CRADLE-VAE: Enhancing Single-Cell Gene Perturbation Modeling with Counterfactual Reasoning-based Artifact Disentanglement

Seungheun Baek[1*], Soyon Park[1*], Yan Ting Chok[1], Junhyun Lee[1], Jueon Park[1], Mogan Gim[2†‡], Jaewoo Kang[13†]

[1]Department of Computer Science, Korea University, Seoul, South Korea
[2]Department of Biomedical Engineering, Hankuk University of Foreign Studies, Yongin, South Korea
[3]AIGEN Sciences, Seoul 04778, South Korea
{sheunbaek, soyon_park, yanting1412, ljhyun33, jueon_park, kangj}@korea.ac.kr, gimmogan@hufs.ac.kr

Abstract

Predicting cellular responses to various perturbations is a critical focus in drug discovery and personalized therapeutics, with deep learning models playing a significant role in this endeavor. Single-cell datasets contain technical artifacts that may hinder the predictability of such models, which poses quality control issues highly regarded in this area. To address this, we propose CRADLE-VAE, a causal generative framework tailored for single-cell gene perturbation modeling, enhanced with counterfactual reasoning-based artifact disentanglement. Throughout training, CRADLE-VAE models the underlying latent distribution of technical artifacts and perturbation effects present in single-cell datasets. It employs counterfactual reasoning to effectively disentangle such artifacts by modulating the latent basal spaces and learns robust features for generating cellular response data with improved quality. Experimental results demonstrate that this approach improves not only treatment effect estimation performance but also generative quality as well.

Code — https://github.com/dmis-lab/CRADLE-VAE

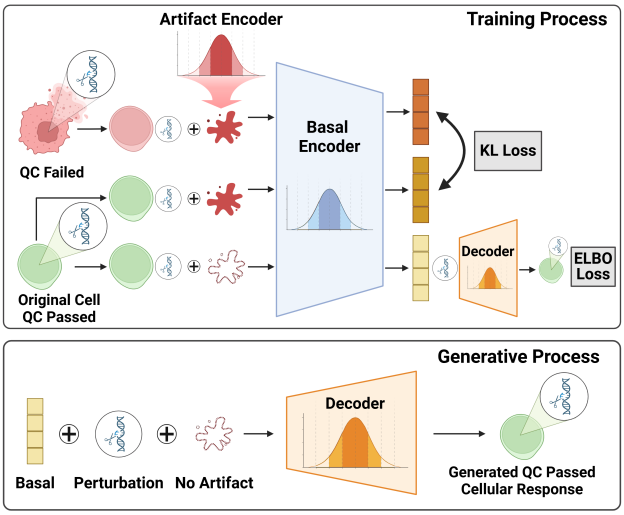

Figure 1: Training and generative process of CRADLE-VAE.

Introduction

Understanding cellular responses to gene perturbations is crucial for identifying potential therapeutic targets. Single-cell technologies such as Perturb-seq (Dixit et al. 2016) have facilitated application of machine learning methodologies in addressing this task due to their high-resolution and high-throughput production of single-cell RNA sequencing (scRNA-seq) data.

Previous works have proposed various computational methods for effectively modeling single-cell gene perturbation outcomes (i.e., treatment effects), mostly involving prediction of scRNA-seq gene expression profiles. One line of work features explicitly modeling the gene-gene relationships, incorporating prior knowledge graphs or networks inferred from the transcriptional data (Roohani, Huang, and Leskovec 2024; Cui et al. 2024). Another centralizes around employing variational autoencoders (VAE) which learn causal representations of single cells through modeling the disentanglement of its perturbation effects (Lopez et al. 2023). SAMS-VAE models the addition of two disentangled factors which are perturbation-independent cell representation (i.e., basal state) and sparse latent effects of gene perturbations (i.e., intervention) (Bereket and Karaletsos 2024).

Despite the endeavor in accurately predicting cellular responses, the quality of training data used in previous works or data generated by their proposed models is not adequately evaluated. scRNA-seq datasets suffer from quality issues which are attributed to the limitations of existing sequencing protocols related to measurement of cells being stressed, broken, or killed. Some data might also correspond to empty droplets or droplets with multiple cells (i.e., doublets) (Ilicic et al. 2016). Conventional quality control (QC) guidelines state that these data are deemed *under-qualified* and the distortions that arise from the limitations of scRNA-seq protocols are said to be *technical artifacts* (Hong et al. 2022).

*These authors contributed equally.
†Corresponding Authors
‡Completed during the author's postdoc at Korea University

A straightforward way to tackle data quality issues caused by the technical artifacts would be resorting to filtering scRNA-seq data based on QC criteria. This method involves excluding QC failed data that may confound downstream analyses and interpretation (10x Genomics 2022). In fact, both the quantity and the quality of scRNA-seq data, from which the model learns the data distribution, strongly influences that of the model performance (Chen et al. 2023). This implies a trade-off between the strictness of gene expression data quality control and the abundance of training data required for effective generalization (Heumos et al. 2023).

Inspired by recent efforts in disentangling the latent gene perturbation effects from the given scRNA-seq data via the VAE framework (Lopez et al. 2023), we propose a similar approach for handling its inherent artifacts as well. Instead of removing the QC failed data samples, we can implement a module that disentangles the inherent technical artifacts from those samples, which ultimately leads to better generative quality while preserving the limited number of scRNA-seq gene expression profiles in the training dataset. This deeply relates to counterfactual reasoning, as our proposed approach not only answers the question *what will the generative outcome be if given this gene perturbation instead?* but also **under this specific gene perturbation, what would the generative outcome have been if technical artifacts had been absent?**

In this work, we propose CRADLE-VAE, a novel VAE framework designed to learn causal representations of scRNA-seq data by utilizing **C**ounterfactual **R**easoning-based **A**rtifact **D**isentang**LE**ment. CRADLE-VAE aims to address quality issues of both training and generated data by disentangling technical artifacts from the natural, perturbation-independent variation in cells through counterfactual reasoning. Specifically, given a QC passed scRNA-seq gene expression profile (i.e., artifact-free) as input, CRADLE-VAE uses an auxiliary loss objective that guides the encoded counterfactual basal state (i.e., artifact-present) towards its reference counterfactual basal state. The latter is constructed as an aggregation of QC failed scRNA-seq data samples under the same gene perturbation treatment.

Our experiments demonstrate that compared with its baselines and ablations, CRADLE-VAE generates gene expression profiles deemed as cellular response predictions that not only showcase superior correlation but also generative quality measured by QC pass rate. To the best of our knowledge, it is the first attempt to model the presence of technical artifacts in scRNA-seq datasets for perturbation response prediction and exploit them leveraging counterfactual reasoning to improve generative quality. The main contributions of this work are summarized as follows:

- We propose CRADLE-VAE, a novel VAE-based cellular response prediction model that addresses quality issues in the realm of scRNA-seq data.
- We introduce an auxiliary loss objective that guides CRADLE-VAE's disentanglement of artifacts during the training process.
- Experimental results show that CRADLE-VAE robustly predicts cellular responses by generating gene expression profiles with higher quality compared to previous methods especially when given unseen perturbations as input.
- Qualitative analysis highlights how our proposed approach contributes to enhancing CRADLE-VAE's disentanglement ability improving its generative quality.

Related Works

Disentanglement in Single-cell Perturbation Response Prediction Recent advancements in single-cell RNA sequencing technologies have significantly enhanced our understanding of cellular responses to chemical and genetic perturbations (Srivatsan et al. 2020; Norman et al. 2019). Due to the complexity of studying the phenotypic effects of cellular perturbations and their underlying factors, previous works have focused on leveraging causal learning which aims to understand the mechanisms by which variables influence each other and predicting the outcome of interventions (Spirtes 2010). CPA utilizes a disentanglement strategy based on adversarial approach (Lotfollahi et al. 2023). Moreover, with VAEs being the primary generative models, studies have focused on disentangling the latent variables that constitute the true distribution of scRNA-seq data. Both sVAE+ (Lopez et al. 2023) and SAMS-VAE (Bereket and Karaletsos 2024) utilize sparse mechanism shifts to disentangle gene perturbations.

Counterfactual Reasoning in Single-cell Perturbation Response Prediction Another line of previous work focuses on employing counterfactual reasoning in predicting the outcomes of single-cell gene perturbations. Counterfactual reasoning helps generative models such as VAEs expand their understanding in causal relationships between different factors such as gene-gene interactions. GraphVCI adopted this concept in enhancing the individuality of cellular responses and dynamically modulating the graph regulatory network structure based on different gene perturbations (Wu et al. 2022). Similarly, CODEX incorporates the counterfactual reasoning approach in predicting the genetically perturbed scRNA-seq data given the unperturbed data (i.e., control expression profile) along with dosage information and specific interventions as input.

None of the previous models have explicitly considered data quality issues caused by scRNA-seq protocols despite being emphasized in biology domain. Our study addresses this by incorporating counterfactual reasoning related to the presence of latent technical artifacts in scRNA-seq data so that the generative model effectively disentangles them during its training process.

Methods

scRNA-seq Dataset

We define a N-sized scRNA-seq dataset $(x_i, p_i, a_i)_{i=1}^{N}$ where each data instance includes a gene expression vector $x_i \in \mathbb{R}^{D_x}$, a gene perturbation vector $p_i \in \{0,1\}^T$ and an artifact presence label $a_i \in \{0,1\}$ where D_x is the total number of genes used in this task, and T is the number of perturbation types. Each bit in p_i specifies whether its corresponding gene was perturbed prior to obtaining x_i. Also,

a_i indicates the presence of technical artifacts in x_i. In our task's context, x_i is the cellular response when given treatment p_i. If x_i passes a predefined quality control criteria, then $a_i = 0$; otherwise, $a_i = 1$.

Quality Control Criteria

We elaborate the process of labeling each expression vector with a_i based on our established quality control (QC) criteria. Having adopted the filtering guidelines provided by Scanpy and 10X Genomics, we established the following six QC sub-criteria: UMI counts, number of features, percent of mitochondrial (mt) reads, percent of hemoglobin reads (hb), percent of ribosomal (rb) reads and doublet detection (Wolf, Angerer, and Theis 2018; 10x Genomics 2022). The first five sub-criteria are determined using data-driven thresholds calculated as scaled median absolute deviation (MAD) (Ocasio et al. 2019; You et al. 2021) while the last criterion is a binary label identified by Scrublet (Wolock, Lopez, and Klein 2019). We used three to five times of the MAD (3σ, 4σ, 5σ) since threshold selection can vary across studies (Ocasio et al. 2019; You et al. 2021), where 3σ represents the strictest QC cut-off, followed by 4σ and 5σ.

Model Architecture

Encoder Module The overall architecture of CRADLE-VAE is shown in Figure 2. During training, the encoder part of CRADLE-VAE takes data instance (x_i, p_i, a_i) as input and encodes it into three different latent representations which are latent basal state embedding $\mathbf{z}_i^b \in \mathbb{R}^{D_z}$, latent perturbation effect embedding $\mathbf{z}_i^p \in \mathbb{R}^{D_z}$ and latent artifact embedding $\mathbf{z}_i^a \in \mathbb{R}^{D_z}$ where D_z is the dimension size of latent subspaces. The objective of this module is to disentangle these three latent variables and learn their individual contributions to the observed true data distribution.

Algorithm 1 shows CRADLE-VAE's encoding process which inherits the formulation basis from Bereket and Karaletsos's work. The latent perturbation effect embedding $\mathbf{z}_i^p$ is an additive composition of global gene-wise perturbation effects, $\mathbf{e}_t$, induced by global sparse latent offsets, $\mathbf{m}_t$, which are sampled from parameterized prior Normal distribution and Bernoulli distribution, respectively (Algorithm 1.2, 3, 7). Similarly, the latent artifact embedding $\mathbf{z}_i^a$ is a multiplication of a_i and $\mathbf{u}$, which is sampled from its own parameterized prior distribution (Algorithm 1.5, 8).

$\mathbf{z}_i^b$ is sampled from a Normal distribution that is parameterized by a neural network $\hat{f}_{enc}$ taking x_i, $\mathbf{z}_i^p$ and $\mathbf{z}_i^a$ as input (Algorithm 1.12). $\mathbf{1}_t$ is the one-hot encoding of the tth gene perturbation treatment while both $\hat{f}_{emb}$ and $\hat{f}_{enc}$ are trainable neural networks.

Decoder Module During training, the decoder part of CRADLE-VAE takes the latent embeddings $(\mathbf{z}_i^b, \mathbf{z}_i^p, \mathbf{z}_i^a)$ as input and samples $\tilde{x}_i$ from a parameterized Gamma-Poisson distribution. Algorithm 2 shows CRADLE-VAE's decoding process where $\hat{f}_{dec}$ is a learnable neural network with final softmax layer that outputs the expected frequency for each gene used for parameterizing the Gamma-Poisson distribution. l_i and θ_d denote the total number of read counts for the ith cell and learnable inverse dispersion used universally across all cells respectively.

Variational Inference Considering the intractability of the data marginal probability $p(X|P, A)$, we define the correlated variational distribution $q(Z|X, P, A)$ by approximating the posterior distribution of latent variables:

$$q(Z^b, M, E, U|X, P, A) = \left(\prod_{t=1}^{T} q(\mathbf{e}_t|\mathbf{m}_t; \phi)q(\mathbf{m}_t; \phi)\right) \times q(\mathbf{u}; \phi)\left(\prod_{i=1}^{N} q(\mathbf{z}_i^b|x_i, p_i, a_i, M, E, U; \phi)\right) \tag{1}$$

for latent basal state embeddings $Z^b \in \mathbb{R}^{N\times D_z}$, global latent perturbation masks $M \in \{0, 1\}^{T\times D_z}$, global latent perturbation embeddings $E \in \mathbb{R}^{T\times D_z}$, global latent artifact embeddings $U \in \mathbb{R}^{1\times D_z}$, gene expression matrix $X \in \mathbb{R}^{N\times D_x}$, gene perturbation matrix $P \in \{0, 1\}^{N\times T}$, and artifact presence labels $A \in \{0, 1\}^N$.

We employ stochastic variational inference (Hoffman et al. 2013) to approximate the posterior distribution $\log p(X|P, A)$. The learnable parameters (θ,ϕ) of CRADLE-VAE are optimized by maximizing the evidence lower bound (ELBO) which is mathematically expressed as below:

$$\mathcal{J}_1(\theta, \phi) = \mathbb{E}_{Z^b,M,E,U\sim q(\cdot|X,P,A;\phi)}\left[\log \frac{p(X, Z^b, M, E, U|P, A; \theta)}{q(Z^b, M, E, U|X, P, A; \phi)}\right] \tag{2}$$

Artifact Disentanglement by Counterfactual Reasoning We propose to exploit the counterfactual outcome of the same gene perturbation treatment as means to reinforce disentanglement of latent variables related to quality degradation caused by technical artifacts. We add the following modifications to CRADLE-VAE's encoding process x_i is a QC passed gene expression profile (i.e., $a_i = 0$).

First, CRADLE-VAE additionally builds a *counterfactual latent artifact embedding* $\mathbf{z}_{i,c}^a = (1 - a_i)\mathbf{u}$ which is opposite to $\mathbf{z}_i^a = a_i\mathbf{u}$ being zero-scaled (Algorithm 1.9). It is then used for sampling the *counterfactual latent basal state embedding* $\mathbf{z}_{i,c}^b$ from a Normal distribution parameterized by $\hat{f}_{enc}$ (Algorithm 1.10). Meanwhile, for each QC passed gene expression profile x_i, we first sample its *counterfactuals* from our dataset that share the same gene perturbation treatment but are QC failed. We then compute their median $\bar{x}_{i,c}$ to feed it along with $\mathbf{z}_i^p$ and $\mathbf{z}_{i,c}^a$ into the neural network $\hat{f}_{enc}$, from where we sample the *reference counterfactual latent basal state embedding* $\bar{\mathbf{z}}_{i,c}^b$ (Algorithm 1.11).

We imposed an auxiliary loss objective that guides $\mathbf{z}_{i,c}^b$ to be aligned with $\bar{\mathbf{z}}_{i,c}^b$. This is done by minimizing the Kullback–Leibler (KL) divergence between the latent basal state embeddings which is mathematically expressed as follows:

$$\mathcal{J}_2(\phi) = -\text{KL}\left[q(Z_c^b|X, P, A; \phi)\|q(\bar{Z}_c^b|\bar{X}, P, A; \phi)\right] \tag{3}$$

We expect the loss objective to provide two benefits for CRADLE-VAE. First, the computed gradients that are back-propagated through $\hat{f}_{enc}$ to $\mathcal{N}(\hat{\mu}, \hat{\sigma})$ exhibit additional supervision to the disentanglement of artifact-related latent

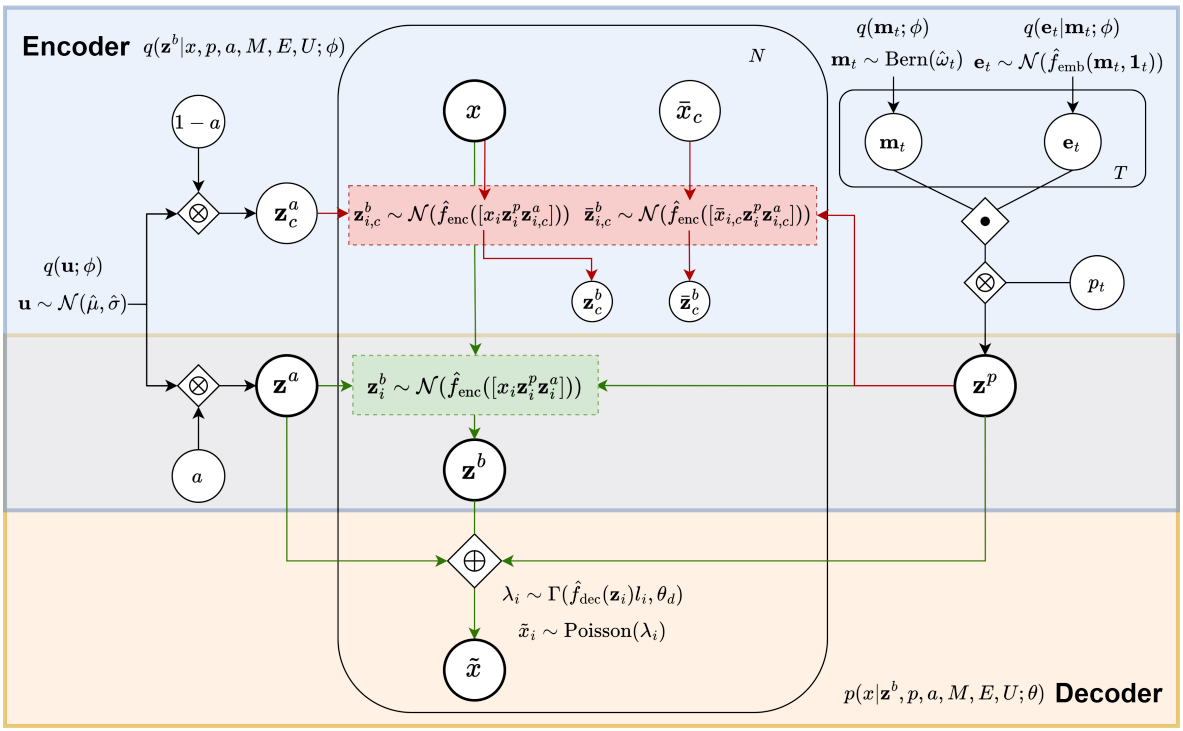

Figure 2: Graphical model of CRADLE-VAE. • represents Hadamard product operation; $\otimes$ represents matrix multiplication operation; $\oplus$ represents vector concatenation.

Algorithm 1: CRADLE-VAE Encoding Process

Require: $X \in \mathbb{R}^{N \times D_x}$, $\bar{X}_c \in \mathbb{R}^{N \times D_x}$, $P \in \{0,1\}^{N \times T}$, $A \in \{0,1\}^N$

for t from 1 to T **do**
  $\mathbf{m}_t \sim \text{Bernoulli}(\hat{\omega}_t)$
  $\mathbf{e}_t \sim \mathcal{N}(\hat{f}_{\text{emb}}(\mathbf{m}_t, \mathbf{1}_t))$
end for
$\mathbf{u} \sim \mathcal{N}(\hat{\mu}, \hat{\sigma})$
for i from 1 to N **do**
  $\mathbf{z}_i^p = \sum_{t=1}^T p_{i,t}(\mathbf{e}_t \odot \mathbf{m}_t)$
  $\mathbf{z}_i^a = a_i \mathbf{u}$
  $\mathbf{z}_{i,c}^a = (1 - a_i)\mathbf{u}$
  $\mathbf{z}_{i,c}^b \sim \mathcal{N}(\hat{f}_{\text{enc}}([x_i \oplus \mathbf{z}_i^p \oplus \mathbf{z}_{i,c}^a]))$
  $\bar{\mathbf{z}}_{i,c}^b \sim \mathcal{N}(\hat{f}_{\text{enc}}([\bar{x}_{i,c} \oplus \mathbf{z}_i^p \oplus \mathbf{z}_{i,c}^a]))$
  $\mathbf{z}_i^b \sim \mathcal{N}(\hat{f}_{\text{enc}}([x_i \oplus \mathbf{z}_i^p \oplus \mathbf{z}_i^a]))$
end for

Algorithm 2: CRADLE-VAE Decoding Process

Require: $\mathbf{z}^b \in \mathbb{R}^{N \times D_z}$, $\mathbf{z}^p \in \mathbb{R}^{N \times D_z}$, $\mathbf{z}^a \in \mathbb{R}^{N \times D_z}$

for i from 1 to N **do**
  $\mathbf{z}_i = [\mathbf{z}_i^b \oplus \mathbf{z}_i^p \oplus \mathbf{z}_i^a]$
  $\lambda_i \sim \Gamma(\hat{f}_{\text{dec}}(\mathbf{z}_i) l_i, \theta_d)$
  $\tilde{x}_i \sim \text{Poisson}(\lambda_i)$
end for

Algorithm 3: CRADLE-VAE Generative Process

Require: $P \in \{0,1\}^{N \times T}$

for t from 1 to T **do**
  $\mathbf{m}_t \sim \text{Bernoulli}(\hat{\omega}_t)$
  $\mathbf{e}_t \sim \mathcal{N}(\hat{f}_{\text{emb}}(\mathbf{m}_t, \mathbf{1}_t))$
end for
$\mathbf{u} \sim \mathcal{N}(\hat{\mu}, \hat{\sigma})$
$\mathbf{z}_i^a = 0\mathbf{u}$
for i from 1 to N **do**
  $\mathbf{z}_i^b \sim \mathcal{N}(0, I)$
  $\mathbf{z}_i^p = \sum_{t=1}^T \mathbf{p}_{i,t}(\mathbf{e}_t \odot \mathbf{m}_t)$
  $\mathbf{z}_i = [\mathbf{z}_i^b \oplus \mathbf{z}_i^p \oplus \mathbf{z}_i^a]$
  $\lambda_i \sim \Gamma(\hat{f}_{\text{dec}}(\mathbf{z}_i) l_i, \theta_d)$
  $\tilde{x}_i \sim \text{Poisson}(\lambda_i)$
end for

variables, facilitating a clearer distinction between QC passed and QC failed cases. Second, the latent basal state embeddings that are encoded by $\hat{f}_{enc}$ help guide the $\hat{f}_{dec}$ to generate the data samples that not only correlate with the true cellular responses but are also more likely to pass the QC criteria. We will explore these benefits later through our quantitative experiments and qualitative analysis.

The overall learning objective that optimizes the trainable parameters θ, ϕ is then defined as follows:

$$\mathcal{J}(\theta, \phi) = \mathcal{J}_1(\theta, \phi) + \alpha \mathcal{J}_2(\phi) \tag{4}$$

where α is the hyperparameter for controlling the alignment

intensity of the auxiliary loss objective.

Generative Process After training, CRADLE-VAE generates its predicted cellular responses by sampling the latent basal state embedding $\mathbf{z}_i^b$ from a normal distribution and combining it with $\mathbf{z}_i^p$ and $\mathbf{z}_i^a$ sampled from the encoder module's parameterized distributions. Finally, $[\mathbf{z}_i^b \oplus \mathbf{z}_i^p \oplus \mathbf{z}_i^a]$ is fed to $\hat{f}_{dec}$, which generates the read counts for each gene (Algorithm 3.11,12). Note that the global latent artifact embedding $\mathbf{u}$ is multiplied by $a_i = 0$ since CRADLE-VAE is used to generate artifact-free gene expression data which is expected to pass the QC criteria (Algorithm 3.6).

Formally, we define the joint probability distribution over the observed and latent variables as:

$$p(X, Z^b, M, E, U|P, A; \theta) = \left(\prod_{t=1}^{T} p(\mathbf{m}_t)p(\mathbf{e}_t)\right) p(\mathbf{u}) \times \left(\prod_{i=1}^{N} p(\mathbf{z}_i^b)p(x_i|\mathbf{z}_i^b, p_i, a_i, M, E, U; \theta)\right) \quad (5)$$

Experiments

Experiment Settings

We evaluated CRADLE-VAE on four Perturb-seq datasets, i.e. Norman dataset (Norman et al. 2019), Dixit dataset (Dixit et al. 2016), Replogle dataset (Replogle et al. 2022), and Adamson dataset (Adamson et al. 2016). We adopted the preprocessing approaches done to Replogle dataset from Lopez et al. and other datasets from Ji et al.. The details of each dataset are shown in Table 1.

Dataset	# of Cells	# of Genes	# of Perts	Perturbation
Norman	111,255	19,018	105 + <u>131</u>	CRISPRa
Dixit	103,420	18,531	10 + <u>45</u>	CRISPR-Cas9
Replogle	118,641	1,187	722	CRISPRi
Adamson	62,623	17,115	90	CRISPRi

Table 1: Summary of Perturb-seq datasets used in our experiments. Notably, Norman and Dixit include multi-gene perturbations which is underlined, while Replogle and Adamson consist of only single-gene perturbations.

We compared CRADLE-VAE against four other causal learning-based VAE models, namely sVAE+ (Lopez et al. 2023), CPA-VAE (Bereket and Karaletsos 2024), SAMS-VAE (Bereket and Karaletsos 2024), and conditional-VAE (Sohn, Lee, and Yan 2015). We additionally considered the variants of CRADLE-VAE trained under different QC threshold settings (3σ,4σ,5σ). Note that we applied the same QC criteria to all data instances partitioned into train, valid and testing purposes.

In our evaluation, we considered the characteristics of data perturbations during the assessment process. For datasets involving multi-gene perturbations, the test set was constructed using combinations not encountered during training, representing approximately 25% of the total possible combinations. Conversely, for datasets involving single perturbations, the evaluation emphasized the models' ability to capture trends in the observed data within the context of single-perturbation scenarios.

To robustly evaluate the models with respect to varying data quality, we trained and evaluated all baseline models with five different random seeds and reported their averaged results. Our main evaluation metric is the Average Treatment Effect Pearson Correlation (ATE-ρ) introduced by Bereket and Karaletsos, that measures the correlation between model-predicted expression values and the experimental data across all genes. We also calculated the R-square score for the estimated average treatment effects as well (ATE-R^2). In addition, we employed the Jaccard similarity between top 50 model-predicted differentially expressed genes and true differentially expressed genes as defined in previous works (Roohani, Huang, and Leskovec 2024).

As our work highlights the importance of addressing quality issues in scRNA-seq data, we formulated an evaluation metric that measures the model's generative quality, denoted as QC Pass Rate (QCPR). QCPR is calculated by dividing the number of generated samples that passed the QC criteria divided by total number of generated samples. Note that the threshold in QC criteria is equally applied for the annotation of Perturb-seq dataset and in the QCPR metric.

Experimental Results

Table 2 shows the quantitative results on the four Perturb-seq datasets. According to the results, CRADLE-VAE overall surpassed all of its baselines in the three evaluation metrics that measure the model's ability to accurately predict cellular responses. Moreover, we achieved the highest QC Pass Rate across all datasets and QC threshold settings, demonstrating its ability to capture the true data distribution of QC passed gene expression profiles due to additional disentanglement of latent artifacts during its training phase. Notably, despite multi-gene perturbation cellular response prediction being more challenging than that of single-gene perturbation, CRADLE-VAE significantly outperforms the second-best model with a large margin, particularly in the Norman and Dixit datasets, both of which contain multi-gene perturbation scRNA-seq data. This highlights CRADLE-VAE's strong generalizability in out-of-distribution (OOD) gene perturbation treatment scenarios.

Ablation Study

To investigate the effects of utilizing causal distribution of artifact disentanglement and our proposed auxiliary loss objective utilizing counterfactual reasoning related to technical artifacts, we conducted experiments on the ablated versions of CRADLE-VAE which are denoted as CRADLE-VAE w/o Causal and CRADLE-VAE w/o CF respectively. The former models the technical artifact as fixed learnable embedding instead of parameterized prior distribution (Algorithm 1.5). The latter removes the KL divergence-based auxiliary loss objective, eliminating the counterfactual reasoning-based approach in aligning the latent basal state embeddings ($\mathcal{J}_2$).

As shown in Table 3, the ablated versions exhibited performance decline, implying the benefits of employing counterfactual reasoning and causal learning. Particularly, we

Dataset		Norman				Dixit			
Model	QC threshold	ATE-ρ	ATE-R^2	Jaccard	QCPR (%)	ATE-ρ	ATE-R^2	Jaccard	QCPR (%)
Conditional VAE	3σ	0.5314± 0.04	0.2766± 0.05	0.2630± 0.02	74.05± 0.28	0.2203± 0.02	0.0434± 0.01	0.0844± 0.01	69.80± 1.48
CPA-VAE		0.5391± 0.08	0.2085± 0.11	0.2408± 0.03	72.53± 0.74	0.3718± 0.05	-0.0250± 0.07	0.1373± 0.01	73.00± 0.44
sVAE+		0.0249± 0.02	-0.0189± 0.01	0.0232± 0.00	75.34± 0.83	0.0259± 0.03	-0.0319± 0.01	0.0310± 0.01	70.77± 0.74
SAMS-VAE		0.4594± 0.03	0.2098± 0.03	0.2362± 0.02	75.18± 0.61	0.0767± 0.06	-0.0213± 0.03	0.0556± 0.02	68.83± 0.75
CRADLE-VAE$_{3\sigma}$		**0.7119± 0.03**	**0.5040± 0.04**	**0.3337± 0.02**	**93.53± 0.64**	**0.6520± 0.02**	**0.3764± 0.03**	**0.4324± 0.04**	**84.83± 1.59**
Conditional VAE	4σ	0.5396± 0.04	0.2855± 0.04	0.2641± 0.02	82.06± 0.36	0.2270± 0.02	0.0448± 0.01	0.0856± 0.01	77.65± 1.31
CPA-VAE		0.5674± 0.08	0.2851± 0.12	0.2442± 0.03	80.16± 0.77	0.3845± 0.05	-0.0054± 0.07	0.1420± 0.01	80.10± 0.43
sVAE+		0.0286± 0.03	-0.0185± 0.01	0.0230± 0.00	82.97± 0.56	0.0220± 0.03	-0.0386± 0.01	0.0313± 0.01	79.26± 0.29
SAMS-VAE		0.4633± 0.03	0.2096± 0.02	0.2376± 0.02	83.20± 0.69	0.0821± 0.06	-0.0220± 0.03	0.0565± 0.02	77.04± 0.78
CRADLE-VAE$_{4\sigma}$		**0.7477± 0.03**	**0.5423± 0.04**	**0.3620± 0.02**	**95.90± 0.34**	**0.6572± 0.03**	**0.3932± 0.04**	**0.4041± 0.04**	**88.18± 0.76**
Conditional VAE	5σ	0.5525± 0.03	0.2990± 0.04	0.2748± 0.02	86.22± 0.40	0.2287± 0.02	0.0459± 0.01	0.0866± 0.01	81.84± 1.18
CPA-VAE		0.5814± 0.08	0.3077± 0.11	0.2543± 0.03	84.30± 0.68	0.3990± 0.04	0.0274± 0.06	0.1461± 0.01	83.36± 0.50
sVAE+		0.0298± 0.03	-0.0187± 0.01	0.0242± 0.01	86.88± 0.49	0.0225± 0.03	-0.0379± 0.01	0.0314± 0.01	83.07± 0.30
SAMS-VAE		0.4732± 0.03	0.2173± 0.02	0.2462± 0.02	87.19± 0.71	0.0885± 0.07	-0.0181± 0.03	0.0566± 0.02	81.27± 0.72
CRADLE-VAE$_{5\sigma}$		**0.7518± 0.03**	**0.5482± 0.04**	**0.3671± 0.02**	**96.62± 0.38**	**0.6258± 0.06**	**0.3239± 0.05**	**0.3493± 0.07**	**91.40± 1.80**
Dataset		**Replogle**				**Adamson**			
Model	QC threshold	ATE-ρ	ATE-R^2	Jaccard	QCPR (%)	ATE-ρ	ATE-R^2	Jaccard	QCPR (%)
Conditional VAE	3σ	0.7022± 0.00	0.4883± 0.01	**0.2688± 0.00**	76.56± 0.26	0.6335± 0.01	0.3954± 0.01	0.3110± 0.01	77.23± 0.44
CPA-VAE		0.5171± 0.01	0.1241± 0.02	0.1438± 0.00	74.83± 0.50	0.5571± 0.02	0.2637± 0.03	0.2123± 0.01	76.64± 0.90
sVAE+		0.5780± 0.01	0.3222± 0.01	0.1565± 0.00	73.89± 0.53	0.5298± 0.02	0.2580± 0.03	0.1778± 0.01	76.27± 0.98
SAMS-VAE		0.6798± 0.03	0.4584± 0.04	0.2404± 0.02	74.96± 0.69	0.3901± 0.01	0.1432± 0.01	0.1846± 0.01	77.34± 0.78
CRADLE-VAE$_{3\sigma}$		**0.7192± 0.01**	**0.5155± 0.01**	0.2667± 0.01	**97.33± 0.04**	**0.7529± 0.01**	**0.5611± 0.02**	**0.3471± 0.01**	**89.92± 0.47**
Conditional VAE	4σ	0.7255± 0.01	0.5233± 0.01	0.2776± 0.00	84.36± 0.24	0.6435± 0.01	0.4059± 0.02	0.3109± 0.01	85.06± 0.52
CPA-VAE		0.5352± 0.01	0.1765± 0.03	0.1494± 0.01	82.92± 0.38	0.5715± 0.02	0.2863± 0.03	0.2103± 0.01	84.80± 0.67
sVAE+		0.6056± 0.01	0.3612± 0.01	0.1661± 0.00	82.15± 0.50	0.5437± 0.02	0.2774± 0.03	0.1773± 0.01	84.66± 0.60
SAMS-VAE		0.7086± 0.03	0.4941± 0.04	0.2516± 0.02	83.01± 0.43	0.3939± 0.01	0.1442± 0.01	0.1808± 0.01	85.36± 0.75
CRADLE-VAE$_{4\sigma}$		**0.7565± 0.01**	**0.5595± 0.01**	**0.2869± 0.01**	**98.10± 0.18**	**0.7636± 0.01**	**0.5770± 0.01**	**0.3367± 0.01**	**93.66± 0.56**
Conditional VAE	5σ	0.7296± 0.01	0.5282± 0.01	0.2793± 0.00	88.45± 0.27	0.6484± 0.01	0.4110± 0.02	0.3110± 0.01	88.75± 0.45
CPA-VAE		0.5380± 0.02	0.1999± 0.03	0.1501± 0.01	87.20± 0.38	0.5758± 0.02	0.2928± 0.04	0.2102± 0.01	88.34± 0.57
sVAE+		0.6137± 0.01	0.3736± 0.01	0.1694± 0.00	86.60± 0.41	0.5488± 0.02	0.2843± 0.03	0.1776± 0.01	88.65± 0.55
SAMS-VAE		0.7167± 0.03	0.4998± 0.03	0.2558± 0.02	87.26± 0.33	0.3952± 0.01	0.1442± 0.01	0.1792± 0.01	89.05± 0.49
CRADLE-VAE$_{5\sigma}$		**0.7638± 0.01**	**0.5719± 0.01**	**0.2931± 0.01**	**98.41± 0.13**	**0.7609± 0.01**	**0.5723± 0.01**	**0.3153± 0.00**	**94.34± 0.42**

Table 2: Quantitative evaluation on Norman, Dixit, Replogle and Adamson dataset across $3\sigma, 4\sigma, 5\sigma$ quality control (QC) thresholds. Note that the QC threshold column refers to the cut-off point – defined as delta-MAD threshold – of the generated data to be included in the evaluation phase. Best results are in bold-faced while second-best ones are underlined.

Model	QC thr.	ATE-ρ	ATE-R^2	Jaccard	QCPR (%)
CRADLE-VAE$_{3\sigma}$	3σ	**0.7119± 0.03**	**0.5040± 0.04**	**0.3337± 0.02**	**93.53± 0.64**
CRADLE-VAE$_{3\sigma}$ w/o CF		0.6505± 0.02	0.4210± 0.02	0.3046± 0.01	91.46± 0.73
CRADLE-VAE$_{3\sigma}$ w/o Causal		0.7018± 0.02	0.4844± 0.02	0.2938± 0.01	92.63± 0.71
CRADLE-VAE$_{4\sigma}$	4σ	**0.7477± 0.03**	**0.5423± 0.04**	**0.3620± 0.02**	**95.90± 0.34**
CRADLE-VAE$_{4\sigma}$ w/o CF		0.7111± 0.03	0.4927± 0.04	0.3240± 0.01	94.24± 0.43
CRADLE-VAE$_{4\sigma}$ w/o Causal		0.7058± 0.03	0.4790± 0.05	0.2946± 0.01	87.90± 5.34
CRADLE-VAE$_{5\sigma}$	5σ	**0.7518± 0.03**	**0.5482± 0.04**	**0.3671± 0.02**	**96.62± 0.38**
CRADLE-VAE$_{5\sigma}$ w/o CF		0.7395± 0.02	0.5315± 0.03	0.3540± 0.01	95.71± 0.49
CRADLE-VAE$_{5\sigma}$ w/o Causal		0.6875± 0.03	0.4402± 0.05	0.3008± 0.03	92.85± 4.06

Table 3: Experimental results on ablated versions of CRADLE-VAE$_\sigma$. Best results are in bold-faced while second-best ones are underlined.

find that modeling the technical artifact as a learnable embedding (CRADLE-VAE w/o Causal) results in a sharper decline, especially at the 5σ QC threshold. While setting a higher QC threshold leads to imbalance between the number of QC passed and failed samples, we speculate that distribution-based artifact modeling is more resilient to such issues compared to its embedding-based version. The effect of removing the counterfactual reasoning (CRADLE-VAE w/o CF) is more profound at the 3σ threshold. This outcome aligns with our assumption that the KL loss objective between the counterfactual latent basal state embeddings aids in the learning of artifact features, particularly when generalization is well-established due to the balanced data instances.

Distributional Generative Quality Analysis

To further analyze CRADLE-VAE's generative quality, we visualized the distributions of actual (Replogle) and model-generated gene counts related to the QC criteria, that results from a specific treatment perturbing the POLD3 gene (Replogle et al. 2022). The rationale behind selecting this particular perturbation is as follows: 1) the number of gene expression profiles treated by this perturbation in the dataset is relatively low (85 compared to the average of 164), 2) only

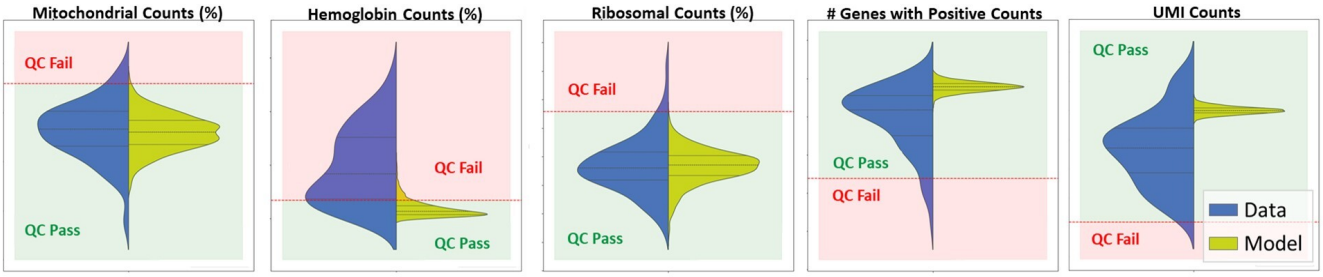

Figure 3: Violin plots showing the data(blue) and model-generated(green) distribution of POLD3-perturbed response for each QC sub-criteria. Red dotted line: predefined QC threshold; green region: QC passed values; red region: QC failed values.

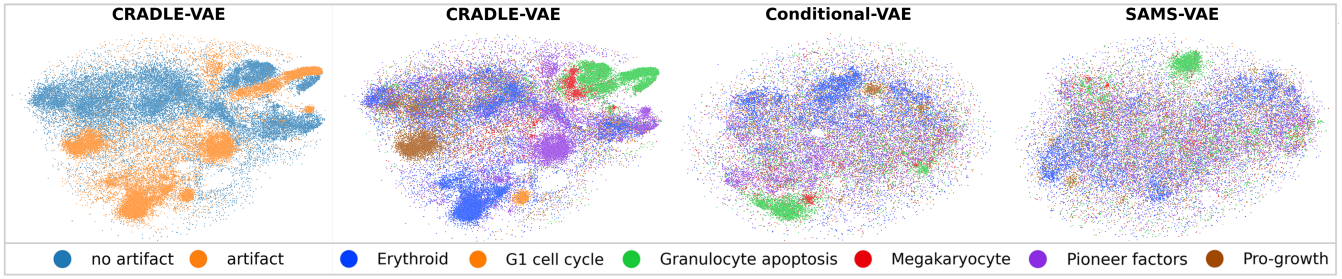

Figure 4: t-SNE plots labelled by the presence of artifacts (left 1) and by perturbation types (right 3) for CRADLE-VAE, conditional-VAE, and SAMS-VAE, respectively.

13% of them passes the QC criteria. This may pose challenges in learning the causal distributions during the training process, especially if the latent effects of technical artifacts are not properly addressed. We expect these challenges to be dealt with the employment of counterfactual reasoning-based artifact disentanglement. Figure 3 shows that CRADLE-VAE exhibits its consistency in robustly generating read counts that satisfy all QC sub-criteria.

We move our focus to a critical sub-criterion responsible for a significant decline in data quality. The distribution of hemoglobin counts in the Replogle dataset predominantly exceed the QC threshold, leading to a high QC failure rate. On the contrary, the distribution generated by CRADLE-VAE is shifted below the threshold, implying a marked enhancement in generative data quality. For both the number of genes with positive counts and UMI count, the violin plots in Figure 3 display a skewed distribution compared to the original data, indicating that our model's generated gene expression profiles yield consistent and higher quality outcomes.

Disentanglement Effect Analysis

We investigated the effects of CRADLE-VAE's disentanglement of two important variables which are perturbation and artifact effects. We utilized t-SNE in visualizing the high-dimensional gene expression profiles generated by CRADLE-VAE, and colored them based on which pathway clusters are relevant to each of their gene perturbations. This aligns with a domain-specific assertion stating that perturbation of genes with similar biological roles are expected to show similar expression patterns. Following the method in (Norman et al. 2019), we grouped them into six pathways for this visualization. As illustrated in Figure 4, CRADLE-VAE appears to form clearer clusters within the same pathway compared to other models, particularly for those related to the pro-growth and megakaryocyte pathways.

Additionally, we examined the disentanglement of artifacts by comparing the same generated data with and without artifacts. In Figure 4, the t-SNE plot within pathways shows distinct clustering based on the presence of artifacts, suggesting that our model successfully disentangles artifact effects. Overall, these suggest that our model can meaningfully separate both latent perturbation and artifact variables, as reflected by the well-defined clusters in the visualizations.

Conclusion

Quality issues in scRNA-seq datasets have been overlooked despite the improvements in predicting cellular responses achieved by previous works. We propose a causal inference-based VAE model CRADLE-VAE which has several advantages. During training, CRADLE-VAE disentangles not only latent perturbation effects but also artifacts that inherently degrade data quality. Additionally, the disentanglement of these artifacts is enhanced by our novel counterfactual reasoning-based approach, using an auxiliary loss objective for aligning counterfactual basal states. As demonstrated in our experiments and analysis, CRADLE-VAE accurately predicts cellular responses with improved generative quality. We expect that CRADLE-VAE addresses the quality issues of both experimentally measured and model-generated single-cell response data upon gene perturbation, eliminating the need of arbitrary quality control standards for scRNA-seq data analysis.

Acknowledgements

This research was supported by the National Research Foundation of Korea [NRF2023R1A2C3004176, RS-2023-00262002], the Ministry of Health & Welfare, Republic of Korea [HR20C0021(3)], ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [IITP-2024- 20200-01819].

This work was supported by Hankuk University of Foreign Studies Research Fund (of 2024).

Figure 2 was created with BioRender.com.

References

10x Genomics. 2022. Common Considerations for Quality Control Filters for Single Cell RNA-seq Data. https://www.10xgenomics.com/analysis-guides/common-considerations-for-quality-control-filters-for-single-cell-rna-seq-data. Accessed: 2024-08-13.

Adamson, B.; Norman, T. M.; Jost, M.; Cho, M. Y.; Nuñez, J. K.; Chen, Y.; Villalta, J. E.; Gilbert, L. A.; Horlbeck, M. A.; Hein, M. Y.; et al. 2016. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7): 1867–1882.

Bereket, M.; and Karaletsos, T. 2024. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. *Advances in Neural Information Processing Systems*, 36.

Chen, H.; Tao, R.; Fan, Y.; Wang, Y.; Wang, J.; Schiele, B.; Xie, X.; Raj, B.; and Savvides, M. 2023. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*.

Cui, H.; Wang, C.; Maan, H.; Pang, K.; Luo, F.; Duan, N.; and Wang, B. 2024. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 1–11.

Dixit, A.; Parnas, O.; Li, B.; Chen, J.; Fulco, C. P.; Jerby-Arnon, L.; Marjanovic, N. D.; Dionne, D.; Burks, T.; Raychowdhury, R.; et al. 2016. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *cell*, 167(7): 1853–1866.

Heumos, L.; Schaar, A. C.; Lance, C.; Litinetskaya, A.; Drost, F.; Zappia, L.; Lücken, M. D.; Strobl, D. C.; Henao, J.; Curion, F.; et al. 2023. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8): 550–572.

Hoffman, M. D.; Blei, D. M.; Wang, C.; and Paisley, J. 2013. Stochastic variational inference. *Journal of Machine Learning Research*.

Hong, R.; Koga, Y.; Bandyadka, S.; Leshchyk, A.; Wang, Y.; Akavoor, V.; Cao, X.; Sarfraz, I.; Wang, Z.; Alabdullatif, S.; et al. 2022. Comprehensive generation, visualization, and reporting of quality control metrics for single-cell RNA sequencing data. *Nature communications*, 13(1): 1688.

Ilicic, T.; Kim, J. K.; Kolodziejczyk, A. A.; Bagger, F. O.; McCarthy, D. J.; Marioni, J. C.; and Teichmann, S. A. 2016. Classification of low quality cells from single-cell RNA-seq data. *Genome biology*, 17: 1–15.

Ji, Y.; Lotfollahi, M.; Wolf, F. A.; and Theis, F. J. 2021. Machine learning for perturbational single-cell omics. *Cell Systems*, 12(6): 522–537.

Lopez, R.; Tagasovska, N.; Ra, S.; Cho, K.; Pritchard, J.; and Regev, A. 2023. Learning causal representations of single cells via sparse mechanism shift modeling. In *Conference on Causal Learning and Reasoning*, 662–691. PMLR.

Lotfollahi, M.; Klimovskaia Susmelj, A.; De Donno, C.; Hetzel, L.; Ji, Y.; Ibarra, I. L.; Srivatsan, S. R.; Naghipourfar, M.; Daza, R. M.; Martin, B.; et al. 2023. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6): e11517.

Norman, T. M.; Horlbeck, M. A.; Replogle, J. M.; Ge, A. Y.; Xu, A.; Jost, M.; Gilbert, L. A.; and Weissman, J. S. 2019. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455): 786–793.

Ocasio, J. K.; Babcock, B.; Malawsky, D.; Weir, S. J.; Loo, L.; Simon, J. M.; Zylka, M. J.; Hwang, D.; Dismuke, T.; Sokolsky, M.; et al. 2019. scRNA-seq in medulloblastoma shows cellular heterogeneity and lineage expansion support resistance to SHH inhibitor therapy. *Nature communications*, 10(1): 5829.

Replogle, J. M.; Saunders, R. A.; Pogson, A. N.; Hussmann, J. A.; Lenail, A.; Guna, A.; Mascibroda, L.; Wagner, E. J.; Adelman, K.; Lithwick-Yanai, G.; et al. 2022. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14): 2559–2575.

Roohani, Y.; Huang, K.; and Leskovec, J. 2024. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6): 927–935.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.

Spirtes, P. 2010. Introduction to causal inference. *Journal of Machine Learning Research*, 11(5).

Srivatsan, S. R.; McFaline-Figueroa, J. L.; Ramani, V.; Saunders, L.; Cao, J.; Packer, J.; Pliner, H. A.; Jackson, D. L.; Daza, R. M.; Christiansen, L.; et al. 2020. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473): 45–51.

Wolf, F. A.; Angerer, P.; and Theis, F. J. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology*, 19: 1–5.

Wolock, S. L.; Lopez, R.; and Klein, A. M. 2019. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell systems*, 8(4): 281–291.

Wu, Y.; Barton, R. A.; Wang, Z.; Ioannidis, V. N.; De Donno, C.; Price, L. C.; Voloch, L. F.; and Karypis, G. 2022. Predicting cellular responses with variational causal inference and refined relational information. *arXiv preprint arXiv:2210.00116*.

You, Y.; Tian, L.; Su, S.; Dong, X.; Jabbari, J. S.; Hickey, P. F.; and Ritchie, M. E. 2021. Benchmarking UMI-based single-cell RNA-seq preprocessing workflows. *Genome Biology*, 22(1): 339.