Conditional Causal Representation Learning for Heterogeneous Single-cell RNA Data Integration and Prediction

Jiayi Dong[1,2] , **Jiahao Li**[1,2] , **Fei Wang**[1,2*]

[1]Shanghai Key Lab of Intelligent Information Processing, Handan Street, Shanghai, China
[2]School of Computer Science and Technology, Fudan University, Handan Street, Shanghai, China
{jiayidong21, lijiahao23}@m.fudan.edu.cn, wangfei@fudan.edu.cn

Abstract

Single-cell sequencing technology provides deep insights into gene activity at the individual cell level, facilitating the study of gene regulatory mechanisms. However, observed gene expression is often influenced by confounding factors such as batch effects, perturbations, and spatial position, which obscure the true gene regulatory network that governs the cell's intrinsic state. To address these challenges, we propose scConCRL, a novel conditional causal representation learning framework designed to extract true gene regulatory relationships independent of confounding information. By considering both fine-grained molecular gene variables and coarse-grained latent domain variables, scConCRL not only uncovers the intrinsic biological signals but also models the complex relationships between these variables. This dual function enables the separation of genuine cellular states from domain information, providing valuable insights for downstream analyses and biological discovery. We demonstrate the effectiveness of our model on multi-domain datasets from different platforms and perturbation conditions, showing its ability to accurately disentangle confounding influences and discover novel gene relationships. Extensive comparisons across various scenarios illustrate the superior performance of scConCRL in several tasks compared to existing methods.

1 Introduction

With the rapid advancement of single-cell RNA sequencing (scRNA-seq) technology, vast datasets are being generated across diverse laboratories and experimental conditions [Jovic *et al.*, 2022]. These datasets often arise from different sequencing methods, perturbation conditions, or disease samples, resulting in significant heterogeneity. Efficiently modeling, integrating, and comparing such multi-domain data has become a major challenge [Stuart *et al.*, 2019; Argelaguet *et al.*, 2021; Hao *et al.*, 2021]. Addressing this issue is crucial for capturing cellular heterogeneity and dynamic changes, ultimately uncovering molecular mechanisms driving phenotypic diversity and disease susceptibility.

Deep generative models have shown great potential in analyzing multi-domain single-cell data but struggle to distinguish biological signals within domains from non-biological signals between domains. Current integration methods typically preserve shared information but often introduce spurious correlations due to a limited identifiable foundation, leading to confusion or loss of biological variations [Aliee *et al.*, 2024]. Consequently, generalization ability is restricted. Causal Representation Learning (CRL), an emerging technique, offers a promising solution by identifying latent factors and causal structures with causal explanations from observational data [Schölkopf *et al.*, 2021]. CRL accurately reflects causal relationships in biological signals and enhances model generalization in unseen scenarios.

In single-cell transcriptomics, gene expression levels represent cellular states but are often confounded by batch effects, drug perturbation, and other covariates. Existing integration methods focus on retaining invariant signals while removing potentially confusing signals [Luecken *et al.*, 2022]. While effective for tasks like cell type annotation, this approach oversimplifies the dynamic cellular generation process and yields inaccurate reconstruction results. Methods like inVAE [Aliee *et al.*, 2024] aim to only separate invariant and domain-specific signals to solve the issue but lack semantic understanding of latent variables and ignore dependencies between them, reducing model interpretability.

Many CRL methods assume independence among latent variables [Tejada-Lapuerta *et al.*, 2023]. However, biological systems often exhibit strong dependencies, such as interactions in gene regulatory networks (GRNs). Ignoring these relationships hinders the model's ability to capture true causal mechanisms and limits its applicability in biological contexts. To effectively apply CRL to single-cell scenarios, it is essential to model these complex interactions and elucidate how gene relationships influence cellular state changes.

To address these challenges, we propose scConCRL, an interpretable conditional causal representation learning framework for integrating and predicting multi-domain scRNA-seq data. Our framework incorporates Structural Equation Modeling (SEM) to explicitly capture relationships between confounding domain variables and intrinsic molecular variables.

*Corresponding author. Our code and supplementary information are available at https://github.com/fdu-wangfeilab/scConCRL.

Genes are modeled as core molecular variables, while batch effects, perturbation conditions, and other domain information are encoded as latent domain variables. By leveraging a conditional disentanglement strategy, scConCRL performs batch effect correction and noise removal without requiring standard references. Furthermore, it models latent domain-molecular variable relationships, enabling robust perturbation predictions and enhanced analysis of complex biological systems.

Our contributions are as follows:

- We propose scConCRL, a causal representation learning framework that identifies both invariant and domain-specific signals, enabling reference-free integration in molecular space and improving generalization across datasets.
- We introduce a strategy that employs Structural Equation Modeling (SEM) to model domain variables (e.g., batch effects, perturbation) and molecular variables, enabling the framework to capture interactions between coarse-grained and fine-grained variables, thereby enhancing interpretability.
- We conduct comprehensive experimental evaluations comparing scConCRL with existing state-of-the-art methods and demonstrate scConCRL's superior performance in several tasks such as GRN inference, data integration, and perturbation prediction.

2 Related Work

2.1 Single-cell Integration

Single-cell multi-domain integration aims to combine data from various technical platforms or experimental conditions to uncover cell types, states, and functions. Traditional methods, such as Canonical Correlation Analysis (CCA) [Hao *et al.*, 2023] and Mutual Nearest Neighbors (MNN) [Haghverdi *et al.*, 2018], align data from two domains by minimizing discrepancies in their latent representations [Haghverdi *et al.*, 2018; Xu *et al.*, 2022]. These approaches are practical and interpretable but face challenges with multi-domain data. Deep learning-based methods [Lopez *et al.*, 2018; Li *et al.*, 2020b; Xiong *et al.*, 2022; Yu *et al.*, 2023] use unsupervised learning for data compression and align multi-domain distributions. However, this often achieves alignment only in the latent space, lacking interpretability and limiting their broader applicability in real-world scenarios.

For molecular-level integration tasks, some deep learning approaches adopt reference-based strategies and map data from different batches to a specific reference batch, but their performance depends heavily on the quality of the reference [Xiong *et al.*, 2022]. In contrast, reference-free methods, such as Scanorama [Hie *et al.*, 2019] and Beaconet [Xu *et al.*, 2024], eliminate batch effects using global distribution information across datasets. While these methods avoid reference bias, their broad constraints can introduce artifacts that distort true biological variations. To address these issues, we propose a reference-free approach that extracts invariant signals and uses mechanism-driven constraints, enabling efficient integration of multi-domain data in the original molecular space while maintaining data quality.

2.2 Single-cell Condition Prediction

Single-cell condition prediction leverages sequencing data to forecast gene expression changes under various conditions. Variational Autoencoder (VAE)-based methods, such as scGen [Lotfollahi *et al.*, 2019], achieve this by learning latent data distributions and identifying distribution shifts. Extensions like trVAE enhance cross-condition predictions by incorporating conditional variables, enabling the modeling of more complex conditions. scPreGAN [Wei *et al.*, 2022] further improves prediction accuracy by capturing shared conditional signals and simulating the perturbation transfer process through adversarial training. Additionally, scShift [Dong and Kluger, 2023] incorporates advancements in causal representation learning to learn conditional and biological patterns, enabling more precise condition prediction. On this basis, scDisInFact [Zhang *et al.*, 2024b] models two different types of domains separately, enabling combined predictions for multiple batches and conditions. Despite significant advances in single-cell condition prediction, challenges remain, particularly regarding the interpretability of models. In this context, scConCRL integrates latent condition and molecular variables, enabling the exploration of how condition influences molecular variables. This framework enhances interpretability and improves condition prediction accuracy.

2.3 Causal Representation Learning

Causal Representation Learning (CRL) aims to identify latent factors and uncover causal structures from observed data, providing interpretable representations. While VAEs excel at representation learning, they struggle to identify true latent factors in nonlinear contexts. Recent improvements, including the use of auxiliary variables and temporal structures, have enhanced VAEs' ability to capture causal mechanisms [Hyvarinen *et al.*, 2019; Khemakhem *et al.*, 2020; Yang *et al.*, 2021]. In multi-domain settings, CRL extracts causal invariances across domains (e.g., different experimental conditions, environments, or datasets), revealing shared mechanisms [Zhang *et al.*, 2024a]. For example, it has been validated on fMRI datasets to identify common causal factors affecting brain activity across conditions. In single-cell analysis, inVAE separated latent representations into invariant and domain-specific components, aiding the study of causal invariance [Aliee *et al.*, 2024]. However, it did not deeply address the interpretation or relationships of latent variables. scConCRL advances this by explicitly modeling relationships between domain-dependent and molecular latent variables, enhancing interpretability and uncovering causal mechanisms in cellular processes.

3 Method

To establish some notation, let $\mathbf{X} \in \mathbb{R}^{n \times m}$ denote a single-cell gene expression matrix, where n represents the number of cells and m represents the number of genes. The matrix $\mathbf{X}$ mixes both biological signals and domain-specific information, such as technical noise or perturbation signals. We

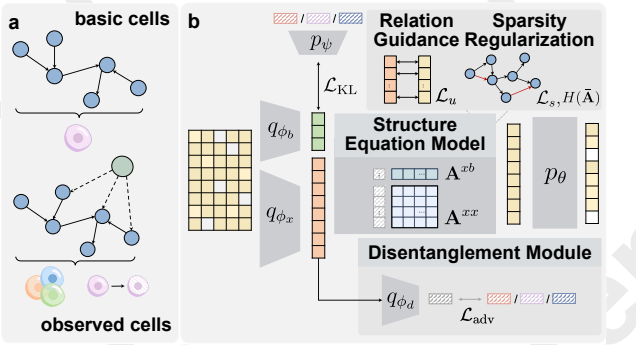

Figure 1: The overall architecture of scConCRL.

denote $\mathbf{B}$ as the variable encoding domain information (e.g., batch or condition). Our goal is twofold: first, to automatically disentangle the k-dimensional latent components, corresponding to batch effects, perturbation signals, and the basal component of cell state signals; and second, to uncover the underlying common GRN, represented by the weighted adjacency matrix $\mathbf{A}^{xx} \in \mathbb{R}^{m \times m}$, which is independent of the batch or condition to which each sample belongs, and the relationship between conditions and genes, captured by the matrix $\mathbf{A}^{xb} \in \mathbb{R}^{m \times k}$.

Confounding factors obscure the discovery of GRNs governing biological processes, leading to discrepancies across GRNs from different domains. scConCRL aims to derive common guidance across domains, inspiring GRN inference and accurately generating cellular information. By incorporating domain-specific factors that affect gene expression, scConCRL accounts for domain information, effectively disentangling true regulatory relationships from domain biases.

scConCRL comprises three main components: a Variational Autoencoder (VAE), a Structural Equation Model (SEM), and a disentanglement module. The encoder maps observed data $\mathbf{X}$ to both molecular and domain-specific latent variables. The SEM and decoder reconstruct the observed data using self-learned relationships. In addition, the disentanglement module encourages the latent representations to be partitioned into distinct variable sets. In this section, we provide a detailed description of each component and the optimization objectives.

3.1 Assumptions on the Generative Process

We aim to model the generative process of multi-domain scRNA-seq by exploring latent causal generative models. The observed data $\mathbf{x}$ is assumed to be generated by underlying latent causal variables, $\mathbf{z}$, with the possibility of complex causal relationships among these variables. In addition, we introduce latent noise variables, $\mathbf{n} = \{\mathbf{n}^b, \mathbf{n}^x\}$, that are exogenous to the system. Specifically, $\mathbf{n}^x$ represents the basal biological factors, and $\mathbf{n}^b$ represents domain-specific factors from batch effects or perturbations.

Furthermore, external domains $\mathbf{b}$, such as technical bias or experimental perturbations, can lead to changes in the distribution of $\mathbf{n}^b$ that influence the data distribution. The introduction of $\mathbf{b}$ allows us to model the effects of these domain-specific factors, offering greater flexibility in adapting to varying conditions. The generative model is defined as follows:

$$
\begin{aligned}
n_i^x &\sim \mathcal{N}(0, \sigma_{n_i}^2), \quad i = 1, ..., m \\
n_j^b &\sim \mathcal{N}(\beta_{j,1}(b), \beta_{j,2}(b)), \quad j = 1, ..., k \\
\mathbf{n} &= \{\mathbf{n}^x, \mathbf{n}^b\} \\
\mathbf{z} &= g(\mathbf{n}) \\
\mathbf{x} &= f(\mathbf{z}, \epsilon)
\end{aligned} \tag{1}
$$

Here, the latent noise variables $\mathbf{n}^x$ and $\mathbf{n}^b$ follow Gaussian distributions, with $\mathbf{n}^b$ being modulated by the observed variable $\mathbf{b}$ through two nonlinear mappings $\beta_{j,1}$ and $\beta_{j,2}$. The choice of a Gaussian distribution is based on the identifiability and expressiveness of latent variables. The latent cell embedding $\mathbf{z}$ is directly generated by the function g, which is defined within a SEM. Subsequently, a nonlinear mapping function f transforms $\mathbf{z}$ into the observed data $\mathbf{x}$, with ϵ representing independent noise. For flexibility, we do not impose a discrete probability distribution on the observed data $\mathbf{x}$.

To estimate the parameters of the scConCRL model, we employ variational inference, which allows us to approximate the posterior distribution of the latent variables. We factorize the posterior distribution as follows:

$$
\begin{aligned}
q(\mathbf{z}, \mathbf{n} | \mathbf{x}, \mathbf{b}) &= q(\mathbf{n} | \mathbf{x}, \mathbf{b}) \delta(\mathbf{z} = g(\mathbf{n})) \\
&= q(\mathbf{z} | \mathbf{x}, \mathbf{b}) \delta(\mathbf{n} = g^{-1}(\mathbf{z}))
\end{aligned} \tag{2}
$$

where $\delta(\cdot)$ is the Dirac delta function. After this factorization, the evidence lower bound (ELBO) can be derived through straightforward computation:

$$
\begin{aligned}
&\mathbb{E}_{q_\chi}[\log p_\theta(\mathbf{x}|\mathbf{b})] \geq \text{ELBO} \\
=&\mathbb{E}_{q_\chi}\Big[\mathbb{E}_{\mathbf{z},\mathbf{n}\sim q_\phi}[\log p_\theta(\mathbf{x}|\mathbf{z},\mathbf{n},\mathbf{b})] \\
&- \text{KL}\big(q_\phi(\mathbf{z},\mathbf{n}|\mathbf{x},\mathbf{b}) \parallel p(\mathbf{z},\mathbf{n}|\mathbf{b})\big)\Big] \\
=&\mathbb{E}_{q_\chi}\Big[\mathbb{E}_{\mathbf{z}\sim q_\phi}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \text{KL}\big(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{b}) \parallel p(\mathbf{z}|\mathbf{b})\big) \\
&- \text{KL}\big(q_\phi(\mathbf{n}|\mathbf{x},\mathbf{b}) \parallel p(\mathbf{n}|\mathbf{b})\big)\Big]
\end{aligned} \tag{3}
$$

According to the above equation, the encoder processes the observed expression data $\mathbf{x}$ and domain information $\mathbf{b}$ to derive the conditional distribution $q_\phi(\mathbf{n}|\mathbf{x},\mathbf{b})$. Another network p_ψ uses the domain information to compute the prior distribution $p(\mathbf{n}^b|\mathbf{b})$. Then, $\mathbf{n}$ is transformed by the layer defined by the SEM into causal representations $\mathbf{z}$, which are subsequently fed into the function f to reconstruct observed data.

3.2 Structural Equation Model with Confounder

The Structural Equation Model (SEM) provides a structured approach to generate observed gene expression data. SEM is widely used in fields like economics and social science to model dependency relationships between variables within a system [Hair Jr *et al.*, 2021]. In its traditional form, a standard SEM for a single domain is defined by the following equations:

$$
\begin{aligned}
\mathbf{z}^s &= \mathbf{A}^T \mathbf{z}^s + \mathbf{n}, \\
\mathbf{z}^s &= (\mathbf{I} - \mathbf{A}^T)^{-1} \mathbf{n} \\
\mathbf{n} &\sim \mathcal{N}(\mathbf{0}, \text{diag}(\sigma_n^2)\mathbf{I})
\end{aligned} \tag{4}
$$

Here, $\mathbf{z}^s$ represents the basic cell embedding data in a single domain, where the matrix $\mathbf{A}$ is a learnable parameter that encodes the relationships between different variables in the observed data, such as gene-gene interactions. $\mathbf{n}$ denotes independent Gaussian variables, and the Gaussian distribution shows good performance in exponential family distributions [Liu *et al.*, 2023]. Previous studies, such as DeepSEM [Shu *et al.*, 2021], have utilized SEM to model gene relationships.

However, when handling multi-domain scRNA-seq data, the relationships modeled by SEM may be confounded by some covariates, particularly due to batch effects [Kaltenpoth and Vreeken, 2023]. Each domain, represented by $\mathbf{b}$, has its own domain-specific network $\mathcal{G}^b$. A common network $\mathcal{G} = \{V, A\}$ is defined as the shared structure $\{\mathcal{G}^b\}_{b=1}^B$ across all domains. This common structure excludes domain-specific edges, helping to reveal the true relationships [Huang *et al.*, 2023].

To control the influence of confounding factors on common structure, we apply the backdoor criterion to enhance causal inference accuracy by controlling for confounding domain variables creating "backdoor paths" between variables, illustrated in Figure 1a. To integrate both true regulatory relationships and domain-specific interference, we use latent variables $\mathbf{z}^b$ to represent domain-specific information associated with domain information $\mathbf{b}$. We then expand the SEM model to include a larger set of variables, $V = \{v^{g_i}\}_{i=0}^m \cup \{v^b\}$, where g denotes gene variables. This expanded model captures both direct gene-gene interactions and domain-specific effects, yielding a unified network structure consistent across domains [Kaltenpoth and Vreeken, 2023]. The SEM model for this extended set of variables is defined as:

$$\bar{\mathbf{A}} = \begin{bmatrix} \mathbf{0} & \mathbf{A}^{xb} \\ \mathbf{0} & \mathbf{A}^{xx} \end{bmatrix},$$
$$\begin{bmatrix} \mathbf{z}^b \\ \mathbf{z}^x \end{bmatrix} = \bar{\mathbf{A}}^T \begin{bmatrix} \mathbf{z}^b \\ \mathbf{z}^x \end{bmatrix} + \begin{bmatrix} \mathbf{n}^b \\ \mathbf{n}^x \end{bmatrix},$$
$$\mathbf{z} = \mathbf{z}^x = g(\mathbf{n}) = [(\mathbf{I} - \bar{\mathbf{A}}^T)^{-1}\mathbf{n}]_{k:k+m} \quad (5)$$

In this formulation, $\bar{A}$ represents the edges between domain variables and genes, while $\mathbf{A}^{xx}$ captures the relationships among the genes themselves. In accordance with the common GRN structure, we set the weight to 0 for the first k columns of $\bar{A}$, corresponding to the edges from gene/domain-related variables to domain-related variables.

Model identifiability refers to the extent to which model parameters can be uniquely determined from the observed data [Khemakhem *et al.*, 2020]. Achieving model identifiability is crucial as it ensures that the SEM efficiently captures the relationships among genes. Based on the assumption of [Khemakhem *et al.*, 2020; Kaltenpoth and Vreeken, 2023], we provide the foundation of identifiability in Appendix A. Further, we promote the adjacency matrix $\bar{A}$ to conform to a directed acyclic graph (DAG) structure. Instead of employing a conventional combinatorial DAG constraint, we adopt a continuous and differentiable constraint function [Zheng *et al.*, 2018]. This function achieves a value of 0 solely when the adjacency matrix $\bar{A}$ represents a DAG. Specifically, we define the constraint function as follows:

$$H(\bar{\mathbf{A}}) \equiv \text{tr}(e^{\bar{\mathbf{A}} \circ \bar{\mathbf{A}}}) - (m + k) = 0 \quad (6)$$

Here $tr(\cdot)$ denotes the trace of a matrix. In addition, we use $\mathcal{L}_s = ||\bar{\mathbf{A}}||_F^2$ to ensure sparsity.

To facilitate the training process, we introduce rough gene relationships as guides on the latent space, incorporating the concept of gene injection into the latent variables [Yang *et al.*, 2021]. Leveraging gene relationships, represented by the adjacency matrix $\mathbf{A}^{xx}$, we impose a specific constraint to minimize the distance between the gene expression data $\mathbf{x}$ and its regression form. This constraint, denoted as $\mathcal{L}_u$, is formulated as follows:

$$\mathcal{L}_u = \mathbb{E}_{q_\chi} \|\mathbf{x} - (\mathbf{A}^{xx})^T \mathbf{x}\|_2^2 \leq \kappa \quad (7)$$

where κ is a small constant.

3.3 Disentangling the Domain and Invariant Factors

To disentangle the latent factors $\mathbf{n}^x$ and $\mathbf{n}^b$, we follow the principles outlined by [Chen *et al.*, 2016; Higgins *et al.*, 2017] promoting independence among the variables. Specifically, we make the posterior distribution of $\mathbf{n}^b$ approximate the prior distribution conditioned on the domain information $\mathbf{b}$. This allows us to replace any potential correlation between $\mathbf{n}^x$ and $\mathbf{n}^b$ with the correlation between $\mathbf{n}^x$ and $\mathbf{b}$.

We use adversarial learning to facilitate the disentangling process. The goal is to minimize the mutual information between $\mathbf{n}^x$ and $\mathbf{b}$, ensuring that domain-specific information contained in $\mathbf{b}$ does not influence the biological factors in $\mathbf{n}^x$. Adversarial learning has proven effective in domains such as domain generalization and factor disentanglement [Zhou *et al.*, 2022].

The primary objective is to make $\mathbf{n}^x$ indistinguishable by a batch discriminator. This discriminator is implemented as a multi-class classifier for categorical domain variables, which predicts the batch or domain label based on the input features and is parameterized by ϕ_d. The training process alternates between updating the feature encoder to learn meaningful features and training the discriminator to distinguish between domains. The adversarial loss, denoted as $\mathcal{L}_{adv}$, is defined as follows:

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}_{\mathbf{n}^x \sim q_{\phi_{\mathbf{n}^x}}} \log q_{\phi_d}(\mathbf{b}|\mathbf{n}^x) \quad \text{(Generator)}$$
$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{\mathbf{n}^x \sim q_{\phi_{\mathbf{n}^x}}} \log q_{\phi_d}(\mathbf{b}|\mathbf{n}^x) \quad \text{(Discriminator)} \quad (8)$$

To tackle the issue of imbalanced sample difficulty in single-cell analysis, we incorporate focal loss at the sample level, which helps the model focus more on harder-to-classify samples [Li *et al.*, 2020a]. At the domain level, we dynamically adjust the loss weight for each domain based on its misclassification rate, using an Exponential Moving Average (EMA), as described by the following equations:

$$r_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbb{I}\left(p_i^{(k)} \leq \max_{k' \neq k} p_i^{(k')}\right) \quad (9)$$
$$\omega_k^{(t)} = \eta\, \omega_k^{(t-1)} + (1 - \eta) r_k \quad (10)$$
$$\omega^{(t)} = \text{softmax}(\omega^{(t)}) \quad (11)$$

N_k is the number of domain k, $p_i^{(k)}$ is the logit of sample i in domain k and $\mathbb{I}$ is the indicator function. In conclusion, adversarial loss drives the model to separate domain-specific information from biological information, resulting in a feature space invariant to domain variations. Minimizing this loss ensures that the learned variables $\mathbf{n}^x$ remain independent of the domain information $\mathbf{b}$.

3.4 Optimization and Inference of scConCRL

The training procedure of our model is formulated as the maximization of ELBO subject to certain constraints, expressed as equations (7), (6), and (8). To tackle this constrained optimization problem, we utilize the augmented Lagrangian algorithm, which leads to the formulation of a new loss function:

$$\begin{aligned}\mathcal{L} = &- \alpha\text{ELBO} + \beta\mathcal{L}_{\text{adv}} + \gamma\mathcal{L}_u \\ &+ \delta\mathcal{L}_s + \lambda H(\bar{\mathbf{A}}) + \frac{\rho}{2}\|H(\bar{\mathbf{A}})\|^2 \end{aligned} \tag{12}$$

Here α, β, γ, λ, and ρ are hyperparameters that control the influence of each constraint term in the optimization process. Of note, in the inference process, we utilize the relationship represented by $\mathbf{A}^{xx}$ to obtain integrated clean scRNA-seq data. For condition prediction, we set the condition domain to generate data under the specified conditions.

4 Experiment

4.1 Setup

Dataset. We simulated scRNA-seq datasets using the scMultiSim R package [Li *et al.*, 2022] with default parameters and generated clean and noisy scRNA-seq data. For real-world datasets, we combined data from mouse Embryonic Stem Cells (**mESC**) from four protocols, alongside four different networks from the BEELINE benchmark [Pratapa *et al.*, 2020] for GRN inference. To evaluate data integration, we used 4 human pancreas datasets (**hPancreas**), 4 mouse cortex datasets (**mCortex**), 8 human Peripheral Blood Mononuclear Cell datasets (**hPBMC**), and 14 human heart datasets (**hHeart**), all with gold standard cell type labels. For perturbation response prediction, we used a human PBMC dataset with seven cell types, including control and interferon-beta (IFN-β)-stimulated cells (**PBMC**). All datasets were preprocessed with normalization, logarithmic transformation, high-variable gene selection, and max-abs scaling. More information can be found in Appendix B.

Performance evaluation. For GRN inference evaluation, given the sparsity and incomplete nature of biological ground truths, we select EPR (Early Precision Rate), AUPRC ratio, and AUROC. To assess data integration performance, we use the scIB score [Luecken *et al.*, 2022], which combines 8 metrics across two aspects: biological conservation and batch correction. The former indicates the representation quality while the latter assesses domain disentanglement. For perturbation response prediction tasks, we evaluate prediction accuracy using R^2 scores [Lotfollahi *et al.*, 2019]. Adjusted R^2 evaluates the relationship between predicted data for a specific cell type and real perturbed data across all cell types. Further details on comparison methods, metrics, and implementation specifics can be found in Appendix C, D, and E.

4.2 GRN inference

GRN inference is one of the real application scenarios of causal structure inference and we evaluate scConCRL on multi-batch mESC dataset to assess its advantages, benchmarking its performance against state-of-the-art methods. We used two sets of highly variable genes (1000 and 2000 HVGs) from mESC datasets and compared scConCRL with two deep learning-based methods (DeepSEM and RegDiffusion [Zhu and Slonim, 2023]) and two traditional methods (PIDC [Chan *et al.*, 2017] and GRNBOOST2 [Moerman *et al.*, 2019]). The performance of scConCRL and the four baseline methods was assessed on true biological networks of varying scales. As shown in Table 1, scConCRL consistently outperforms the other methods, achieving higher accuracy in both non-cell-type-specific and cell-type-specific GRN inference. It excels with domain adaptation mitigating dataset domain shift issues, which affects the performance of deep learning baselines (DeepSEM and RegDiffusion). PIDC and GRNBOOST2 perform well when handling 1,000 HVGs but perform poorly when handling 2,000 HVGs, which highlights that scConCRL is a better choice in handling complex scenarios. Under the 2,000 HVGs setting, we further compared the performance of combining data from all four batches with using any single batch of data. We evaluated the four baseline methods on each single batch (labeled "Batch 0-3") and on the complete dataset (labeled "All"). For these baseline methods, combining the four batches does not improve performance compared to using individual batches, primarily due to variations in data quality and noise levels (Supplementary Tables 2-3). In contrast, scConCRL consistently outperforms using any single batch, demonstrating the importance of eliminating batch-related noise and bias for multi-batch data downstream analysis.

4.3 Integration in molecular space

Data integration was conducted to disentangle true biological variability from confounding variation. We have validated the model's ability to integrate data in the molecular space across multiple datasets, which requires not only achieving distributional alignment but also preserving high-quality data. We compared scConCRL with other end-to-end molecular-level integration methods, as well as denoising + batch-correction two-step strategies which can recover data quality. As shown in the results of Table 2 and Supplementary Table 4, scConCRL significantly outperforms all other methods across all datasets. Specifically, scConCRL leads by 20-30% in batch correction on the hPBMC, mCortex, and hHeart datasets, and by 5% on the hPancreas dataset. In terms of biological conservation, scConCRL excels beyond all end-to-end molecular-level integration methods, preserving biological signals better than other end-to-end methods. Overall, two-step strategies can better preserve and restore biological signals, while their fusion results are less satisfactory. Seurat performs well on the hPancreas dataset but falls short on other datasets, indicating its limited applicability to large-scale datasets. The MNN and SCALEX methods may be constrained by the chosen reference batch, limiting their performance. Reference-free methods Beaconet and Scanorama fail to outperform reference-based methods due

Method	Non Spec			Spec			STR			lofgof		
	EPR	AUPRC R.	AUROC	EPR	AUPRC R.	AUROC	EPR	AUPRC R.	AUROC	EPR	AUPRC R.	AUROC
1000 HVGs												
PIDC	2.8817	1.8668	**0.6113**	1.2406	1.2500	0.6055	5.0771	3.2920	0.6676	1.3872	1.4123	0.6035
GRNBOOST2	2.5718	1.5749	0.5809	1.1396	1.1873	0.5572	3.9712	2.5503	0.6316	1.5569	1.4953	0.6074
DeepSEM	2.8701	1.6027	0.5627	1.1685	1.1960	0.5623	**5.4415**	3.2936	0.6607	1.3950	1.3846	0.5732
RegDiffusion	3.2303	1.7525	0.5667	1.0034	1.0759	0.5028	5.3598	3.2888	0.6597	1.1721	1.1919	0.5136
scConCRL	**3.5386**	**2.0316**	0.5947	**1.2836**	**1.2898**	**0.6101**	**5.4415**	**3.5424**	**0.6956**	**1.6512**	**1.5167**	**0.6179**
2000 HVGs												
PIDC	3.4309	1.7206	0.5789	1.1258	1.1290	0.5559	4.7304	2.6867	0.6324	1.2092	1.2359	0.5393
GRNBOOST2	2.2946	1.3605	0.5532	1.0870	1.1241	0.5372	2.9336	1.6689	0.5799	1.3873	1.3382	0.5585
DeepSEM	2.8982	1.4037	0.5409	1.1255	1.1571	0.5483	4.7029	2.1031	0.5752	1.3663	1.3122	0.5512
RegDiffusion	3.5156	1.5565	0.5355	0.8701	1.0363	0.5029	4.9916	2.3251	0.5979	1.1114	1.1210	0.5141
scConCRL	**3.9699**	**1.8833**	**0.5797**	**1.1548**	**1.1835**	**0.5649**	**5.4026**	**2.9010**	**0.6418**	**1.5097**	**1.4081**	**0.5816**

Table 1: GRN inference performance comparison of different methods in mESC dataset with 1000 HVGs and 2000 HVGs. "Non Spec" denotes the ground truth dataset from non-specific ChIP-seq, "Spec" represents the ground truth dataset from cell-type-specific ChIP-seq data, "STR" is the STRING network, and "lofgof" is the lofgof network. The numbers in **bold** indicate the best performance, while the underlined ones denote the second best.

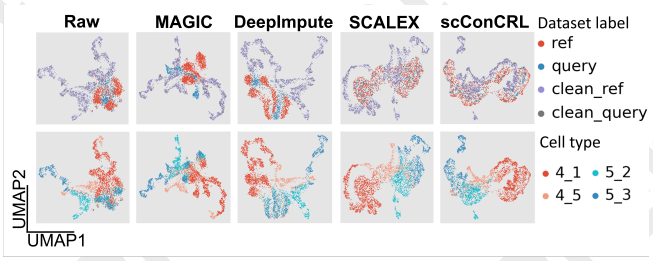

Figure 2: UMAP visualization comparing the reference data (clean_ref), query data (clean_query), predicted results of the reference data (ref), and the predicted results of the query data (query) from different methods. The top plot is color-coded by dataset labels, while the bottom plot is color-coded by cell type labels.

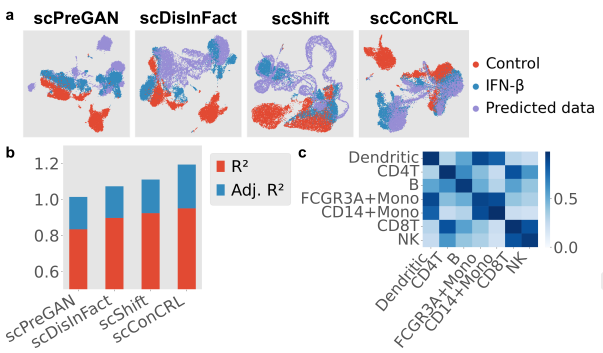

Figure 3: Evaluation of scConCRL's accuracy in predicting perturbation responses. a: UMAP visualization comparison of gene expression of PBMC dataset under different conditions. b: R^2 and adjusted R^2 for the mean of gene expression in predicting responses versus actual responses for top-100 DEGs in the PBMC dataset. c: Heatmap of detailed R^2 results on the PBMC dataset.

to weaker constraints. Supplementary Figures 1-4 confirmed scConCRL's dual strengths of batch correction and biological conservation. scConCRL demonstrates superior performance in data integration in the molecular space. Moreover, an ablation study based on the task (Appendix F and Supplementary Table 1) proves the effectiveness of our design.

Further, we tested the ability of our scConCRL algorithm for online integration analysis, where the model is trained on reference datasets and applied to infer data from a new query dataset. For each experiment, a single-batch subset was used as the query dataset, with the remaining datasets as the reference. We compared scConCRL to the SCALEX algorithm using the scIB score. In this experiment, batches were simplified into "reference" and "query" categories. As shown in Table 3, scConCRL outperforms SCALEX in both batch correction and biological feature preservation, especially achieving a 3% improvement in batch correction metrics. To assess the impact of data integration on data quality, we performed experiments on simulated datasets and visualized the integrated datasets alongside clean datasets, including both query and reference data. We compared scConCRL with two widely used denoising methods, MAGIC and DeepImpute, as well as SCALEX. scConCRL outperforms them, with seamless integration of the predicted reference (red) and query (blue) data, and the integrated dataset closely approximating the clean dataset. This demonstrates scConCRL's superior ability to denoise and integrate both observed and unseen data, restoring true expression levels (Figure 2).

4.4 Perturbation prediction

When it comes to the condition domain, scConCRL excels at disentangling the condition information and predicting perturbation responses, a crucial intervention task for analyzing drug effects. A typical perturbation dataset includes control data and real perturbed data, and models are used to predict the perturbed data. We trained models separately for each cell type using the rest and inferred unseen perturbed data based on the control data. Using the PBMC dataset, we compared the performance of scConCRL with baselines including scPreGAN [Wei *et al.*, 2022], scDisInFact [Zhang *et al.*, 2024b], and scShift [Dong and Kluger, 2023]. We visualized the distribution of control data, real perturbed data, and predicted perturbation data in Figure 3a and Supplementary Figure 5. scConCRL shows the best performance, with the predicted data closely matching the real perturbed data and clearly separated from the control data. In contrast, scPreGAN produces two separate prediction clusters, distanced from real data. scDisInFact was directionally correct but

Method	hPancreas			hPBMC			mCortex			hHeart		
	Bio.	Batch	Total	Bio.	Batch	Total	Bio.	Batch	Total	Bio.	Batch	Total
Raw	0.4932	0.3161	0.4223	0.3854	0.4277	0.4108	0.2866	0.4536	0.3534	0.5333	0.4704	0.5082
denoise + batch effect removal method Combat												
MAGIC	0.7436	0.3164	0.5727	0.6318	0.3665	0.5257	0.6894	0.3842	0.5673	0.7216	0.3743	0.5827
DeepImpute	0.6897	0.4700	0.6018	0.6167	0.4775	0.5610	0.6142	0.4055	0.5308	0.6442	0.4875	0.5815
SAVERX	0.7417	0.5046	0.6469	0.6849	0.4480	0.5901	0.6992	0.4343	0.5932	0.6433	0.4815	0.5786
Bis	0.7163	0.5036	0.6312	0.6582	0.4698	0.5828	0.6625	0.4646	0.5833	0.5621	0.4855	0.5314
scVI	0.7235	0.4594	0.6179	0.6761	0.4638	0.5912	0.6770	0.4529	0.5873	0.7076	0.4431	0.6018
integration in molecular space												
Seurat	0.7448	0.6742	0.7166	0.4198	0.3195	0.3797	0.4837	0.3990	0.4499	0.4130	0.4734	0.4371
MNN	0.7177	0.5056	0.6329	0.6531	0.5004	0.5920	0.6649	0.4658	0.5853	0.6886	0.4877	0.6082
SCALEX	0.6870	0.5685	0.6396	0.6404	0.4895	0.5800	0.6595	0.5704	0.6239	0.6597	0.4891	0.5914
Beaconet	0.7080	0.5280	0.6360	0.6330	0.4924	0.5768	0.6489	0.5543	0.6110	0.6622	0.5487	0.6168
Scanorama	0.7223	0.4866	0.6281	0.6413	0.4162	0.5513	0.6578	0.4443	0.5724	0.7219	0.4835	0.6265
scConCRL	**0.7579**	**0.7492**	**0.7544**	**0.6700**	**0.6984**	**0.6814**	**0.6873**	**0.7378**	**0.7075**	**0.7326**	**0.7031**	**0.7208**

Table 2: Integration performance comparison of different methods across four datasets.

Method	hPancreas			hPBMC			mCortex			hHeart		
	Bio.	Batch	Total	Bio.	Batch	Total	Bio.	Batch	Total	Bio.	Batch	Total
SCALEX	0.6624	0.4759	0.5505	0.6527	0.5558	0.5946	0.6737	0.5292	0.5871	0.6553	0.5796	0.6099
scConCRL	0.7072	0.5152	0.5920	0.6628	0.5988	0.6244	0.6740	0.5593	0.6051	0.6664	0.6237	0.6408

Table 3: Online molecular-level integration performance comparison of different methods across datasets.

Method	AUPRC	AUROC	EP
scDisInFact	0.1412	0.4978	0.1266
scConCRL	0.8222	0.9471	0.7388

Table 4: Comparison of methods using AUPRC, AUROC, EP.

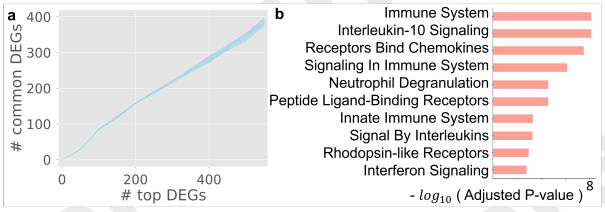

Figure 4: Performance of key gene detection. a: Number of common DEGs for ground truth and inference results of scConCRL with ten experiments. b: Highly correlated response pathways obtained from gene enrichment analysis using the top 100 predicted DEGs from the PBMC dataset with the Reactome 2022 gene database.

struggled to match real perturbed data. scShift's predictions are more scattered, further limiting its performance. We also calculated R^2 and adjusted R^2 to quantify performance. The average results per cell type are shown in Figure 3b, and the overall similarity between predicted and actual perturbed data is presented in Figure 3c. These results further demonstrate scConCRL's superior performance.

4.5 Key gene detection

In the scConCRL framework, the matrix $\mathbf{A}^{xb}$ captures the relationship between conditions and genes, identifying condition-related key genes (CKGs) [Zhang *et al.*, 2024b]. To evaluate the performance of scConCRL versus scDisInFact, we assessed CKG detection accuracy on the PBMC dataset. First, we used the Wilcoxon rank-sum test to identify the differentially expressed genes (DEGs) between real perturbed and control data, and then the ground truth was subsequently established based on p-values. For scConCRL, the absolute values of $\mathbf{A}^{xb}$ are the CKG score for each gene, reflecting the likelihood of a gene being a CKG. As shown in Table 4, scConCRL consistently outperforms scDisInFact, demonstrating superior performance in CKG detection and highlighting the enhanced interpretability. To further validate scConCRL, we examined the overlap between predicted DEGs and ground truth. As shown in Figure 4a, scConCRL achieves nearly 70% overlap, indicating high accuracy. Pathway analysis of the top 100 predicted DEGs (Figure 4b) revealed strong associations with immune system functions and signaling pathways involved in immune modulation and viral defense.

5 Conclusion

In this paper, we introduce scConCRL, conditional causal representation learning that tackles the challenges of integrating and analyzing heterogeneous single-cell data. By modeling the relationships between latent domain variables and gene variables through SEM, scConCRL captures both coarse- and fine-grained variables' interactions. Evaluations on multi-domain scRNA-seq datasets demonstrate its superior performance in several tasks including GRN inference, data integration, perturbation prediction, and key gene detection. scConCRL reveals complex dependencies, offering an interpretable solution for understanding cellular diversity and dynamic gene regulatory mechanisms.

References

[Aliee *et al.*, 2024] Hananeh Aliee, Ferdinand Kapl, Duy Pham, Batuhan Cakir, Takahiro Jimba, James Cranley, Sarah A Teichmann, Kerstin B Meyer, Roser Vento-Tormo, and Fabian J Theis. invae: Conditionally invariant representation learning for generating multivariate single-cell reference maps. *bioRxiv*, pages 2024–12, 2024.

[Argelaguet *et al.*, 2021] Ricard Argelaguet, Anna SE Cuomo, Oliver Stegle, and John C Marioni. Computational principles and challenges in single-cell data integration. *Nature biotechnology*, 39(10):1202–1215, 2021.

[Chan *et al.*, 2017] Thalia E Chan, Michael PH Stumpf, and Ann C Babtie. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems*, 5(3):251–267, 2017.

[Chen *et al.*, 2016] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

[Dong and Kluger, 2023] Mingze Dong and Yuval Kluger. Deep identifiable modeling of single-cell atlases enables zero-shot query of cellular states. *bioRxiv*, 2023.

[Haghverdi *et al.*, 2018] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421–427, 2018.

[Hair Jr *et al.*, 2021] Joseph F Hair Jr, G Tomas M Hult, Christian M Ringle, Marko Sarstedt, Nicholas P Danks, Soumya Ray, Joseph F Hair, G Tomas M Hult, Christian M Ringle, Marko Sarstedt, et al. An introduction to structural equation modeling. *Partial least squares structural equation modeling (PLS-SEM) using R: a workbook*, pages 1–29, 2021.

[Hao *et al.*, 2021] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

[Hao *et al.*, 2023] Yuhan Hao, Tim Stuart, Madeline H Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, and Rahul Satija. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, 2023.

[Hie *et al.*, 2019] Brian Hie, Bryan Bryson, and Bonnie Berger. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nature biotechnology*, 37(6):685–691, 2019.

[Higgins *et al.*, 2017] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

[Huang *et al.*, 2023] Zenan Huang, Haobo Wang, Junbo Zhao, and Nenggan Zheng. idag: Invariant dag searching for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19169–19179, 2023.

[Hyvarinen *et al.*, 2019] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

[Jovic *et al.*, 2022] Dragomirka Jovic, Xue Liang, Hua Zeng, Lin Lin, Fengping Xu, and Yonglun Luo. Single-cell rna sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*, 12(3):e694, 2022.

[Kaltenpoth and Vreeken, 2023] David Kaltenpoth and Jilles Vreeken. Nonlinear causal discovery with latent confounders. In *International Conference on Machine Learning*, pages 15639–15654. PMLR, 2023.

[Khemakhem *et al.*, 2020] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[Li *et al.*, 2020a] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020.

[Li *et al.*, 2020b] Xiangjie Li, Kui Wang, Yafei Lyu, Huize Pan, Jingxiao Zhang, Dwight Stambolian, Katalin Susztak, Muredach P Reilly, Gang Hu, and Mingyao Li. Deep learning enables accurate clustering with batch effect removal in single-cell rna-seq analysis. *Nature communications*, 11(1):2338, 2020.

[Li *et al.*, 2022] Hechen Li, Ziqi Zhang, Michael Squires, Xi Chen, and Xiuwei Zhang. scmultisim: simulation of multi-modality single cell data guided by cell-cell interactions and gene regulatory networks. *Research Square*, 2022.

[Liu *et al.*, 2023] Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifiable latent polynomial causal models through the lens of change. *arXiv preprint arXiv:2310.15580*, 2023.

[Lopez *et al.*, 2018] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

[Lotfollahi *et al.*, 2019] Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

[Luecken *et al.*, 2022] Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.

[Moerman *et al.*, 2019] Thomas Moerman, Sara Aibar Santos, Carmen Bravo González-Blas, Jaak Simm, Yves Moreau, Jan Aerts, and Stein Aerts. Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35(12):2159–2161, 2019.

[Pratapa *et al.*, 2020] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 17(2):147–154, 2020.

[Schölkopf *et al.*, 2021] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[Shu *et al.*, 2021] Hantao Shu, Jingtian Zhou, Qiuyu Lian, Han Li, Dan Zhao, Jianyang Zeng, and Jianzhu Ma. Modeling gene regulatory networks using neural network architectures. *Nature Computational Science*, 1(7):491–501, 2021.

[Stuart *et al.*, 2019] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *cell*, 177(7):1888–1902, 2019.

[Tejada-Lapuerta *et al.*, 2023] Alejandro Tejada-Lapuerta, Paul Bertin, Stefan Bauer, Hananeh Aliee, Yoshua Bengio, and Fabian J Theis. Causal machine learning for single-cell genomics. *arXiv preprint arXiv:2310.14935*, 2023.

[Wei *et al.*, 2022] Xiajie Wei, Jiayi Dong, and Fei Wang. scpregan, a deep generative model for predicting the response of single-cell expression to perturbation. *Bioinformatics*, 38(13):3377–3384, 2022.

[Xiong *et al.*, 2022] Lei Xiong, Kang Tian, Yuzhe Li, Weixi Ning, Xin Gao, and Qiangfeng Cliff Zhang. Online single-cell data integration through projecting heterogeneous datasets into a common cell-embedding space. *Nature Communications*, 13(1):6118, 2022.

[Xu *et al.*, 2022] Xinyi Xu, Xiaokang Yu, Gang Hu, Kui Wang, Jingxiao Zhang, and Xiangjie Li. Propensity score matching enables batch-effect-corrected imputation in single-cell rna-seq analysis. *Briefings in Bioinformatics*, 23(4):bbac275, 2022.

[Xu *et al.*, 2024] Han Xu, Yusen Ye, Ran Duan, Yong Gao, Yuxuan Hu, and Lin Gao. Beaconet: A reference-free method for integrating multiple batches of single-cell transcriptomic data in original molecular space. *Advanced Science*, page 2306770, 2024.

[Yang *et al.*, 2021] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9593–9602, 2021.

[Yu *et al.*, 2023] Xiaokang Yu, Xinyi Xu, Jingxiao Zhang, and Xiangjie Li. Batch alignment of single-cell transcriptomics data using deep metric learning. *Nature communications*, 14(1):960, 2023.

[Zhang *et al.*, 2024a] Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36, 2024.

[Zhang *et al.*, 2024b] Ziqi Zhang, Xinye Zhao, Mehak Bindra, Peng Qiu, and Xiuwei Zhang. scdisinfact: disentangled learning for integration and prediction of multi-batch multi-condition single-cell rna-sequencing data. *Nature Communications*, 15(1):912, 2024.

[Zheng *et al.*, 2018] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

[Zhou *et al.*, 2022] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022.

[Zhu and Slonim, 2023] Hao Zhu and Donna K Slonim. From noise to knowledge: Probabilistic diffusion-based neural inference of gene regulatory networks. *bioRxiv*, pages 2023–11, 2023.