ISMB/Bioinformatics, 2026, pp. 1–9
DOI added during production
Advance Access Publication Date: Day Month Year
Proceedings Paper

Topology Matters: The Trade-off Between Wasserstein Critics and Discriminators in Single-Cell Data Integration

Kendall Reid [1,*] Genevieve Stein-O'Brien [1] and Erhan Guven[2]

[1]Neuroscience, Johns Hopkins University, Baltimore, 21218, MD, USA
[2]Johns Hopkins University Applied Physics Laboratory, Laurel, 20707, MD, USA

[*]Corresponding author. kreid20@jh.edu

Abstract

Motivation: High-throughput gene expression measurements are biased by technical and biological confounding variables, which obscure the true biological signals of interest. Adversarial autoencoders are a popular solution that corrects for these confounding effects, often relying on discriminator networks that approximate the Jensen-Shannon divergence. However, previous research has established that the Jensen-Shannon divergence suffers from vanishing gradients when distributions do not overlap, leading to poor results. We identify that this is a critical vulnerability in gene expression deconfounding, as samples from distinct experimental batches often occupy disjoint regions of the high-dimensional space. In contrast, the Wasserstein distance remains a valid metric with informative gradients even for disjoint distributions. While both approaches appear in the literature, no study has rigorously isolated the adversarial objective to systematically evaluate its impact on batch alignment, biological conservation, and scalability across varying dataset complexities.

Results: We introduce a multi-class reference-based Wasserstein critic to systematically benchmark adversarial objectives. In fundamental binary integration tasks, we confirm that the Wasserstein critic yields superior mixing in scenarios with disjoint support. However, extensive reference sensitivity analysis reveals that this performance relies on a topologically dense reference batch. Consequently, we define the topological boundaries for adversarial alignment: the Wasserstein critic is superior for reference-mapping tasks involving dense anchors and disjoint query batches, whereas the standard discriminator approximates a global centroid more effectively in fragmented scenarios.

Availability and Implementation: Source code is available at `https://github.com/kreid415/wasserstein-critic-deconfounding`**. Data are available at** `https://figshare.com/articles/dataset/Benchmarking_atlas-level_data_integration_in_single-cell_genomics_-_integration_task_datasets_Immune_and_pancreas_/12420968/1`**.**

Contact: kreid20@jh.edu.

Supplementary information: No supplementary data submitted.

Keywords Adversarial Deconfounding, Single-cell Integration, Wasserstein Critic, Batch Effect Correction

Introduction

High-throughput gene expression experiments have revolutionized our understanding of molecular biology and cellular processes by providing a snapshot of a biological system at a point in time. Unfortunately, these measurements are subject to significant bias that can obscure the biological signals we wish to study. These biases stem from both the experimental conditions of the measurement, such as the technology used (Tung et al., 2017), and biological variables (Stegle et al., 2015), such as the age or disease status of the sample. This problem is compounded across experiments and laboratory conditions, making it challenging to compare gene expression results without controlling for these confounding variables (Leek et al., 2010; Goh et al., 2022; Yu et al., 2024).

A popular class of algorithms for removing confounding bias is latent-space models trained with adversarial losses that encourage learned representations free of information from confounding variables (Dincer et al., 2020; He et al., 2025). These techniques typically rely on discriminator networks that classify samples based on a confounding variable, thereby approximating the Jensen-Shannon (JS) divergence between distributions. However, a fundamental theoretical limitation of the JS divergence is that it fails to provide a meaningful gradient when two distributions have non-overlapping support (Arjovsky et al., 2017). This is a critical vulnerability in gene expression deconfounding, where samples from distinct experimental batches often occupy disjoint regions of the high-dimensional latent space. In this scenario, a standard discriminator can perfectly separate the batches, leading to gradient saturation and training instability that prevents the autoencoder from learning to align the data.

This instability has been addressed in the broader deep learning literature by replacing the discriminator with a Wasserstein critic—a network constrained to approximate the Earth Mover's (Wasserstein) distance (Arjovsky et al., 2017). Unlike the JS

Received: Date Month Year. **Revised received:** Date Month Year. **Accepted:** Date Month Year
© The Author(s) 2026. Copyright and licence statements to be updated by the publisher during production.

divergence, the Wasserstein distance remains a continuous and differentiable metric even when distributions are non-overlapping, providing robust gradients that drive alignment regardless of the initial separation. While Wasserstein-based approaches have been proposed for biological batch correction (Wang et al., 2023, 2022), they often include additional steps that obscure the contribution of the adversarial mechanism. To date, a rigorous isolation of the adversarial mechanism—comparing the standard discriminator against a Wasserstein critic while holding the autoencoder backbone constant—has not been performed.

In this work, we strip away these auxiliary constraints to investigate the Wasserstein objective in its purest form and seek to isolate the specific contribution of the adversarial mechanism itself. To enable this rigorous comparison in realistic multi-batch settings, we implement a *multi-headed critic* formulation that generalizes the Wasserstein distance to multiple batches. This formulation allows us to explicitly and simultaneously minimize the Earth Mover's distance of each source batch to a biologically grounded reference. By holding the autoencoder backbone constant, we control for architectural variables and directly test the efficacy of the objective function. We demonstrate that while replacing the standard Jensen-Shannon discriminator with this Wasserstein critic improves performance in disjoint settings, it introduces a topological bottleneck in scaling. Our results reveal a fundamental trade-off: the Wasserstein objective dominates in scenarios with severe, disjoint batch effects and overlapping cell-types, whereas the discriminator scales more effectively in fragmented atlases with disjoint cell-types.

Related Work

The problem of experimental integration in gene expression data is one of the most pressing problems in computational biology (Luecken et al., 2025), giving rise to a suite of solutions tailored to different use cases (Tran et al., 2020). These methods can be broadly categorized into statistical approaches and reference-based mapping (Luecken et al., 2022). Statistical methods, such as LIMMA (Ritchie et al., 2015), ComBat (Johnson et al., 2006), and scVI (Lopez et al., 2018), model the gene expression data as coming from a shared underlying process and regress out technical factors. In contrast, reference-based methods attempt to map query datasets into the space of a reference dataset. Popular reference methods include Mutual Nearest Neighbor (MNN) (Haghverdi et al., 2018), Harmony (Korsunsky et al., 2019), and Seurat (Stuart et al., 2019), which rely on iterative geometric alignment to bridge batch differences.

Adversarial learning offers a powerful alternative for reference-based mapping, allowing for end-to-end differentiable deconfounding within deep learning architectures. Early approaches, such as AD-AE (Dincer et al., 2020), utilize standard discriminator networks to remove confounding information from the latent space. Recent works have begun to incorporate Wasserstein-based objectives. For example, ResPAN (Wang et al., 2022) combines a Wasserstein critic with Random Walk Mutual Nearest Neighbors (RW-MNN) to guide alignment. Similarly, IMAAE (Wang et al., 2023) employs a Wasserstein-based adversarial autoencoder to tackle scenarios where cell types are disjoint across batches. Notably, the authors of IMAAE explicitly attribute the failure of reference-based alignment in these settings to a "lack of constraints" when the anchor batch is sparse, motivating their decision to enforce a global Gaussian prior. Consequently, the specific contribution of the *Wasserstein critic itself*—as a stable replacement for the discriminator in pure adversarial deconfounding—has not been isolated or rigorously evaluated. In this work, we bridge this gap by stripping away these additions and systematically comparing the Wasserstein critic against the standard discriminator within a controlled, state-of-the-art autoencoder framework.

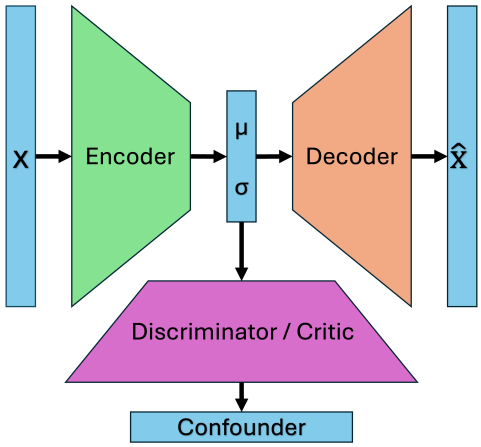

Figure 1 Experimental Architecture. To isolate the impact of the adversarial objective, we employ a fixed VAE backbone. We systematically swap the adversarial regularizer between a standard discriminator and Wasserstein critic to compare their efficacy in removing confounding information v.

Materials and Methods

To rigorously isolate the impact of the adversarial objective, we employ a fixed Variational Autoencoder (VAE) backbone adapted from scCRAFT (He et al., 2025). We hold the encoder, decoder, and auxiliary loss functions constant across all experiments, varying only the adversarial regularization mechanism (standard discriminator vs. Wasserstein critic). The total objective consists of four components: (1) a reconstruction loss, (2) a prior regularization loss, (3) the adversarial deconfounding loss (variable), and (4) a triplet loss to preserve local structure. A simplified architecture is shown in Figure 1.

Variational Autoencoder and Reconstruction Loss

The input to the model is log1p-transformed gene expression data, while the prediction target is the raw gene count values. To account for the sparsity of genomic data, the model assumes the data x follows a negative binomial (NB) distribution.

The encoder network, $q_\phi(z|x)$, is a three-layer feedforward network ($\rho \to 1024 \to 512$) projecting to a latent dimension of size 256, where ρ is the number of genes. This network predicts the vectors μ and σ that parameterize the latent normal distribution $Z \sim N(\mu, \sigma)$. Neural network batch normalization layers and ReLU activations are applied after each dense layer to ensure training stability.

The decoder reconstructs the expression profile by combining the sampled latent code z with the known confounding batch ID v. This is achieved via two parallel networks:

$$\hat{z} = \mathrm{ReLU}(G_1(z, v) + G_2(v))$$

where G_1 maps the concatenated latent and confounding vectors back to gene space ($z \oplus v \to 512 \to 1024 \to \rho$), and G_2 captures baseline batch-specific effects ($v \to 512 \to 1024 \to \rho$). This combined representation $\hat{z}$ is fed into a final likelihood network to predict the NB parameters (mean μ and dispersion r).

The reconstruction objective combines the NB negative log-likelihood with a cosine similarity loss to prevent the model from overfitting to high-count genes:

$$\mathcal{L}_{\text{NB}} = -\log P_{\text{NB}}(x|\mu, r) \tag{1}$$

$$\mathcal{L}_{\cos} = \mathbb{E}\left[1 - \frac{\langle \mathcal{X}, G(\mathcal{Z}, v)\rangle}{\parallel \mathcal{X}\parallel_2 \parallel G(\mathcal{Z}, v)\parallel_2}\right] \tag{2}$$

These terms are summed with the Kullback-Leibler (KL) divergence, which regularizes the latent space toward a standard normal prior:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{NB}} + \lambda_{\cos}\mathcal{L}_{\cos} + \beta_{\text{KL}}\text{KL}\left(q_\phi(z \mid x) \parallel N(0, I)\right) \tag{3}$$

In our experiments, we set the hyperparameters $\lambda_{\cos} = 20$ and $\beta_{\text{KL}} = 0.005$, identical to the original implementation (He et al., 2025).

Baseline: Discriminator-Based Deconfounding

To serve as a baseline for comparison, we implement the standard adversarial deconfounding approach using a discriminator network $h_\upsilon : \mathcal{Z} \to \mathcal{V}$. This network acts as a classifier that attempts to predict the confounding variable v (e.g., batch ID) from the latent embedding z. The discriminator is trained to minimize the prediction error (typically cross-entropy), effectively approximating the Jensen-Shannon divergence between batch distributions. The objective function for the discriminator is:

$$\min_{\upsilon} \mathbb{E}_{x\sim\mathcal{D}, z\sim q_\phi(z|x)}\left[L_{\text{CE}}(h_\upsilon(z), v)\right] \tag{4}$$

The discriminator architecture consists of a two-layer Multi-Layer Perceptron (MLP) mapping from the latent dimension to a vector of probabilities for each batch ($256 \to 128 \to 128 \to k$), with ReLU activations following the first two layers and a sigmoid or softmax following the final linear layer.

Multi-Headed Wasserstein Critic Formulation

To address the gradient instability of the discriminator, we replace it with a Wasserstein critic. To avoid the quadratic computational cost ($O(K^2)$) of pairwise matching in multi-batch settings, we formulate a *reference-based multi-headed critic*. We designate one batch as the *reference* (r) and align all other source batches ($i \neq r$) to this biological ground truth. We perform a sensitivity analysis for the choice of reference and report the results below.

We modify the critic network $C_\rho : \mathcal{Z} \to \mathbb{R}^K$ to output a *vector* of scores. The i-th component, $C_\rho(z)_i$, represents the transport cost between batch i and the reference. The critic maximizes the sum of pairwise Wasserstein distances:

$$L_C = -\frac{1}{K-1}\sum_{\substack{i=1\\ i\neq r}}^{K}\left(\mathbb{E}_{z\sim P_i}[C_\rho(z)_i] - \mathbb{E}_{z\sim P_r}[C_\rho(z)_i]\right) \tag{5}$$

Simultaneously, the encoder (generator) minimizes this distance to confuse the critic:

$$L_G = \frac{1}{K-1}\sum_{\substack{i=1\\ i\neq r}}^{K}\left(\mathbb{E}_{z\sim P_i}[C_\rho(z)_i] - \mathbb{E}_{z\sim P_r}[C_\rho(z)_i]\right) \tag{6}$$

We apply the Gradient Penalty (GP) (Gulrajani et al., 2017) individually for each head to enforce the Lipschitz constraint locally. We calculate this penalty on points interpolated strictly between the source batch and the reference batch to enforce the constraint along the relevant transport manifold. Crucially, to ensure this penalty remains well-defined, we employ a stratified sampling strategy that forces every mini-batch to contain samples from the reference distribution. This prevents the instability that arises if a mini-batch randomly lacks reference samples. We use a 10:1 (critic:encoder) update ratio, as the Wasserstein objective provides stable gradients even near optimality.

Auxiliary Regularization and Total Loss

To mitigate latent over-correction (where biological distinctions are erased), the backbone includes a dual-resolution triplet loss. This enforces that cells of the same type remain closer in latent space than cells of different types:

$$\mathcal{L}_{\text{triplet}} = \max(\parallel z_r - z_p\parallel_2^2 - \parallel z_r - z_n\parallel_2^2 + \text{margin}\,, 0) \tag{7}$$

Final Objective. The total loss function for our experiments combines the fixed backbone losses with the variable adversarial term:

$$\mathcal{L}_{\text{Total}} = \underbrace{\mathcal{L}_{\text{VAE}} + \mathcal{L}_{\text{triplet}}}_{\text{Fixed Backbone}} - \underbrace{\lambda_{\text{adv}}\mathcal{L}_{\text{adversary}}}_{\text{Variable}} \tag{8}$$

Implementation, Training, and Model Selection

We utilized three standard benchmarks (Luecken et al., 2022): **Pancreas** (9 batches), **Lung** (2 batches), and **Immune** (4 batches). All models were implemented in PyTorch and trained on a single NVIDIA 4090 GPU using the Adam optimizer ($\beta_1 = 0.5, \beta_2 = 0.9$) with a batch size of 1024. To ensure stability, the Gradient Penalty weight was fixed at $\lambda_{GP} = 10$ (Gulrajani et al., 2017), and we employed a warmup period (first 10 epochs) where $\lambda_{\text{adv}} = 0$ to allow for valid reconstruction before alignment. We used stratified sampling to ensure equal representation of all batches within each mini-batch, preventing mode collapse in smaller experimental batches. Models were trained and evaluated using a nested cross-validation scheme (5 outer folds, 3 inner folds), where inner loops ran for 40 epochs and outer loops for 80 epochs. Hyperparameters were selected to maximize the composite integration score (iLISI − cLISI) on held-out data.

Experiments and Datasets

We designed a comprehensive experimental framework to isolate the behavioral differences between the Jensen-Shannon (discriminator) and Wasserstein (critic) objectives. Using standard single-cell integration benchmarks (Luecken et al., 2022) and a fixed VAE backbone (He et al., 2025), we conducted three targeted experiments to assess performance across varying levels of complexity:

1. **Binary Integration (Fundamental Baseline):** We first restricted the problem to the two largest batches in each dataset. This simplifies the multi-class problem to a binary classification task, allowing us to directly test the fundamental hypothesis that the Wasserstein critic will outperform the discriminator in integrating data from different experimental technologies.
2. **Multi-Batch Scalability:** To understand how the adversaries scale with problem complexity, we incrementally increased the number of batches from 2 to K (adding batches in descending order of size). The reference batch for these experiments was fixed to the largest batch in each dataset.
3. **Reference Sensitivity Analysis:** To evaluate the robustness of the multi-headed critic formulation, we exhaustively trained models using every possible batch as the reference target. This experiment determines whether the critic's performance is dependent on the specific topology of the chosen reference.

Confounding Scenarios. For every experiment, we evaluated two distinct confounding scenarios to test robustness against correlation:

- **Balanced (Ideal):** We intersected the cell types across all batches, keeping only the shared types. In this scenario, batch and cell type are independent, which we hypothesize will be an easier task.
- **Unbalanced (Realistic):** We used the raw datasets where cell type proportions vary significantly between batches. This introduces a strong correlation between batch and biology, a scenario known to cause mode collapse in standard adversarial training.

For every combination of dataset, experiment, and scenario, we trained both a discriminator-based and a Wasserstein critic-based VAE and evaluated performance on held-out test data. For statistical comparisons, we applied the Corrected Paired t-test (Nadeau and Bengio, 2003) to account for the variance introduced by cross-validation overlap.

Data Preprocessing

Preprocessing. Prior to analysis, we applied standard quality control filters: cells expressing fewer than 300 genes and genes expressed in fewer than 5 cells were removed. Feature selection was performed by identifying the top 2,000 highly variable genes (HVGs) using the Seurat v3 dispersion method. For the VAE input, count data were library-size normalized (scaled to 10^4 reads per cell) and log1p-transformed ($\log(x+1)$).

A summary of the dataset compositions, including the number of cells and distinct sequencing technologies (batches), is provided in Table 1.

Evaluation Framework

To strictly quantify the trade-off between removing confounding artifacts and preserving biological signal, we utilize a suite of complementary metrics. We categorize these into local and global measures to ensure our evaluation captures both fine-grained manifold alignment and macro-scale cluster structure.

Local Metrics (Primary). We prioritize the Local Inverse Simpson's Index (LISI) (Korsunsky et al., 2019) as our primary evaluation tool. Unlike global summary statistics, LISI provides a cell-specific measure of integration quality, allowing us to assess how well the model aligns distributions within the local neighborhood of each cell:

- **iLISI (Integration):** Measures the effective number of batches in a cell's local neighborhood. Higher values indicate superior mixing of distinct experimental batches.
- **cLISI (Conservation):** Measures the effective number of cell types in a neighborhood. We normalize this metric between 0 and 1, such that lower values (closer to 0) indicate superior conservation, signifying that distinct cell types remain separated and are not erroneously merged.

LISI metrics were calculated with a perplexity of 30 and $k = 90$ nearest neighbors (Korsunsky et al., 2019).

Global Metrics (Secondary). To assess macro-scale alignment, we employ standard metrics calculated using the `scib` Python framework (Luecken et al., 2022). To ensure consistent interpretability, we utilize the scaled versions of these metrics, where *higher values always indicate better performance*:

- **ASW Batch (Integration):** The Batch Silhouette Width measures the degree of batch mixing. We report the absolute scaled value $(1 - |\mathrm{ASW}_{\mathrm{batch}}|)$, normalized to $[0, 1]$. A score of 1 indicates perfect batch mixing, while 0 indicates complete separation.
- **ASW Cell Type (Conservation):** The Biological Silhouette Width measures the compactness of cell-type clusters. We report the standard silhouette width normalized to $[0, 1]$, where 1 indicates perfectly distinct cell types, and 0 indicates overlapping biological populations.
- **ARI (Adjusted Rand Index):** Measures the agreement between latent clustering and true cell type labels (0 = random, 1 = perfect match).

Generalization to Unseen Data. A critical component of our study is evaluating whether the adversarial alignment generalizes to new data. Unlike standard benchmarks that often evaluate on the training set, we compute all reported metrics on *held-out test data*. Specifically, we project test cells into the learned latent space and construct a joint k-nearest neighbor graph containing both training and test cells. Metrics are then computed specifically for the test nodes, measuring how well unseen query cells integrate into the established reference manifold.

Results

The Wasserstein Critic Achieves Superior Local Integration with Negligible Conservation Costs in the Binary Case

We first explore the utility of the Wasserstein critic in the simplest deconfounding scenario: integrating two distinct batches. This reduces the problem to binary classification and the corresponding binary Wasserstein formulation. We restricted the analysis to the two largest batches for each dataset (Pancreas, Immune, and Lung) under both balanced and unbalanced confounding scenarios. Hyperparameters were optimized using nested cross-validation, and performance was evaluated using five metrics covering both integration and biological conservation (Figure 2). Statistical significance was assessed using corrected paired t-tests (Nadeau and Bengio, 2003)(Table 2).

Table 1 Detailed composition of benchmark datasets, including cell counts and cell type diversity per batch.

Pancreas Dataset (9 Batches)			Lung Dataset (2 Batches)			Immune Dataset (4 Batches)		
Tech	**Cells**	**Types**	**Protocol**	**Cells**	**Types**	**Chemistry**	**Cells**	**Types**
CEL-Seq	1,004	13	10x Chromium	22,771	16	10X	8,829	10
CEL-Seq2	2,285	13	Drop-Seq	6,485	11	smart-seq2	1,022	4
Fluidigm C1	638	13				10x Chromium v2	12,928	16
inDrop-Seq V1	1,937	14				10x Chromium v3	10,727	12
inDrop-Seq V2	1,724	14						
inDrop-Seq V3	3,605	14						
inDrop-Seq V4	1,303	14						
SMARTer	1,492	4						
Smart-Seq V2	2,394	13						

Integration Performance. The Wasserstein critic demonstrated a strong advantage in local batch mixing. The critic achieved higher iLISI scores across all scenarios, indicating superior removal of local batch effects. In the Immune dataset, the critic performed nearly as well as the discriminator, with similar significant gains observed in the Lung dataset and the Pancreas Balanced scenario. Interestingly, global integration measures (ASW Batch) did not show the same effect; the discriminator performed comparably or slightly better (e.g., Lung and Pancreas). This discrepancy suggests that while the discriminator effectively aligns global distributions, the critic better resolves fine-grained batch structures within local cell neighborhoods.

Biological Conservation. A consistent trade-off was observed in biological conservation, with the aggressive batch mixing of the Wasserstein critic resulting in a minor reduction in cell-type purity. In terms of local neighborhood structure (cLISI), the critic exhibited slightly higher scores compared to the discriminator, indicating that distinct cell types were less strictly separated in the local manifold. This difference was statistically significant in the Immune, Lung, and Pancreas unbalanced datasets. However, the magnitude of this shift was negligible (typically < 0.02), and in the Pancreas balanced dataset, the difference did not reach statistical significance. This pattern was corroborated by the global ASW Cell Type metric, where the critic yielded consistently lower scores, reflecting slightly less compact clusters. This reduction was statistically significant across most scenarios, though it appeared only as a trend in the Pancreas Unbalanced and Immune balanced scenarios. Importantly, despite these statistical differences, the absolute conservation scores remained high across all datasets, suggesting that the critic's substantial gains in integration quality come at a minimal cost to biological fidelity. The Adjusted Rand Index (ARI), which measures clustering agreement, showed similar patterns with the discriminator consistently outperforming the critic in the Pancreas and Immune datasets. In the Lung dataset, differences in ARI were statistically insignificant.

Sensitivity to Confounding Structure. We examined whether the correlation between batch and biological variables (Unbalanced scenarios) disproportionately affected either method. Contrary to the theoretical expectation that correlated confounding would universally degrade performance, we found no systematic pattern across datasets. For example, while the discriminator's iLISI performance notably dropped in the Immune dataset, it actually improved in the Pancreas dataset. Similarly, the Wasserstein critic showed variable sensitivity: it maintained high performance in the Immune and Pancreas datasets but suffered a significant drop in the Lung dataset. Despite this heterogeneity, the critic maintained its relative advantage, outperforming the discriminator in local integration (iLISI) across all individual scenarios, regardless of the confounding structure.

Summary. Overall, these results highlight a distinct behavior between the two adversaries. The discriminator acts as a conservative regularizer, preserving high within-cluster density (high ASW Cell Type) but failing to fully integrate local neighborhoods. In contrast, the Wasserstein critic serves as a stronger adversary, forcing the encoder to thoroughly mix batches. This yields significant improvements in integration (iLISI) at the cost of a minor, consistent reduction in cell-type compactness, though the downstream effect on clustering (ARI) varies by biological context.

Scalability and Stability in Multi-Batch Integration

To see how well these methods scaled to larger numbers of batches, we incrementally increased the problem complexity from the binary baseline ($K = 2$) up to the full dataset ($K = 9$ for Pancreas, $K = 4$ for Immune). We excluded the Lung dataset from this analysis as it contains only two batches. This experiment assesses whether the multi-headed Wasserstein formulation maintains its advantage as the number of alignment tasks grows (Figure 3).

Performance Decay. As expected, integration difficulty increases with the number of batches. We observed a decrease in performance across all scenarios for both methods, reflecting the inherent challenge of aligning multiple disjoint distributions into a shared low-dimensional manifold.

Scalability Depends on Reference Density. Our multi-batch experiments reveal that the choice of optimal adversary depends on the intrinsic structure of the dataset and the density of the reference batch. In the Immune dataset, the Wasserstein critic consistently outperformed the discriminator as batches were added. We attribute this to the severity of the batch effects and the high density of the batches. The reference batch provided a comprehensive support for the biological manifold, allowing the critic to robustly pull disjoint batches together without losing signal.

Conversely, in the Pancreas dataset ($K = 9$), the discriminator eventually overtook the critic. We hypothesize this highlights a trade-off in our reference-based multi-headed formulation. As K increases, our critic must solve $K - 1$ distinct transport problems to a single reference batch. The Pancreas dataset consists of many fragmented batches with significantly smaller sample sizes. In this case, the multi-headed critic struggles because the chosen reference frequently lacks the full spectrum of cell types present in the query batches. Forcing alignment to a sparse reference inherently compresses the latent space, erasing biological distinctions found in the larger atlas. The standard discriminator, which aggregates all cells into a global

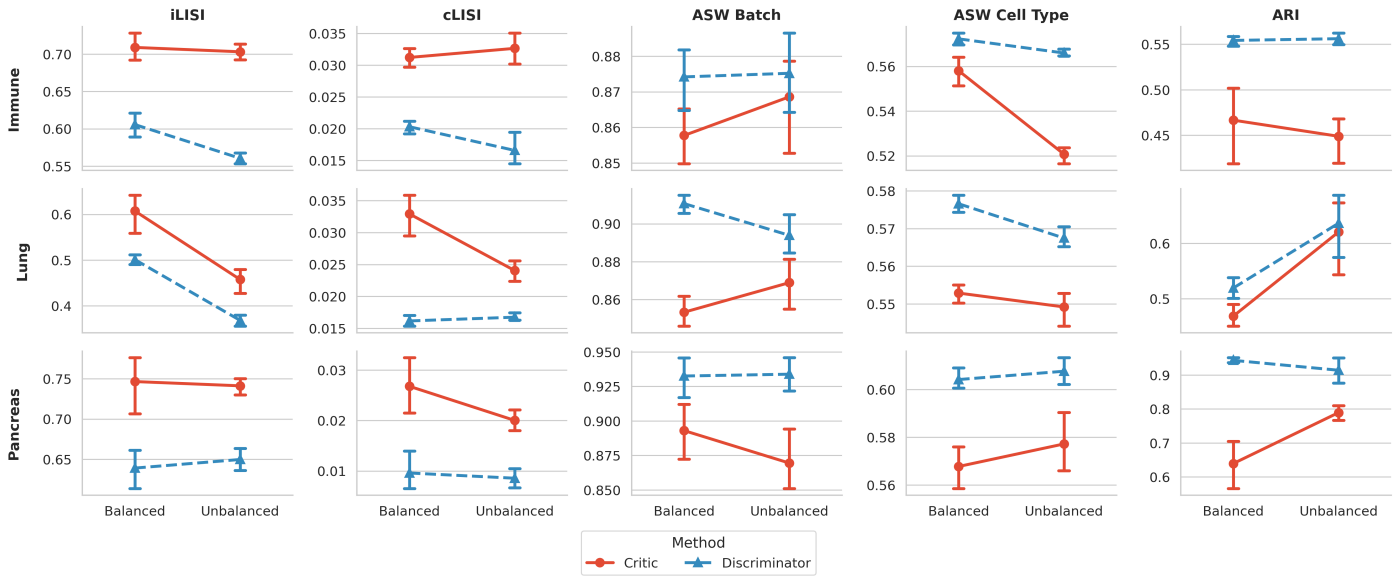

Figure 2 Performance comparison of adversarial deconfounding methods across datasets and batch scenarios. We evaluated the Wasserstein critic (Red circles, solid line) against a standard discriminator (Blue triangles, dashed line) across three datasets (Pancreas, Immune, Lung). Performance was assessed using five metrics. **Integration metrics:** iLISI and ASW Batch (higher is better) measure the removal of batch effects. **Conservation metrics:** ASW Cell Type and ARI (higher is better) measure the retention of biological signal, while cLISI (lower is better) measures the purity of cell-type neighborhoods. The x-axis denotes the confounding scenario (Balanced vs. Unbalanced). Points represent the mean score across 5-fold cross-validation, and error bars indicate the 95% confidence interval.

Table 2 Statistical Comparison: Wasserstein Critic vs. Discriminator across Balanced and Unbalanced Scenarios.

		Balanced Scenario			Unbalanced Scenario		
Dataset	Metric	Critic (SD)	Discrim (SD)	p-value	Critic (SD)	Discrim (SD)	p-value
Pancreas							
	iLISI	0.747 (0.045)	0.639 (0.032)	0.0814	0.741 (0.013)	0.650 (0.019)	0.0016
	ASW Batch	0.893 (0.025)	0.933 (0.018)	0.0375	0.870 (0.029)	0.934 (0.016)	0.0279
	cLISI	0.027 (0.008)	0.010 (0.005)	0.0863	0.020 (0.003)	0.009 (0.003)	0.0072
	ASW Cell	0.568 (0.011)	0.604 (0.006)	0.0220	0.577 (0.015)	0.608 (0.008)	0.0590
	ARI	0.640 (0.090)	0.944 (0.009)	0.0066	0.790 (0.029)	0.915 (0.048)	0.0253
Immune							
	iLISI	0.709 (0.023)	0.606 (0.021)	0.0069	0.703 (0.013)	0.561 (0.009)	6.27e-04
	ASW Batch	0.858 (0.010)	0.874 (0.011)	0.1350	0.869 (0.017)	0.875 (0.014)	0.7093
	cLISI	0.031 (0.002)	0.020 (0.001)	0.0030	0.033 (0.003)	0.017 (0.003)	0.0036
	ASW Cell	0.558 (0.008)	0.572 (0.004)	0.1053	0.521 (0.005)	0.566 (0.002)	1.57e-04
	ARI	0.467 (0.052)	0.554 (0.008)	0.0621	0.449 (0.033)	0.556 (0.008)	0.0113
Lung							
	iLISI	0.608 (0.056)	0.501 (0.015)	0.0278	0.458 (0.034)	0.368 (0.016)	0.0029
	ASW Batch	0.853 (0.010)	0.911 (0.006)	0.0042	0.869 (0.017)	0.894 (0.013)	0.0303
	cLISI	0.033 (0.004)	0.016 (0.001)	0.0046	0.024 (0.002)	0.017 (0.001)	0.0085
	ASW Cell	0.553 (0.003)	0.577 (0.003)	6.82e-04	0.549 (0.006)	0.568 (0.003)	0.0139
	ARI	0.468 (0.025)	0.519 (0.024)	0.0925	0.621 (0.090)	0.637 (0.071)	0.8536

*Corrected for CV overlap (Nadeau & Bengio, 2003)

classification task, appears more robust to this sparsity issue because it utilizes the union of all batches rather than relying on a single reference anchor.

Practical Recommendation. Our findings suggest a clear heuristic for practitioners: Use the Wasserstein critic when integrating large, disjoint datasets (e.g., cross-species or multi-modal integration), but prefer the standard discriminator for atlasing projects involving many small, fragmented sequencing runs.

Hyperparameter Sensitivity and Stability

A major practical limitation of adversarial deconfounding is the difficulty of tuning the regularization strength λ. To evaluate the stability of the two objectives, we analyzed the impact of varying λ across orders of magnitude on both integration (iLISI) and conservation (cLISI) (Figure 4).

Robustness to Regularization. Both the Wasserstein critic and discriminator exhibit broad optimal operating windows on the iLISI metric with the critic performing consistently higher than

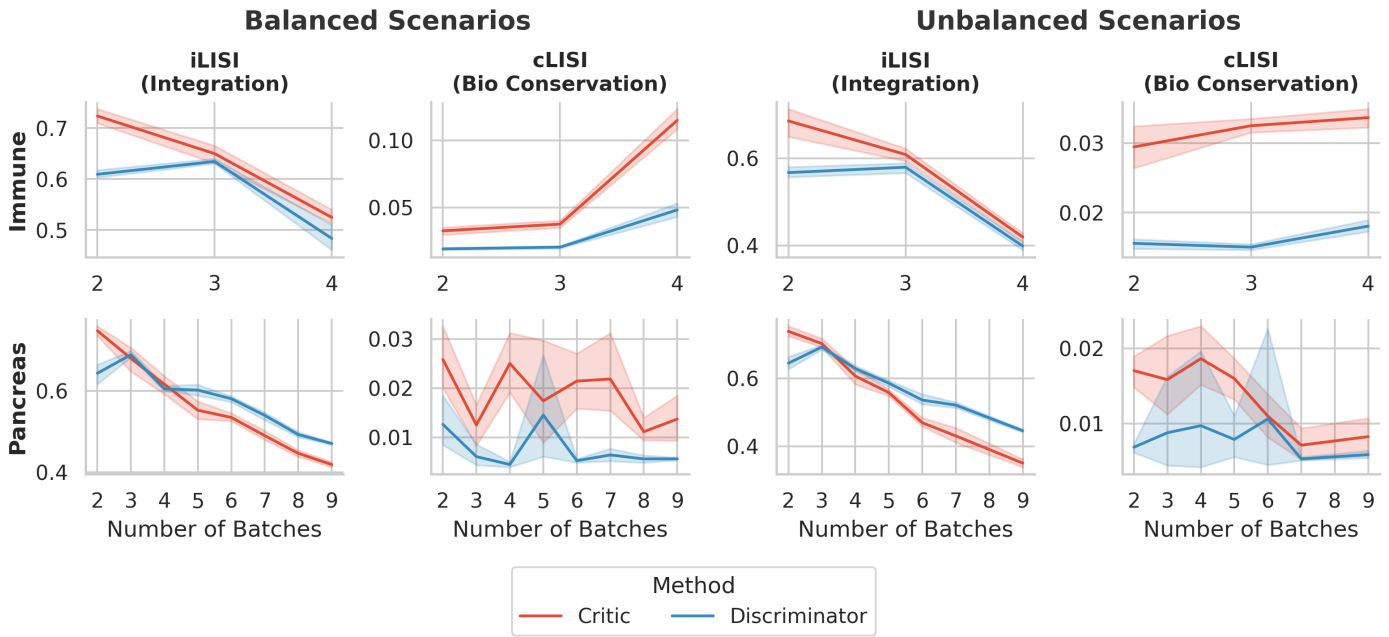

Figure 3 Impact of batch complexity on deconfounding performance. We evaluated the stability of the Wasserstein critic (Red line) and discriminator (Blue line) as the number of distinct batches increased (x-axis). The experiment was conducted across the Immune and Pancreas datasets under both Balanced and Unbalanced confounding scenarios. Performance is reported for **iLISI** (Integration, higher is better) and **cLISI** (Biological Conservation, lower is better). Solid lines represent the mean score, and error bars indicate the 95% confidence interval.

the discriminator. In the cLISI metric, the discriminator shows a broad flat curve that is insensitive to the choice of adversarial coefficient, whereas the critic gradually increases, highlighting the tradeoff between integration and biological preservation.

Pareto Efficiency. Crucially, at low regularization strengths ($\lambda = 0.01$), the Wasserstein critic achieves a strict Pareto improvement over the discriminator on the Pancreas and Immune datasets: it yields significantly higher integration scores (iLISI) while maintaining equivalent biological conservation (cLISI). This suggests that in these cases, the Wasserstein objective is fundamentally more efficient.

Impact of Reference Batch Selection

A unique architectural feature of our multi-headed critic is its alignment to a specific reference manifold. While this avoids the $O(K^2)$ complexity of all-versus-all matching, it introduces a dependency on the quality of the chosen reference batch. To quantify this, we exhaustively trained models using every possible batch as the reference target and analyzed the variance in performance (Figure 5).

Sensitivity to Batch Completeness. The performance variance across reference batches reflects the underlying sensitivity to reference density. As shown in Figure 5, in both balanced scenarios there is a low variance in the iLISI performance. The unbalanced Immune experiments continue to show a low variance attributable to the high cell-type overlap across batches without artificial balancing. In comparison, the Pancreas dataset has a high degree of cell-type differences across its many batches which creates a high variance in iLISI. Inspection reveals that these performance drops occur exclusively when the reference batch lacks representation of key cell types found in other batches which forces the encoder to erase biological signals that are present in the query but absent in the reference.

Practical Guidelines. These results suggest a clear heuristic for applying the Wasserstein critic: select the reference batch with the highest number of cells and most diverse cell-type composition.

Discussion

Adversarial deconfounding has become a cornerstone of deep learning-based batch correction, yet the choice of the adversarial objective itself has largely remained an implementation detail rather than a subject of rigorous study. In this work, we demonstrated that this choice is not merely technical but fundamental to the quality of integration. Our experiments reveal the following key results:

Mechanism Matters. In scenarios with strong confounding or disjoint distributions, the discriminator frequently failed to mix local neighborhoods effectively. We hypothesize that this occurs because the discriminator quickly converges to a "perfect classifier" in the early stages of training, providing gradients that saturate and vanish before the autoencoder can align the distributions. In contrast, the Wasserstein critic provided a robust, persistent gradient that forced mixing even when batches were initially far apart.

The Integration-Conservation Trade-off. While the Wasserstein critic significantly outperformed the discriminator in removing batch effects (higher iLISI), it incurred a consistent penalty in biological conservation (higher cLISI). Future work might explore adaptive weighting schemes, such as an annealing schedule that relaxes the adversarial penalty (λ_{adv}) once a sufficiency threshold for mixing is reached. This could potentially capture the best of both worlds, achieving the mixing of the critic with the conservation of the discriminator.

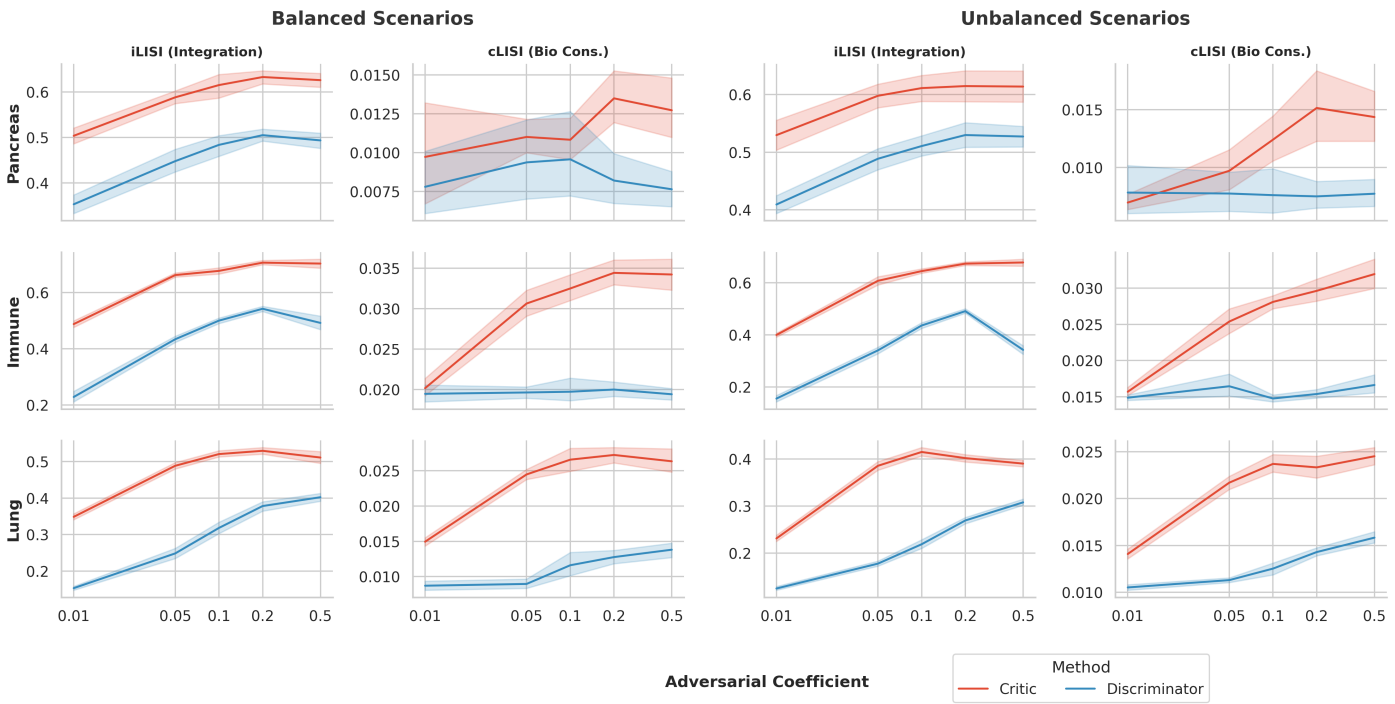

Figure 4 Sensitivity analysis of the adversarial coefficient across datasets. We evaluated the impact of varying the adversarial regularization strength (λ) on the performance of the Wasserstein critic (Red line) and discriminator (Blue line). The evaluation spans three datasets (Pancreas, Immune, Lung) under both Balanced and Unbalanced confounding scenarios. Performance is reported for **iLISI** (Integration, higher is better) and **cLISI** (Biological Conservation, lower is better). The x-axis represents the adversarial coefficient. Solid lines represent the mean score across validation folds, and error bars indicate the 95% confidence interval.

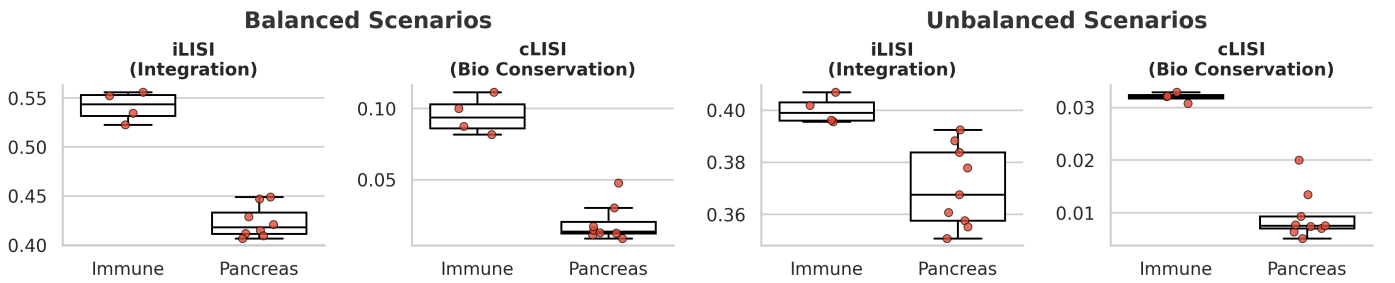

Figure 5 Sensitivity of the Wasserstein critic to reference batch selection. To assess the sensitivity to reference batch, we evaluated the performance of the Wasserstein critic using every possible batch as the reference target. The experiment spans two datasets under both Balanced and Unbalanced confounding scenarios. Box plots summarize the distribution of performance scores, where each individual point (Red) represents the result from a specific reference batch choice. Performance is reported for **iLISI** (Integration, higher is better) and **cLISI** (Biological Conservation, lower is better). A tighter distribution indicates that the method's performance is robust regardless of which batch is chosen as the reference.

Scalability Depends on Reference Density. In the Immune dataset, where batches were large and dense, the critic dominated because the reference provided comprehensive support for the transport map. Conversely, in the Pancreas dataset, which consists of many fragmented batches with small sample sizes, the discriminator overtook the critic, corroborating the results of (Wang et al., 2023). Our results empirically define the topological boundary of reference-based alignment– while the standard discriminator implicitly approximates a barycenter by aggregating all batches, the reference-based critic explicitly reveals the cost of forcing alignment to a topologically sparse anchor. This suggests the critic is the superior choice for atlas extension (mapping to a dense reference), while the discriminator remains optimal for *de novo* atlas assembly from fragmented data.

From Star Topology to Wasserstein Barycenters. This sparsity limitation is inherent to the *star topology* induced by our multi-headed formulation, where a single reference batch acts as the central hub for $K - 1$ pairwise transport maps. While this topology avoids the quadratic complexity of all-vs-all matching, it is topologically fragile: if the reference is incomplete, the entire alignment suffers. The theoretical solution to this limitation is the *Wasserstein barycenter* (Agueh and Carlier, 2011). Unlike our current approach which fixes the center to an existing batch, a barycenter approach would learn a "virtual" center—the geometric

median of all batch distributions. Implementing a barycenter objective would allow the model to align all batches to a shared, optimal manifold containing the union of all biological states, effectively combining the robust gradients of the Wasserstein critic with the global robustness of the discriminator.

Limitations and Future Directions. To isolate the impact of the adversarial objective, we deliberately restricted our analysis to a single, state-of-the-art VAE backbone. While this controlled design was necessary to validate the mechanism, future work should evaluate whether these findings generalize to other architectures and models. Additionally, our current multi-headed formulation addresses a single categorical confounder. Extending this to handle multiple simultaneous confounders or continuous covariates would significantly broaden the applicability of the Wasserstein objective.

Finally, the most promising frontier for this approach lies in *multi-modal integration*. Data modalities such as scATAC-seq are inherently sparser and more disjoint than transcriptomics, making the vanishing gradient problem even more likely. Adapting the Wasserstein critic framework to align distinct modalities (e.g., RNA and Chromatin accessibility) could unlock significant performance gains where standard discriminators struggle most.

References

M. Agueh and G. Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2), 2011.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 2017.

A. B. Dincer, J. D. Janizek, and S.-I. Lee. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics*, 36, 2020.

W. W. B. Goh, C. H. Yong, and L. Wong. Are batch effects still relevant in the age of big data? *Trends in biotechnology*, 2022.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 2017.

L. Haghverdi, A. T. L. Lun, M. D. Morgan, and J. C. Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5), 2018.

C. He, P. Filippidis, S. H. Kleinstein, and L. Guan. Partially characterized topology guides reliable anchor-free scRNA-integration. *Communications Biology*, 8(1), 2025.

W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1), 2006.

I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-R. Loh, and S. Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12), 2019.

J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10), 2010.

R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12), 2018.

M. D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, and F. J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1), 2022.

M. D. Luecken, S. Gigante, D. B. Burkhardt, R. Cannoodt, D. C. Strobl, N. S. Markov, L. Zappia, G. Palla, W. Lewis, D. Dimitrov, M. E. Vinyard, D. S. Magruder, M. F. Mueller, A. Andersson, E. Dann, Q. Qin, D. J. Otto, M. Klein, O. B. Botvinnik, L. Deconinck, K. Waldrant, S. N. Yasa, A. Szałata, A. Benz, Z. Li, B. Rieck, C. Ahlmann-Eltze, E. da Veiga Beltrame, C. Bravo González-Blas, A. T. Chen, B. DeMeo, C. Ergen, S. Floc'hlay, A. Gayoso, S. Hicks, Y. Ji, V. Kleshchevnikov, G. La Manno, M. G. Lombardo, R. Lopez, D. Righelli, H. Sarkar, V. Svensson, A. Tong, G. Xing, C. Xu, J. M. Bloom, A. O. Pisco, J. Saez-Rodriguez, D. Wulsin, L. Pinello, Y. Saeys, F. J. Theis, S. Krishnaswamy, and Open Problems Jamboree Members. Defining and benchmarking open problems in single-cell analysis. *Nature Biotechnology*, 43(7), 2025.

C. Nadeau and Y. Bengio. Inference for the generalization error. *Machine Learning*, 52(3), 2003.

M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), 2015.

O. Stegle, S. A. Teichmann, and J. C. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3), 2015.

T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of Single-Cell data. *Cell*, 177(7), 2019.

H. T. N. Tran, K. S. Ang, M. Chevrier, X. Zhang, N. Y. S. Lee, M. Goh, and J. Chen. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, 21(1), 2020.

P.-Y. Tung, J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K. Pritchard, and Y. Gilad. Batch effects and the effective design of single-cell gene expression studies. *Sci Rep*, 7, 2017.

X. Wang, C. Zhang, L. Wang, and P. Zheng. Integrating multiple single-cell rna sequencing datasets using adversarial autoencoders. *International Journal of Molecular Sciences*, 24(6), 2023.

Y. Wang, T. Liu, and H. Zhao. Respan: a powerful batch correction model for scrna-seq data through residual adversarial networks. *Bioinformatics*, 38(16), 2022.

Y. Yu, Y. Mai, Y. Zheng, and L. Shi. Assessing and mitigating batch effects in large-scale omics studies. *Genome Biology*, 25, 2024.