decoder; each block applies a linear transformation, followed by layer normalization and a SiLU nonlinearity. Let $\mathbf{S} = \mathrm{scale}(\mathbf{X}_{\log})$ and $\hat{\mathbf{S}}$ denote the decoder outputs stacked across cells; reconstructions in log space are obtained by inverse scaling, $\mathbf{X}'_{\log} = \mathrm{invscale}(\hat{\mathbf{S}})$.

2.4. Gene Scaling

We apply a robust, monotone per-gene scaling with three steps:

$$\begin{aligned}
\text{(Winsorize)} \quad & \bar{x}_{ij} = \min(\max(x_{ij}, q_j^{(2)}), q_j^{(99.5)}),\\
\text{(Standardize)} \quad & z_{ij} = \frac{\bar{x}_{ij} - \mu_j}{\sigma_j}, \qquad \sigma_j \leftarrow \max(\sigma_j, \varepsilon),\\
\text{(Rescale)} \quad & s_{ij} = 2 \cdot \frac{z_{ij} - z_j^{\min}}{z_j^{\max} - z_j^{\min}} - 1,\\
& z_j^{\max} - z_j^{\min} \geq \varepsilon.
\end{aligned}$$

We denote the scaled matrix by $\mathbf{S}$ and use $\mathrm{scale}(\cdot)$ / $\mathrm{invscale}(\cdot)$ for forward/inverse transforms. Here $\varepsilon > 0$ is an arbitrary small constant used for numerical stability (we use $\varepsilon = 10^{-8}$). We define $z_j^{\min} = \min_i z_{ij}$ and $z_j^{\max} = \max_i z_{ij}$.

2.5. Masked Denoising Objective

For each mini-batch $\mathcal{B}$ of cells, we sample a binary mask $\mathbf{M} \in \{0,1\}^{|\mathcal{B}|\times G}$ with different rates for zeros vs. non-zeros in logcounts:

$$\begin{aligned}
m_{ij} &\sim \mathrm{Bernoulli}(\pi_{ij}),\\
\pi_{ij} &= \begin{cases} p_{\text{zero}} p_{\text{bio}}(i,j) & x_{ij} = 0,\\ p_{\text{nonzero}} & x_{ij} > 0, \end{cases}
\end{aligned}$$

We use $(p_{\text{zero}}, p_{\text{nonzero}}) = (0.01, 0.30)$ and fill masked non-zeros with zero in the scaled space.

Noise for masked zeros. For masked zero entries we inject noise (in log space) using the observed count range per gene. Let c_{ij} denote the (raw) count and $c_j^{\max} = \max_i c_{ij}$.

$$\begin{aligned}
\eta_{ij} &\sim \mathrm{Uniform}(0, \nu), \quad \nu = 0.2,\\
\tilde{x}_{ij} &= \log_2(1 + \eta_{ij} c_j^{\max}),
\end{aligned}$$

Table 1. Summary of the experimental scRNA-seq datasets used for benchmarking. N denotes the number of samples (cells), M the initial number of features (genes) prior to preprocessing, and K the number of distinct cell clusters reported in the original study. Datasets are available at: `[link omitted for double-blind review]`

Dataset	N (Cells)	M (Genes)	K	Origin	Reference
Blakeley	30	16,862	3	Human blastocyst	(Blakeley et al., 2015)
Li	561	55,186	7	Human cell lines	(Li et al., 2017)
Pollen	301	23,730	11	Human cell lines	(Pollen et al., 2014)
Darmanis	420	21,516	8	Human cortex	(Darmanis et al., 2015)
Usoskin	622	19,532	4	Mouse lumbar	(Usoskin et al., 2015)

then apply $\mathrm{scale}(\cdot)$.

Loss. Let $r_{ij} = \hat{s}_{ij} - s_{ij}$ denote the reconstruction residual in the scaled space. For a mini-batch, let $\mathcal{Z} = \{(i,j) : x_{ij} = 0\}$ and $\mathcal{N} = \{(i,j) : x_{ij} > 0\}$ denote observed zeros and non-zeros, respectively. The weighted masked reconstruction loss is

$$\mathcal{L}_{\text{mask}} = \frac{w_{\text{bio}} \sum_{(i,j)\in\mathcal{Z}} m_{ij}\, r_{ij}^2 + w_{\text{nz}} \sum_{(i,j)\in\mathcal{N}} m_{ij}\, r_{ij}^2}{w_{\text{bio}} \sum_{(i,j)\in\mathcal{Z}} m_{ij} + w_{\text{nz}} \sum_{(i,j)\in\mathcal{N}} m_{ij}},$$

with weights $(w_{\text{bio}}, w_{\text{nz}}) = (2.0, 1.0)$. The biozero regularizer encourages predicted values at likely biological zeros to be close to the gene-wise scaled zero value:

$$\mathcal{L}_{\text{bio}} = \frac{\sum_{(i,j)\in\mathcal{Z}} p_{\text{bio}}(i,j)\,(\hat{s}_{ij} - s_{0j})^2}{\sum_{(i,j)\in\mathcal{Z}} p_{\text{bio}}(i,j)},$$
$$s_{0j} = [\mathrm{scale}(\mathbf{0})]_j.$$

We minimize the combined objective $\mathcal{L} = \mathcal{L}_{\text{mask}} + \gamma\, \mathcal{L}_{\text{bio}}$.

MaskClass$_{\text{best}}$ vs. MaskClass$_{\text{balanced}}$. MaskClass$_{\text{best}}$ sets $\gamma = 0$, while MaskClass$_{\text{balanced}}$ sets $\gamma = 1$; all other parameters are identical.

2.6. Biozero Probability Estimation

We estimate p_{bio} from observed counts c_{ij}. The counts follow a zero-inflated negative binomial (NB) model with a latent count t_{ij} and dropout indicator d_{ij}:

$$t_{ij} \sim \mathrm{NB}(\mu_{ij}, \phi_j),$$
$$d_{ij} \sim \mathrm{Bernoulli}(\delta_{ij}),$$
$$c_{ij} = (1 - d_{ij})\, t_{ij}.$$

Here $\mu_{ij} = \ell_i \mu_j$ with ℓ_i a cell-wise library-size factor and ϕ_j a gene-wise dispersion estimated by a (shrinkage-regularized) method-of-moments estimator on non-zero counts. Writing $f_{ij}(0) = \Pr(t_{ij} = 0 \mid \mu_{ij}, \phi_j)$, we have $\Pr(c_{ij} = 0) = \delta_{ij} + (1 - \delta_{ij}) f_{ij}(0)$ and for an observed zero,

$$p_{\text{bio}}(i,j) = \Pr(t_{ij} = 0 \mid c_{ij} = 0) = \frac{(1 - \delta_{ij}) f_{ij}(0)}{\delta_{ij} + (1 - \delta_{ij}) f_{ij}(0)}.$$

We set $p_{\text{bio}}(i,j) = 0$ when $c_{ij} > 0$. We compute $\ell_i = \big(\sum_j c_{ij}\big) / \mathrm{median}_k\big(\sum_j c_{kj}\big)$ and model δ_{ij} as a smooth decreasing function of $\log \mu_{ij}$ (fit to the observed zero rates). Moment-based dispersion estimates can be numerically unstable in highly sparse, zero-inflated regimes; we mitigate this with shrinkage and lower bounds, and note that more robust dispersion fits (e.g., `glmGamPoi`) are a possible extension.

Cell sparsity adjustment. We compute each cell's observed zero fraction

$$z_i = \frac{1}{G} \left|\{j \in \{1, \ldots, G\} : c_{ij} = 0\}\right|,$$

rescale z_i to $[0,1]$ using the 5th–95th percentile range, and shrink

$$p_{\text{bio}} \leftarrow p_{\text{bio}}(1 - \alpha z_i), \qquad \alpha = 0.3.$$

2.7. Training Details

We optimize with Adam (learning rate 10^{-4}, weight decay 0), batch size 32, for 300 epochs on GPU; random seeds are fixed for reproducibility. We employ a transductive setting: the model is trained on the full observed dataset to reconstruct masked entries, with no held-out set for the reconstruction task.

2.8. Evaluation Metrics

(i) Denoising accuracy (mean squared error; MSE). On synthetic data ($N = 5000$ cells), we compare the reconstruction $\mathbf{X}'_{\log}$ to $\mathbf{X}^{\star}_{\log}$ and report

$$\mathrm{MSE} = \frac{1}{NG} \sum_{i,j} \left(x^{\star}_{ij} - x'_{ij}\right)^2.$$

We additionally report biozero-MSE over entries with $x^{\star}_{ij} = 0$ and dropout-MSE over entries with $x_{ij} = 0$ and $x^{\star}_{ij} > 0$. Each experiment was repeated multiple times with different random seeds (10 repeats for baselines; 5 repeats for MaskClass due to GPU budget); we report mean and standard deviation.

(ii) Runtime scaling. To evaluate scalability, we ran each method on each synthetic cell count $N \in \{1\text{k}, 5\text{k}, 10\text{k}, 15\text{k}, 20\text{k}, 25\text{k}, 50\text{k}, 100\text{k}\}$ and report the average wall-clock runtime per dataset.

(iii) Downstream clustering. We apply principal component analysis (PCA) to $\mathbf{X}'_{\log}$ and keep $D = \max\{2, \min(50, N, G)\}$ components, then run k-means with k equal to the number of ground-truth labels and report Adjusted Rand Index (ARI), normalized mutual information (NMI), purity, and adjusted silhouette width (ASW). Let n_{ij} be the number of cells with true label i assigned to cluster j, $a_i = \sum_j n_{ij}$, and $b_j = \sum_i n_{ij}$. Define $T = \sum_{ij} \binom{n_{ij}}{2}$, $A = \sum_i \binom{a_i}{2}$, and $B = \sum_j \binom{b_j}{2}$. Then

$$\text{ARI} = \frac{T - \frac{AB}{\binom{N}{2}}}{\frac{1}{2}(A+B) - \frac{AB}{\binom{N}{2}}},$$

$$\text{NMI}(U, V) = \frac{2I(U;V)}{H(U) + H(V)},$$

$$\text{Purity}(U, V) = \frac{1}{N} \sum_k \max_j |c_k \cap t_j|,$$

$$\text{ASW} = \frac{1}{N} \sum_i s(i).$$

Here $s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$, where $a(i)$ is the mean distance from cell i to cells in its assigned cluster and $b(i)$ is the mean distance to the nearest other cluster.

2.9. Algorithm Outline

Figure 1 provides a high-level view of the MaskClass pipeline, from biozero probability estimation to masked corruption and reconstruction. Algorithm 1 (Algorithm 1) gives the complete training and imputation procedure, including mask sampling, corruption, and the loss minimized by $\text{MaskClass}_{\text{best}}$ and $\text{MaskClass}_{\text{balanced}}$.

Algorithm 1 MaskClass training and imputation.

Input: logcounts $\mathbf{X}_{\log}$, counts $\mathbf{C}$, scaler $\text{scale}/\text{invscale}$, and hyperparameters $p_{\text{zero}}, p_{\text{nonzero}}, \nu, \alpha, \gamma$, and epochs E.
Compute biozero probabilities $p_{\text{bio}}(i, j)$ from counts via the model in Section 2.6.
Compute $z_i = \frac{1}{G} |\{j : c_{ij} = 0\}|$ and set $p_{\text{bio}} \leftarrow p_{\text{bio}}(1 - \alpha z_i)$.
Scale $\mathbf{S} \leftarrow \text{scale}(\mathbf{X}_{\log})$.
for epoch $= 1$ to E **do**
  for mini-batch $\mathcal{B}$ **do**
    Sample $m_{ij} \sim \text{Bernoulli}(\pi_{ij})$ where $\pi_{ij} = p_{\text{zero}} p_{\text{bio}}(i, j)$ if $x_{ij} = 0$ and $\pi_{ij} = p_{\text{nonzero}}$ if $x_{ij} > 0$.
    Corrupt inputs: masked non-zeros $\tilde{s}_{ij} \leftarrow 0$; masked zeros $\tilde{x}_{ij} = \log_2(1 + \eta_{ij} c_j^{\max})$ with $\eta_{ij} \sim \text{Uniform}(0, \nu)$ and $c_j^{\max} = \max_i c_{ij}$, then $\tilde{s}_{ij} \leftarrow \text{scale}(\tilde{x}_{ij})$.
    Forward pass: $\hat{\mathbf{s}}_i = g_\phi(f_\theta(\tilde{\mathbf{s}}_i))$.
    Minimize $\mathcal{L} = \mathcal{L}_{\text{mask}} + \gamma \mathcal{L}_{\text{bio}}$ ($\text{MaskClass}_{\text{best}}$: $\gamma = 0$; $\text{MaskClass}_{\text{balanced}}$: $\gamma = 1$).
  end for
end for
Reconstruct $\mathbf{X}'_{\log} \leftarrow \text{invscale}(\hat{\mathbf{S}})$ and keep observed non-zeros fixed.
Output: reconstructed logcounts $\mathbf{X}'_{\log}$.

2.10. Computational Considerations

The masked loss computes MSE only over masked entries, preventing trivial identity mapping. Robust scaling constrains dynamic range and stabilizes optimization, while inverse scaling restores original log-count units for evaluation.

3. Experiments and Results

3.1. Denoising Accuracy on Synthetic Data ($N = 5{,}000$)

We evaluate denoising accuracy on the four Splatter simulations described in Section 2.2 with $N = 5{,}000$ cells, comparing reconstructed logcounts $\mathbf{X}'_{\log}$ against the synthetic ground truth $\mathbf{X}^{\star}_{\log}$ (Section 2.8). Unless stated otherwise, **MaskClass** refers to the balanced configuration ($\text{MaskClass}_{\text{balanced}}$; γ=1). Table 2 summarizes mean squared error (MSE), with parenthesized values reporting dropout-MSE over entries that are observed as zero but non-zero in the ground truth. MaskClass achieves the lowest MSE across all scenarios, outperforming the strongest baseline (DCA) by 43–60% in overall MSE and 20–49% in dropout-MSE. Figure 2 visualizes all three error components (overall, dropout, and biozero). While MaskClass yields the best overall and dropout-MSE, more conservative methods achieve lower biozero-MSE (e.g., Baseline trivially attains zero biozero-MSE by leaving zeros unchanged), highlighting a trade-off between recovering dropouts and preserving true zeros.

Software and Data

If a paper is accepted, we strongly encourage the publication of software and data with the camera-ready version of the paper whenever appropriate. This can be done by including a URL in the camera-ready copy. However, **do not** include URLs that reveal your institution or identity in your submission for review. Instead, provide an anonymous URL or upload the material as "Supplementary Material" into the OpenReview reviewing system. Note that reviewers are not required to look at this material when writing their review.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

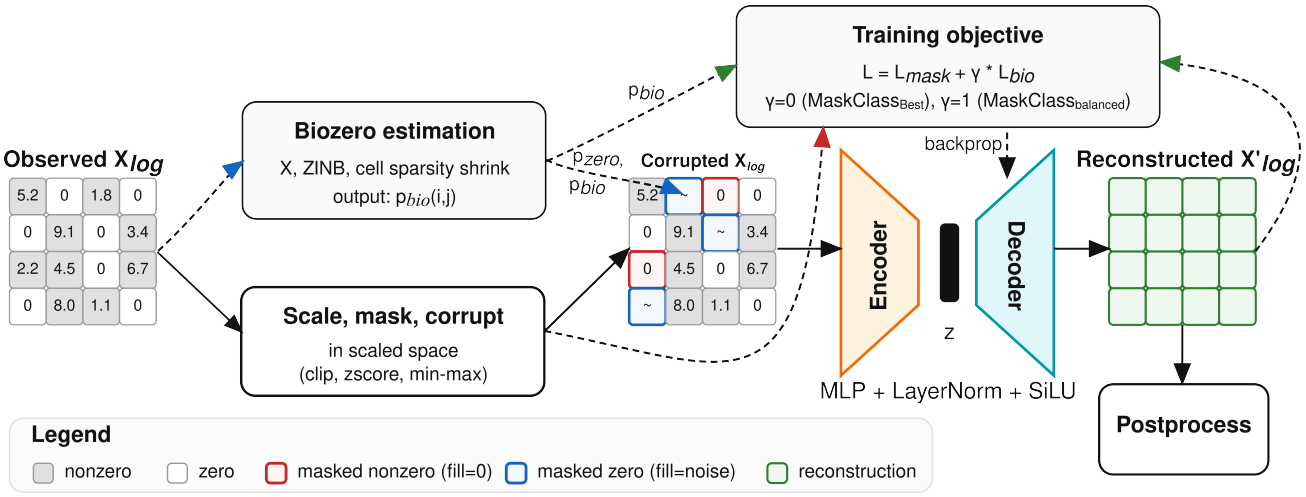

Figure 1. **MaskClass overview.** Starting from logcounts $\mathbf{X}_{\log}$ and counts $\mathbf{C}$, we estimate a biozero probability $p_{\text{bio}}(i, j)$ for observed zeros. During self-supervised training, non-zeros are masked at a fixed rate, while zeros are masked with probability proportional to p_{bio}; masked zeros are corrupted by injecting noise and masked non-zeros are replaced by zero in the scaled space. A symmetric autoencoder reconstructs the scaled matrix, and parameters are learned by minimizing a weighted masked reconstruction loss plus an optional biozero regularizer (enabled in $\text{MaskClass}_{\text{balanced}}$).

Table 2. Denoising accuracy (MSE) on synthetic data with $N = 5{,}000$ cells. Values are mean (and $\pm$ std when non-zero) across repeats; parenthesized values report dropout-MSE. Lower is better.

Method	Sim-Equal	Sim-Unequal	Sim-Rare	Sim-Batch
Baseline	6.955 (14.009)	6.799 (14.128)	7.596 (15.234)	6.857 (14.399)
MAGIC	4.372 ± 0.002 (6.198 ± 0.002)	4.083 ± 0.002 (5.584 ± 0.002)	4.454 ± 0.002 (5.965 ± 0.002)	3.932 ± 0.001 (5.263 ± 0.001)
AutoClass	4.545 ± 0.026 (6.445 ± 0.020)	4.253 ± 0.016 (5.850 ± 0.020)	4.722 ± 0.025 (6.251 ± 0.033)	4.124 ± 0.020 (5.605 ± 0.026)
ccImpute	4.759 (9.582)	4.755 ± 0.007 (9.869 ± 0.014)	4.296 ± 0.024 (8.614 ± 0.047)	4.174 ± 0.005 (8.763 ± 0.010)
DCA	1.755 ± 0.016 (2.715 ± 0.021)	1.498 ± 0.010 (2.212 ± 0.015)	1.712 ± 0.010 (2.602 ± 0.019)	1.376 ± 0.006 (2.005 ± 0.008)
MaskClass	$\mathbf{0.728 \pm 0.017}$ ($\mathbf{1.417 \pm 0.036}$)	$\mathbf{0.630 \pm 0.012}$ ($\mathbf{1.254 \pm 0.026}$)	$\mathbf{0.688 \pm 0.005}$ ($\mathbf{1.323 \pm 0.009}$)	$\mathbf{0.788 \pm 0.013}$ ($\mathbf{1.600 \pm 0.029}$)

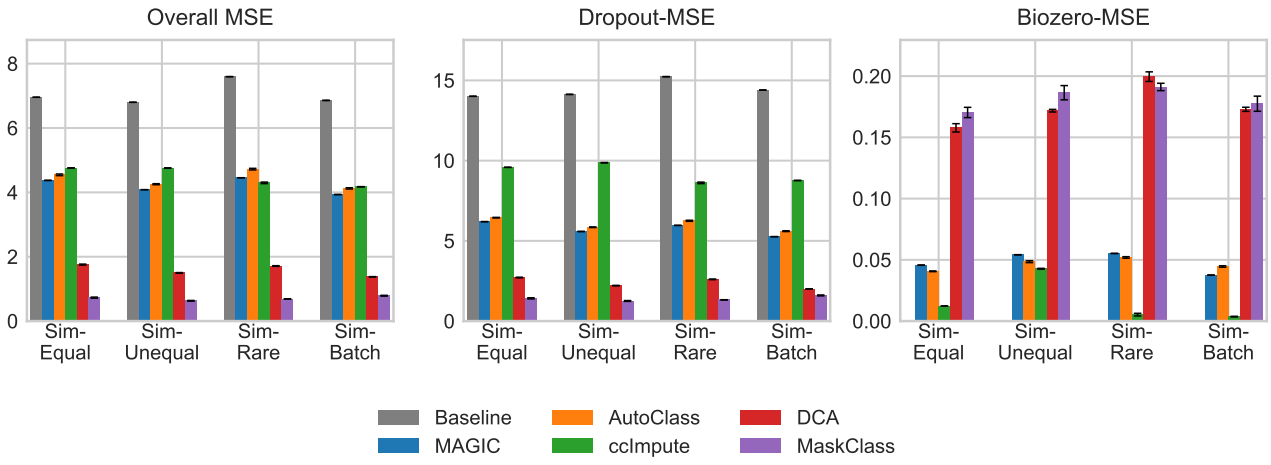

Figure 2. Bar plots (mean $\pm$ std) of denoising error components on synthetic data with $N = 5{,}000$ cells. We report overall MSE, dropout-MSE over entries with $x_{ij} = 0$ and $x^\star_{ij} > 0$, and biozero-MSE over entries with $x^\star_{ij} = 0$. Lower is better.

References

Blakeley, P., Fogarty, N. M. E., del Valle, I., Wamaitha, S. E., Hu, T. X., Elder, K., Snell, P., Christie, L., Robson, P., and Niakan, K. K. Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development*, 142(18):3151–3165, 2015.

Darmanis, S., Sloan, S. A., Zhang, Y., Duh, M., Caneda, C., Shuer, L. M., Hayden Gephart, M. G., Barres, B. A., and Quake, S. R. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, 112(23):7285–7290, 2015.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. Single-cell rna-seq denoising using a deep count autoencoder. *Nature Communications*, 10: 390, 2019. doi: 10.1038/s41467-018-07931-2.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M., and Zhang, N. R. Saver: gene expression recovery for single-cell rna sequencing. *Nature Methods*, 15(7):539–542, 2018. doi: 10.1038/s41592-018-0033-z.

Jiang, R., Sun, T., Song, D., and Li, J. J. Statistics or biology: the zero-inflation controversy about scrna-seq data. *Genome Biology*, 23(1):31, 2022. doi: 10.1186/s13059-022-02601-5.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., and Hemberg, M. Sc3: consensus clustering of single-cell rna-seq data. *Nature Methods*, 14(5):483–486, 2017. doi: 10.1038/nmeth.4236.

Li, H., Courtois, E. T., Sengupta, D., Tan, Y., Chen, K. H., Goh, J. J. L., Kong, S. L., Chua, C., Hon, L. K., Tan, W. S., et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature genetics*, 49(5):708–718, 2017.

Li, H., Brouwer, C. R., and Luo, W. A universal deep neural network for in-depth cleaning of single-cell rna-seq data. *Nature Communications*, 13:1901, 2022. doi: 10.1038/s41467-022-29576-y.

Li, W. V. and Li, J. J. An accurate and robust imputation method *scImpute* for single-cell rna-seq data. *Nature Communications*, 9:997, 2018. doi: 10.1038/s41467-018-03405-7.

Lun, A. T., McCarthy, D. J., and Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. *F1000Research*, 5:2122, 2016.

Malec, M., Kurban, H., and Dalkilic, M. ccimpute: an accurate and scalable consensus clustering based algorithm to impute dropout events in the single-cell rna-seq data. *BMC Bioinformatics*, 23:291, 2022. doi: 10.1186/s12859-022-04814-8.

Pollen, A. A., Nowakowski, T. J., Chen, J., Retallack, H., Sandoval-Espinosa, C., Nicholas, C. R., Shuga, J., Liu, S. J., Oldham, M. C., Diaz, A. A., et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnology*, 32(10):1053–1058, 2014.

Svensson, V. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, 2020. doi: 10.1038/s41587-019-0379-5.

Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lönnerberg, P., Lou, D., Hjerling-Leffler, J., Haeggström, J., Kharchenko, O., Kharchenko, P. V., et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nature Neuroscience*, 18(1):145–153, 2015.

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., Bierie, B., Mazutis, L., Wolf, G., Krishnaswamy, S., and Peér, D. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3): 716–729.e27, 2018. doi: 10.1016/j.cell.2018.05.061.

Zappia, L., Phipson, B., and Oshlack, A. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., and others. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017. doi: 10.1038/ncomms14049.