Universiteit Leiden
The Netherlands

Opleiding Informatica

Evaluation of Causal Inference Methods for Gene Perturbation Prediction

Reconstructing and extending perturbation prediction metrics

Luca Adrianus Verheul

Supervisors:
Saber Salehkaleybar & Gijs van Seeventer

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

28/04/2026

Abstract

Single-cell CRISPR perturbation screens enable systematic discovery of gene regulatory interactions, but the perturbation space remains sparse: many genes (and conditions) are never assayed experimentally. This motivates models that can *predict unseen perturbation effects* from limited interventional data.

In this thesis we evaluate **Bicycle** [RCM+24], a causal model that combines a latent dynamical system with intervention-aware parameterization to infer a *cyclic* gene regulatory network and generalize to held-out perturbations. We reproduce the Bicycle training and evaluation pipeline on three biological contexts from the Perturb-CITE-seq study of Frangieh et al. [FMT+21], namely, Control, IFN-γ, and Co-culture.

Beyond the standard *interventional mean absolute error* (I-MAE) used in the Bicycle paper, we adopt two additional metrics from the *Virtual Cell Challenge* [Arc]: (1) the *Differential Expression Score* (DES), measuring recovery of differential gene sets, and (2) the *Perturbation Discrimination Score* (PDS), measuring whether perturbation signatures remain distinguishable even when effect sizes differ.

Our results show that I-MAE, DES, and PDS are correlated but not redundant: models with similar I-MAE can differ substantially in DES/PDS, indicating that I-MAE alone may fail to reflect biological correctness of predicted perturbation response signatures. We further observe a consistent model ranking across all metrics and discuss implications for evaluation practice in perturbation modeling.

Contents

1 Introduction

Single-cell CRISPR perturbation technologies such as Perturb-seq have enabled systematic, high-throughput interrogation of gene regulatory networks (GRNs). Instead of assaying one perturbation at a time, these platforms simultaneously profile thousands of cells exposed to diverse gene knock-outs or knock-downs, providing a uniquely detailed view into how gene expression changes under intervention [D+16, A+16, D+17]. However, the perturbation space is always incomplete: many genes are not perturbed in a given experiment, and testing all perturbations in all contexts is infeasible. This motivates the need for computational models capable of **predicting the effect of unseen perturbations** using only a subset of observed interventions.
The recently proposed **Bicycle** model (Intervention-Based Causal Discovery with Cycles) addresses this challenge by combining causal inference with a latent dynamical systems model to infer regulatory interactions directly from observational and interventional single-cell data [RCM+24]. Bicycle learns a **cyclic** causal graph and uses it to generalize across perturbations, enabling predictions for conditions not encountered during training. Cycles are particularly important in biology, where feedback regulation is common (e.g., signaling loops and autoregulation).

1.1 Thesis scope and research questions

This thesis evaluates Bicycle on a subset of the Perturb-CITE-seq dataset of Frangieh et al. [FMT+21], extending the original evaluation by incorporating three complementary metrics used in the **Virtual Cell Challenge**:

- the *Interventional Mean Absolute Error* (I-MAE) already considered in Bicycle,
- the *Differential Expression Score* (DES),
- the *Perturbation Discrimination Score* (PDS).

These metrics assess distinct aspects of model performance, and let us test whether I-MAE alone is a reliable indicator of biologically meaningful prediction quality, definitions, are provided in Section 3.
Concretely, we address the following questions:

1. **Reproducibility:** Can we reproduce the quantitative behavior and ranking trends reported for Bicycle on the Frangieh data, and what deviations arise in absolute I-MAE?
2. **Metric agreement:** How strongly do I-MAE, DES, and PDS correlate across contexts and training runs, and where do they disagree?
3. **Metric validity:** In what ways can I-MAE be misleading as a *standalone* performance metric for perturbation prediction?

1.2 Datasets

We evaluate Bicycle across three biologically distinct contexts from the Frangieh dataset:

- **Control**: baseline T-cell culture without cytokine stimulation or without tumor-cell co-culture.

- **IFN-γ**: interferon-γ stimulation, inducing an immune-activated transcriptional state and altering baseline pathway activity.
- **Co-culture**: T cells cultured with melanoma cells, introducing tumor–immune interaction signals and additional environmental variation.

These contexts differ in baseline pathway activation and regulatory state. As a result, the same genetic perturbation can yield context-dependent downstream expression changes. For example, IFN-γ activates immune pathways and can shift regulatory sensitivity. Co-culture introduces tumor–immune signaling cues.

1.3 Contributions

The main contributions of this thesis are:

- A reconstruction of the Bicycle pipeline for the Frangieh contexts with careful reporting of preprocessing, training, and evaluation decisions.
- An extended evaluation that combines I-MAE with two Virtual Cell Challenge metrics (DES, PDS) to provide a multidimensional view of performance.
- An empirical analysis and discussion of why I-MAE may not be a sufficient proxy for biologically meaningful perturbation prediction quality.

2 Background & Preliminaries

This section discusses the methodology of the Bicycle model based directly on the original paper [RCM+24]. We also introduce the causal and statistical concepts needed for interpreting Bicycle's training objective and evaluation.

2.1 Single-cell perturbation screens and notation

We consider a dataset of cells indexed by $n \in \{1, \dots, N\}$. Each cell is measured across G selected genes. Let $x_n \in \mathbb{N}^G$ denote the observed unique molecular identifier (UMI) counts for cell n. Cells are grouped into *contexts* (Control / IFN-γ / Co-culture) and into *perturbation conditions*. A perturbation condition i targets a set of genes $P^i \subseteq \{1, \dots, G\}$.
Throughout, we distinguish (i) *directly perturbed* genes $g \in P^i$ and (ii) *non-target* (or *downstream*) genes $g \notin P^i$. Bicycle's interventional evaluation focuses on the non-target set because predicting downstream consequences is the goal.

2.2 Causal graphs and interventions

Bicycle is motivated primarily by a *stochastic dynamical systems* view of gene regulation, rather than by starting from a classical structural causal model (SCM) formulation [Pea09]. In the original paper, the latent gene-expression state is modeled through a continuous-time stochastic process whose steady-state distribution defines the distribution of latent expression states across cells

[RCM+24]. More specifically, Bicycle builds on a linear stochastic differential equation (SDE) / Ornstein–Uhlenbeck framework, which is well suited to gene regulatory networks because it can represent feedback-rich systems while still admitting a tractable steady-state solution.
The causal interpretation is introduced at the level of the interaction matrix. Bicycle assigns the regulatory coefficient matrix β causal semantics: the coefficient β_{hg} is interpreted as the direct effect of gene h on the transcription rate of gene g in the latent dynamical system [RCM+24]. This is consistent with recent work on cyclic causal discovery, where causal relations are defined in systems that may contain feedback loops rather than being restricted to directed acyclic graphs (DAGs) [B+21]. This distinction matters in biology, since gene regulatory networks often contain reciprocal regulation and feedback.
Interventions are incorporated by modifying the dynamical parameters only for directly perturbed genes. Concretely, for a perturbation condition i targeting genes P^i, Bicycle constructs condition-specific parameters α^i, β^i, and σ^i, while retaining the unperturbed parameters for genes not directly targeted by the intervention [RCM+24]. This design implements the *independent causal mechanisms* principle as an inductive bias: most mechanisms are shared across environments, and only the mechanisms associated with directly intervened-upon genes are allowed to change [RCM+24, SLB+21].

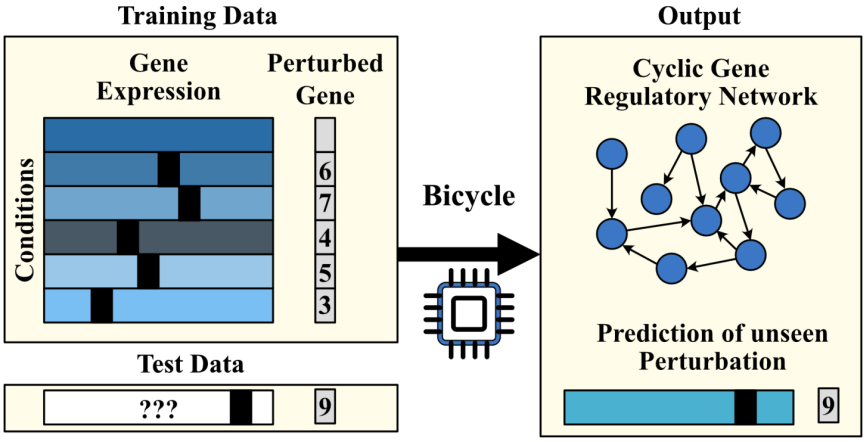

Figure 1: Bicycle uses data from multiple interventions to estimate an underlying causal graph and to predict responses to unseen interventions. The model is built from intervention-specific stochastic dynamics whose steady-state latent distributions are linked to observed count data [RCM+24].

2.3 Latent Dynamical System

Let $z \in \mathbb{R}^G$ denote the latent, noise-free gene expression state. Following Bicycle [RCM+24], gene expression is modeled as a multivariate Ornstein–Uhlenbeck (OU) process:

$$dz_g(t) = \left(\alpha_g + \sum_{g \neq h} \beta_{hg} z_h(t) - z_g(t) \right) dt + \sum_h \sigma_{gh}\, dW_h(t),$$

where α_g is the basal transcription rate, β_{hg} is the direct causal effect of gene h on g, and $W_h(t)$ are Wiener processes. Define the drift matrix $B = I_G - \beta^\top$, where $I_G \in \mathbb{R}^{G\times G}$ is the identity matrix and $\beta^\top$ is the transpose of the regulatory coefficient matrix $\beta \in \mathbb{R}^{G\times G}$ with entries $(\beta)_{hg} = \beta_{hg}$. When B has eigenvalues with positive real parts, this system admits a closed-form steady-state distribution $p_{\text{ss}}(z)$, i.e., the distribution approached as $t \to \infty$ [WU45]:

$$p_{\text{ss}}(z) = \mathcal{N}(\bar{z}, \omega) \quad \text{with} \quad \bar{z} = B^{-1}\alpha,$$

and covariance ω given by the Lyapunov equation:

$$B\,\omega + \omega B^\top = \sigma\sigma^\top.$$

This formulation links continuous-time dynamics to cyclic causal structure.

2.4 Observation Model

Observed single-cell RNA-seq counts are *compositional*: for a given cell n, sequencing produces a finite total number of captured molecules (UMIs), so only relative abundances across genes are directly observable. Bicycle therefore models the observed counts $x_n \in \mathbb{N}^G$ conditionally on the latent state z_n using a Multinomial likelihood,

$$x_n \mid z_n, N_n \sim \text{Multinomial}(N_n, \pi_n), \qquad \pi_n = \text{Softmax}(z_n),$$

where $N_n = \sum_{g=1}^{G} x_{ng}$ is the observed library size and $z_n \in \mathbb{R}^G$ denotes the latent log-proportion state. Equivalently,

$$p(x_n \mid z_n) = \text{Multinomial}(x_n; \pi_n), \qquad \pi_n = \text{Softmax}(z_n).$$

In this view, z_n captures the underlying biological state, while the Multinomial distribution captures sampling variability induced by finite sequencing depth.

2.5 Perturbation Modeling via Independent Causal Mechanisms

A key innovation of Bicycle is incorporating perturbation-specific modifications to the dynamical system parameters. For a perturbation condition i targeting genes P^i, Bicycle defines perturbation-specific parameters

$$\alpha_g^i, \quad \beta_{hg}^i, \quad \sigma_{gh}^i,$$

but **only** for genes directly targeted by the intervention. For all other genes $g \notin P^i$, parameters remain unchanged:

$$\alpha_g^i = \alpha_g, \quad \beta_{hg}^i = \beta_{hg}, \quad \sigma_{gh}^i = \sigma_{gh}.$$

This realizes the **Independent Causal Mechanisms** (ICM) principle, which posits that the causal generative process is composed of autonomous modules that remain invariant across environments unless directly intervened upon [PJS17, SLB+21]. In Bicycle, this becomes an explicit inductive bias: most parameters are shared across perturbed and unperturbed contexts, and changes are localized to the mechanism of the directly perturbed gene [RCM+24].

2.6 Variational Inference

Because the latent states z_n are not observed, the exact posterior distribution over z_n given the count data is not tractable. Bicycle therefore uses variational inference: it replaces the intractable posterior with a tractable approximation $q(z_n)$, chosen as an independent Gaussian distribution with a separate mean and standard deviation for each latent coordinate and each cell. The model parameters and the variational parameters are then optimized jointly by maximizing a variational Evidence Lower Bound (ELBO), augmented with structural penalties that encode the steady-state dynamical assumptions of Bicycle and encourage sparsity in the inferred regulatory graph.
Concretely, the objective combines four components: (i) a data-fit term that rewards accurate reconstruction of the observed counts, (ii) a KL term that keeps the latent state close to the intervention-specific steady-state Gaussian prior, (iii) a penalty enforcing approximate consistency with the Lyapunov equation implied by the OU dynamics, and (iv) an ℓ_1-penalty on β that promotes a sparse interaction matrix. The objective displayed below is reproduced from the Bicycle paper, with only minor notational changes.

Steady-State distribution of latent expression

$$\mathcal{L}^{\gamma,\xi,\lambda}(x,\alpha,\beta,\sigma,\hat{\alpha},\hat{\sigma},\mu_z,\sigma_z) = \frac{1}{N}\sum_{n=1}^{N}\Big[\langle \log p(x_n \mid z_n)\rangle_{q(z_n|\mu_{z_n},\sigma_{z_n})} - \gamma \mathrm{KL}\big(q(z \mid \mu_{z_n},\sigma_{z_n}) \,\|\, \mathcal{N}(z;\bar{z}^{i_n},\omega^{i_n})\big)\Big] - \frac{\xi}{I}\sum_{i=0}^{I}\left\|B^i\omega^i + \omega^i(B^i)^T - \sigma^i(\sigma^i)^T\right\|_2^2 - \lambda\,\|\beta\|_1 \tag{1}$$

Log-Likelihood of observed counts

Sparsity

Δ between LHS and RHS of Lyapunov Eq.

Here, x denotes the observed count data; $\alpha, \beta, \sigma, \hat{\alpha}, \hat{\sigma}$ are the model parameters; and μ_z, σ_z are the variational parameters. Thus, $\mathcal{L}^{\gamma,\xi,\lambda}$ defines the criterion that is optimized jointly over the model and variational parameters during training.
The first term is the expected log-likelihood of the observed counts under the variational distribution and measures how well the model explains the data. The second term regularizes the latent representation by encouraging it to remain close to the steady-state distribution associated with the intervention applied to cell n. The third term penalizes violations of the Lyapunov equation, thereby tying the learned covariance structure to the assumed linear dynamics. The final term encourages sparsity in the regulatory coefficients, which improves interpretability and helps prevent overfitting. The hyperparameters γ, ξ, and λ control the relative weight of these components.

2.7 How I-MAE is defined in Bicycle

Bicycle evaluates generalization using *Interventional Mean Absolute Error* (I-MAE, defined in Section 3.1) on held-out perturbations. The key point is that Bicycle does *not* simply output a mean expression vector per perturbation. Instead, the metric evaluates how well the model predicts each gene d in each cell, *conditioned on all other genes* $\neg d$ in that cell [RCM+24].

In words, Bicycle approximates the conditional distribution $p(x_{nd} \mid x_{n\neg d})$ via a variational approximation over the latent variables. The I-MAE is then computed on *non-target* genes $d \in G \setminus P^i$ for each held-out intervention i. This makes I-MAE *graph-dependent* because, in each held-out intervention condition P_i, the model predicts the expression of every non-directly perturbed gene $d \in G \setminus P_i$ from the observed expression of the remaining genes in the same cell. The learned regulatory graph determines which conditional dependencies are used for these predictions, so I-MAE is not merely an unconditional "mean profile" error.

2.8 Reconstructing Bicycle and reproducibility considerations

In addition to implementing Bicycle as described in [RCM+24], we attempted to *closely reproduce* the I-MAE values reported in the original paper.

Across repeated runs, we matched the hyperparameters, training schedule, loss-weighting coefficients, sparsity penalty grid for λ, and preprocessing choices to the extent they were documented in [RCM+24] and the official code. We also matched the reported gene subset and the 90/10 perturbation split protocol (55 train perturbations, 6 test).

Overall, we were able to closely recreate the reported I-MAE behavior (including qualitative rankings across contexts and the spread across runs), but we did not obtain a perfect quantitative match to the absolute I-MAE values reported by [RCM+24]; see footnote [1]). [1]This is plausible given that several training details of the original Perturb-seq runs (e.g., epoch budget and optimizer settings) are not fully specified in the paper. Secondly our reproduction differs in a few implementation choices.

Concretely, we trained for up to 40,000 epochs (plus 5,000 epochs of latent pretraining) with early stopping, used Adam with a learning rate of 10^{-3} and batch size 1024, explored a wider regularization sweep (including spectral and Lyapunov scaling), and used a full-rank covariance factorization (rank = 61) for numerical stability, whereas the paper motivates a low-rank covariance parameterization for large real-world datasets. In our experiments, I-MAE values tended to fall near the lower end of the ranges shown in the paper's boxplots, and this discrepancy persisted across contexts and random seeds.

This gap is plausible because I-MAE is sensitive to small details of (i) conditional inference, (ii) how control cells are handled, (iii) normalization/log-transform choices applied after predicting compositional probabilities, and (iv) the number of runs. Minor differences that are conceptually equivalent can still shift absolute I-MAE. Importantly, the *relative* behavior of Bicycle—context difficulty ranking and generalization trends—was consistent with [RCM+24], so the reconstructed pipeline remains a valid foundation for the extended evaluation with I-MAE, DES, and PDS.

[1]Code for our reproduction is available on GitHub.

3 Motivation

Most prior work evaluates perturbation prediction using a reconstruction error metric such as the Interventional Mean Absolute Error (I-MAE). While simple and interpretable, I-MAE alone does not capture biologically meaningful structure.
In the **Virtual Cell Challenge**, this motivated the introduction of two additional metrics—the Differential Expression Score (DES) and the Perturbation Discrimination Score (PDS) [Arc]—designed to evaluate perturbation modeling beyond reconstruction error.

3.1 Interventional Mean Absolute Error (I-MAE)

The I-MAE is defined as the average absolute error between predicted and true expression under held-out perturbations, evaluated on *non-target* genes. In Bicycle, predictions are obtained from a conditional distribution and converted into normalized log-expression values [RCM+24].
Although useful for quantifying reconstruction accuracy, I-MAE has several limitations as a *sole* metric:

- **Many genes do not change:** If only a small subset of genes responds to the perturbation, a model can obtain a low average error by predicting the many unaffected genes well. This can happen even if it misses the key responders.
- **Averaging can hide missed perturbation effects:** Two predictions can have similar MAE even when one fails to capture the perturbation response itself. This occurs because MAE only measures absolute deviation from the observed expression values: a prediction that stays close to the overall or control-like mean can obtain a similar error, even if it misses which genes should increase or decrease under the perturbation.
- **No notion of distinguishability:** MAE does not test whether the predicted response for perturbation p is more similar to the true response of p than to the true response of another perturbation. As a result, a model can predict overly similar perturbation signatures for different interventions and still obtain a moderate MAE, because the error is averaged over genes rather than evaluating whether perturbations remain identifiable from their predicted responses.

3.2 Differential Expression Score (DES)

Differential expression (DE) refers to genes whose expression differs significantly between perturbed and control cells under a statistical test. DES evaluates whether the model correctly predicts **which genes become significantly differentially expressed** after a perturbation. For each perturbation k, predicted and true DE genes are identified by comparing perturbed cells to controls. The comparison is comprised of a Wilcoxon test with multiple-testing correction and an FDR threshold as in the Virtual Cell Challenge definition) [Arc]. Here the Wilcoxon rank-sum test is a nonparametric test comparing expression distributions between perturbed and control groups, and FDR is the false discovery rate controlled via multiple-testing correction across genes. Here k indexes perturbations (i.e., target genes) in the evaluation set; $G_{k,\text{true}}$ and $G_{k,\text{pred}}$ denote the

differentially expressed (DE) gene sets for perturbation k. The per-perturbation score is

$$\mathrm{DES}_k = \frac{|G_{k,\mathrm{pred}} \cap G_{k,\mathrm{true}}|}{|G_{k,\mathrm{true}}|}.$$

Without a size constraint, DES would allow a trivial overprediction strategy: by calling many genes differentially expressed, a model could increase overlap with the true DE set even if its predictions were biologically uninformative. The evaluation therefore constrains the evaluated predicted DE set to match the size of the true DE set [Arc]. As a result, DES focuses on *which* genes are recovered as differentially expressed, rather than on absolute effect magnitude across the full profile.

3.3 Perturbation Discrimination Score (PDS)

PDS evaluates whether a model predicts a *distinctive perturbation signature* for each intervention. More precisely, it asks whether the prediction for perturbation p is most similar to the *true* signature of the same perturbation, rather than to the signatures of other perturbations [Arc].
Let K denote the number of perturbations being evaluated. For each perturbation $k \in \{1, \ldots, K\}$, define:

- $y_k \in \mathbb{R}^G$: the true pseudobulk expression profile, obtained by averaging the $\log(1+x)$-normalized expression values across all cells under perturbation k;
- $\hat{y}_k \in \mathbb{R}^G$: the corresponding predicted pseudobulk expression profile;
- $y_{\mathrm{ntc}} \in \mathbb{R}^G$ and $\hat{y}_{\mathrm{ntc}} \in \mathbb{R}^G$: the true and predicted non-targeting control pseudobulk profiles.

From these, define the true and predicted *delta signatures* for perturbation k as

$$\delta_k = y_k - y_{\mathrm{ntc}}, \qquad \hat{\delta}_k = \hat{y}_k - \hat{y}_{\mathrm{ntc}}.$$

Now fix a predicted perturbation p. For every true perturbation $t \in \{1, \ldots, K\}$, define the Manhattan (ℓ_1) distance

$$d_{pt} = \|\hat{\delta}_p - \delta_t\|_1.$$

These distances measure how similar the predicted signature of perturbation p is to each true perturbation signature. Following the Virtual Cell Challenge definition, the directly targeted gene is excluded from this distance calculation [Arc].
Next, sort the distances $d_{p1}, \ldots, d_{pK}$ in ascending order and let $r_p \in \{1, \ldots, K\}$ denote the rank of the correct match $t = p$. Thus:

- $r_p = 1$ if the prediction for perturbation p is closest to its own true signature;
- larger r_p means that one or more incorrect perturbation signatures are closer.

The per-perturbation PDS is then defined as

$$\mathrm{PDS}_p = 1 - \frac{r_p - 1}{K}.$$

This score lies in $[0, 1]$, where 1 is best. The overall PDS is the mean of PDS_p over all evaluated perturbations p.
Intuitively, PDS rewards models that recover the *pattern* of the perturbation response well enough to distinguish one perturbation from another, even when absolute effect sizes are not perfectly matched.

3.4 Why All Three Metrics Are Needed

The three performance metrics capture complementary information:

- **I-MAE**: overall conditional reconstruction accuracy,
- **DES**: whether the correct DE genes were identified,
- **PDS**: whether perturbations remain distinguishable.

A model may have low I-MAE but fail to capture the correct biological response (low DES), or may capture the response structure but not its magnitude (high PDS, moderate I-MAE). Evaluating all three metrics provides a multidimensional assessment of perturbation prediction quality.

4 Related Work

We position Bicycle and our evaluation approach in the broader literature on (i) single-cell perturbation screens, (ii) perturbation response prediction, (iii) causal discovery with cycles, and (iv) benchmarking and metric design.

4.1 Single-cell perturbation datasets

Early pooled CRISPR + scRNA-seq methods such as Perturb-seq enabled scalable mapping from genetic interventions to transcriptional states [D+16, A+16, J+16]. Subsequent improvements increased throughput and expanded modalities (e.g., CITE-seq proteins) [M+19, DRB+21].
Frangieh et al. introduced *Perturb-CITE-seq* to profile both RNA and surface proteins in patient-derived melanoma–TIL co-cultures [FMT+21]. A key feature of this study is explicit *context dependence*: the same genetic perturbation can induce different transcriptional programs depending on immune pressure and culture condition (Control vs IFN-γ vs Co-culture). This makes the dataset a challenging and realistic testbed for out-of-distribution (OOD) perturbation prediction.

4.2 Perturbation response prediction models

A major line of work focuses on predicting expression changes under unseen perturbations without explicitly recovering a causal GRN. Representative approaches include variational autoencoder (VAE) models such as scGen [LWT19] (and extensions for compositional/complex settings such as CPA [LKSDD+23]), conditional generative models such as PerturbNet [YQSW25], graph-based deep learning approaches for single- and multi-gene perturbations such as GEARS [RHL24], and causal meta-learning approaches such as CaML [NMR+23].
Bicycle differs by aiming to infer an *interpretable* regulatory matrix β grounded in a steady-state dynamical system and by encoding interventions through ICM-style parameter invariance.

4.3 Causal discovery with cycles and interventions

Many causal discovery methods assume acyclicity, including NOTEARS [ZARX18]. We highlight NOTEARS in particular because Rohbeck et al. use it as a baseline when evaluating Bicycle on

both synthetic and Perturb-seq benchmarks [RCM+24]. Interventional extensions exist, but explicit acyclicity constraints preclude modeling feedback. Cyclic causal discovery has been studied in equilibrium and dynamical settings [MH13, B+21, SF25]. Bicycle contributes by combining cyclic structure with intervention-specific mechanism changes and a count-based likelihood appropriate for scRNA-seq.

4.4 Benchmarks and evaluation metrics

Benchmarking perturbation predictors is non-trivial: different metrics reward different aspects of performance. The Virtual Cell Challenge proposes an evaluation framework aimed at context generalization and uses DES and PDS alongside MAE [Arc]. This thesis follows that philosophy: I-MAE is informative but insufficient as a single "leaderboard metric".
Related work suggests two recurring themes [RCM+24, NMR+23, Arc, L+25]:

- **Causal structure matters** for out-of-distribution generalization under intervention, particularly in biological systems with feedback [RCM+24, NMR+23].
- **Metric choice matters** because different metrics can change or even invert model rankings [Arc, L+25].

Our experiments therefore emphasize multi-metric evaluation.

5 Experiments

5.1 Dataset and splits

We use the Frangieh Perturb-CITE-seq dataset as processed in the Bicycle paper [RCM+24, FMT+21]. Following [RCM+24], we restrict to the same subset of $G = 61$ genes and evaluate each biological context separately. In this processed subset, each perturbation corresponds to a single target gene from the $G = 61$-gene panel. For readability, we refer to perturbations by their target-gene index (1–61). The train/test split is therefore a split over perturbations (targets), not over measured genes in the expression vector.
Perturbation split. We follow the reported protocol: 90% of perturbations are used for training, 10% are held out for testing. The held-out set is unseen during training.
Controls. Control (non-targeting) cells are used to define baseline expression and are necessary for computing delta signatures (PDS) and DE testing (DES). In our pipeline we ensure that each context has an explicit control subset aligned with the perturbation condition metadata. Table 1 summarizes the resulting dataset sizes and split counts per context.

Table 1: Summary of the Frangieh Perturb-CITE-seq subsets used in this thesis, reported separately for the three biological contexts. #Cells is the total number of cells in the context-specific dataset after preprocessing; #Control is the number of non-targeting control cells; #Perturbed is the number of cells assigned to a gene-targeting perturbation; #Perts is the number of distinct perturbation conditions (target genes) present in that context; and #Train/#Test give the number of perturbation conditions assigned to the training and held-out test split, respectively. The split is performed over perturbations rather than over genes in the expression vector or over individual cells.

Context	#Cells	#Control	#Perturbed	#Perts	#Train	#Test
Control	21,155	13,142	8,013	61	55	6
IFN-γ	36,312	23,771	12,541	61	55	6
Co-culture	31,694	19,667	12,027	61	55	6

5.2 Preprocessing

We match the Bicycle paper's preprocessing as closely as possible.

- **Bicycle training:** trained on raw UMI counts with a Multinomial likelihood.
- **Metric computation:** For DES/PDS computation we follow the Virtual Cell Challenge metric definitions, which operate on normalized log-expression and pseudobulk delta signatures based on $\log(1+x)$-transformed, library-size-normalized counts.

5.3 Model training and hyperparameters

We train Bicycle with the hyperparameters described in [RCM+24].

- $\gamma = 1.0$, $\xi = 1.0$.
- $\lambda \in \{0.01, 0.1, 1.0\}$ tuned on training perturbations.
- Optimizer: Adam.
- Batch size = 1024, learning rate = 10^{-3}, number of epochs = 40000 (plus latent pretraining for 5000 epochs), early stopping enabled with patience = 500 and $\Delta_{\min} = 0.01$; number of seeds/runs: 5.

5.4 Evaluation protocol

We evaluate on held-out perturbations for each context and report performance across repeated training seeds.

5.4.1 Metric 1: I-MAE

We compute I-MAE as in the Bicycle paper: for each held-out perturbation i and each cell n, we evaluate a conditional distribution for each non-target gene $d \in G \setminus P^i$, convert to expected normalized log-expression, and compute MAE.

5.4.2 Metric 2: DES

For each perturbation k, we compare perturbed cells to non-targeting controls using a Wilcoxon rank-sum test per gene, apply multiple-testing correction, define the true/predicted DE gene sets at the chosen FDR, and compute overlap as in the Virtual Cell Challenge rubric [Arc].

5.4.3 Metric 3: PDS

We compute pseudobulk expression per perturbation and controls, compute delta signatures, and rank predicted deltas by similarity to true deltas using Manhattan (ℓ_1) distance; the normalized rank yields PDS [Arc].

5.4.4 Baselines for I-MAE

In addition to BICYCLE, we report two point-prediction baselines. We evaluate all methods under the same I-MAE functional: for each perturbed cell n under intervention i, and for each non-target gene $d \in \mathcal{G} \setminus \mathcal{P}_i$, we compare the true normalized log-expression y_{nd} to a predicted value $\hat{y}_{nd}$. For BICYCLE, $\hat{y}_{nd}$ is obtained from the model-induced conditional expectation (as described in Section 2.7). For baselines, $\hat{y}_{nd}$ is defined directly as a fixed reference profile (i.e., we bypass conditional inference but still compute the same per-gene absolute error).

- **WT-Baseline ("no perturbation effect").**

$$\hat{y}_{nd} := \mu_d^{\text{ctrl}},$$

  where μ_d^{ctrl} is the mean normalized log-expression of gene d over non-targeting control cells in the same biological context. This tests whether any method improves over predicting the control state for every perturbed cell.

- **VCC cell-mean baseline ("global mean").**

$$\hat{y}_{nd} := \mu_d^{\text{train}},$$

  where μ_d^{train} is the mean normalized log-expression of gene d over all training cells (controls + training perturbations) in the same biological context. This tests whether models outperform a context-specific global mean, analogous in spirit to the Virtual Cell Challenge cell-mean reference.

Because these baselines ignore perturbation identity, they can score deceptively well on reconstruction-style metrics when perturbation effects are weak or sparse, motivating complementary metrics such as DES and PDS.

5.4.5 Metric agreement analysis

To quantify agreement between metrics across contexts and repeated runs, we compute pairwise correlations using both Pearson's r (linear association) and Spearman's ρ (rank association). Since I-MAE is a loss (lower is better) while DES and PDS are scores (higher is better), negative correlations between I-MAE and DES/PDS are expected.

5.5 Robustness and sensitivity analysis

DES and PDS are not determined by the data alone: their values depend on the evaluation pipeline used to define differential expression and perturbation similarity. In this thesis, we therefore fix one such pipeline and use it unchanged across all contexts and runs.
Specifically, DES is computed using gene-wise two-sided Wilcoxon rank-sum tests with tie correction, followed by Benjamini–Hochberg correction and an FDR cutoff of 0.05 [Arc, BH95]. PDS is computed by forming perturbation-minus-control pseudobulk delta signatures and ranking predicted signatures against true signatures using Manhattan (ℓ_1) distance [Arc]. For both metrics, the expression values entering the evaluation are the normalized log-expression values

$$L_n = \sum_{g=1}^{G} x_{ng}, \qquad \tilde{x}_{ng} = \frac{x_{ng}}{L_n}, \qquad y_{ng} = \log(1 + \tilde{x}_{ng}).$$

Thus, the concrete choices fixed in our evaluation are: the DE test, the multiple-testing correction, the FDR threshold, the PDS similarity measure, and the expression transformation used before DE testing and pseudobulk construction. Our robustness discussion should therefore be read as a statement about sensitivity to changes in this evaluation pipeline, not as a claim that DES or PDS have a unique, method-independent definition.

5.5.1 DES sensitivity

DES depends on how the "true" DE set $G_{k,\text{true}}$ is defined by the DE-testing pipeline (test statistic, multiple-testing correction, and FDR threshold). Changing the FDR threshold changes $|G_{k,\text{true}}|$ and therefore changes the DES values. We therefore interpret DES as a metric that is meaningful *relative to a fixed DE-calling procedure* (in our case, Wilcoxon + BH at FDR = 0.05), rather than as an absolute quantity comparable across different DE pipelines.

5.5.2 PDS sensitivity

PDS depends on the similarity measure used to rank delta signatures and on whether signatures are scaled or standardized before similarity computation. Recent work notes that distance metrics and scaling can shift PDS rankings across methods [L+25]. For this thesis, we therefore treat PDS as a complement to (not a replacement for) I-MAE and DES, emphasizing agreement patterns and rank stability rather than absolute values.

5.6 Main results

5.6.1 Aggregate I-MAE performance by context

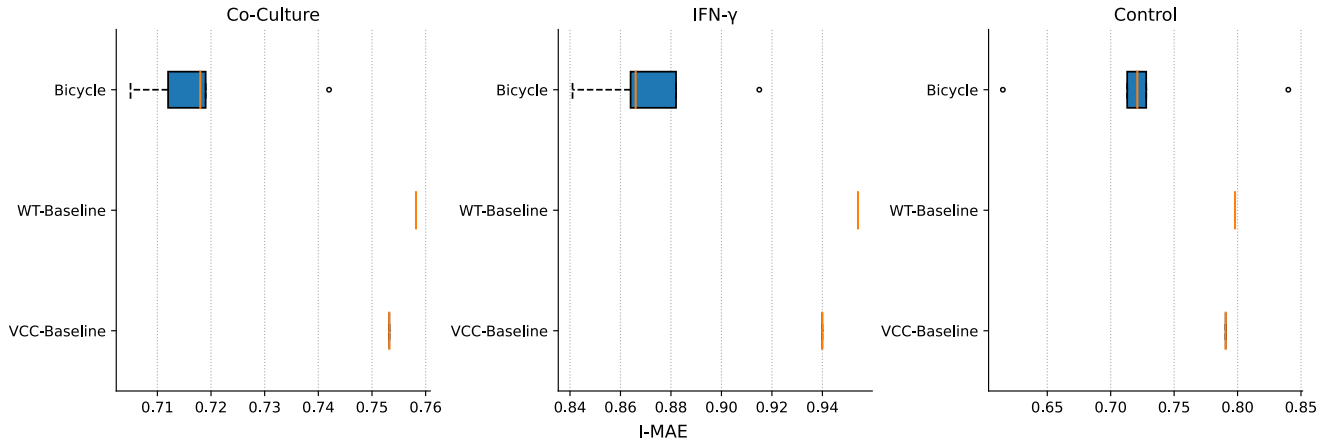

Figure 2: Comparison of out-of-distribution generalization in terms of the interventional mean absolute error (I-MAE) across three biological contexts. Co-culture: lymphocytes co-cultured with melanoma cells. IFN-γ: lymphocytes stimulated by interferon-γ. Control: lymphocytes cultured in neutral medium. Methods shown are *VCC-Baseline*, *WT-Baseline*, and *Bicycle*; baseline definitions are given in Section 5.4.4. Box plots show the distribution of I-MAE over repeated runs (minimum, lower quartile, median, upper quartile, maximum, and outliers).

Table 2: Aggregate performance on the held-out test perturbations (target genes 56–61, 1-indexed) for the three biological contexts. Bicycle results are shown as mean ± standard deviation across repeated runs. "I-MAE (WT)" is the control-mean baseline, which predicts the context-specific non-targeting control mean for each perturbed cell. "I-MAE (VCC)" is the Virtual Cell Challenge-style cell-mean baseline, which predicts the mean expression over all training cells in the same context. DES and PDS are the corresponding Bicycle scores on the same held-out perturbations. Lower I-MAE indicates better performance; higher DES and PDS indicate better performance.

Context	I-MAE ↓	I-MAE (WT) ↓	I-MAE (VCC) ↓	DES ↑	PDS ↑
Control	0.723 ± 0.080	0.797	0.79	0.969 ± 0.000	0.389 ± 0.007
IFN-γ	0.874 ± 0.027	0.955	0.94	0.938 ± 0.000	0.374 ± 0.004
Co-culture	0.719 ± 0.014	0.758	0.753	0.954 ± 0.000	0.590 ± 0.000

Table 3: Aggregate performance across runs (mean ± std). Train aggregates over the 55 training perturbations (target genes 1–55, 1-indexed).

Context	DES ↑	PDS ↑
Control	0.962 ± 0.002	0.558 ± 0.001
IFN-γ	0.970 ± 0.003	0.540 ± 0.002
Co-culture	0.985 ± 0.000	0.546 ± 0.000

Dataset difficulty ranking. On the held-out test interventions (the 6 unseen perturbations targeting genes 56–61), IFN-γ is consistently the most challenging context, with the highest I-MAE and the lowest DES/PDS. Control and Co-culture are easier test contexts for held-out perturbation prediction, showing lower I-MAE and higher DES/PDS. Co-culture performs best on I-MAE and PDS, while Control is marginally best on DES. Overall, this points to context-dependent differences in prediction difficulty rather than an artifact of any single metric.

5.6.2 Correlation between metrics

To test whether the three metrics are redundant, we compute pairwise correlations across all runs (and all contexts). Note that I-MAE is a loss (lower is better), while DES and PDS are scores (higher is better), so negative correlations between I-MAE and the other metrics are expected.

Table 4: Correlations between metrics over all runs and contexts.

	Pearson r	Spearman ρ
I-MAE vs DES	-0.510	-0.712
I-MAE vs PDS	-0.234	-0.589
DES vs PDS	0.531	0.684

Interpretation. First, runs with lower I-MAE tend to have higher DES and PDS (negative correlations), but the relationship is not proportional: improvements in I-MAE do not translate linearly into improvements in DE overlap (DES) or signature ranking (PDS). Second, the fact that Spearman correlations are stronger than Pearson correlations indicates that *rank agreement* across metrics is more reliable than *value agreement*: a run that ranks well on one metric typically ranks well on the others, even if the numerical gaps differ. Finally, DES and PDS are positively correlated but not interchangeable, motivating the qualitative checks in Section 5.7.1 where we examine failure modes that can look acceptable under I-MAE but still produce biologically unfaithful signatures.

5.7 Qualitative analysis and case studies

5.7.1 When I-MAE can be misleading

I-MAE averages absolute errors over all non-target genes, so it can look favorable even when the predicted perturbation *signature* is biologically incorrect. The three failure modes below make explicit the limitations already discussed in Section 3.1: averaging can mask errors on responsive

genes, mean-like predictions can appear accurate under MAE, and distinct perturbations can collapse to similar predicted responses. These cases explain why DES and PDS remain informative even when I-MAE appears acceptable.

(1) Error averaging can hide signature mistakes. Because I-MAE aggregates over many genes, errors on a relatively small set of responsive genes can be diluted by many near-unchanged genes. In this regime, two runs can have similar I-MAE while differing substantially in whether they capture the correct direction and ranking of expression changes. PDS is designed to be sensitive to this by comparing the full ranked delta signature to ground truth.

(2) "Smooth" predictions can look good under I-MAE. A predictor that regresses toward a context-specific mean profile can achieve a decent mean absolute error even if it underestimates perturbation-specific effects (i.e., it predicts changes that are too small). This can yield acceptable I-MAE while producing weak or non-specific delta signatures, which is penalized by PDS and can also reduce DES when true DE effects are not recovered.

(3) Correct mean-level reconstruction does not imply correct DE sets. A model may match overall expression levels well but still fail to recover which genes are truly differentially expressed under a perturbation (wrong DE set, or missing key DE genes). DES directly tests overlap of DE gene sets and can therefore disagree with I-MAE in such cases.
Overall, these cases explain why the correlations between metrics are imperfect (Table 4): I-MAE, DES, and PDS are related, but each detects distinct aspects of biological faithfulness.

6 Implementation and Reproducibility

This section documents the practical decisions required to reproduce Bicycle training and to compute I-MAE/DES/PDS consistently across contexts. In perturbation modeling, these details are not cosmetic: metric values can shift meaningfully with small changes in normalization, control selection, and pseudobulk construction.

6.1 Data handling and context isolation

We treat each biological context (Control, IFN-γ, Co-culture) as a separate dataset and train a separate model per context. This mirrors the protocol in [RCM+24] and avoids mixing context effects into the learning objective.
Within each context, we ensure that:

- Control (non-targeting) cells are explicitly identified and retained for both DE testing (DES) and delta-signature computation (PDS).
- The perturbation label per cell is consistent across training, evaluation, and metric code.
- The held-out set contains perturbations that are unseen during training (90/10 split).

6.2 Compute environment and stability considerations

Training Bicycle involves optimizing per-cell variational parameters and can exhibit run-to-run variance. We therefore report results over repeated runs where feasible and summarize performance with mean ± std. We also observed occasional outliers in I-MAE, consistent with the behavior noted in [RCM+24]. When using Bicycle in practice, a multi-seed aggregation (e.g., reporting the median prediction/score across several independent training seeds) can provide a more robust estimate than a single run.

7 Conclusions and Further Research

7.1 Conclusions

This thesis evaluated the Bicycle model on three biological contexts derived from the Frangieh Perturb-CITE-seq data, using a held-out perturbation split as in the original protocol. While we did not reproduce the exact numerical I-MAE values reported in [RCM+24], we recovered the same qualitative behavior: performance is context-dependent, and the relative difficulty ordering is consistent across metrics.

A central empirical takeaway is that I-MAE is a useful but incomplete summary of perturbation-prediction quality. In our experiments, lower I-MAE usually coincides with higher DES and PDS, indicating broad agreement between the metrics. However, this agreement is far from perfect, and is stronger for *ordering* runs than for quantifying the size of performance differences (Table 4). Consequently, I-MAE alone can miss biologically relevant variation: runs with comparable reconstruction error can still differ in which DE genes they recover and in whether their predicted perturbation signatures remain distinguishable. DES and PDS therefore add complementary information rather than merely repeating what is already captured by I-MAE.

Finally, comparing contexts shows that generalization difficulty is not uniform: IFN-γ is the most challenging setting under all three metrics on the held-out perturbations, whereas Control and Co-culture are easier. This reinforces the importance of reporting results per context and using multiple metrics when judging biological faithfulness of perturbation predictions.

7.2 Limitations

Several limitations should be considered.

- **Metric dependence:** DES and PDS are not uniquely defined by the data; they require explicit design choices, including the DE test (Wilcoxon rank-sum with tie correction), the multiple-testing procedure (Benjamini–Hochberg), the FDR threshold (0.05), and the PDS similarity/ranking criterion (Manhattan/ℓ_1 distance). Alternative but reasonable choices (e.g., different FDR thresholds or distance metrics) can change absolute scores and, in some cases, the relative ranking of methods. We therefore emphasize agreement patterns across metrics rather than relying on a single headline number.

- **Data dependence:** All experiments were conducted on the processed Perturb-CITE-seq subset used in prior work, restricted to $G = 61$ genes and evaluated per biological context. This reduced panel simplifies the prediction task and may overstate performance compared to

genome-wide settings, where modeling gene–gene dependencies, sparsity, and sequencing depth variation is more challenging. Consequently, quantitative results should be interpreted as evidence within this benchmark rather than as a direct estimate of genome-scale performance.

7.3 Further research

We suggest several directions.

1. **Metric robustness studies:** Systematically vary DE testing methods and similarity measures in PDS to assess stability of conclusions.

2. **Broader baselines:** Compare Bicycle against strong perturbation predictors under the same train/test perturbation split and the same DES/PDS implementation, including GEARS [RHL24], PerturbNet [YQSW25], and CaML [NMR+23]. This tests whether the interpretability of an explicit GRN (β) trades off against predictive performance when judged by signature-level metrics beyond I-MAE.

3. **Beyond steady state:** Extend Bicycle-style causal modeling to explicitly time-resolved perturbation trajectories in settings where the dataset includes multiple post-perturbation time points rather than only a single steady-state-like measurement.

References

[A+16] Britt Adamson et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882.e21, 2016.

[Arc] Arc Institute. Virtual cell challenge — evaluation and scoring. https://virtualcellchallenge.org/evaluation. Accessed: 2025-12-14.

[B+21] Stephan Bongers et al. Foundations of structural causal models with cycles and latent variables. *arXiv preprint arXiv:2106.10890*, 2021.

[BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[D+16] Atray Dixit et al. Perturb-seq: Dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2016.

[D+17] Paul Datlinger et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, 14(3):297–301, 2017.

[DRB+21] Paul Datlinger, Andre F Rendeiro, Thorina Boenke, Martin Senekowitsch, Thomas Krausgruber, Daniele Barreca, and Christoph Bock. Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nature Methods*, 18(6):635–642, 2021.

[FMT+21] Chris J Frangieh, Johannes C Melms, Pratiksha I Thakore, Kathryn R Geiger-Schuller, Patricia Ho, Adrienne M Luoma, Brian Cleary, Livnat Jerby-Arnon, Shruti Malu, Michael S Cuoco, et al. Multimodal pooled perturb-cite-seq screens in patient models define mechanisms of cancer immune evasion. *Nature Genetics*, 53(3):332–341, 2021.

[J+16] D. A. Jaitin et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell*, 167(7):1883–1896.e15, 2016.

[L+25] Qiyuan Liu et al. Effects of distance metrics and scaling on the perturbation discrimination score. *arXiv preprint arXiv:2511.16954*, 2025.

[LKSDD+23] Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L. Ibarra, Sanjay R. Srivatsan, Mohsen Naghipourfar, Riza M. Daza, Beth Martin, Jay Shendure, Jose L. McFaline-Figueroa, Pierre Boyeau, F. Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J. Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023.

[LWT19] Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scgen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, 2019.

[M+19] Eleni P Mimitou et al. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nature Methods*, 16(5):409–412, 2019.

[MH13] Joris Mooij and Tom Heskes. Cyclic causal discovery from continuous equilibrium data. *arXiv preprint arXiv:1309.6849*, 2013.

[NMR+23] Hamed Nilforoshan, Michael Moor, Yusuf Roohani, Yining Chen, Anja Šurina, Michihiro Yasunaga, Sara Oblak, and Jure Leskovec. Zero-shot causal learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2301.12292.

[Pea09] Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

[PJS17] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, Cambridge, MA, 2017.

[RCM+24] Martin Rohbeck, Brian Clarke, Katharina Mikulik, Alexandra Pettet, Oliver Stegle, and Kai Ueltzh"offer. Bicycle: Intervention-based causal discovery with cycles. In *Proceedings of the 3rd Conference on Causal Learning and Reasoning (CLeaR)*, volume 236 of *Proceedings of Machine Learning Research*, pages 209–242, 2024.

[RHL24] Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42:927–935, 2024.

[SF25] Muralikrishnna G. Sethuraman and Faramarz Fekri. Differentiable cyclic causal discovery under unmeasured confounders, 2025.

[SLB+21] Bernhard Sch"olkopf, 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.

[WU45] M. C. Wang and G. E. Uhlenbeck. On the theory of the brownian motion ii. *Reviews of Modern Physics*, 17(2-3):323–342, 1945.

[YQSW25] Hengshi Yu, Weizhou Qian, Yuxuan Song, and Joshua D. Welch. Perturbnet predicts single-cell responses to unseen chemical and genetic perturbations. *Molecular Systems Biology*, 21(8):960–982, 2025.

[ZARX18] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.