RECLAIM: Cyclic Causal Discovery Amid Measurement Noise

Muralikrishnna G. Sethuraman Faramarz Fekri
School of Electrical and Computer Engineering, Georgia Institute of Technology

Abstract

Uncovering causal relationships is a fundamental problem across science and engineering. However, most existing causal discovery methods assume acyclicity and direct access to the system variables—assumptions that fail to hold in many real-world settings. For instance, in genomics, cyclic regulatory networks are common, and measurements are often corrupted by instrumental noise. To address these challenges, we propose RECLAIM, a causal discovery framework that natively handles both cycles and measurement noise. RECLAIM learns the causal graph structure by maximizing the likelihood of the observed measurements via expectation-maximization (EM), using residual normalizing flows for tractable likelihood computation. We consider two measurement models: (i) Gaussian additive noise, and (ii) a linear measurement system with additive Gaussian noise. We provide theoretical consistency guarantees for both the settings. Experiments on synthetic data and real-world protein signaling datasets demonstrate the efficacy of the proposed method.

1 Introduction

Understanding cause-effect relationships among variables is a central problem in science and engineering Sachs et al. (2005); Friedman et al. (2000). Causal models provide a mechanistic view of a system, enabling prediction of its behavior under unseen interventions. Causal relations are typically encoded as a directed graph (DG), where edges represent causal dependencies, and learning them is naturally posed as a graph structure learning problem.

Causal discovery methods broadly fall into three categories: (i) *constraint-based*, (ii) *score-based*, and (iii) *hybrid*. Constraint-based methods Spirtes et al. (2000); Triantafillou and Tsamardinos (2015) identify causal graphs consistent with the conditional independence constraints observed in data, but scale poorly as the number of required independence tests grows exponentially. Score-based methods Hauser and Bühlmann (2012) instead optimize a scoring function such as the Bayesian Information Criterion (BIC) over the space of graphs; greedy search is typically employed since the search space grows super-exponentially with the number of nodes. Hybrid methods Tsamardinos et al. (2006); Solus et al. (2021) combine elements of both approaches.

With few exceptions, most causal discovery methods rely on two key assumptions: (i) the underlying causal graph is acyclic (a directed acyclic graph, or DAG), and (ii) the system variables are directly observed. While these assumptions simplify the learning problem, they are often violated in practice. Variables of interest are frequently *latent*—for instance, a person's beliefs cannot be measured directly, but surveys provide noisy proxies of the underlying state. Furthermore, feedback loops are prevalent in biological systems Freimer et al. (2022), and such systems often exhibit structured measurement noise such as "dropout." Imposing

these assumptions in such settings can lead to misleading causal conclusions, limiting the practical applicability of existing methods.

To address these challenges, we propose RECLAIM, a novel causal discovery framework that learns latent causal structure while jointly modeling directed cycles and measurement noise. We consider two measurement noise models: (i) Gaussian additive noise, and (ii) a linear measurement system with Gaussian noise. Using interventional data, we first estimate the noise distribution parameters, then employ an expectation-maximization (EM) procedure to learn the causal graph via likelihood maximization.

1.1 Related Works

Cyclic causal discovery. The prior work on cyclic causal discovery spans across constraint-based (Richardson, 1996), ICA-based (Lacerda et al., 2012), and score-based (Huetter and Rigollet, 2020; Améndola et al., 2020) approaches. Several methods also leverage interventional data for structure recovery (Hyttinen et al., 2012; Huetter and Rigollet, 2020). Most relevant to our work, Sethuraman et al. (2023) proposed a differentiable framework for nonlinear cyclic graphs that models data likelihood directly, bypassing acyclicity constraints. Sethuraman and Fekri (2025) later extended this to account for latent confounders.

Causal discovery from indirect measurements. Under indirect measurements, causal discovery broadly splits into two paradigms. *Causal representation learning* (CRL) aims to jointly recover latent variables and their causal structure from high-dimensional observations under unknown transformations. Identifiability has been studied under parametric assumptions such as linear models (Squires et al., 2023; Buchholz et al., 2023) and polynomial transformations (Ahuja et al., 2024), as well as in fully nonparametric settings (von Kügelgen et al., 2023). *Causal discovery under measurement noise*, by contrast, assumes a known structural relationship between latents and observations, and focuses solely on recovering the causal graph over the latent variables. Harris and Drton (2013); Yoon et al. (2020) established consistency of the PC algorithm under Gaussian measurement error, and Saeed et al. (2020) extended this to a general constraint-based framework accommodating additive noise and dropouts. In the cyclic regime, Sethuraman et al. (2024) address dropout noise by treating zeroed-out variables as missing data.

1.2 Contributions

In this work, we address two central limitations in causal discovery: (i) inability to handle measurement error, and (ii) restriction to acyclic graphs. Our main contributions are:

- We introduce RECLAIM, a differentiable causal discovery framework that natively handles nonlinear cyclic relationships in the presence of additive Gaussian measurement error.
- We consider two measurement noise models: Gaussian additive noise and linear measurement systems, and provide a consistent estimator for the noise variance when appropriate interventions are available.
- We establish that exact maximization of the proposed score function under appropriate interventions identifies the interventional Markov equivalence class of the ground-truth graph.

- We conduct extensive experiments comparing RECLAIM against state-of-the-art causal discovery methods on both synthetic and real-world datasets, and show that increasing the number of measurements results in improved recovery performance.

1.2.1 Organization.

The remainder of the paper is organized as follows. Section 2 introduces the problem setup and the measurement models under consideration. Section 3 presents the technical details of RECLAIM. Section 4 evaluates its effectiveness on synthetic and real-world benchmarks. Section 5 concludes the paper.

2 Problem Setup

2.1 Structural Causal Model

Let $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ denote a directed graph with vertex set $\mathcal{X} = \{X_1, \ldots, X_d\}$ and edge set $\mathcal{E} \subseteq \mathcal{X} \times \mathcal{X}$. Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be the associated random (endogenous) variable. The edge $(X_i, X_j) \in \mathcal{E}$ denotes a directed edge originated from the node X_i to the node X_j, with slight abuse of notation we also denote this edge as $X_i \to X_j$. Following Bollen (1989); Pearl (2009), we utilize the framework of structural causal models (SCM) to represent the functional dependencies in the causal graph. For each $i \in \mathcal{X}$, we have the following *structural equation*:

$$X_i = f_i(\mathrm{pa}_{\mathcal{G}}(X_i)) + Z_i, \tag{1}$$

where f_i, called the *causal mechanism*, encodes the functional dependency between the parents and the children. $\mathrm{pa}_{\mathcal{G}}(X_i) \triangleq \{X_j \in \mathcal{X} \mid X_j \to X_i \in \mathcal{E}\}$ denotes the *parent set* of a node $X_i \in \mathcal{X}$. The set of variables $\boldsymbol{Z} = (Z_1, \ldots, Z_d)$ represent the *exogenous noise* within the system. We assume that the exogenous noise variables are independent of each other, i.e., $Z_i \perp Z_j$ for $i, j \in [d]$ and $i \neq j$. This assumption, also known as *causal sufficiency*, excludes hidden confounding and selection bias. Let $f = (f_1, \ldots, f_d)$ be the combined causal mechanism, combining eq. (1) for all $i \in [d]$, we obtain the following:

$$\boldsymbol{X} = f(\boldsymbol{X}) + \boldsymbol{Z}. \tag{2}$$

However, it is not guaranteed that eq. (2) has a solution, primarily due to the (potential) presence of cycles in $\mathcal{G}$. As a result, restrictions are necessary on the causal mechanism to ensure the system attains equilibrium (see appendix A for a detailed discussion). To that end, following Sethuraman et al. (2023), we make the following assumption on the map $(\mathrm{id} - f) : \boldsymbol{X} \mapsto \boldsymbol{Z}$, which we call the *forward map*.

Assumption 1. *The forward map* $(id - f) : \boldsymbol{X} \mapsto \boldsymbol{Z}$ *is a* diffeomorphism[1].

Finally, the SCM provides us with a probability density for the exogenous noise variables, which we denote as p_Z. Given p_Z, we can obtain the probability density of the endogenous variables using the change of variable property of probability densities. That is,

$$p_{\mathcal{G}}(\boldsymbol{X}) = p_Z\big((\mathrm{id} - f)(\boldsymbol{X})\big)\big|J_{(\mathrm{id}-f)}(\boldsymbol{Z})\big|, \tag{3}$$

where $J_{(\mathrm{id}-f)}(\boldsymbol{X})$ is the Jacobian matrix of the forward map evaluated at $\boldsymbol{X}$.

[1] A function f is a diffeomorphism if f^{-1} exists, and both f and f^{-1} are differentiable.

Interventions. We consider *surgical interventions* Pearl (2009) within our framework. Under surgical interventions, when a subset of nodes $\mathcal{X}_I \subseteq \mathcal{X}$ are (surgically) intervened, we obtain a mutilated graph, denoted as $\mathrm{do}(I)(\mathcal{G})$, where all the incoming edges to the intervened nodes are removed (see fig. 1). We identify an intervention by the index set $I \subseteq [d]$ of the nodes intervened in $\mathcal{X}$. Given a set of intervened nodes $\mathcal{X}_I \subseteq \mathcal{X}$, let $\mathbf{U} \in \{0,1\}^{d\times d}$ be a diagonal masking matrix with $U_{kk} = 1$ if $X_k \in \mathcal{X}_I$, and 0 otherwise. The structural equations in eq. (2) are then modified as follows:

$$\boldsymbol{X} = \mathbf{U}f(\boldsymbol{X}) + \mathbf{U}\boldsymbol{Z} + \boldsymbol{C}, \tag{4}$$

where $\boldsymbol{C} \in \mathbb{R}^d$ denotes the values of the intervened nodes. Consequently, under the intervention I, the probability density of the observations is given by

$$p_{\mathrm{do}(I)(\mathcal{G})}(\boldsymbol{X}) = p_C(\boldsymbol{X}_I)p_Z\big(\boldsymbol{Z}_{[d]\setminus I}\big)\big|J_{(\mathrm{id}-\mathbf{U}f)}(\boldsymbol{Z})\big|, \tag{5}$$

where p_C denotes the interventional density, and $\boldsymbol{Z}_{[d]\setminus I} = \big[(\mathrm{id} - \mathbf{U}f)(\boldsymbol{X})\big]_{[d]\setminus I}$. In our experiments, we assume that the intervened nodes are sampled from a distribution with known variance $\sigma_{\mathcal{I}}^2$.

2.2 Measurement System

In practice, the causal variables $\mathcal{X}$ are latent and only indirectly observed; they pass through a measurement system, yielding *measured* variables $\mathcal{Y} = \{Y_1, \ldots, Y_p\}$ (with $\boldsymbol{Y}$ being the corresponding random variable). Each Y_j represents a coarsened measurement of $\boldsymbol{X}$ given by,

$$Y_j = g_j(\boldsymbol{X}, \varepsilon_j), \quad j = 1, \ldots, p, \tag{6}$$

where $g = (g_1, \ldots, g_p)$ denotes the *coarsening mechanism* (also referred as the *measurement channel*), and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_d)$ represents any external factor that can affect the measurement system. We assume that $p \geq d$ (otherwise, identifiability is ill-posed), and $\varepsilon_i \perp \varepsilon_j$, for $i, j \in [p]$ and $i \neq j$. This results in a combined measurement graph, denoted as $\mathcal{G}_m$, whose vertex set contains both the latent variables $\mathcal{X}$ and the measured variables $\mathcal{Y}$, i.e., $\mathcal{X} \cup \mathcal{Y}$, see fig. 1.

In our work, we deal with two different types of measurement system:

1. **Gaussian additive noise**: In this case, $p = d$, $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, and

$$Y_i = X_i + \varepsilon_i, \quad i = 1, \ldots, d. \tag{7}$$

   Given the latent variables $\boldsymbol{X}$, the density of the measurements is also Gaussian, i.e., $\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}, \mathbf{D}_\varepsilon)$, where $\mathbf{D}_\varepsilon = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_d^2)$. We refer to $p_{Y|X}(\boldsymbol{Y} \mid \boldsymbol{X})$ as the *measurement channel density*.

2. **Linear measurement system with Gaussian noise**: Similar to the previous setting, we have $\varepsilon_j \sim \mathcal{N}(\mathbf{0}, \sigma_j^2)$. However, in this case, p can be larger than d, i.e., $p \geq d$, and

$$Y_j = \sum_{i=1}^{d} A_{ji} X_i + \varepsilon_j. \tag{8}$$

   The matrix $\mathbf{A} = [A_{ji}]$, with $i \in [d]$ and $j \in [p]$, is referred to as the *measurement matrix*, and is assumed to be known and full (column) rank. Given the latent variables $\boldsymbol{X}$, the conditional density of $\boldsymbol{Y}$ is again Gaussian, i.e., $\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x} \sim \mathcal{N}(\mathbf{A}\boldsymbol{x}, \mathbf{D}_\varepsilon)$, where $\mathbf{D}_\varepsilon = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_p^2)$.

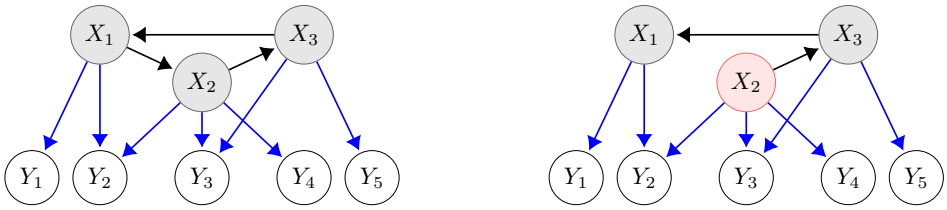


Figure 1: (a) Illustration of a causal graph $\mathcal{G}_m$ encoding both the system variables $\mathcal{X}$ and the measured variables $\mathcal{Y}$ for a linear measurement system. (b) Mutilated graph, $\text{do}(I)(\mathcal{G}_m)$, resulting from a surgical intervention on X_2.

Goal. Given a set of interventions $\mathcal{I} = \{I_k\}_{k=1}^K$, *our goal is to learn structure of* $\mathcal{G}$ *by maximizing the log-likelihood of the coarsened measurements.* We present our approach towards addressing this problem in the next section.

3 RECLAIM Causal Discovery Framework

Let $\boldsymbol{\theta}, \boldsymbol{\phi}$ represent the parameters of the latent variable density $p_{\mathcal{G}}(\boldsymbol{X})$ and the conditional measurement density $p_{Y|X}(\boldsymbol{Y} \mid \boldsymbol{X})$, respectively. For an interventional experiment $I \subseteq [d]$, we aim to learn the causal graph structure $\mathcal{G}$ by maximizing the likelihood of the observed variables. However, this maximization is generally intractable, since

$$p(\boldsymbol{Y} \mid \boldsymbol{\theta}, \boldsymbol{\phi}, I) = \int p_{\text{do}(I)(\mathcal{G})}(\boldsymbol{X} \mid \boldsymbol{\theta})\, p_{Y|X}(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\phi})\, d\boldsymbol{X}$$

does not admit a closed-form solution in general.

To address this intractability, RECLAIM adopts an EM-based approach that maximizes a tractable lower bound on $\log p(\boldsymbol{Y} \mid \boldsymbol{\theta}, \boldsymbol{\phi})$. We now describe the framework in detail, beginning with defining the score function to be maximized, followed by an overview of the EM-based optimization procedure. We then detail the technical components of the expectation step: estimating the measurement process parameters, computing the latent likelihood, and sampling latents conditioned on the observations. Finally, we establish the theoretical guarantees of RECLAIM, showing that exact maximization of the score function recovers the true causal structure up to Markov equivalence.

3.1 Score Function

Given a family of interventions $\mathcal{I} = \{I_k\}_{k=1}^K$ and a causal graph $\mathcal{G}$, we define the *score function* as the regularized log-likelihood of the observed variables under each interventional regime. Following Sethuraman and Fekri (2025), this takes the form:

$$\mathcal{S}_{\mathcal{I}}(\mathcal{G}) \triangleq \sup_{\boldsymbol{\theta}, \boldsymbol{\phi}} \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{Y} \sim p^{(k)}} \log p(\boldsymbol{Y} \mid \boldsymbol{\theta}, \boldsymbol{\phi}, I_k) - \lambda |\mathcal{G}|, \tag{9}$$

where $p^{(k)}$ denotes the data-generating density under the k-th interventional experiment I_k, and $\lambda|\mathcal{G}|$ is an ℓ_0-type sparsity penalty on the edge set of $\mathcal{G}$.

In practice, we assume access to a collection of finite samples per interventional experiments, $\mathcal{D}_{\mathcal{I}} = \{\{\boldsymbol{y}^{(k,\ell)}\}_{\ell=1}^{n_k}\}_{k=1}^K$. Thus, the expectation in eq. (9) is replaced with sample mean.

Finally, we maximize the score function over the space of graphs $\mathcal{G}$ resulting in the following score function:

$$\tilde{\mathcal{S}}_{\mathcal{I}}(\mathcal{D}_{\mathcal{I}}) \triangleq \sup_{\mathcal{G},\boldsymbol{\theta},\boldsymbol{\phi}} \sum_{k=1}^{K} \sum_{\ell=1}^{n_k} \log p(\boldsymbol{y}^{(k,\ell)} \mid \boldsymbol{\theta}, \boldsymbol{\phi}, I_k) - \lambda |\mathcal{G}|. \tag{10}$$

3.2 Optimizing The Score Via Penalized Expectation-Maximization

As discussed earlier, the intractability of eq. (10) stems from integrating out the latent variables, precluding a closed-form solution. To circumvent this, we adopt the penalized EM framework of Chen et al. (2014). Beginning from an initial guess $\boldsymbol{\theta}^0$ for the latent SCM parameters (encompassing both the neural network weights and the graph adjacency), the algorithm alternates between the two steps below until a convergence criterion is met:

E-step: Given the current parameter estimate $\boldsymbol{\theta}^t$ and the observations, form the surrogate objective $\mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t)$ as the expected complete-data log-likelihood:

$$\mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t) = \sum_{k=1}^{K} \sum_{\ell=1}^{n_k} \mathbb{E}_{\boldsymbol{X} \sim p(\cdot \mid \boldsymbol{y}^{(k,\ell)}, \boldsymbol{\theta}^t, \hat{\boldsymbol{\phi}}, I_k)} \Big[\log p\big(\boldsymbol{X}, \boldsymbol{y}^{(k,\ell)} \mid \boldsymbol{\theta}, \hat{\boldsymbol{\phi}}, I_k\big) \Big]. \tag{11}$$

where $\hat{\boldsymbol{\phi}}$ denotes the estimated parameters of the measurement channel (see section 3.3 for details), and

$$\log p\big(\boldsymbol{X}, \boldsymbol{y}^{(k,\ell)} \mid \boldsymbol{\theta}, \hat{\boldsymbol{\phi}}, I_k\big) = \log p_{\mathrm{do}(I_k)(\mathcal{G})}(\boldsymbol{X} \mid \boldsymbol{\theta}) + \log p\big(\boldsymbol{y}^{(k,\ell)} \mid \boldsymbol{X}, \hat{\boldsymbol{\phi}}\big). \tag{12}$$

M-step: Update the parameters by solving the penalized maximization of the surrogate:

$$\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} \; \mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t) - \lambda \mathcal{R}(\boldsymbol{\theta}), \tag{13}$$

where $\mathcal{R}(\boldsymbol{\theta})$ is a sparsity-promoting regularizer on the graph.

The maximization in the M-step is carried out with stochastic gradient ascent. A key observation is that the surrogate $\mathcal{Q}$ serves as a lower bound (up to a constant) on the marginal log-likelihood:

$$\sum_{k=1}^{K} \sum_{\ell=1}^{n_k} \log p\Big(\boldsymbol{y}^{(k,\ell)} \mid \boldsymbol{\theta}, \hat{\boldsymbol{\phi}}, I_k\Big) \;\geq\; \mathcal{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t) - \text{const}. \tag{14}$$

Consequently, each M-step improves a valid lower bound on the observed-data log-likelihood rather than optimizing it directly. A formal derivation of eq. (14) and a convergence analysis are deferred to appendix B.

3.3 Estimation of Measurement System Parameters

Computing the surrogate objective $\mathcal{Q}$ in eq. (11) requires evaluating the measurement channel density, whose parameters are a priori unknown and must therefore be estimated from data. In the two measurment processes under consideration (Gaussian additive noise and linear measurement system), the unknown parameters correspond to the variances of the additive noise, which fully identify the measurement channel. Since the latent variables $\boldsymbol{X}$ are not directly observed, estimating these parameters requires additional structure on the data-generating mechanism. We formalize this as the following condition on the intervention family $\mathcal{I}$.

Condition 2 (Measurement Channel Identifiability)**.** *An intervention family* $\mathcal{I} = \{I_k\}_{k=1}^K$ *is* measurement channel identifiable *if, for each node* $X_i \in \mathcal{X}$*, there exists an intervention* $I_k \in \mathcal{I}$ *such that* $X_i \in \mathcal{X}_{I_k}$*.*

Intuitively, condition 2 requires that every latent variable is targeted by at least one intervention, ensuring that its variance is known under that interventional regime and can therefore be used to identify the measurement noise. We now show that this condition is sufficient to identify the measurement parameters for each of the three noise models under consideration.

3.3.1 Gaussian Additive Noise.

Since $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$ are independent, the marginal variance of Y_i decomposes as $\sigma_{Y_i}^2 = \sigma_{X_i}^2 + \sigma_i^2$, where $\sigma_{Y_i}^2$ can be estimated directly from observations. While $\sigma_{X_i}^2$ is generally unknown under observational data, condition 2 guarantees the existence of an intervention $I_k \in \mathcal{I}$ under which X_i is intervened upon, fixing its variance to the known interventional variance $\sigma_{\mathcal{I}}^2$. The measurement noise variance is then recovered as $\sigma_i^2 = \sigma_{Y_i}^2 - \sigma_{\mathcal{I}}^2$.

3.3.2 Linear Measurement System with Gaussian Noise.

In this setting, each observed variable Y_j is a linear mixture of all latent variables (cf. eq. (8)), making direct variance decomposition infeasible. Notably, condition 2 does not require a single intervention targeting all nodes simultaneously. Instead, we exploit the structure of the measurement matrix $\mathbf{A}$ via the following proposition.

Proposition 3. *Let* $\mathbf{A} \in \mathbb{R}^{p \times d}$ *be the measurement matrix with* $\operatorname{rank}(\mathbf{A}) = d$*, let* $\boldsymbol{a}_i$ *denote the* i*-th column of* $\mathbf{A}$*, and let* $\mathbf{A}_{-i}$ *denote the measurement matrix excluding the* i*-th column. Then, there exists a vector* $\boldsymbol{t} \in \mathbb{R}^p$ *such that* $\mathbf{A}_{-i}^\top \boldsymbol{t} = \mathbf{0}$ *and* $\boldsymbol{a}_i^\top \boldsymbol{t} \neq 0$*.*

We defer the proof to appendix C.1. The key idea is to project the observations onto a direction $\boldsymbol{t}_i$ that isolates the contribution of X_i. Specifically, choosing $\boldsymbol{t}_i$ such that $\mathbf{A}_{-i}^\top \boldsymbol{t}_i = \mathbf{0}$ and $\boldsymbol{a}_i^\top \boldsymbol{t}_i \neq 0$, and letting $\zeta_i = \boldsymbol{t}_i^\top \boldsymbol{Y}$, we obtain:

$$\zeta_i = \boldsymbol{t}_i^\top \boldsymbol{Y} = \boldsymbol{t}_i^\top \mathbf{A} \boldsymbol{X} + \boldsymbol{t}_i^\top \boldsymbol{\varepsilon} = (\boldsymbol{t}_i^\top \boldsymbol{a}_i) X_i + \boldsymbol{t}_i^\top \boldsymbol{\varepsilon}.$$

Since $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$ are independent, the variance of ζ_i decomposes as:

$$\sigma_{\zeta_i}^2 = (\boldsymbol{t}_i^\top \boldsymbol{a}_i)^2 \sigma_{X_i}^2 + \boldsymbol{t}_i^\top \mathbf{D}_\varepsilon \boldsymbol{t}_i. \tag{15}$$

Condition 2 guarantees the existence of an intervnetion I_k containing X_i, and $\sigma_{X_i}^2 = \sigma_{\mathcal{I}}^2$. Letting $\boldsymbol{t}_i^{\odot} \triangleq \boldsymbol{t}_i \odot \boldsymbol{t}_i$, $b_i = \sigma_{\zeta_i}^2 - (\boldsymbol{t}_i^\top \boldsymbol{a}_i)^2 \sigma_{\mathcal{I}}^2$, and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_p^2)$, eq. (15) reduces to the linear equation $(\boldsymbol{t}_i^{\odot})^\top \boldsymbol{\sigma}^2 = b_i$. Aggregating sufficiently many such projection vectors into a matrix $\mathbf{T}_2 = [(\boldsymbol{t}^{(k)})_k^\top]$, the noise variances are recovered by solving the convex least-squares problem:

$$\min_{\boldsymbol{\sigma}^2 \geq 0} \left\| \mathbf{T}_2 \boldsymbol{\sigma}^2 - \boldsymbol{b} \right\|_2^2. \tag{16}$$

This approach requires $\mathbf{T}_2$ to have rank p, which may not always be achievable. However, the following theorem shows that this is a generic property of the measurement matrix.

Theorem 4. *Let* $\mathbf{A} \in \mathbb{R}^{p \times d}$ *with* $p \geq d$ *and* $\operatorname{rank}(\mathbf{A}) = d$*. For each* $i \in [d]$*, let* $S_i(\mathbf{A}) \triangleq \mathcal{N}(\mathbf{A}_{-i}^\top)$*, and define the admissible projection set as*

$$\mathcal{T}(\mathbf{A}) \triangleq \bigcup_{i=1}^{d} \{\boldsymbol{t} \in S_i(\mathbf{A}) : \boldsymbol{a}_i^\top \boldsymbol{t} \neq 0\}.$$

Let $\bar{\mathcal{A}}$ denote the set of full-rank matrices $\mathbf{A}$ for which no choice of $\boldsymbol{t}^{(1)}, \ldots, \boldsymbol{t}^{(p)} \in \mathcal{T}(\mathbf{A})$ yields $\operatorname{rank}(\mathbf{T}_2) = p$. Then $\bar{\mathcal{A}}$ has zero Lebesgue measure.

We defer the proof to appendix C.2. The projection vectors $\boldsymbol{t}$'s are sampled to construct $\mathbf{T}_2$ using the procedure detailed in algorithm D.1 in appendix D.1.

3.4 Computing Likelihood of Latent Variables

From eqs. (11) and (12), computing the surrogate loss function $\mathcal{Q}$ requires evaluating log likelihood of the latent variables, $\log p_{\mathrm{do}(I_k)(\mathcal{G})}(\boldsymbol{X} \mid \boldsymbol{\theta})$. We describe how this is computed by detailing the SCM parameterization and the associated log-likelihood.

Modeling the Causal Mechanism. We model each causal mechanism f_i in eq. (1) using a contractive neural network, i.e., a neural network with Lipschitz constant less than one, which is enforced during training by rescaling layer weights by their spectral norm Behrmann et al. (2019). Contractivity guarantees, via the Banach fixed point theorem Rudin (1953a), that the map $(\mathrm{id} - \mathbf{U}f)$ is invertible (ensuring that assumption 1 is satisfied) and the log-determinant of its Jacobian is well-defined.

To prevent spurious self-loops and encourage sparsity, we introduce a dependency mask $\mathbf{M} \in \{0,1\}^{d \times d}$ with zero diagonal, whose entries are treated as Bernoulli random variables sampled via the Gumbel-softmax distribution Jang et al. (2016), $\mathbf{M} \sim p_M(\mathbf{M} \mid \boldsymbol{\theta})$. The causal mechanism then takes the form:

$$[\mathrm{F}_{\boldsymbol{\theta}}(\boldsymbol{X})]_i = [\mathrm{NN}_{\boldsymbol{\theta}}(\mathbf{M}_{*,i} \odot \boldsymbol{X})]_i, \tag{17}$$

where $\mathbf{M}_{*,i}$ denotes the i-th column of $\mathbf{M}$. and the sparsity penalty is $\mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{M} \sim p_M(\cdot \mid \boldsymbol{\theta})} \|\mathbf{M}\|_1$.

Log-Determinant of the Jacobian. Given the contractive neural network parameterization in eq. (17), the log-likelihood in eq. (11) requires evaluating the log-determinant of the Jacobian of $(\mathrm{id} - \mathbf{U}f)$. Naïve computation of the log-determinant of the Jacobian function has a computation cost that is $\mathcal{O}(d^3)$. However, since the causal mechanism f is contractive, the log-determinant admits the convergent power series expansion Behrmann et al. (2019):

$$\log \left| \det J_{(\mathrm{id} - \mathbf{U}f)}(\boldsymbol{X}) \right| = -\sum_{m=1}^{\infty} \frac{1}{m} \mathrm{Tr}\Big\{ J^m_{\mathbf{U}f}(\boldsymbol{X}) \Big\}. \tag{18}$$

The terms in the power series only depend on the trace of the powers of the Jacobian matrix, and hence reduces the computation cost to $\mathcal{O}(d^2)$. However, Truncating this series introduces bias, which we correct using the Russian roulette estimator Chen et al. (2019), and reduce the per-iteration cost using the Hutchinson trace estimator Hutchinson (1989). This results in the final unbiased estimator given by eq. (29) in appendix D.2. We defer the details to appendix D.2.

3.5 Sampling Latents Given the Observations

With the measurement parameters $\hat{\boldsymbol{\phi}}$ estimated and the latent likelihood in hand, the remaining challenge in evaluating $\mathcal{Q}$ is computing the posterior expectation over $\boldsymbol{X}$. Since the nonlinear causal mechanisms preclude a closed-form posterior, we employ *Sampling Importance Resampling* (SIR) Smith and Gelfand (1992) to draw approximate posterior samples.

We draw S proposal samples $\{\boldsymbol{x}^{(s)}\}_{s=1}^{S}$ from a Gaussian proposal $q(\boldsymbol{X} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}_y, \mathbf{D}_\varepsilon)$, where the proposal mean $\boldsymbol{\mu}_y$ depends on the measurement model:

$$\boldsymbol{\mu}_y = \begin{cases} \boldsymbol{y} & \text{additive Gaussian and dropout,} \\ (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \boldsymbol{y} & \text{linear measurement system.} \end{cases} \tag{19}$$

Each sample is then assigned an importance weight proportional to:

$$w^{(s)} \propto \frac{p(\boldsymbol{x}^{(s)} \mid \boldsymbol{\theta})\, p(\boldsymbol{y} \mid \boldsymbol{x}^{(s)}, \boldsymbol{\theta})}{q(\boldsymbol{x}^{(s)} \mid \boldsymbol{y})}, \tag{20}$$

and a final set of samples is obtained by resampling from $\{\boldsymbol{X}^{(s)}\}_{s=1}^{S}$ according to the normalized weights $\{\tilde{w}^{(s)}\}_{s=1}^{S}$.

3.6 Consistency Under Exact Maximization

Having established the practical components of RECLAIM, we now turn to its theoretical justification. The following theorem shows that exact maximization of the score function in eq. (9) recovers the true causal structure up to $\mathcal{I}$-Markov equivalence. Specifically, the graph $\hat{\mathcal{G}}$ obtained by maximizing eq. (9) lies in the same general Markov equivalence class as the ground-truth graph $\mathcal{G}^*$ Bongers et al. (2021) across all interventional settings $I_k \in \mathcal{I}$.

Theorem 5. *Let $\mathcal{I} = \{I_k\}_{k=1}^{K}$ be a family of interventional targets satifying condition 2, let $\mathcal{G}^*$ denote the ground truth directed graph, $p^{(k)}$ denote the data generating distribution for $I_k \in \mathcal{I}$, and $\hat{\mathcal{G}} := \arg\max_{\mathcal{G}} \mathcal{S}(\mathcal{G})$. Then, under assumptions C.8 to C.11 and C.13, and for a suitably chosen $\lambda > 0$, we have that $\hat{\mathcal{G}} \equiv_{\mathcal{I}} \mathcal{G}^*$. That is, $\hat{\mathcal{G}}$ is $\mathcal{I}$-Markov equivalent to $\mathcal{G}^*$.*

Here, we provide an overview of the key assumptions needed for theorem 5, as well as proof sketch below. See appendix C.3 for the full assumptions as well as the complete proof. Assumption C.8 ensures that the data-generating distribution lies within the model class, assumption C.9 guarantees that all the statistical independencies observed in the data is a result of σ-separation in the data generating graph (Forré and Mooij, 2017). Assumptions C.10 and C.11 prevent the score from diverging to infinity. Finally, assumption C.13 ensures that the map from latent distribution to the observed distribution is injective.

Proof (sketch). Building on the characterization of general directed Markov equivalence class by Bongers et al. (2021), extended to the interventional setting, we show that any graph outside this equivalence class has a strictly lower score than the ground truth graph $\mathcal{G}^*$. This follows from the following two facts: (i) certain independencies present in the data are not captured by graphs outside the equivalence class, and (ii) each latent distribution results in a unique observed distribution. Combined with the expressiveness of the model class, this prevents such graphs from fitting the data properly. □

4 Experiments

We evaluate RECLAIM on both synthetic and real-world datasets against three state-of-the-art baselines: NODAGS-Flow Sethuraman et al. (2023), DCDI Brouillard et al. (2020), and Anchored-CI Saeed et al. (2020). Together, these baselines span the key axes of comparison—NODAGS-Flow handles cycles and interventions but not measurement error; DCDI handles interventions but assumes acyclicity and ignores measurement error; and Anchored-CI models measurement error but assumes acyclicity and operates on observational data only. Since Anchored-CI returns a CPDAG representing a Markov equivalence

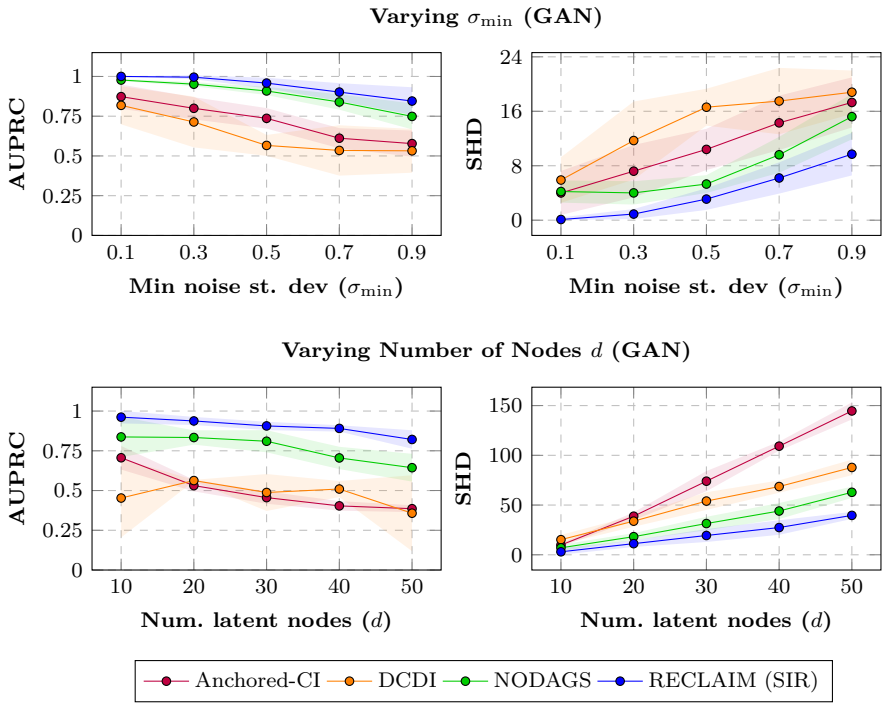


Figure 2: Performance comparison with varying minimum noise standard deviation (σ) (top), and varying number of latent nodes (d) (bottom) for *Gaussian additive noise* system. Shaded regions show ± 1 standard deviation over 10 trials.

class, we evaluate it on observational data and report the best score across all graphs in the equivalence class.

4.1 Synthetic Experiments

In all our synthetic experiments, cyclic graphs were generated using the Erdős-Rényi (ER) random graph model with the expected outgoing edge density set to 2. RECLAIM and the baselines were evaluated on nonlinear SCMs with the latent variables generated using the following structural equations:

$$\boldsymbol{x} = \tanh\big(\mathbf{W}^\top \boldsymbol{x}\big) + \boldsymbol{z},$$

where the weighted adjacency matrix is faithful to the graph generated using the ER model, and $0.2 \leq |W_{ij}| \leq 0.9$. The matrix $\mathbf{W}$ was rescaled to ensure that the causal mechanism remains contractive. The training dataset consists of observational data and single node interventions over all the nodes in the graph, i.e., $\mathcal{I} = \emptyset \cup \{\{i\}\}_{i=1}^{d}$. For each interventional experiment $I_k \in \mathcal{I}$, the intervened nodes $X_i \in \mathcal{X}_{I_k}$ were sampled from $\mathcal{N}(0, 1)$, with $N_k = 1000$ samples per intervention.

We report performance using two metrics: AUPRC (higher is better) and SHD (lower is better). AUPRC summarizes edge recovery quality across thresholds, while SHD counts the number of edge additions, deletions, and reversals needed to recover the ground truth graph. Since SHD requires a binary adjacency matrix, we threshold the estimated edge probabilities at 0.8.

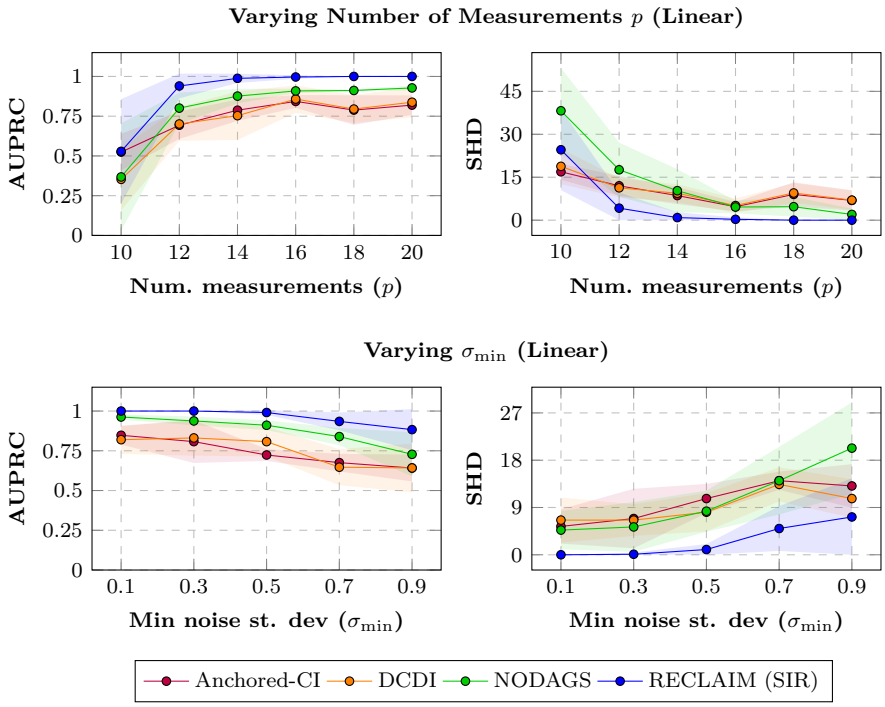


Figure 3: Performance comparison with varying number of measurements (p) (top), and varying minimum noise standard deviation (σ) (bottom) for *Linear measurement* system. Shaded regions show ± 1 standard deviation over 10 trials.

4.1.1 Gaussian Additive Noise.

We now consider the Gaussian additive noise setting, where measurements are corrupted as described in eq. (7). We evaluate all methods as a function of the minimum measurement noise scale $\sigma_{\min}$ and the number of latent nodes d, probing robustness and scalability respectively, with results shown in fig. 2.

Sensitivity to measurement noise scale. We fix $d = 10$ and vary $\sigma_{\min}$ between 0.1 and 0.9, with $\sigma_{\max} = \sigma_{\min} + 0.3$. As shown in fig. 2 (top), all methods degrade as $\sigma_{\min}$ increases, but RECLAIM degrades more gracefully, the performance gap over the baselines widens with noise scale, highlighting the benefit of explicitly modeling measurement error. NODAGS-Flow matches the performance of RECLAIM when $\sigma_{\min} = 0.1$ as the influence of noise is very low at that scale.

Impact of number of latent nodes. We fix $\sigma_{\min} = 0.5$ and $\sigma_{\max} = 1$, and vary d between 10 and 50. As shown in fig. 2 (bottom), performance degrades for all methods as d increases, but the gap between cyclic methods (RECLAIM and NODAGS-Flow) and acyclic methods (DCDI and Anchored-CI) widens substantially, suggesting that the cost of misspecifying graph structure compounds with graph size.

4.1.2 Linear Measurement System.

We now consider the linear measurement system setting, where measurements follow eq. (8) with $A_{ij} \sim \mathcal{N}(0, 1.5)$. We evaluate all methods as a function of the number of measurements

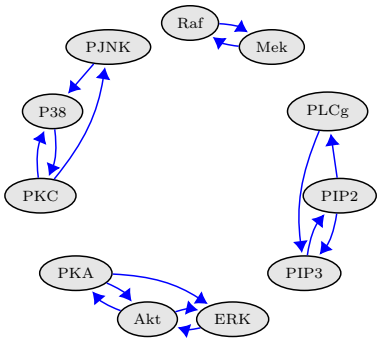


Figure 4: Estimated graph learnt from Sachs et al. Sachs et al. (2005) dataset

p and the minimum noise scale $\sigma_{\min}$, probing the effect of measurement redundancy and noise robustness respectively, with results shown in fig. 3.

Impact of number of measurements. We fix $d = 10$, $\sigma_{\min} = 0.3$, $\sigma_{\max} = 0.6$, and vary p between 10 and 20. As shown in fig. 3 (top), all methods perform poorly at $p = d = 10$, but RECLAIM recovers sharply as p increases, achieving near-perfect recovery for $p \geq 12$. This suggests that even modest measurement redundancy is sufficient for RECLAIM to disentangle the latent structure, whereas the baselines continue to struggle across the full range of p.

Sensitivity to measurement noise scale. We fix $p = 15, d = 10$ and vary $\sigma_{\min}$ between 0.1 and 0.9, with $\sigma_{\max} = \sigma_{\min} + 0.3$. As shown in fig. 3 (bottom), the overall trend mirrors the Gaussian additive noise setting (all methods degrade with increasing noise), but RECLAIM degrades more gracefully than the baselines. Notably, RECLAIM is more robust here than in the Gaussian additive noise setting (AUPRC of 0.88 vs 0.84 at $\sigma_{\min} = 0.9$), which we attribute to the additional structure provided by the linear measurement system allowing for more accurate noise parameter estimation.

4.2 Real-World Protein Signaling Dataset

We evaluate RECLAIM and the baselines on the Sachs et al. Sachs et al. (2005) protein signaling dataset, which consists of fluorescence intensity measurements of phosphorylated proteins and phospholipid components in human immune system cells across 13 interventional environments. These measurements are noisy proxies for true protein levels, corrupted by instrument and antibody staining noise Krutzik et al. (2004). Importantly, 11 proteins and phospholipids are measured, making this a natural testbed for causal discovery under measurement noise.

Table 1: Performance comparison with respect to SHD on Sachs et al. Sachs et al. (2005) dataset

Method	SHD
RECLAIM	19
NODAGS	22
DCDI	21
Anchored-CI	19

We train RECLAIM on the first 9 interventional environments. Since the true variance of the intervened proteins is unavailable, we treat the measurement noise variance as a learnable parameter. The estimated signaling network after 200 epochs is shown in fig. 4, with SHD scores relative to the Sachs et al. Sachs et al. (2005) ground-truth reported in table 1.

The Raf↔Mek cycle recovered by RECLAIM is consistent with the well-documented negative feedback from ERK to Raf-1 Sturm et al. (2011), which DAG-constrained methods structurally cannot recover. We note that the ground-truth graph from Sachs et al Sachs et al. (2005) is itself a DAG, despite feedback loops being known to exist in these systems Sturm et al. (2011)—meaning our SHD scores are conservative with respect to RECLAIM's true recovery performance.

Additional experiments. Additionally, we also provide results in appendix E for the following settings: (i) varying number of cycles in latent graph, (ii) varying degree on nonlinearity of latent SCM, and (iii) varying sparsity of the latent graph.

5 Discussion and Conclusion

In this work, we introduced RECLAIM, a differentiable causal discovery framework that jointly handles cyclic causal graphs, interventional data, and measurement noise in a unified probabilistic framework. Unlike existing methods that assume acyclicity or direct observation of system variables, RECLAIM operates on coarsened measurements across two noise settings: additive Gaussian, and linear measurement systems. We establish consistency of the graph estimator in the large-sample regime and demonstrate through experiments that RECLAIM outperforms state-of-the-art baselines, with particular gains in cyclic and high-noise regimes.

A key limitation is the reliance on interventional data for measurement noise parameter estimation—insufficient interventional coverage can degrade both noise estimation and graph recovery. Future directions include relaxing this requirement, extending support to soft and unknown interventional targets, non-Gaussian exogenous noise, and jointly learning the measurement system from data.

Acknowledgment

This material is based on work supported by the National Science Foundation (NSF) under Grant no. 2502298.

References

Kartik Ahuja, Amin Mansouri, and Yixin Wang. Multi-domain causal representation learning via weak distributional invariances. In *International Conference on Artificial Intelligence and Statistics*, pages 865–873. PMLR, 2024.

Carlos Améndola, Philipp Dettling, Mathias Drton, Federica Onori, and Jun Wu. Structure learning for cyclic linear causal models. In *Conference on Uncertainty in Artificial Intelligence*, pages 999–1008. PMLR, 2020.

Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *International Conference on Machine Learning*, pages 573–582. PMLR, 2019.

P Billingsley. Probability and measure, anniversary edition, wiley, hoboken. *New Jersey*, 1979(1986):2012, 1995.

Kenneth A Bollen. *Structural equations with latent variables*. John Wiley & Sons, 1989.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *Advances in Neural Information Processing Systems*, 36: 45419–45462, 2023.

Lin S Chen, Ross L Prentice, and Pei Wang. A penalized em algorithm incorporating missing data mechanism for gaussian parameter estimation. *Biometrics*, 70(2):312–322, 2014.

Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019.

Gerald B Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 1999.

Patrick Forré and Joris M Mooij. Markov properties for graphical models with cycles and latent variables. *arXiv preprint arXiv:1710.08775*, 2017.

Jacob W Freimer, Oren Shaked, Sahin Naqvi, Nasa Sinnott-Armstrong, Arwa Kathiria, Christian M Garrido, Amy F Chen, Jessica T Cortez, William J Greenleaf, Jonathan K Pritchard, et al. Systematic discovery and perturbation of regulatory genes in human t cells reveals the architecture of immune networks. *Nature Genetics*, 54(8):1133–1144, 2022.

Nir Friedman. The bayesian structural em algorithm. In *Conference on Uncertainty in Artificial Intelligence*, 1998. URL https://api.semanticscholar.org/CorpusID:447055.

Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. In *Proceedings of the fourth annual international conference on Computational molecular biology*, pages 127–135, 2000.

Naftali Harris and Mathias Drton. Pc algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(11), 2013.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1):2409–2464, 2012.

Jan-Christian Huetter and Philippe Rigollet. Estimation rates for sparse linear cyclic causal models. In *Conference on Uncertainty in Artificial Intelligence*, pages 1169–1178. PMLR, 2020.

Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. Learning linear cyclic causal models with latent variables. *The Journal of Machine Learning Research*, 13(1):3387–3439, 2012.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Peter O Krutzik, Jonathan M Irish, Garry P Nolan, and Omar D Perez. Analysis of protein phosphorylation and cellular signaling events by flow cytometry: techniques and clinical applications. *Clinical immunology*, 110(3):206–221, 2004.

Gustavo Lacerda, Peter L Spirtes, Joseph Ramsey, and Patrik O Hoyer. Discovering cyclic causal models by independent components analysis. *arXiv preprint arXiv:1206.3273*, 2012.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of machine learning research*, 21(99):1–108, 2020.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Thomas Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pages 454–461, 1996.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill Book Company, Inc., New York-Toronto-London, 1953a.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill Book Company, Inc., New York-Toronto-London, 1953b.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Basil Saeed, Anastasiya Belyaeva, Yuhao Wang, and Caroline Uhler. Anchored causal inference in the presence of measurement error. In *Conference on uncertainty in artificial intelligence*, pages 619–628. PMLR, 2020.

Muralikrishnna G Sethuraman and Faramarz Fekri. Differentiable cyclic causal discovery under unmeasured confounders. *arXiv preprint arXiv:2508.08450*, 2025.

Muralikrishnna G Sethuraman, Romain Lopez, Rahul Mohan, Faramarz Fekri, Tommaso Biancalani, and Jan-Christian Hütter. Nodags-flow: Nonlinear cyclic causal structure learning. In *International Conference on Artificial Intelligence and Statistics*, pages 6371–6387. PMLR, 2023.

Muralikrishnna G Sethuraman, Razieh Nabi, and Faramarz Fekri. Missnodag: Differentiable cyclic causal graph learning from incomplete data. *arXiv preprint arXiv:2410.18918*, 2024.

Adrian FM Smith and Alan E Gelfand. Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, 46(2):84–88, 1992.

Liam Solus, Yuhao Wang, and Caroline Uhler. Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika*, 108(4):795–814, 2021.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International conference on machine learning*, pages 32540–32560. PMLR, 2023.

E Sturm, E González-Alfonso, S Veilleux, J Fischer, J Graciá-Carpio, S Hailey-Dunsheath, A Contursi, A Poglitsch, A Sternberg, R Davies, et al. Massive molecular outflows and negative feedback in ulirgs observed by herschel-pacs. *The Astrophysical Journal Letters*, 733(1):L16, 2011.

Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *The Journal of Machine Learning Research*, 16(1):2147–2205, 2015.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36:48603–48638, 2023.

C. F. Jeff Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95 – 103, 1983. doi: 10.1214/aos/1176346060. URL https://doi.org/10.1214/aos/1176346060.

Grace Yoon, Raymond J Carroll, and Irina Gaynanova. Sparse semiparametric canonical correlation analysis for data of mixed types. *Biometrika*, 107(3):609–625, 2020.

Appendices

The appendices are organized as follows. Appendix A discusses the solvability of the SCM considered in the work. Appendix B analyses the convergence of the EM algorithm in RECLAIM. Appendix C contains all the proofs. Appendix D provides additional details on the implementation of RECLAIM and the baselines. Finally, appendix E provides additional experimental results and training time comparison between RECLAIM and the baselines.

A Solvability of Cyclic SCMs

Consider the cyclic SCM defined below:

$$\boldsymbol{X} = f(\boldsymbol{X}) + \boldsymbol{Z}. \tag{21}$$

In general, it is not guaranteed that eq. (21) has solution, this is primarily due to the (potential) presence of cycles in $\mathcal{G}$. In order to see why this is true, note that, we can view eq. (21) as the unique fixed point (if it exists) to the following discrete time dynamical system:

$$\boldsymbol{X}(t+1) = f\big(\boldsymbol{X}(t)\big) + \boldsymbol{Z}, \tag{22}$$

where $\boldsymbol{X}(t)$ denotes the value of the random variable $\boldsymbol{X}$ at time t. In the following example, we illustrate how the choice of f dictates whether or not eq. (22) attains a unique fixed point.

Example A.1. *Consider $X_1, X_2 \in \mathbb{R}$ given by the following structural equations:*

$$X_1(t+1) = aX_2(t) + Z_1, \quad X_2(t+1) = bX_1(t) + Z_2.$$

Figure 5 shows an illustration of the evolution of trajectories for two different choices of the parameters (a, b). From the right plot we can see that when we set $a = 1.2$, and $b = 2.3$, the systems is unstable and trajectory diverges. On the other hand, when $a = 0.7$ and $b = -0.8$ (left plot), the system is stable and reaches a stationary point.

To ensure eq. (22) admits a unique fixed point, we must impose additional restrictions on the causal mechanism f. One sufficient condition is to require f to be *contractive*.

Definition A.2 (Contractive function)**.** *A function $f : \mathbb{R}^d \to \mathbb{R}^d$ is said to be contractive if there exists a constant $L < 1$ such that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,*

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

The constant L is called the Lipschitz constant.

When the causal mechanism f is contractive, Banach fixed point theorem gaurantees that eq. (22) has a unique fixed point (Rudin, 1953b, theorem 9.23). Consequently, the causal mechanisms considered in this work are restricted to be contractive and neural networks are used to model the causal mechanism in practice. Contractivity is achieved via spectral scaling of the layer weights as described in section 3.4 in the main paper.

B Convergence of EM algorithm

Here, we provide the convergence analysis of RECLAIM's EM-based score maximization algorithm. Let $\boldsymbol{\Theta}^{(t)} = (\boldsymbol{\theta}^{(t)}, \hat{\boldsymbol{\phi}})$. For ease of notation, we focus on a single interventional

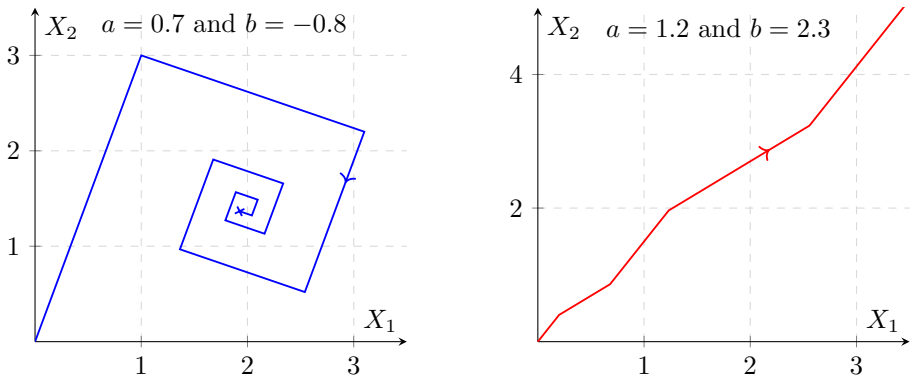


Figure 5: Illustration of state evolution for different choice of a and b for the system defined in example A.1. (Left) Choosing $a = 0.7$ and $b = -0.8$ results in a stable system and the system reaches a fixed point. (Right) Choosing $a = 1.2$ and $b = 2.3$ results in an unstable system and the trajectory diverges.

experiment $I_k \in \mathcal{I}$. However, the final result would still hold even when we consider the sum of likelihood over all the interventional experiments. Our analysis relies on the convergence of the EM algorithm Wu (1983); Friedman (1998), with the crux of it relying on establishing that the total log-likelihood of the observed variables either increases or stays the same in each iteration of the algorithm. That is,

$$\sum_{\ell=1}^{n_k} \log p(\boldsymbol{y}^{(k,\ell)} \mid \boldsymbol{\Theta}^{t+1}, I_k) \geq \sum_{\ell=1}^{n_k} \log p(\boldsymbol{y}^{(k,\ell)} \mid \boldsymbol{\Theta}^{t}, I_k) \tag{23}$$

To that end, note that

$$\begin{aligned}\sum_{i=1}^{n_k} \log p(\boldsymbol{y}^{(k,\ell)} \mid \boldsymbol{\Theta}, I_k) = &\sum_{i=1}^{n_k} \log p(\boldsymbol{y}^{(k,\ell)}, \boldsymbol{X}^{(k,\ell)} \mid \boldsymbol{\Theta}, I_k) \\ &- \sum_{i=1}^{n_k} \log p(\boldsymbol{X}^{(k,\ell)} \mid \boldsymbol{y}^{(k,\ell)}, \boldsymbol{\Theta}, I_k)\end{aligned}$$

Taking expectation with respect $\boldsymbol{X}^{(k,\ell)} \mid \boldsymbol{y}^{(k,\ell)}$ on both side, we get

$$\begin{aligned}\sum_{i=1}^{n_k} \log p(\boldsymbol{y}^{(k,\ell)} \mid \boldsymbol{\Theta}, I_k) &= \sum_{i=1}^{n_k} \mathbb{E}_{\boldsymbol{X}^{(k,\ell)} \mid \boldsymbol{y}^{(k,\ell)}; \boldsymbol{\Theta}^t} \log p(\boldsymbol{y}^{(k,\ell)} \mid \boldsymbol{\Theta}, I_k) \\ &= \underbrace{\sum_{i=1}^{n_k} \mathbb{E}_{\boldsymbol{X}^{(k,\ell)} \mid \boldsymbol{y}^{(k,\ell)}; \boldsymbol{\Theta}^t} \log p(\boldsymbol{y}^{(k,\ell)}, \boldsymbol{X}^{(k,\ell)} \mid \boldsymbol{\Theta}, I_k)}_{=Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t)} \\ &\quad - \sum_{i=1}^{n_k} \mathbb{E}_{\boldsymbol{X}^{(k,\ell)} \mid \boldsymbol{y}^{(k,\ell)}; \boldsymbol{\Theta}^t} \log p(\boldsymbol{X}^{(k,\ell)} \mid \boldsymbol{y}^{(k,\ell)}, \boldsymbol{\Theta}, I_k).\end{aligned} \tag{24}$$

The first term on the RHS in the above equation is nothing but $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$. This is maximized in the M-step, i.e., $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{t+1}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$. On the other hand,

$$\begin{aligned}\sum_{i=1}^{n_k} \mathbb{E}_{\boldsymbol{X}^{(k,\ell)} \mid \boldsymbol{y}^{(k,\ell)}; \boldsymbol{\Theta}^t} \log \frac{p(\boldsymbol{X}^{(k,\ell)} \mid \boldsymbol{y}^{(k,\ell)}, \boldsymbol{\Theta}^{t+1}, I_k)}{p(\boldsymbol{X}^{(k,\ell)} \mid \boldsymbol{y}^{(k,\ell)}, \boldsymbol{\Theta}^{t}, I_k)} = \\ - D_{KL}\Big(p(\boldsymbol{X}^{(k,\ell)} \mid \boldsymbol{y}^{(k,\ell)}, \boldsymbol{\Theta}^t, I_k) \big\| p(\boldsymbol{X}^{(k,\ell)} \mid \boldsymbol{y}^{(k,\ell)}, \boldsymbol{\Theta}^{t+1}, I_k)\Big) \leq 0.\end{aligned}$$

Thus,

$$\sum_{i=1}^{n_k} \mathbb{E}_{\boldsymbol{X}^{(k,\ell)}|\boldsymbol{y}^{(k,\ell)};\boldsymbol{\Theta}^t} \log p(\boldsymbol{X}^{(k,\ell)} \mid \boldsymbol{y}^{(k,\ell)}, \boldsymbol{\Theta}^{t+1}, I_k)$$

$$\leq \sum_{i=1}^{n_k} \mathbb{E}_{\boldsymbol{X}^{(k,\ell)}|\boldsymbol{y}^{(k,\ell)};\boldsymbol{\Theta}^t} \log p(\boldsymbol{X}^{(k,\ell)} \mid \boldsymbol{y}^{(k,\ell)}, \boldsymbol{\Theta}^{t}, I_k). \quad (25)$$

From combining eqs. (24) and (25), we can see that at the end of the M-step eq. (23) is satisfied. Thus, RECLAIM reaches a stationary point of the optimization objective.

C Proofs

All the proofs are contained in this section. Starting with proof of proposition 3 in appendix C.1. Appendix C.2 contains the proof of theorem 4. Finally, appendix C.3 contains the proof for theorem 5.

C.1 Proof of Proposition 3

Proof. The claim of the proposition is equivalent to the following statement:

$$\mathcal{N}(\mathbf{A}_{-i}^\top) \not\subset \boldsymbol{a}_i^\perp.$$

Since $\text{rank}(\mathbf{A}) = d$, the vector $\boldsymbol{a}_i \notin \text{span}(\{\boldsymbol{a}_j\}_{j\neq i})$ as the columns of $\mathbf{A}$ are linearly independent. Furthermore, the null space of $\mathbf{A}_{-i}^\top$ is the orthogonal complement of the span of the columns of $\mathbf{A}_{-i}$, i.e., $\mathcal{N}(\mathbf{A}_{-i}^\top) = \text{span}(\{\boldsymbol{a}_j\}_{j\neq i})^\perp$.

Now, *let us assume, for the sake of contradiction that* for each $\boldsymbol{t} \in \mathbb{R}^p$ such that $\mathbf{A}_{-i}^\top \boldsymbol{t} = \mathbf{0}$, we have $\boldsymbol{a}_i^\top \boldsymbol{t} = 0$. This would mean that

$$\mathcal{N}(\mathbf{A}_{-i}^\top) \subset \boldsymbol{a}_i^\perp.$$

Taking orthogonal complement on both sides, we get

$$(\boldsymbol{a}_i^\perp)^\perp \subset \mathcal{N}(\mathbf{A}_{-i}^\top)^\perp.$$

Note that, $(\boldsymbol{a}_i^\perp)^\perp = \text{span}(\boldsymbol{a}_i)$, and also that $\mathcal{N}(\mathbf{A}_{-i}^\top)^\perp = \text{span}(\{\boldsymbol{a}_j\}_{j\neq i})$. That is, $\text{span}(\boldsymbol{a}_i) \subset \text{span}(\{\boldsymbol{a}_j\}_{j\neq i})$. This is a contradiction. Thus, there exists a vector $\boldsymbol{t} \in \mathbb{R}^p$ such that $\boldsymbol{a}_i^\top \boldsymbol{t} \neq 0$ and $\mathbf{A}_{-i}^\top \boldsymbol{t} = \mathbf{0}$. □

C.2 Proof of Theorem 4

Before we prove the theorem above, we state a technical lemma from functional analysis which is utilized in the proof of the theorem.

Lemma C.1. *Let $P : \mathbb{R}^n \to \mathbb{R}$ be a non-trivial polynomial, then the zero set of P has zero Lebesgue measure.*

See Folland (1999) for more details on lemma C.1. We are now ready to prove theorem 4

Theorem 4. We divide the proof into two parts. In part 1, we show the existence of a measurement matrix $\mathbf{A}_0$ that is not a part of $\bar{\mathcal{A}}$. Then, in part 2, using the result from part 1 we then conclude that the set $\bar{\mathcal{A}}$ has zero Lebesgue measure.

Part 1: Existence of a measurement matrix $\mathbf{A}_0$ such that $\mathbf{T}_2$ is full rank

For $p \geq d$, let us define $\mathbf{A}_0$ as follows:

$$\mathbf{A}_0 = \begin{bmatrix} \mathbf{I}_d \\ \mathbf{0} \end{bmatrix}, \quad \text{i.e., } \boldsymbol{a}_{0,i} = \varepsilon_i \text{ for } i = 1, \ldots, d.$$

For $i = 1, \ldots, d$, let $\boldsymbol{t}^{(i)} = \varepsilon_i$, and for $i = d+1, \ldots, p$, let $\boldsymbol{t}^{(i)} = \varepsilon_1 + \varepsilon_i$. It is easy to check that $\{\boldsymbol{t}^{(1)}, \ldots, \boldsymbol{t}^{(p)}\} \subseteq \mathcal{T}(\mathbf{A}_0)$. Thus, $\mathbf{T}_2$ takes the following form:

$$\mathbf{T}_2 = \begin{bmatrix} \mathbf{I}_d & \mathbf{0} \\ \mathbf{E}_{1,p-d} & \mathbf{I}_{p-d} \end{bmatrix}$$

where $\mathbf{E}_{1,p-d} \in \mathbb{R}^{p-d \times d}$ denotes a matrix with ones in the first column and zeros everywhere else. Clearly, $\text{rank}(\mathbf{T}_2) = p$, i.e., $\mathbf{T}_2$ is full rank. Thus, we have shown the existence of a non-trivial measurement matrix $\mathbf{A}_0 \notin \bar{\mathcal{A}}$.

Part 2: $\bar{\mathcal{A}}$ has zero Lebesgue measure

Since $\text{rank}(\mathbf{A}) = d$, we know that $\text{rank}(\mathbf{A}_{-i}) = d-1$, and thus, $\mathbf{A}_{-i}^\top \in \mathbb{R}^{(d-1)\times p}$ has rank $d-1$. Then, there exists a subset of rows $J_i \subset \{1, \ldots, p\}$, with $|J_i| = d-1$ such that the square matrix

$$\mathbf{M}_i(\mathbf{A}) \triangleq (\mathbf{A}_{-i})_{J_i,*} \in \mathbb{R}^{(d-1)\times(d-1)}$$

is invertible.

Let $\boldsymbol{t}$ be such that $\mathbf{A}_{-i}^\top \boldsymbol{t} = 0$ for some $i \in [d]$. Then,

$$\mathbf{M}_i(\mathbf{A})^\top \boldsymbol{t}_{J_i} + \mathbf{R}_i(\mathbf{A})^\top \boldsymbol{t}_{[p]\setminus J_i} = 0,$$

where $\mathbf{R}_i(\mathbf{A})$ denotes the submatrix of $\mathbf{A}$ formed by rows not included in J_i. Since $\mathbf{M}_i(\mathbf{A})$ is invertible, we have

$$\boldsymbol{t}_{J_i} = -\Big(\mathbf{M}_i(\mathbf{A})^\top\Big)^{-1} \mathbf{R}_i(\mathbf{A})^\top \boldsymbol{t}_{[p]\setminus J_i}.$$

Thus, every admissible $\boldsymbol{t}$ has to have the following form:

$$\boldsymbol{t}(\mathbf{A}, \boldsymbol{u}) = \begin{bmatrix} -\Big(\mathbf{M}_i(\mathbf{A})^\top\Big)^{-1} \mathbf{R}_i(\mathbf{A})^\top \boldsymbol{u} \\ \boldsymbol{u} \end{bmatrix}, \quad \boldsymbol{u} \in \mathbb{R}^{p-d}.$$

Moreover,

$$\mathbf{M}_i(\mathbf{A})^{-1} = \frac{\text{adj}(\mathbf{M}_i(\mathbf{A}))}{\det(\mathbf{M}_i(\mathbf{A}))}$$

Thus, the each coordinate of $\boldsymbol{t}(\mathbf{A}, \boldsymbol{u})$ are: linear in $\boldsymbol{u}$, divided $\det(\mathbf{M}_i(\mathbf{A}))$, and multiplied by polynomial functions of entries of $\mathbf{A}$. Upon taking Hadamard product with itself, each entry of $\boldsymbol{t}^{\odot}$ is a polynomial function of $\boldsymbol{u}$ and the entries of $\mathbf{A}$ divided by the determinant of the minor matrices.

Suppose we choose p such vectors $\boldsymbol{t}^{(1)}, \ldots, \boldsymbol{t}^{(p)}$ (each possibly corresponding to different column of $\mathbf{A}$). The determinant of the resulting $\mathbf{T}_2$, $\det(\mathbf{T}_2)$ is also polynomial in $\boldsymbol{u}^{(\ell)}$, for $\ell \in [p]$; and polynomial in entries of $\mathbf{A}$ along with a division of the determinant of the minor matrices. That is,

$$\det(\mathbf{T}_2) = \frac{\text{Polynomial in } \mathbf{A}, \boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(p)}}{\prod_{\ell=1}^{p} \det(\mathbf{M}_\ell(\mathbf{A}))^{k_\ell}}$$

Let us define

$$D(\mathbf{A}) \triangleq \prod_{\ell=1}^{p} \det(\mathbf{M}_\ell(\mathbf{A}))^{k_\ell}$$

and

$$P(\mathbf{A}, \boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(p)}) \triangleq D(\mathbf{A}) \cdot \det(\mathbf{T}_2).$$

Note that, since the minor matrices are invertible

$$P(\mathbf{A}, \boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(p)}) = 0 \iff \det(\mathbf{T}_2) = 0.$$

In *Part 1*, we showed the existence of an $\mathbf{A}_0$ and $\boldsymbol{u}_0^{(1)}, \ldots, \boldsymbol{u}_0^{(p)}$ such that $P(\mathbf{A}_0, \boldsymbol{u}_0^{(1)}, \ldots, \boldsymbol{u}_0^{(p)}) \neq 0$. Thus the polynomial P is non-trivial. Hence, from lemma C.1, the zero set of P has zero Lebesgue measure. Thus, $\bar{\mathcal{A}}$—which is a subset of the zero set of P—has zero Lebesgue measure. □

C.3 Proof of Theorem 5

In this subsection we provide the proof of consistency of RECLAIM under exact maximization. We start by reviewing the relevant definitions and prior results required to prove theorem 5.

C.3.1 Preliminaries

Consider a directed graph $\mathcal{G} = (\mathcal{X}, \mathcal{E})$. A *path* π between nodes X_i and X_k is a sequence of nodes $(X_{i_0}, X_{i_1}, \ldots, X_{i_n})$, with $X_{i_0} = X_i$, $X_{i_n} = X_k$, and every two consecutive nodes in the sequence are connected by an edge, i.e., either $X_{i_j} \to X_{i_{j+1}} \in \mathcal{E}$ or $X_{i_j} \leftarrow X_{i_{j+1}} \in \mathcal{E}$ for all $j = 0, \ldots, n-1$, with X. A path is *directed* if all the edges from X_i to X_k are oriented the same way. A cycle through node X_i consists of a directed path from X_i to a node X_j and the directed edge $X_j \to X_i$. For a node X_i, the set of *ancestors* is defined as $\text{an}_\mathcal{G}(X_i) := \{X_j \in X \mid \text{a directed path exists between } X_j \text{ and } X_i\}$. Similarly, the *descendant* of a node is defined as $\text{de}_\mathcal{G}(X_i) := \{X_j \in X \mid \text{a directed path exists between } X_i \text{ and } X_j\}$. The *strongly connected component* of a node X_i is given by the intersection of the ancestors and the descendants of X_i, i.e., $\text{sc}_\mathcal{G}(X_i) = \text{an}_\mathcal{G}(X_i) \cap \text{de}_\mathcal{G}(X_i)$. We can apply these definitions to subsets $\mathcal{X}_U$ by taking the union over all the elements in the subset, i.e., $\text{de}_\mathcal{G}(\mathcal{X}_U) = \cup_{i \in U} \text{de}_\mathcal{G}(X_i)$. Finally, a node X_i is called a *collider* in a path π if it satisfies the following two conditions: (i) it is a non-endpoint node, and (ii) the subpath (X_k, X_i, X_j) is of the form $X_k \to X_i \leftarrow X_j$.

In a cyclic graph, the notion of σ-separation was introduced by Forré and Mooij (2017) to relate structural properties in the graph to independencies observed in the generated distribution.

Definition C.2 (σ-separation)**.** *Let $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ be a directed graph and let $\mathcal{X}_C \subseteq \mathcal{X}$ be a subset of nodes. A path $\pi = (X_{i_0}, X_{i_1}, \ldots, X_{i_n})$ is said to be σ-blocked given $\mathcal{X}_C$ if*

1. *the first node of π, $X_{i_0} \in \mathcal{X}_C$ or its last node $X_{i_n} \in \mathcal{X}_C$, or*
2. *π contains a collider $X_j \notin \mathcal{X}_C$,*
3. *π contains a non-collider $X_j \in \mathcal{X}_C$ that points towards a neighbor that is not in the same strongly connected component as X_j in $\mathcal{G}$, i.e., $X_k \leftarrow X_j \in \pi$ and $X_k \notin sc_\mathcal{G}(X_j)$, or $X_j \to X_\ell \in \pi$ and $X_\ell \notin sc_\mathcal{G}(X_j)$.*

The path π is said to be σ-open given $\mathcal{X}_C$ if it is not σ-blocked. Two subsets of nodes $\mathcal{X}_A, \mathcal{X}_B \subseteq \mathcal{X}$ is said to be σ-separated given $\mathcal{X}_C$ if all the paths between X_a and X_b, where $a \in A$ and $b \in B$, are σ-blocked given $\mathcal{X}_C$, and is denoted by

$$\mathcal{X}_A \overset{\sigma}{\underset{\mathcal{G}}{\perp}} \mathcal{X}_B \mid \mathcal{X}_C.$$

When the graph is cyclic, σ-separation defined above reduces to the standard d-separation. We node define the property that connects σ-separations with distributional independencies.

Definition C.3 (General directed global Markov Property Forré and Mooij (2017))**.** *Let $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ be a directed graph and p denote the probability density of the observations $\mathcal{X}$. The probability density p satisfies the* general directed global Markov property *if for $\mathcal{X}_A, \mathcal{X}_B, \mathcal{X}_C \subseteq \mathcal{X}$*

$$\mathcal{X}_A \overset{\sigma}{\underset{\mathcal{G}}{\perp}} \mathcal{X}_B \mid \mathcal{X}_C \implies \boldsymbol{X}_A \underset{p}{\perp} \boldsymbol{X}_B \mid \boldsymbol{X}_C,$$

where $\boldsymbol{X}_A \perp_p \boldsymbol{X}_B \mid \boldsymbol{X}_C$, denotes conditionally independence of $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ given $\boldsymbol{X}_C$ with respect to p.

Joint causal modeling of interventions. We adopt the *joint causal inference* (JCI) framework proposed by Mooij et al. (2020) to unify multiple interventional settings into a single representative system. This is done by augmenting the system with context variables $\mathcal{C}_\mathcal{I} = (\boldsymbol{C}_1, \ldots, \boldsymbol{C}_K)$, where $\boldsymbol{C}_k$ corresponds to the k-th interventional setting I_k, we call this the *meta system*. We consider the system to be under the k-th interventional setting when $\boldsymbol{C}_j = \emptyset$ for $j \neq k$ and $\boldsymbol{C}_k = \boldsymbol{\xi}_{I_k}$ where $\boldsymbol{\xi}_{I_k} \in \mathbb{R}^{|I_k|}$. Additionally, we construct an augmented graph $\mathcal{G}^\mathcal{I}$ consisting of both the latent system variables $\mathcal{X}$ and the context variables $\mathcal{C}_\mathcal{I}$. The children of each context variable $\boldsymbol{C}_k$ are the interventional targets in I_k, i.e., $\mathrm{ch}_{\mathcal{G}^\mathcal{I}}(\boldsymbol{C}_k) = \mathcal{X}_{I_k}$. Note that there are no edges going from variables in $\mathcal{X}$ to the context variables $\mathcal{C}_\mathcal{I}$, see fig. 6. Finally, given a family of interventional targets $\mathcal{I} = \{I_k\}_{k=1}^K$, and the corresponding context variables $\mathcal{C}_\mathcal{I}$, the structural equations governing the meta system is given by

$$\tilde{f}_i(\mathrm{pa}_{\mathcal{G}^\mathcal{I}}(X_i), Z_i) = \begin{cases} (\boldsymbol{C}_k)_i & \text{if , } \exists k \in [K] \text{s.t. } \boldsymbol{C}_k \neq \emptyset, X_i \in \mathcal{X}_{I_k}, \\ f_i(\mathrm{pa}_\mathcal{G}(X_i)) + Z_i, & \text{otherwise.} \end{cases}$$

Note that $\mathrm{pa}_{\mathcal{G}^\mathcal{I}}(X_i)$ includes both latent system variables and the context variables. The meta-SCM above defines a probability distribution over the joint meta-system variables:

$$p_{\mathcal{G}^\mathcal{I}}(\boldsymbol{X}, \mathcal{C}_\mathcal{I}) = p_{\mathcal{G}^\mathcal{I}}(\mathcal{C}_\mathcal{I}) p_{\mathcal{G}^\mathcal{I}}(\boldsymbol{X} \mid \mathcal{C}_\mathcal{I}), \tag{26}$$

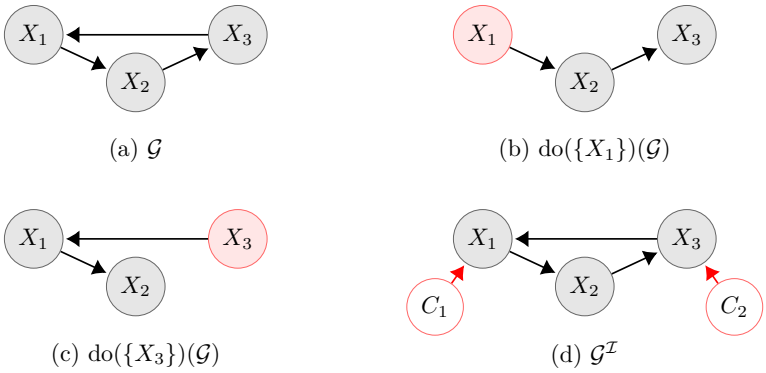
where $p_{\mathcal{G}^\mathcal{I}}(\mathcal{C}_\mathcal{I})$ is called the *context distribution* and as noted by Mooij et al. (2020), the behavior of the system is unaffected by the context distribution. Furthermore, when $\boldsymbol{C}_k = \emptyset$ for all $k \in [K]$, $p_{\mathcal{G}^\mathcal{I}}(\boldsymbol{X} \mid \mathcal{C}_\mathcal{I})$ becomes the observational data, and

$$p_{\mathcal{G}^\mathcal{I}}(\boldsymbol{X} \mid \boldsymbol{C}_k = \boldsymbol{\xi}_{I_k}, \boldsymbol{C}_{-k} = \emptyset) = p_{\mathrm{do}(I_k)(\mathcal{G})}(\boldsymbol{X}).$$

Recall, for the interventional setting I_k, probability density governing the latents $\boldsymbol{X}$ is given by eq. (5) which we repeat here for convenience

$$p_{\mathrm{do}(I)(\mathcal{G})}(\boldsymbol{X}) = p_C(\boldsymbol{X}_I) p_Z\big(\boldsymbol{Z}_{[d]\setminus I}\big) \big| J_{(\mathrm{id} - \mathbf{U}f)}(\boldsymbol{Z}) \big|.$$

Definition C.4. *Let $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ be a directed graph, and $\mathcal{I} = \{I_k\}_{k=1}^K$ be a family of interventional experiments. Let $\mathcal{M}_\mathcal{I}(\mathcal{G})$ denote the set of positive densities $p_{\mathcal{G}^\mathcal{I}} : \mathbb{R}^{2d} \to \mathbb{R}$ such that $p_{\mathcal{G}^\mathcal{I}}$ is given by eq. (26) for all $f : \mathbb{R}^d \to \mathbb{R}^d$, with $f_i(\boldsymbol{X}) = f_i(pa_\mathcal{G}(X_i))$, such that the resulting forward map $(id - f)$ is a diffeomorphism, for all $(\sigma^2_{Z,1}, \ldots, \sigma^2_{Z,d})$ such that $\boldsymbol{Z} \sim \mathcal{N}\big(0, Diag(\sigma^2_{Z,1}, \ldots, \sigma^2_{Z,K})\big)$.*



(a) $\mathcal{G}$ (b) $\text{do}(\{X_1\})(\mathcal{G})$ (c) $\text{do}(\{X_3\})(\mathcal{G})$ (d) $\mathcal{G}^{\mathcal{I}}$

Figure 6: Illustration of the augmented graph $\mathcal{G}^{\mathcal{I}}$ corresponding the set of interventional targets $\mathcal{I} = \{\emptyset, \{1\}, \{3\}\}$. (b) and (c) represent the mutilated graph obtained after interventions on X_1 and X_3 respectively. The augmented graph is the union of $\mathcal{G}$, $\text{do}(\{X_1\})(\mathcal{G})$, and $\text{do}(\{X_3\})(\mathcal{G})$ along with the context variables C_1 and C_2.

$\mathcal{M}_{\mathcal{I}}(\mathcal{G})$ denotes the set of interventional density generated by $\mathcal{G}^{\mathcal{I}}$. We now show that these densities satisfy the general directed global Markov property.

Proposition C.5. *Let $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ be a directed graph, and $\mathcal{I} = \{I_k\}_{k=1}^K$ be a family of interventional experiments, let $p \in \mathcal{M}_{\mathcal{I}}(\mathcal{G})$, then p satisfies the general directed global Markov property relative to $\mathcal{G}^{\mathcal{I}}$.*

Proof. For a directed graph $\mathcal{G}$ and a choice of $f : \mathbb{R}^d \to \mathbb{R}^d$ such that the forward map $(\text{id} - f)$ is a diffeomorphism, the structural equations are uniquely solvable with respect to each strongly connected component of $\mathcal{G}$. Moreover, since the addition of context variable doesn't introduce any cycles, from (Bongers et al., 2021, Theorem A.21), the distribution $p_{\mathcal{G}^{\mathcal{I}}}$ is unique and it satisfies the general directed global Markov property. □

Given a family of interventional experiments, we now define a notion of an equivalence class of directed (cyclic) graphs based on the set of distributions induced by them.

Definition C.6 ($\mathcal{I}$-Markov Equivalence Class)**.** *Two directed graphs $\mathcal{G}$ and $\mathcal{G}'$ are $\mathcal{I}$-Markov equivalent if and only if $\mathcal{M}_{\mathcal{I}}(\mathcal{G}) = \mathcal{M}_{\mathcal{I}}(\mathcal{G}')$, denoted as $\mathcal{G} \equiv_{\mathcal{I}} \mathcal{G}'$. The set of all directed graphs that are $\mathcal{I}$-Markov equivalent to $\mathcal{G}$ is the $\mathcal{I}$-Markov equivalence class of $\mathcal{G}$, denoted as $\mathcal{I}$-MEC$(\mathcal{G})$.*

C.3.2 Proof of Theorem 5

We now prove the main theoretical contribution of this work. Recall the score function defined in section 3.1

$$\mathcal{S}_{\mathcal{I}}(\mathcal{G}) \triangleq \sup_{\boldsymbol{\theta}, \boldsymbol{\phi}} \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{Y} \sim p^{(k)}} \log p(\boldsymbol{Y} \mid \boldsymbol{\theta}, \boldsymbol{\phi}, I_k) - \lambda |\mathcal{G}|,$$

where $p^{(k)}$ is the true data-generating distribution for $\mathcal{I} = \{I_k\}_{k=1}^K$, and $(\boldsymbol{\theta}, \boldsymbol{\phi})$ represents the model parameters (latent SCM and measurement process). In the context of the meta system, the score function above is equivalent to the following score:

$$\mathcal{S}_{\mathcal{I}}(\mathcal{G}) \triangleq \sup_{\boldsymbol{\theta}, \boldsymbol{\phi}} \mathbb{E}_{(\boldsymbol{Y}, \boldsymbol{C}) \sim p^*_{\mathcal{I}}} \log p_{\mathcal{G}^{\mathcal{I}}}(\boldsymbol{Y}, \boldsymbol{C} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) - \lambda |\mathcal{G}|,$$

where $p^*_{\mathcal{I}}$ denotes the joint ground-truth distribution for the observed and the context variables, and $p_{\mathcal{G}^{\mathcal{I}}}(\boldsymbol{Y}, \boldsymbol{C} \mid \boldsymbol{\theta})$ is given by

$$p_{\mathcal{G}^{\mathcal{I}}}(\boldsymbol{Y}, \boldsymbol{C} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) = \int_{\boldsymbol{X}} p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\phi}) p_{\mathcal{G}^{\mathcal{I}}}(\boldsymbol{X}, \boldsymbol{C} \mid \boldsymbol{\theta})\, d\boldsymbol{X}. \tag{27}$$

Note that $p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\phi})$ is defined by the measurement process $g(\boldsymbol{X}, \boldsymbol{\varepsilon})$.

We define $\mathcal{P}_{\mathcal{I}}(\mathcal{G})$ as the set of all distributions $p_{\mathcal{G}^{\mathcal{I}}}(\boldsymbol{X}, \boldsymbol{C} \mid \boldsymbol{\theta})$ that can be expressed by the model specified by eqs. (2) and (5) in the main paper, and eq. (26). That is, $\mathcal{P}_{\mathcal{I}}(\mathcal{G}) := \{p \mid \exists \boldsymbol{\theta} \text{ s.t } p = p_{\mathcal{G}^{\mathcal{I}}}(\cdot \mid \boldsymbol{\theta})\}$. Thus, it is clear that $\mathcal{P}_{\mathcal{I}}(\mathcal{G}) \subseteq \mathcal{M}_{\mathcal{I}}(\mathcal{G})$. Similarly, we define the of observed distributions expressable by $\mathcal{G}$, denoted as $\mathcal{P}_{Y,\mathcal{I}}(\mathcal{G}) = \{p \mid \exists (\boldsymbol{\theta}, \boldsymbol{\phi}) \text{ s.t. } p = p_{\mathcal{G}^{\mathcal{I}}}(\cdot \mid \boldsymbol{\theta}, \boldsymbol{\phi})\}$. In the following proposition we show that $\mathcal{P}_{Y,\mathcal{I}}(\mathcal{G})$ is uniquely determined by a $p \in \mathcal{P}_{\mathcal{I}}(\mathcal{G})$ when $\mathcal{I}$ satisfies condition 2.

Proposition C.7. *Let $\mathcal{I} = \{I_k\}_{k=1}^K$ be a family of interventional experiments satisfying condition 2. Then for a fixed set of measurement parameters $\boldsymbol{\phi}$, under both Gaussian additive noise (GAN) measurement process and linear measurement system, the map $T : \mathcal{P}_{\mathcal{I}}(\mathcal{G}) \to \mathcal{P}_{Y,\mathcal{I}}(\mathcal{G})$ given by eq. (27) is injective.*

Proof. Under condition 2, the measurement noise variance is fully identifiable (see section 3.3 for more details). For sake of simplicity, let us assume that the noise variances are equal (the conclusion even otherwise) and set to σ^2. Taking the characteristic functions of $Tp = Tp'$ and dividing by the non-vanishing Gaussian factor $e^{-\sigma^2 \|\boldsymbol{\xi}\|^2/2}$ yields $\mathbb{E}_p[e^{-i\boldsymbol{\xi}^\top h(\boldsymbol{X})}] = \mathbb{E}_{p'}[e^{-i\boldsymbol{\xi}^\top h(\boldsymbol{X})}]$ for all $\boldsymbol{\xi}$. Here, $f(\boldsymbol{X}) = \boldsymbol{X}$ for GAN system and $f(\boldsymbol{X}) = \mathbf{A}\boldsymbol{X}$ for linear measurement system. Thus, from the uniqueness of the characteristic function Billingsley (1995), $f_{\#}p = f_{\#}p'$. Since f is injective, $p = p'$. □

Theorem 5 relies on the following set of assumptions. The first one ensures that the model is capable of representing the ground truth distribution.

Assumption C.8 (Sufficient Capacity)**.** *The joint ground truth distribution $p^*_{\mathcal{I}}$ is such that $p^*_{\mathcal{I}} \in \mathcal{P}_{\mathcal{I}}(\mathcal{G}^*)$, where $\mathcal{G}^*$ is the ground truth latent graph.*

In other words, there exists a $\boldsymbol{\theta}$ such that $p^*_{\mathcal{I}} = p_{\mathcal{G}^{\mathcal{I}}}(\cdot \mid \boldsymbol{\theta})$. The second assumption generalizes the notion of faithfulness assumption to the interventional setting.

Assumption C.9 ($\mathcal{I}$-σ-faithfulness)**.** *For any subset of nodes $A, B, D \subseteq \mathcal{X} \cup \mathcal{C}_{\mathcal{I}}$, and $I_k \in \mathcal{I}$*

$$A \underset{\mathcal{G}^{\mathcal{I}}}{\overset{\sigma}{\not\perp}} B \mid D \implies \boldsymbol{A} \underset{p_{\mathcal{G}^{\mathcal{I}}}}{\not\perp} \boldsymbol{B} \mid \boldsymbol{C}.$$

The above assumption implies that any conditional independency observed in the data must imply a σ-separation in the corresponding interventional ground truth graph.

Assumption C.10 (Strict positivity)**.** *For all $(\boldsymbol{X}, \boldsymbol{C})$ and $p_{\mathcal{G}^{\mathcal{I}}} \in \mathcal{M}_{\mathcal{I}}(\mathcal{G})$, $p_{\mathcal{G}^{\mathcal{I}}}(\boldsymbol{X}, \boldsymbol{C}) > 0$.*

Assumption C.11 (Finite differential entropy)**.** *For a family of interventional targets $\{X_{\mathcal{I}_m}\}_{m=0}^M$,*

$$\left|\mathbb{E}_{p^*_{\mathcal{I}}} \log p^*_{\mathcal{I}}(\boldsymbol{Y}, \boldsymbol{C})\right| < \infty.$$

The last two assumptions ensure that the scenario where $\mathcal{S}(\mathcal{G}^*)$ and $\mathcal{S}(\mathcal{G})$ are both infinity is avoided. This is formalized in the lemma below taken from Brouillard et al. (2020).

Lemma C.12 (Finiteness of the score function Brouillard et al. (2020))**.** *Under the assumptions C.8 and C.11, the score function is finite, that is, $|\mathcal{S}_{\mathcal{I}}(\mathcal{G})| < \infty$.*

We finally make the assumption that the parameter space of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are compact.

Assumption C.13 (Parameter compactness)**.** *The parameters of the measurement process $\boldsymbol{\phi}$ belong to a compact metric space Φ.*

For the both Gaussian additive noise (GAN) and linear measurement system, the parameters correspond to the variance of the Gaussian distribution, $\boldsymbol{\phi} = (\sigma_1^2, \ldots, \sigma_p^2)$. For the case of GAN system, $p = d$. The implication of the above assumption is that $\sigma_j^2 \leq \sigma_{\max}^2$, i.e., it is bounded above. Additionally, note that since we restrict the causal mechanism to *contractive*, the parameters of latent SCM also belong to a compact space.

From the results of Brouillard et al. (2020), we can now express the difference in score function between $\mathcal{G}^*$ and $\mathcal{G}$ as the minimization of KL diverengence plus the difference in the regularization terms.

Lemma C.14 (Rewritting the score function Brouillard et al. (2020))**.** *Under assumptions C.8 and C.11, we have*

$$\mathcal{S}(\mathcal{G}^*) - \mathcal{S}(\mathcal{G}) = \inf_{\boldsymbol{\theta}, \boldsymbol{\phi}} D_{KL}(p_{\mathcal{I}}^* \| p_{\mathcal{G}^{\mathcal{I}}}(\cdot \mid \boldsymbol{\theta}, \boldsymbol{\phi})) + \lambda(|\mathcal{G}| - |\mathcal{G}^*|).$$

Furthermore, Sethuraman and Fekri (2025) showed that the KL divergence appearing the in lemma above is strictly positive when the latent variables are directly observed.

Lemma C.15 (Lemma A.18, Sethuraman and Fekri (2025))**.** *Let $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ be a directed graph, for a set of interventional targets $\mathcal{I} = \{I_k\}_{k=1}^K$, and $p^* \notin \mathcal{M}_{\mathcal{I}}(\mathcal{G})$, then*

$$\inf_{p \in \mathcal{M}_{\mathcal{I}}(\mathcal{G})} D(p^* \| p) > 0.$$

We now extend this result to the case of indirect measurements.

Lemma C.16. *Let $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ be a directed graph, for a set of interventional targets $\mathcal{I} = \{I_k\}_{k=1}^K$ satisfying condition 2, and $p_Y^* \notin \mathcal{M}_{Y,\mathcal{I}}(\mathcal{G})$, then*

$$\inf_{p \in \mathcal{M}_{Y,\mathcal{I}}(\mathcal{G}_m)} D(p_Y^* \| p_Y) > 0.$$

Proof. Since $\mathcal{I}$ satisfies condition 2, the parameters of the measurement process is fully identified. Moreover, from proposition C.7, the map $T : \mathcal{P}_{\mathcal{I}}(\mathcal{G}) \to \mathcal{P}_{Y,\mathcal{I}}(\mathcal{G})$. Thus, $T^{-1} p_Y^* \notin \mathcal{P}_{\mathcal{I}}(\mathcal{G})$. Thus, from lemma C.15, no sequence $\{p^{(k)}\}_{k=1}^{\infty} \subset \mathcal{M}_s i(\mathcal{G})$ exists such that $\lim_{k\to\infty} p^{(k)} = T^{-1} p_Y^*$. Therefore, there exists no sequence $\{p_Y^{(k)}\}_{k=1}^{\infty} \subset \mathcal{M}_{Y,si}(\mathcal{G})$ such that $\lim_{k\to\infty} p_Y^{(k)} = p_Y^*$. Thus,

$$\inf_{p \in \mathcal{M}_{Y,\mathcal{I}}(\mathcal{G}_m)} D(p_Y^* \| p_Y) > 0.$$

This proves the lemma. □

We are now ready to prove theorem 5. Recall,

Theorem 5.. *Let $\mathcal{I} = \{I_k\}_{k=1}^K$ be a family of interventional targets satifying condition 2, let $\mathcal{G}^*$ denote the ground truth directed graph, $p^{(k)}$ denote the data generating distribution for $I_k \in \mathcal{I}$, and $\hat{\mathcal{G}} := \arg\max_{\mathcal{G}} \mathcal{S}(\mathcal{G})$. Then, under assumptions C.8 to C.11 and C.13, and for a suitably chosen $\lambda > 0$, we have that $\hat{\mathcal{G}} \equiv_{\mathcal{I}} \mathcal{G}^*$. That is, $\hat{\mathcal{G}}$ is $\mathcal{I}$-Markov equivalent to $\mathcal{G}^*$.*

Proof. The proof is a direct extension of Theorem 2 in Sethuraman and Fekri (2025), with lemma C.16 substituting lemma C.15. Which we present here for self-containment.

It is sufficient to show that for $\mathcal{G} \notin \mathcal{I}\text{-MEC}(\mathcal{G}^*)$, the score function of $\mathcal{G}$ is strictly lower than the score function of $\mathcal{G}^*$, i.e., $\mathcal{S}(\mathcal{G}^*) > \mathcal{S}(\mathcal{G})$. Since $\mathcal{G} \notin \mathcal{I}\text{-MEC}(\mathcal{G}^*)$ and $p^*_{\mathcal{I}} \in \mathcal{M}_{\mathcal{I}}(\mathcal{G}^*)$ (by assumption C.8), there must exist subsets of nodes $A, B, D \subseteq \mathcal{X} \cup \mathcal{C}_{\mathcal{I}}$ such that either:

$$A \overset{\sigma}{\underset{\mathcal{G}}{\perp}} B \mid D \quad \text{and} \quad A \overset{\sigma}{\underset{\mathcal{G}^*}{\not\perp}} B \mid D, \tag{C1}$$

or

$$A \overset{\sigma}{\underset{\mathcal{G}}{\not\perp}} B \mid D \quad \text{and} \quad A \overset{\sigma}{\underset{\mathcal{G}^*}{\perp}} B \mid D, \tag{C2}$$

If no such subsets exist, then $\mathcal{G}$ and $\mathcal{G}^*$ impose the same σ -separation constraints and thus induce the same set of distributions. This would imply that $\mathcal{G} \in \mathcal{I}\text{-MEC}(\mathcal{G}^*)$, contradicting our assumption. Since $p^*_{\mathcal{I}} \in \mathcal{M}_{\mathcal{I}}(\mathcal{G}^*)$, in the case of (C1), it must be true that $\boldsymbol{A} \not\perp_{p^*_{\mathcal{I}}} \boldsymbol{B} \mid \boldsymbol{D}$ (assumption C.9). Therefore $p^*_{\mathcal{I}}$ doesn't satisfy the general directed Markov property with respect to $\mathcal{G}^{\mathcal{I}}$ and hence $p^*_{\mathcal{I}} \notin \mathcal{M}_{\mathcal{I}}(\mathcal{G})$. For (C2), if $p^*_{\mathcal{I}} \in \mathcal{M}_{\mathcal{I}}(\mathcal{G})$, then from assumption C.9, it must be true that $\boldsymbol{A} \not\perp_{p^*_{\mathcal{I}}} \boldsymbol{B} \mid \boldsymbol{C}$. However, $p^*_{\mathcal{I}} \in \mathcal{M}_{\mathcal{I}}(\mathcal{G}^*)$, and proposition C.5 implies that $\boldsymbol{A} \perp_{p^*_{\mathcal{I}}} \boldsymbol{B} \mid \boldsymbol{C}$, resulting in a contradiction. Therefore, $p^*_{\mathcal{I}} \notin \mathcal{M}_{\mathcal{I}}(\mathcal{G})$. Morevover, from proposition C.7 we can conclude that $p^*_{Y,\mathcal{I}} \notin \mathcal{M}_{Y,\mathcal{I}}(\mathcal{G})$.

For convenience, let

$$\eta(\mathcal{G}) := \inf_{\boldsymbol{\theta}} D_{KL}(p^*_{Y,\mathcal{I}} \| p_{Y,\mathcal{G}^{\mathcal{I}}}(\cdot \mid \boldsymbol{\theta}, \boldsymbol{\phi}^*)).$$

Note that

$$\eta(\mathcal{G}) = \inf_{\boldsymbol{\theta}} D_{KL}(p^*_{Y,\mathcal{I}} \| p_{Y,\mathcal{G}^{\mathcal{I}}}(\cdot \mid \boldsymbol{\theta}, \boldsymbol{\phi}^*)) \geq \inf_{p \in \mathcal{M}_{Y,\mathcal{I}}(\mathcal{G})} D_{KL}(p^{(k)}_Y \| p_Y) > 0,$$

where we use lemma C.16 for the final inequality. Thus, from Lemma C.14

$$\mathcal{S}(\mathcal{G}^*) - \mathcal{S}(\mathcal{G}) = \eta(\mathcal{G}) + \lambda(|\mathcal{G}| - |\mathcal{G}^*|).$$

Following Brouillard et al. (2020), we now show that by choosing λ sufficiently small, the above equation is stictly positive. Note that if $|\mathcal{G}| \geq |\mathcal{G}^*|$ then $\mathcal{S}(\mathcal{G}^*) - \mathcal{S}(\mathcal{G}) > 0$. Let $\mathbb{G}^+ := \{\mathcal{G} \mid |\mathcal{G}| < |\mathcal{G}^*|\}$. Choosing λ such that $0 < \lambda < \min_{\mathcal{G} \in \mathbb{G}^+} \frac{\eta(\mathcal{G})}{|\mathcal{G}^*| - |\mathcal{G}|}$ we see that:

$$\begin{aligned}
&\lambda < \min_{\mathcal{G} \in \mathbb{G}^+} \frac{\eta(\mathcal{G})}{|\mathcal{G}^*| - |\mathcal{G}|} \\
\iff &\lambda < \frac{\eta(\mathcal{G})}{|\mathcal{G}^*| - |\mathcal{G}|} \quad \forall \mathcal{G} \in \mathbb{G}^+ \\
\iff &\lambda(|\mathcal{G}^*| - |\mathcal{G}|) < \eta(\mathcal{G}) \quad \forall \mathcal{G} \in \mathbb{G}^+ \\
\iff &0 < \eta(\mathcal{G}) + \lambda(|\mathcal{G}| - |\mathcal{G}^*|) = \mathcal{S}(\mathcal{G}^*) - \mathcal{S}(\mathcal{G}) \quad \forall \mathcal{G} \in \mathbb{G}^+.
\end{aligned}$$

Thus, every graph outside of the general directed Markov equivalence class of $(\mathcal{G}^*)^{\mathcal{I}}$ has a strictly lower score. □

D Implementation Details

In this section we provide further technical details on the implementation of RECLAIM and the baselines.

Algorithm D.1 Sampling Projection Vectors for Noise Estimation

Require: Measurement matrix $\mathbf{A} \in \mathbb{R}^{p \times d}$, intervention family $\mathcal{I}$ satisfying condition 2, thresholds $\epsilon_{\text{sig}} > 0$, $\delta > 0$, target number of vectors m
Ensure: Projection matrix $\mathbf{T}_2 \in \mathbb{R}^{m \times p}$

Initialize $\mathbf{T}_2 \leftarrow [\,]$
for $i = 1, \ldots, d$ **do**
  Compute $\mathbf{M}_i \leftarrow \mathbf{A}_{-i}^\top \in \mathbb{R}^{(d-1)\times p}$
  Compute SVD: $\mathbf{M}_i = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$
  Extract basis $\mathbf{B}_i \in \mathbb{R}^{p \times r}$ from columns of $\mathbf{V}$ corresponding to zero singular values, where $r = p - d + 1$
  while fewer than $\lfloor m/d \rfloor$ vectors collected for node i **do**
    Sample $\boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r)$
    Set $\boldsymbol{t} \leftarrow \mathbf{B}_i \boldsymbol{u}$
    if $|\boldsymbol{a}_i^\top \boldsymbol{t}| < \epsilon_{\text{sig}}$ **then** ▷ weak signal
      reject and **continue**
    end if
    Normalize: $\boldsymbol{t} \leftarrow \boldsymbol{t}/\|\boldsymbol{t}\|$
    if $\exists\, (\boldsymbol{t}')^{\odot} \in \mathbf{T}_2$ s.t. $\cos\left(\boldsymbol{t}^{\odot}, (\boldsymbol{t}')^{\odot}\right) > 1 - \delta$ **then** ▷ insufficient diversity
      reject and **continue**
    end if
    Append $(\boldsymbol{t}^{\odot})^\top$ to $\mathbf{T}_2$
  end while
end for
return $\mathbf{T}_2$

D.1 Projection Vector Sampling for Measurement Noise Estimation

The overall projection vector sampling algorithm for noise variance estimation is summarized in algorithm D.1. We start with an empty $\mathbf{T}_2$ matrix and iterating through each node in the latent graph. For each $i \in [d]$, let $\mathbf{M}_i$ be the measurement matrix $\mathbf{A}$ excluding the i-th column. We then compute its singular value decomposition (SVD) and extract the columns of $\mathbf{V}$ matrix corresponding to zero singular values, denoted as $\mathbf{B}_i$. These columns represent the basis vectors of the null space of $\mathbf{M}_i$. A random vector $\boldsymbol{u}$ is then generated to obtain the projection matrix $\boldsymbol{t} = \mathbf{B}_i \boldsymbol{u}$. The vector $\boldsymbol{t}$ is appended to $\mathbf{T}_2$ if it provided sufficient signal and is sufficiently diverse from the existing projection vectors already in $\mathbf{T}_2$. The process concludes once we sample the required number of rows for $\mathbf{T}_2$. Theorem 4 guarantees that the above procedure would conclude and the results matrix $\mathbf{T}_2$ would be full (column) rank.

D.2 Log-determinant of the Jacobian Computation

Computing the log-determinant of the Jacobian computation poses a significant challenge when evaluation the log-density of the latent variables. To overcome this issue, we exploit the power-series expansion of $\log(1-x)$ to obtain the following:

$$\log \left| \det J_{(\text{id}-\mathbf{U}f)}(\boldsymbol{X}) \right| = -\sum_{m=1}^{\infty} \frac{1}{m} \text{Tr}\Big\{ J_{\mathbf{U}f}^m(\boldsymbol{X}) \Big\}. \tag{28}$$

This brings the complexity of down to $\mathcal{O}(d^2)$. Further improvement can be made by using *Hutchinson trace estimator* Hutchinson (1989)

$$\text{Tr}\Big\{ J_{\mathbf{U}f}^m(\boldsymbol{X}) \Big\} = \mathbb{E}_{\mathbf{W}} \Big[\mathbf{W}^\top J_{\mathbf{U}f}^m(\boldsymbol{X}) \mathbf{W} \Big],$$

where $\mathbb{E}[\mathbf{W}] = 0$ and $\mathbb{E}[\mathbf{W}\mathbf{W}^\top] = \mathbf{I}$. The above estimator only depends on a vector-Jacobian product which can be efficiently computed using an auto-differentiation library (such as PyTorch), often reducing the complexity to $\mathcal{O}(d)$. In practice, the power series is truncated to a finite number of terms. This, however, introduces bias in the log-determinant of the Jacobian estimate. To improve on this, following Chen et al. (2019), the series cut-off n is randomly sampled, $n \sim p_N$, where p_N is a distribution over natural numbers $\mathbb{N}$. The individual terms in eq. (28) are then reweighted by the inverse probability of ending there. Finally, we have the following unbiased estimator:

$$\log \left| \det J_{(\mathrm{id}-\mathbf{U}f)}(\boldsymbol{X}) \right| = -\mathbb{E}_{n,\mathbf{W}} \left[\sum_{m=1}^{n} \frac{\mathbf{W}^\top J^m_{\mathbf{U}f}(\boldsymbol{X})\mathbf{W}}{m \cdot P(N \geq n)} \right]. \tag{29}$$

D.3 RECLAIM and Baselines Code Details

RECLAIM. Our framework was built using the `Pytorch` library in Python and the code is provided as a part of the supplementary materials.

We follow the setup of Sethuraman et al. (2023), employing neural networks (NNs) with dependency masks parameterized by a Gumbel-softmax distribution. The log-determinant of the Jacobian is computed using a power series expansion combined with the Hutchinson trace estimator. To mitigate bias from truncating the power series expansion, the number of terms is sampled from a Poisson distribution, as detailed in section 3 and appendix D.2. The final objective is optimized using the Adam optimizer Kingma and Ba (2015).

The learning rate in all our experiments was set to 10^{-2}. The neural network models used in our experiments contained one multi-layer perceptron layer and `tanh` activation. The graph sparsity regularization constant λ was set to 10^{-3} for all the experiments. The models were trained and evaluated on NVIDIA RTX6000 GPUs.

Baselines. For NODAGS-Flow, we used the code provided by authors Sethuraman et al. (2023) available at https://github.com/Genentech/nodags-flows. The default values were set for the hyperparameters. For DCDI, we used the codebase provided by the authors Brouillard et al. (2020), available at https://github.com/slachapelle/dcdi. The default hyperparameters were used while training and evaluating the model. For Anchored-CI, we implemented Algorithm 1 in Saeed et al. (2020) using an implementation of PC algorithm available in `CausalDag` library in Python.

E Additional Experiments

In all the experiments below, we fix $d = 10$, $\sigma_{\min} = 0.3$, and $\sigma_{\max} = 0.6$. For linear measurement noise process, the number of measurements was fixed to $p = 15$.

Varying number of cycles. We evaluate the sensitivity of RECLAIM to the number of cycles in the graph. The number of cycles was varied between 0 and 8 in steps of 2. The results are summarized in fig. 7. The two cyclic graph methods (RECLAIM and NODAGS-Flow) remain stable across the entire range of varying cycles, whereas, the DAG methods (Anchored-CI and DCDI) exhibit a decreasing trend for both the noise processes. Overall RECLAIM outperforms all the baselines.

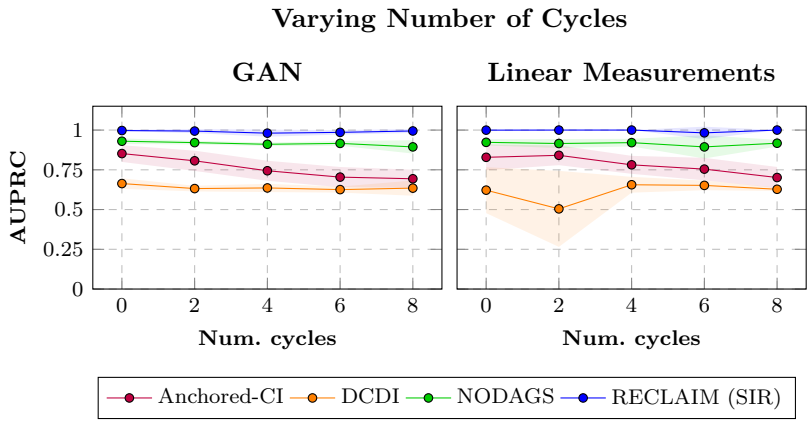


Figure 7: Performance comparison with varying number of cycles in the latent variable graph.

Varying degree of nonlinearity. We evaluate the sensitivity of RECLAIM to the degree of nonlinearity of the latent data generation process. The latent variables are sampled from the following SCM:

$$\boldsymbol{X} = (1-\beta)\mathbf{W}^\top \boldsymbol{X} + \beta \tanh \mathbf{W}^\top \boldsymbol{X} + \boldsymbol{Z},$$

where β controls the degree on nonlinearity. $\beta = 0$ corresponds to fully linear SCM and $\beta = 1$ corresponds to fully nonlinear SCM. The results are summarized in fig. 8. As seen from the figure, RECLAIM exhibits robustness to nonlinearity in the data and achieves the best performance when compared to the baselines for all values of β.

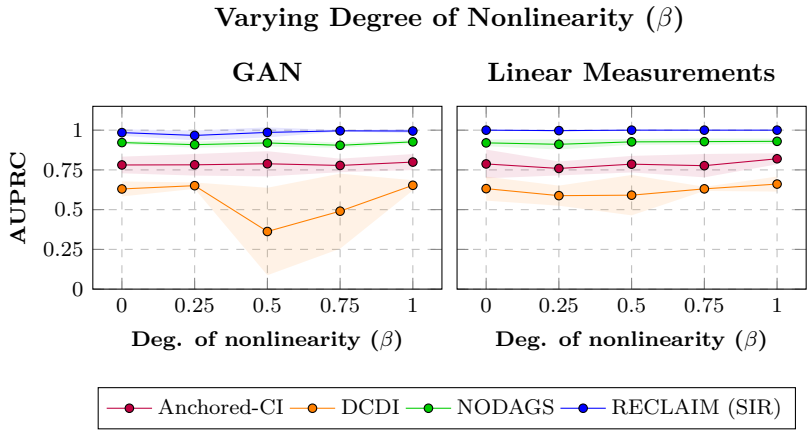


Figure 8: Performance comparison with varying degree of nonlinearity in the latent variable SCM.

Varying latent graph sparsity. We evaluate the sensitivity of RECLAIM to the sparsity of latent variable graph. The number of outgoing edge density of the latent variable graph was varied between 1 and 4. The results are summarized in fig. 9. RECLAIM remains robust with respect to the recovery performance even as the latent graphs become more dense. NODAGS-Flow exhibits a decreasing trend in AUPRC score, while the DAG methods consistently score below the cyclic methods.

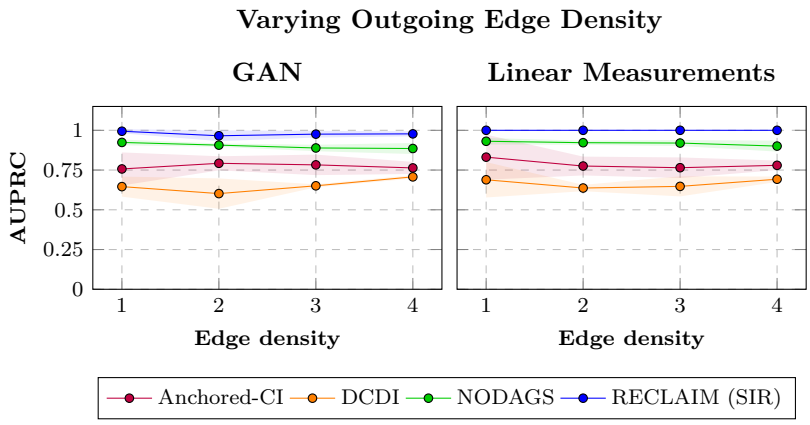


Figure 9: Performance comparison with varying latent graph sparsity.