Wasserstein Critics Outperform Discriminators in Adversarial Deconfounding of Gene Expression Data

1st Kendall Reid
Bioinformatics
Johns Hopkins University
Baltimore, MD
kreid20@jh.edu
https://orcid.org/0000-0002-2105-6471

2nd Erhan Guven
Applied Physics Lab
Johns Hopkins University
Baltimore, MD
eguven2@jhu.edu

***Abstract*—High-throughput gene expression measurements are biased by technical and biological confounding variables, which obscure true biological signals. A common deep-learning-based solution involves training latent-space models with adversarial regularizers to ignore information from confounding variables. These methods rely on discriminator-networks which are unstable to train, especially when confounding effects are strong, leading to issues like vanishing gradients. Inspired by Wasserstein GANs, we propose replacing the standard discriminator with a Wasserstein critic for this deconfounding task. We systematically compare the performance of a discriminator-based approach against a Wasserstein critic-based approach on a state-of-the-art deconfounding autoencoder architecture. We evaluate these methods on standard single-cell integration benchmark datasets and demonstrate that the Wasserstein critic significantly outperforms the discriminator in integration quality with marginal loss of biological fidelity.**

I. 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, like the technology used [1], and biological variables [2], like 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 [3].

A popular class of algorithms for removing confounding bias are latent-space models trained with adversarial losses that encourage learned representations free of information from confounding variables [4], [5]. These techniques rely on a well-trained discriminator networks that distinguishes samples based on a confounding variable of interest, similar to how Generative Adversarial Networks (GANs) learn to generate data by distinguishing between real and generated data. While effective when tuned well, these methods can be unstable to train and are heavily dependent on the correct hyperparameters. These issues are exacerbated when the impact of the confounding variable is strong, such as when two different technologies are used for sequencing, which can cause the discriminator network to greatly overpower the latent-space generating network, resulting in disappearing gradients and failed training. This problem has been addressed in GANs by using Wasserstein critics —networks trained to approximate the Wasserstein distance between two distributions —instead of classification-based discriminators [6]. While Wasserstein-based approaches have been proposed for batch correction [7], a systematic comparison of the standard discriminator against a Wasserstein critic for deconfounding on a state-of-the-art architecture has not been performed.

In this work, we systematically compare the performance of the discriminator and Wasserstein critic based approaches for sequencing technology integration. We use standard integration datasets [8] to train adversarial deconfounding networks based a state-of-the-art architecture [5]. We evaluate the resulting latent spaces based on their integration quality and preservation of biological information using the integration local inverse Simpson's index (iLISI) and the cell-type local inverse Simpson's index (cLISI), respectively [9]. Our experiments show that the Wasserstein critic improves integration performance compared to the discriminator, with only a minor loss in cell-type distinction.

II. Related Work

The problem of batch effects and confounding variables in gene expression data has given rise to a suite of solutions to match different use cases [10]. These methods can be broken down into two main classes of algorithms, statistical methods and reference-based methods [11]. Statistical methods assume that the sampled gene-expression data comes from the same underlying data generating process, and by fitting that underlying distribution, we can correct for the undesirable variables. Well known examples of statistically-based methods are LIMMA [12], COMBAT [13], Liger [14], and scVI [15].

In comparison, reference-based methods take a dataset as reference and attempt to map all other datasets into the same reference data space. Popular reference-based methods include Mutual Nearest Neighbor (MNN) [16], Batch Balanced kNN (BBKNN) [17], Harmony [18], and Seurat [19].

The use of reference-based mapping has been well explored in the field of deep-learning based gene-expression modeling. A common approach is to apply adversarial regularizers to remove information about confounding variables from the learned latent space [4]. This idea is attractive because the deconfounding process is differentiable, allowing for this technique to be easily applied to any gene-expression deep learning model and trained in an end-to-end fashion. Deep learning-based methods include iMAP [20], MMD-ResNet [21], and ResPAN [7]. This idea has persisted, with more advanced architectures developed with specialized loss functions to encourage the preservation of biological information [5].

Each of these methods relies on adversarial training to evaluate how well the latent spaces of each dataset aligns with the reference. The use of discriminatory adversarial networks was initially explored in generative adversarial networks (GANs), which use dueling generator-discriminator networks to generate realistic data points from a target dataset [22]. One of the primary issues that arises when training GANs occurs when either the discriminator or the generator becomes more potent than the other. This imbalance causes a loss of meaningful gradients and results in failed convergence. To alleviate this issue, Wasserstein GANs were introduced, which replace the discriminator with Wasserstein critics, allowing gradients to flow when the discriminator overpowers the generator resulting in more stable training and improved results [6]. We hypothesize that in cases in which confounding variables have strong effects, like when technologies differ, the Wasserstein critic will outperform the discriminator in adversarial deconfounding.

III. Materials and Methods

The baseline model that we use is the model developed by He *et al.* with the only modification being the replacement of the discriminator with a Wasserstein critic. We summarize the baseline model and modifications here. The model consists of four loss components: (1) a reconstruction loss, (2) a prior regularization loss, (3) a domain adaptation loss based on a GAN discriminator or Wasserstein critic, and (4) a triplet loss to prevent discriminator over-correction. A simplified architecture can be seen in 1.

A. Variational Autoencoder and Reconstruction Loss

To account for the sparsity of genomic data, a variational autoencoder (VAE) model is trained to recreate the gene expression data x, where x is assumed to come from a zero-inflated negative binomial distribution (ZINB). The approximate posterior, or encoder network, $q_\phi(z|x)$, is a three-layer feedforward network with layer size ($\rho \rightarrow 1024 \rightarrow 512$), where ρ is the number of genes in the expression profile X. This network learns to predict vectors μ and σ of length 256

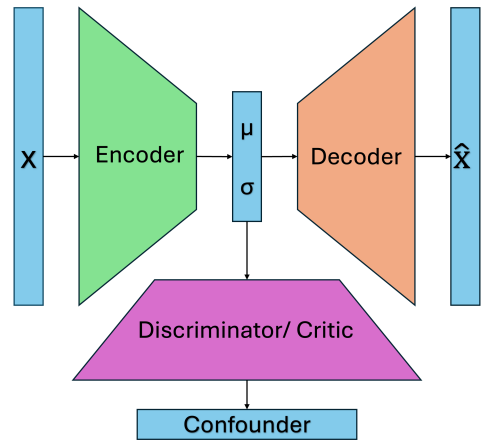

Fig. 1. The baseline model is a variational autoencoder (VAE) that learns a latent space representation. An adversarial regularizer, which we compare as either Discriminator or Wasserstein Critic, is trained on this latent space to remove information about the confounding variable (e.g., technology).

that parameterize a normal distribution: $Z \sim N(\mu, \sigma)$. Batch normalization and ReLU activations follow each dense layer for stability and non-linearity.

The sampled latent values z are combined with the true confounding ID variable v using two different decoder networks, yielding a combined latent representation:

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

G_1 is a feedforward network that maps dimensionality of z plus the number of unique confounding keys back to the number of genes ($\dim(Z) + \text{unique}(v) \rightarrow 512 \rightarrow 1024 \rightarrow \rho$). G_2 takes only the confounding key and maps back to the number of genes ($\text{unique}(v) \rightarrow 512 \rightarrow 512 \rightarrow 1024 \rightarrow \rho$). In both cases, each layer but the final layer is followed by a ReLU activation. This combined representation is then fed into the likelihood network that learns to predict the mean μ and dispersion r of a zero-inflated negative binomial that represents the underlying data shown below:

$$\mathcal{L}_{\text{ZINB}} = P_{\text{ZINB}}(x|\mu, r, \pi) = \begin{cases} \pi + (1-\pi)\left(\frac{r}{r+\mu}\right)^r & \text{if } x = 0, \\ (1-\pi)\frac{\Gamma(x+r)}{\Gamma(x+1)\Gamma(r)}\left(\frac{r}{r+\mu}\right)^r\left(\frac{\mu}{r+\mu}\right)^x & \text{if } x > 0. \end{cases} \quad (1)$$

The ZINB loss is prone to overly focusing on individual genes, so He *et al.* augment the ZINB reconstruction loss with the additional reconstruction cosine loss:

$$\mathcal{L}_{\cos} = \mathbb{E}\left[1 - \frac{\langle \mathcal{X}, G(\mathcal{Z}, v)\rangle}{\| \mathcal{X}\|_2 \, \| G(\mathcal{Z}, v)\|_2}\right] \quad (2)$$

Combining 1 and 2 yields the reconstruction loss:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{ZINB}} + \mathcal{L}_{\cos} \quad (3)$$

The reconstruction loss is combined with a regularization term to ensure that the latent variables approximate the prior,

which is assumed to be a standard normal distribution. This yields the formula:

$$\mathcal{L}(x) = \mathcal{L}_{\text{recon}} + \underbrace{\mathrm{KL}\left(q_\phi(z \mid x) \,\|\, p(z)\right)}_{\text{regularization}} \tag{4}$$

B. Adversarial Deconfounding

To remove information about the target confounding variable from the latent space, a discriminator is trained to produce an adversarial loss. The adversarial network, $h_\upsilon : \mathrm{Z} \rightarrow \mathrm{C}$, learns to predict the confounding variables given the sampled latent variables. This network is trained to optimize the following objective:

$$\min_{\upsilon} \mathbb{E}\left[L(h_\upsilon(x), c)\right] \tag{5}$$

where L is the adversarial loss. In the case of the discriminator, this is any differentiable loss function for a given confounder. In the case of our experiments, we use the binary cross-entropy between two experimental technologies.

The encoder network aims to create a latent space that maximizes equation 5, which is achieved by adding the following term to the original loss shown in equation 4:

$$\mathcal{L}(x) = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{KL}} - \underbrace{\lambda L(h_\upsilon(z), c)}_{\text{deconfounding}} \tag{6}$$

where λ is a hyperparameter. The discriminator maps from the latent space to a one-dimensional value. The discriminator uses layers that map from $(256 \rightarrow 128 \rightarrow 1)$ with ReLU activations following the first two layers.

C. Wasserstein Critic

Wasserstein GANs (WGANs) [6] make a number of changes to improve stability, which we summarize here. Firstly, the discriminator in a traditional GAN uses a sigmoid function in the final layer to output a prediction $\sigma(x) \in (0, 1)$. The sigmoid function saturates at 0 and 1, which can lead to vanishing gradients and failed training. The WGAN removes the sigmoid function and instead outputs a real value of how realistic the input data is. The binary cross function loss is no longer applicable, so instead the Wasserstein distance is used, which measures the difference between two distributions regardless of whether they overlap. For a critic C and generator G, the Wasserstein loss is:

$$\max\left[C(x) - C(G(z))\right].$$

The discriminator aims to maximize the Wasserstein distance, yielding the following loss function:

$$L_C = -[C(x) - C(G(z))].$$

Simultaneously, the generator aims to maximize the score given to its generated data shown in the following loss:

$$L_G = -C(G(z)).$$

Due to the critic's unbounded score, the gradient can grow infinitely large, leading to gradient explosion. To avoid this, we impose a 1-Lipschitz constraint on the gradient. This was originally achieved with gradient clipping [6], but has been improved with a regularization strategy [23] to encourage 1-Lipschitz gradients shown below:

$$L_{GP} = \lambda(||\nabla C(\hat{x})||_2 - 1)^2$$

D. Adversarial Deconfounding with Wasserstein Critic

We implement the Wasserstein loss into our adversarial deconfounding equation 4 as follows. For a given confounder with two possible classes, we arbitrarily assign one class as the reference and the other as the target. We separate these samples into two datasets. The critic network C_ρ aims to assign a positive score to the latent representation of all samples from the target dataset and a negative score to all samples from the reference, maximizing the Wasserstein distance, yielding the loss shown below:

$$L_C = -[C_\rho(z_t) - C_\rho(z_s)].$$

Simultaneously, the encoder aims to create indistinguishable latent representations by maximizing the score for the reference dataset and minimizing the score for the target dataset. This is shown in the following loss:

$$L_G = -[C(z_s) - C(z_t)].$$

Plugging this into equation 4 and including the gradient regularization yields:

$$\mathcal{L}_W = \lambda([C(z_s) - C(z_t)] + \lambda_{GP}(||\nabla C(\hat{x})||_2 - 1)^2$$

E. Dual-resolution triplet loss

To mitigate any latent over-correction by the discriminator, He *et al.* introduce the following loss:

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

where z_r, z_p, z_n are the latent embeddings of the reference, positive, and negative cells. Positive cells are defined as belonging to the same cell type, whereas negative cells belong to a different cell type.

Combining 6 with 7 yields our final simplified loss:

$$\mathcal{L}(x) = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{KL}} - \mathcal{L}_{\text{deconfound}} + \mathcal{L}_{\text{triplet}} \tag{8}$$

All code to reproduce these experiments are available at https://github.com/kreid415/wasserstein-critic-deconfounding.

F. Training Details

In order to jointly train our encoder and adversarial networks, we alternate training the encoder and adversarial network. We use a learning rate of 0.001 for all networks with the ADAM optimizer. For the GAN, we alternate between training the generator and the discriminator at every iteration. For the Wasserstein critic, we train the critic for 10 steps for every step of the generator. We train for a total of 500 epochs. To allow for the network to learn an initial representation, a 50-epoch warm-up is used in which the deconfounding loss is set to zero. Each mini-batch is stratified to have roughly the same number of cells for each cell type. To balance the different loss

functions, we use weighted losses. The triplet loss weight and cosine similarity loss weight are fixed at $\lambda_T = 1, \lambda_{cos} = 20$, as per the original implementation.

To select the deconfounding loss weight, we perform nested k-fold cross-validation stratified by cell type with a hyperparameter grid search over [0.01, 0.05, 0.1, 0.2, 0.5]. We choose these coefficients based on the scale of the original results scCRAFT [5].We perform this search with 5 outer-folds and 3 inner-folds and evaluate our performance using a composite score, iLISI - cLISI, which aims to maximize the iLISI score and minimize the cLISI score.

G. Experiments

For each of the datasets listed below, we trained a ZINB-VAE with adversarial regularization to ignore the technology confounder using a discriminator and a Wasserstein critic. For each method, we measured (1) how well the confounder was removed, and (2) how well the latent spaces preserved biological information. The evaluation metrics for these goals are elaborated below. We compare each of these methods using a paired t-test on the results from held-out test data.

H. Dataset

To evaluate our approach, we use the Immune, Pancreas, and Lung standard integration benchmark datasets from [8]. For each dataset we create a subset of the data from the largest two sequencing technologies by cell count to align with the Wasserstein critic's binary constraints. We perform quality control by filtering cells with less than 300 genes and filtering genes with less than 5 cells. We then normalize counts per cell to $1e4$, convert to $\log 1$p, and reduce the number of genes modeled by restricting to the top 2000 highly variable genes. A summary description of the dataset sizes and representative sequencing technologies is presented below:

TABLE I
COMBINED CELL COUNTS BY DATASET AND TECHNOLOGY.

Source	Cell Count
Pancreas Dataset	
inDrop-Seq	3,605
Smart-Seq	2,394
Total	5,999
Lung Dataset	
10x Chromium	22,771
Drop-Seq	6,485
Total	29,256
Immune Dataset	
10x Chromium v2	12,928
10x Chromium v3	10,727
Total	23,655
Grand Total	58,910

I. Evaluation

To evaluate our approach, we use the Local Inverse Simpson's Index (LISI), which measures the diversity of categories in the local neighborhood of a cell's representational embedding. We use this to calculate both iLISI (integration LISI, for deconfounding) and cLISI (cell-type LISI, for cell type biological information preservation). The local neighborhood is defined using a perplexity of 30 and a k-nearest neighbor graph using k=90 neighbors, which are the standard suggested values [18].

For robust evaluation in our nested cross-validation, we generate embeddings for both the training and held-out test sets. We then build the kNN graph on the combined embeddings of all cells (train and test). Finally, we calculate the iLISI and cLISI scores for only the subset of cells belonging to the test set, which assesses how well they integrate into the established training data structure.

IV. RESULTS

To understand the difference between the discriminator and critic approaches, we performed both qualitative and quantitative analyses. Qualitatively, we inspected the latent representation of the cells from the Pancreas data set and compared each method to see how well confounding variable effects are removed on held-out test data. In order to visualize the latent space for these models, we used the uniform manifold approximation and projection (UMAP) algorithm [24] to project from our latent space into a 2-dimensional representation. Figure 2 shows a clear separation of confounding variable (technology) that does not correspond to the biological cell type groupings. In comparison, Figure 3 shows that in both the case of the discriminator and the critic, the technologies are aligned. This is reflected in the cell type coloring, where the clusters are dominated by the cell type and not disjointed by the confounding effect.

Quantitatively, we compared the differences between both the iLISI and cLISI metrics for the critic and discriminator across all three datasets using a paired t-test to measure statistical significance. The quantitative metrics show mixed results when comparing the discriminator and the critic as shown in 4. In all cases, the normalized iLISI value is higher in the critic than in the discriminator. This indicates that the samples are mixed more thoroughly by the critic than the discriminator. On the other hand, the normalized cLISI value is also higher in the critic than in the discriminator. This indicates that the cell types are less well defined in the critic than the discriminator. This is supported by the paired t-test shown in Table II, which shows statistically significant differences in mean between the two approaches, with the critic having a higher mean iLISI and cLISI. Notably, both the critic and the discriminator have markely worse performance on the Lung dataset, which has the highest degree of technology class imbalance.

V. DISCUSSION

Gene expression data offers a lens into the inner workings of the cell and its molecular mechanisms. Unsupervised learning algorithms like adversarial deconfounding variational autoencoders are a promising direction towards identifying unique

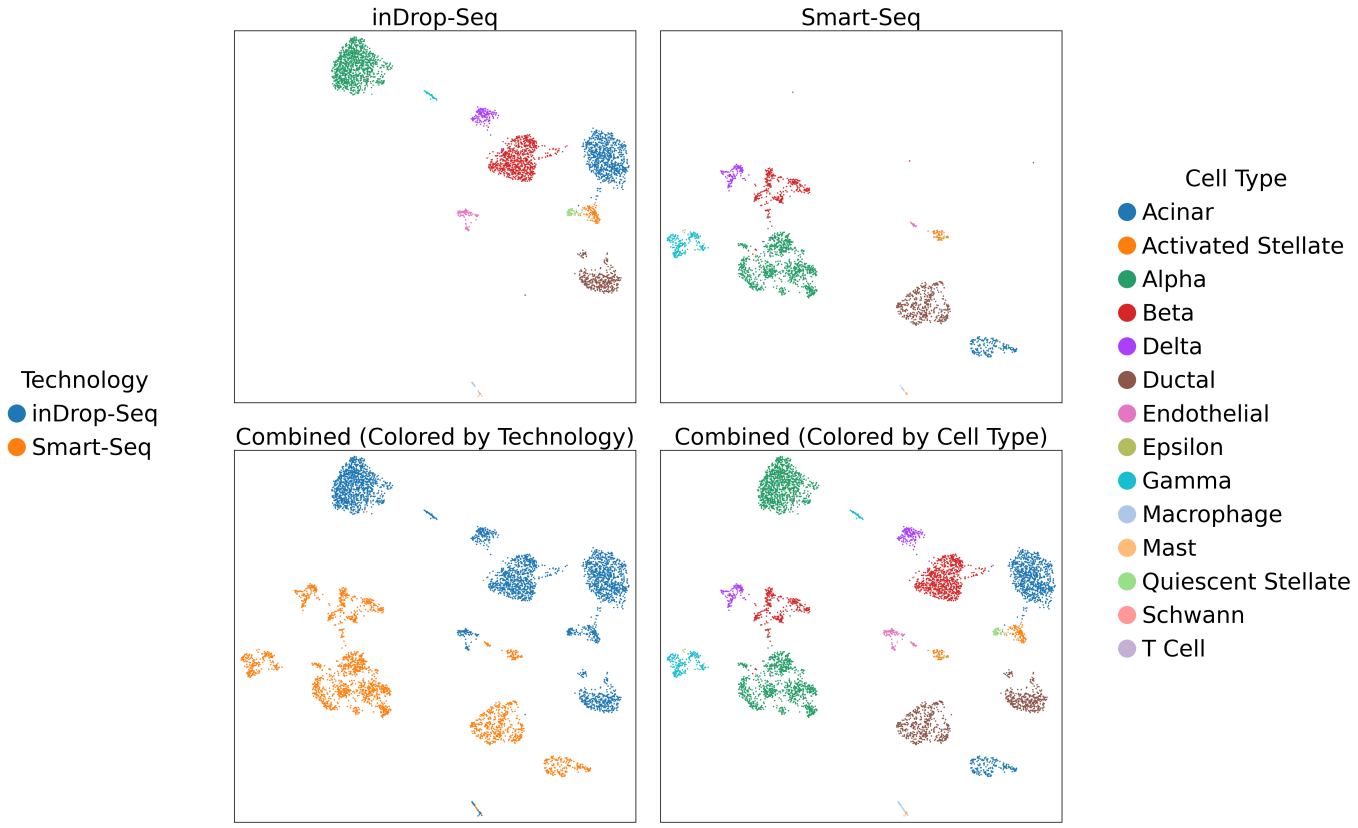

Fig. 2. UMAP visualization of the baseline gene expression for the Pancreas dataset. (Top) Visualization of the two technologies used for training colored by cell type. (Bottom) Combined visualization of the two technologies. (Bottom left) Cells are colored by sequencing technology, revealing a strong batch effect (clear separation). (Bottom Right) The same latent space is colored by cell type, showing that the biological groupings are obscured by the technology confounder.

TABLE II
PAIRED T-TEST RESULTS (CRITIC VS. DISCRIMINATOR)

Metric	Critic Mean (SD)	Disc. Mean (SD)	Mean Diff	95% CI of Diff	t-stat	p-value
Pancreas Dataset						
ILISI	0.7098 (0.0364)	0.5709 (0.0299)	0.1389	(0.0519, 0.2259)	4.4343	0.011384
CLISI	0.0191 (0.0017)	0.0100 (0.0056)	0.0091	(0.0015, 0.0166)	3.3313	0.029070
Immune Dataset						
ILISI	0.7133 (0.0202)	0.3472 (0.0577)	0.3661	(0.2835, 0.4487)	12.3024	0.000251
CLISI	0.0399 (0.0027)	0.0174 (0.0041)	0.0225	(0.0143, 0.0308)	7.5684	0.001634
Lung Dataset						
ILISI	0.3836 (0.0158)	0.2533 (0.0092)	0.1302	(0.1050, 0.1555)	14.3414	0.000137
CLISI	0.0314 (0.0031)	0.0130 (0.0011)	0.0184	(0.0134, 0.0235)	10.1118	0.000538

signals and processes from confounding variables. In this work, we show that the choice of adversarial deconfounding regularizer greatly impacts the resulting representation. In the binary integration case, the use of Wasserstein critics greatly improves upon the baseline discriminator approach with marginal loss of biological fidelity.

Qualitatively, both algorithms appeared to function similarly. They clearly mixed the two different technology sources, resulting in clusters that are driven by the biological signal of cell-type. Quantitatively, we see a minor tradeoff between approaches. While the critic outperformed the discriminator in integration, it sacrifices a small amount of biological consistency as cell types become mixed as well. We argue that the gains in integration greatly outweigh the loss of fidelity; however, we acknowledge that there may be situations in which this is not the case. Ultimately, in comparing these approaches, we have gained a better understanding of the tradeoffs between these two methods and when they may be best suited for a given scenario. Both methods struggle when there is a large imbalance in the number of samples across technologies, however the critic-based approach still performs significantly better.

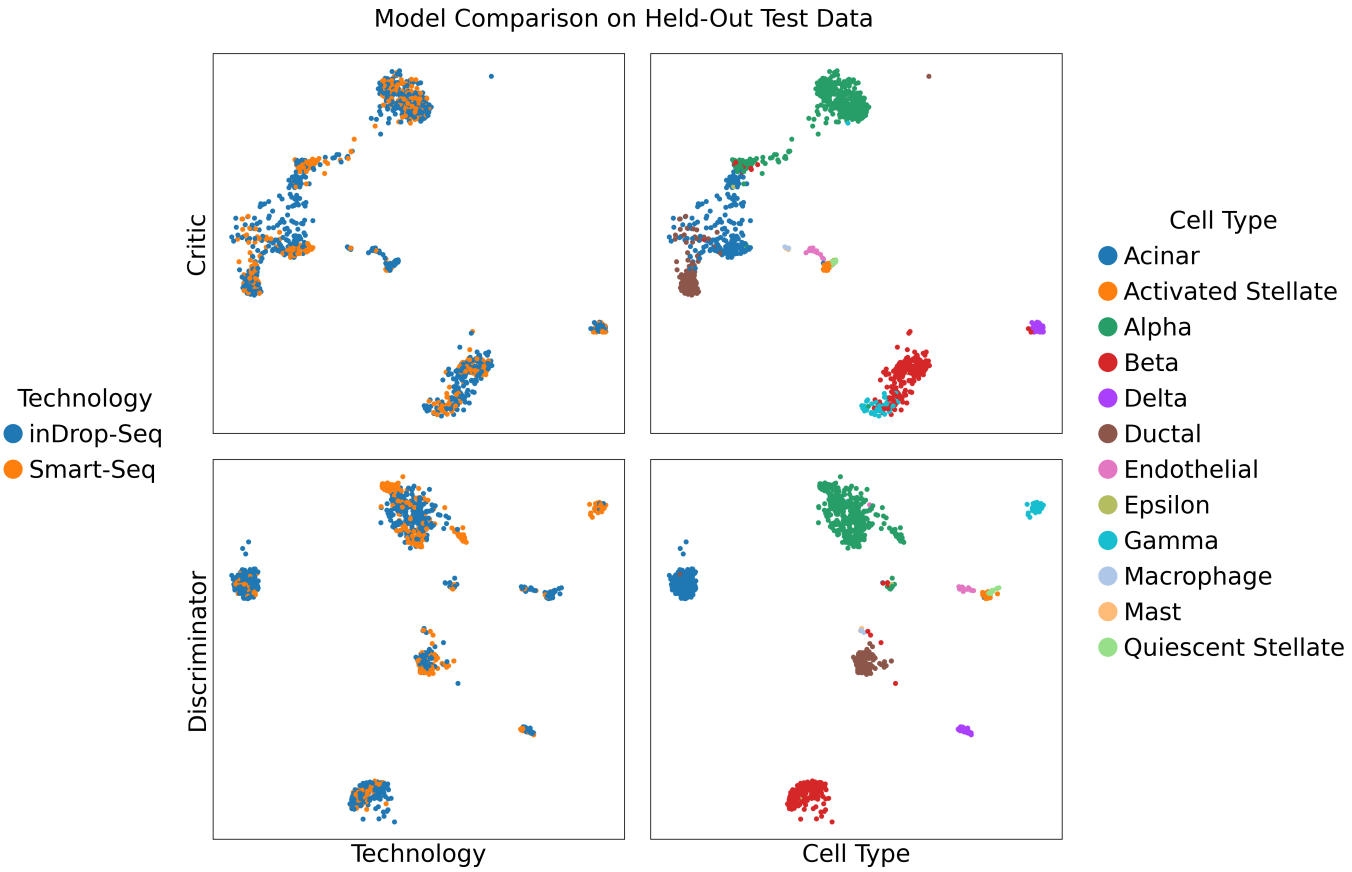

Fig. 3. UMAP visualizations of deconfounded latent spaces (Pancreas dataset). This figure compares the latent spaces generated by the (Top Row) Wasserstein Critic and the (Bottom Row) Discriminator. (Left Column) Cells colored by technology. Both methods show successful integration, with the two technologies well-mixed (compare to Fig. 2). (Right Column) Cells colored by cell type. Both methods now reveal clear clusters corresponding to biological cell types

Future work can aim to improve upon the limitations of the current formulation. This formulation of the Wasserstein critic only works for binary classification; however, adapting this to multiclass classification would allow for integrating many confounding variables at a time. In addition, the current formulation only works for a single confounder at a time, which is limiting for large multi-variable experiments. A potential extension of this work could be the multimodal integration of multi-omic data. Finally, this experiment was limited to discrete variables, so developing an alternative regularization for continuous variables would allow for a broader range of variables to be controlled for and tested.

References

[1] P.-Y. Tung *et al.*, "Batch effects and the effective design of single-cell gene expression studies," en, *Sci Rep*, vol. 7, p. 39 921, Jan. 2017.

[2] O. Stegle, S. A. Teichmann, and J. C. Marioni, "Computational and analytical challenges in single-cell transcriptomics," *Nature Reviews Genetics*, vol. 16, no. 3, pp. 133–145, Mar. 2015.

[3] J. T. Leek *et al.*, "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nature Reviews Genetics*, vol. 11, no. 10, pp. 733–739, Oct. 2010.

[4] A. B. Dincer, J. D. Janizek, and S.-I. Lee, "Adversarial deconfounding autoencoder for learning robust gene expression embeddings," *Bioinformatics*, vol. 36, no. Supplement_2, pp. i573–i582, Dec. 2020, ISSN: 1367-4803. DOI: 10 . 1093 / bioinformatics / btaa796. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/Supplement_2/i573/50693499/btaa796.pdf.

[5] C. He *et al.*, "Partially characterized topology guides reliable anchor-free scRNA-integration," *Communications Biology*, vol. 8, no. 1, p. 561, Apr. 2025.

[6] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17, Sydney, NSW, Australia: JMLR.org, 2017, pp. 214–223.

[7] Y. Wang, T. Liu, and H. Zhao, "Respan: A powerful batch correction model for scrna-seq data through residual adversarial networks," *Bioinformatics*, vol. 38, no. 16, pp. 3942–3949, Jun. 2022, ISSN: 1367-4803.

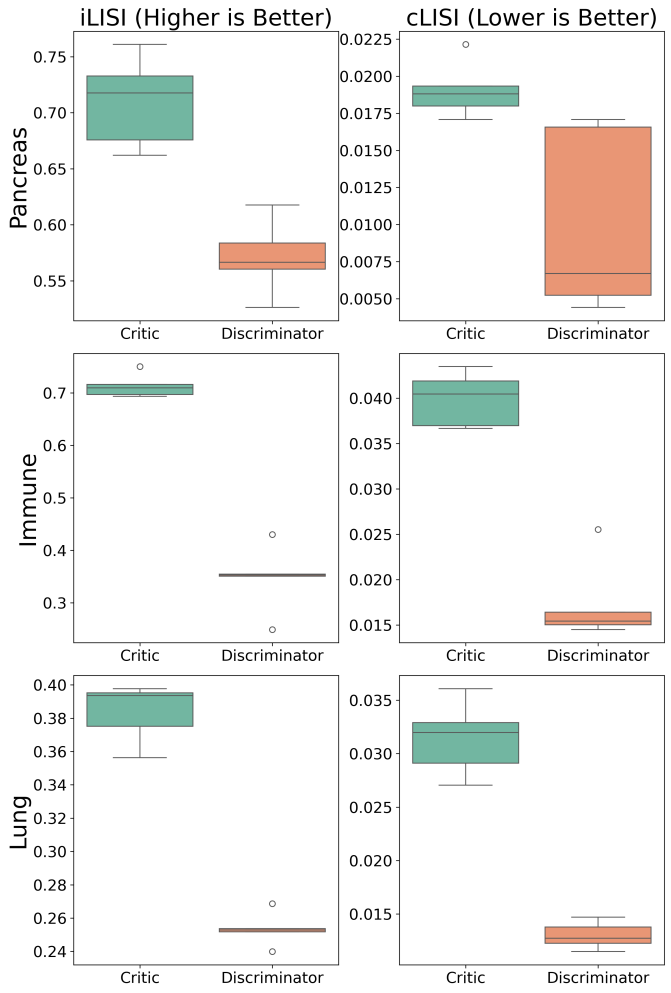

Fig. 4. Quantitative comparison of integration (iLISI) and biological preservation (cLISI). This figure summarizes the performance metrics from the 5-fold cross-validation across all three datasets. (Left) Integration LISI (iLISI), where a score of 1.0 indicates perfect integration. The Wasserstein Critic consistently achieves a higher iLISI score. (Right) Cell-type LISI (cLISI), where a score of 0.0 indicates perfect biological separation. The Discriminator achieves a marginally better (lower) cLISI score in all cases.

DOI: 10.1093/bioinformatics/btac427. eprint: https://academic.oup.com/bioinformatics/article-pdf/38/16/3942/49889943/btac427.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btac427.

[8] M. D. Luecken *et al.*, "Benchmarking atlas-level data integration in single-cell genomics," *Nature Methods*, vol. 19, no. 1, pp. 41–50, Jan. 2022.

[9] I. Korsunsky *et al.*, "Fast, sensitive and accurate integration of single-cell data with harmony," *Nature Methods*, vol. 16, no. 12, pp. 1289–1296, Dec. 2019.

[10] H. T. N. Tran *et al.*, "A benchmark of batch-effect correction methods for single-cell RNA sequencing data," *Genome Biology*, vol. 21, no. 1, p. 12, Jan. 2020.

[11] M. D. Luecken *et al.*, "Benchmarking atlas-level data integration in single-cell genomics," *Nature Methods*, vol. 19, no. 1, pp. 41–50, Jan. 2022.

[12] M. E. Ritchie *et al.*, "Limma powers differential expression analyses for rna-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, e47–e47, Jan. 2015, ISSN: 0305-1048. DOI: 10.1093/nar/gkv007. eprint: https://academic.oup.com/nar/article-pdf/43/7/e47/7207289/gkv007.pdf. [Online]. Available: https://doi.org/10.1093/nar/gkv007.

[13] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, Apr. 2006, ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxj037. eprint: https://academic.oup.com/biostatistics/article-pdf/8/1/118/25435561/kxj037.pdf. [Online]. Available: https://doi.org/10.1093/biostatistics/kxj037.

[14] J. D. Welch *et al.*, "Single-Cell multi-omic integration compares and contrasts features of brain cell identity," *Cell*, vol. 177, no. 7, 1873–1887.e17, Jun. 2019.

[15] R. Lopez *et al.*, "Deep generative modeling for single-cell transcriptomics," *Nature Methods*, vol. 15, no. 12, pp. 1053–1058, Dec. 2018.

[16] L. Haghverdi *et al.*, "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors," *Nature Biotechnology*, vol. 36, no. 5, pp. 421–427, May 2018.

[17] A. B. Rosenberg *et al.*, "Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding," *Science*, vol. 360, no. 6385, pp. 176–182, 2018. DOI: 10.1126/science.aam8999. eprint: https://www.science.org/doi/pdf/10.1126/science.aam8999. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aam8999.

[18] I. Korsunsky *et al.*, "Fast, sensitive and accurate integration of single-cell data with harmony," *Nature Methods*, vol. 16, no. 12, pp. 1289–1296, Dec. 2019.

[19] T. Stuart *et al.*, "Comprehensive integration of Single-Cell data," *Cell*, vol. 177, no. 7, 1888–1902.e21, Jun. 2019.

[20] F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY: Large-scale single-cell gene expression data analysis," *Genome Biology*, vol. 19, no. 1, p. 15, Feb. 2018.

[21] U. Shaham *et al.*, "Removal of batch effects using distribution-matching residual networks," *Bioinformatics*, vol. 33, no. 16, pp. 2539–2546, Apr. 2017, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx196. eprint: https://academic.oup.com/bioinformatics/article-pdf/33/16/2539/49040696/bioinformatics_33_16_2539.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/btx196.

[22] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, ISSN: 0001-0782. DOI: 10.1145/3422622. [Online]. Available: https://doi.org/10.1145/3422622.

[23] I. Gulrajani *et al.*, "Improved training of wasserstein gans," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Long Beach, California, USA: Curran Associates Inc., 2017, pp. 5769–5779, ISBN: 9781510860964.

[24] J. Healy and L. McInnes, "Uniform manifold approximation and projection," *Nature Reviews Methods Primers*, vol. 4, no. 1, p. 82, Nov. 2024.