Root Cause Analysis of Measurement and Mechanistic Anomalies

Hendrik Suhr, David Kaltenpoth and Jilles Vreeken
CISPA Helmholtz Center for Information Security
{hendrik.suhr,david.kaltenpoth,jv}@cispa.de

Abstract

Root cause analysis of anomalies aims to identify how and why a sample deviates from the normal process. Existing methods primarily focus on telling which features are responsible, ignoring that anomalies can arise through two fundamentally different processes: measurement errors, where the sample is generated normally but one or more values is recorded incorrectly, and mechanism shifts, where the causal process that generated the sample was changed. While measurement errors can often be safely corrected, mechanistic anomalies require careful consideration. In this paper, we formally define a causal model that explicitly captures both types by treating outliers as latent interventions on latent ("true") and observed ("measured") variables and show under which conditions the distinction is possible. Based on this model, we develop an efficient inference procedure for localizing root causes and distinguishing anomaly types. Experiments on synthetic and real-world data show that our method provides state-of-the-art and highly robust performance in both root cause localization and classification of anomaly types.

1 Introduction

Real-world datasets are often noisy and contain anomalies—samples generated by processes that deviate from the norm. These anomalies can compromise model reliability, lead to biased conclusions, or misguide scientific and engineering insights. Importantly, not all anomalies are alike: some result from measurement or data collection errors, such as sensor malfunctions or misrecorded values, while others reflect genuine deviations in the underlying causal mechanisms, indicating that the process itself has changed. We call these fundamental types of anomalies *measurement anomalies* and *mechanistic anomalies*. In many applications, it is important to tell these two apart: an anomalous pressure reading in an industrial plant may indicate a faulty sensor, which calls for recalibration or replacement, or a genuine change in the underlying physical process, which may require operational intervention.

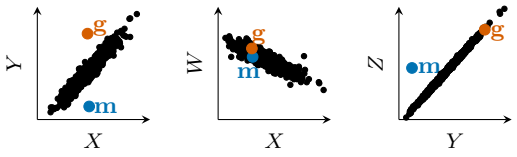


Figure 1: Different outlier types under the causal DAG $W \leftarrow X \rightarrow Y \rightarrow Z$.

We give an example using the DAG $W \leftarrow X \rightarrow Y \rightarrow Z$ in Fig. 1. In the left plot, points g and m stand out. For neither point is it immediately clear if this is due to a measurement error or change in the mechanism generating the point. We can gain more insight by examining additional causal relationships. The middle plot shows that neither point stands out for $X \rightarrow W$. The right-hand plot, however, reveals that point m violates both $X \rightarrow Y$ and $Y \rightarrow Z$ in a way that is consistent with having measured too low a value for Y. In contrast, as the red point (g) does follow the regular relationships between $X \rightarrow W$ and $Y \rightarrow Z$, this indicates this anomaly is most likely generated by a mechanism shift between $X \rightarrow Y$.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Explainable anomaly detection (XAD) typically relies on feature-attribution methods that explain an anomaly score with respect to a trained model, quantifying which inputs the model associates most strongly with anomalousness [26, 41]. Root cause analysis (RCA) of anomalies instead seeks to identify the variables that actually cause the deviation in the underlying data-generating mechanism, by incorporating causal assumptions and reasoning [4, 12, 31, 32].

However, existing methods have thus far blurred the distinction between measurement and mechanistic outliers, despite its practical importance. To bridge this gap, we propose a causal formalization of measurement and mechanistic anomalies. Our framework allows us to not only identify which variables caused the anomaly, but in addition, also determine whether an anomaly arises from a measurement error or a mechanism change.

Specifically, we interpret mechanistic and measurement anomalies as different interventions on the underlying causal graph $\mathcal{G}$. The key difference from existing RCA methods is that we consider interventions on latent ("real") and observed ("measured") variables, respectively. Similar to existing RCA methods [32], we leverage the Independence of Causal Mechanisms [38, 48], and in particular the *Sparse Mechanism Shift* hypothesis (SMS) [36, 49]. This posits that explanations that require *multiple* mechanisms to fail *jointly* are less likely. In Fig. 1, this implies that interpreting $\mathbf{m}$ as a measurement outlier (one failure) is more plausible than concerted mechanism changes in both $X \to Y$ and $Y \to Z$.

Since clean training data is often unavailable in practice, we depart from existing methods and consider the unsupervised in-sample setting, where anomalies must be identified directly within the given dataset. Our contributions are threefold: (i) We formalize mechanistic and measurement anomalies within a causal framework. (ii) We characterize identifiability of the reduced intervention probabilities, the global probabilities with which different mechanistic and measurement interventions occur. (iii) We develop a practical algorithm that takes as input a known or learned causal DAG, achieves state-of-the-art performance on synthetic and real-world datasets, and discovers meaningful anomaly types in a case study on real-world data.

2 Related Work

We situate our work within the literature on causal reasoning for outlier analysis; an overview of unsupervised outlier detection and feature attribution is provided in Appx. A.8.

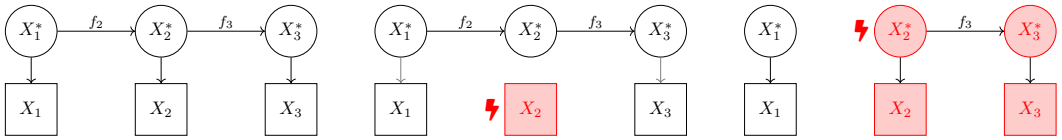
Root Cause Analysis (RCA) aims to identify the causal factors that most directly explain a target outcome of interest, ranging from treatment effects to system failures [17, 20, 21, 39, 51–53].

Recent work has instantiated RCA for anomaly detection. Budhathoki et al. [4] assume a known structural causal model (SCM) and attribute feature-level contributions using Shapley values, which was adopted in subsequent RCA work [51, 53]. Han et al. [12] propose a causal VAE for anomaly localization. Orchard et al. [32] assume a polytree-structured DAG and identify root causes via marginal outlier scores. However, these approaches primarily identify *where* anomalies originate, without explaining *how* they occur. We address this gap by extending the causal framework to jointly model the two principal categories of measurement and mechanistic outliers.

In the time series literature, heuristic distinctions between "sensor" and "process" anomalies have been proposed [2, 13, 19, 46]. These methods rely on correlation-based criteria or domain knowledge rather than an explicit causal formalization. The closest such approach distinguishes "cyber" and "measurement" anomalies using Granger causality [54], which captures predictive temporal dependencies rather than interventional causal structure. The method is further restricted to time series and assumes sufficient samples per failure mode. In contrast, our approach detects, localizes, and classifies heterogeneous failure types even with as little as a single sample per type on i.i.d. data.

3 Causal Model

Throughout, let $\mathbf{X}^* = (X_1^*, \dots, X_d^*)$ denote latent causal state variables and $\mathbf{X} = (X_1, \dots, X_d)$ their observed readouts. Most samples follow an anomaly-free process; a small fraction deviate from it and are called *outliers* or *anomalies*. Bold uppercase letters denote variable sets, and bold lowercase letters denote their samples.



(a) Clean process (b) Measurement Outlier (c) Mechanistic Outlier

Figure 2: Different outlier types induce different downstream behavior. ϟ indicate interventions that correspond to measurement outliers (intervention on observed variables) and mechanistic outliers (intervention on latent variables).

3.1 Anomaly-free Causal Model

We first describe the anomaly-free process. Let $\mathbf{V} = \mathbf{X}^* \cup \mathbf{X}$ and let $\mathcal{G} = (\mathbf{V}, E)$ be a DAG. We write $\mathrm{Pa}_j^* \subset \mathbf{X}^*$ for the latent parents of X_j^*, $\mathrm{Pa}_j^* = \{X_k^* : X_k^* \to X_j^* \in E\}$, and $\mathrm{Pa}_j := \{X_k \in \mathbf{X} : X_k^* \in \mathrm{Pa}_j^*\}$ for their observed counterparts. Each X_j has exactly one parent, X_j^* (see Fig. 2a).

Unless stated otherwise, sink nodes are defined with respect to the latent subgraph $\mathcal{G}[\mathbf{X}^*]$, i.e., they are latent variables with no children in $\mathbf{X}^*$. As standard causal modeling conditions, we assume causal sufficiency and faithfulness over $\mathbf{X}^*$ [35]:

1. *Causal Sufficiency*: There are no hidden confounders affecting two or more variables of $\mathbf{X}^*$.
2. *Causal Faithfulness*: All independencies in the data are reflected by d-separation in the graph.

In the anomaly-free regime, we model the data by a structural causal model [SCM; 35]. Each latent X_j^* is generated by a structural assignment with independent noise U_j, and each observation is a faithful measurement of the corresponding latent state, i.e. $X_j^* = f_j(\mathrm{Pa}_j^*, U_j)$, $X_j = X_j^*$. Under this clean model, the joint distribution factorizes in causal order as

$$P(\mathbf{X}^*, \mathbf{X}) = \prod_{j=1}^{d} P(X_j^* \mid \mathrm{Pa}_j^*)\, P(X_j \mid X_j^*)\,.$$

Thus, in the absence of anomalies, the observed X_j accurately reflect the latent causal process encoded by $\mathcal{G}$. Anomalies are modeled as sparse sample-level deviations from this clean process.

3.2 Mechanistic and Measurement Anomalies

We model two kinds of deviations from the clean process: perturbations of latent causal mechanisms and perturbations of their observed measurements.

Definition 3.1 (Mechanistic Anomalies)**.** A mechanistic anomaly at variable j is modeled by replacing the clean structural assignment of X_j^* with an outlier variable $O_j^{mech} \sim P_{O_j^{mech}}$. With latent indicator $Z_j \sim \mathrm{Bernoulli}(\pi_j^{mech})$,

$$X_j^* = (1 - Z_j) f_j(\mathrm{Pa}_j^*, U_j) + Z_j\, O_j^{mech}, \qquad z_j \in \{0, 1\}\,.$$

When $Z_j = 1$, the latent causal state itself is perturbed, and the effect may propagate.

Definition 3.2 (Measurement Anomalies)**.** A measurement anomaly at variable j is modeled by replacing only the observed readout X_j with an outlier variable $O_j^{meas} \sim P_{O_j^{meas}}$. With latent indicator $W_j \sim \mathrm{Bernoulli}(\pi_j^{meas})$,

$$X_j = (1 - W_j) X_j^* + W_j O_j^{meas}, \qquad w_j \in \{0, 1\}\,.$$

When $W_j = 1$, the latent causal state X_j^* remains unchanged, so the perturbation does not propagate.

The replacement variables O_j^{mech} and O_j^{meas} can be viewed as additional exogenous variables, analogous to the usual noise variables in an SCM. We take $(\mathbf{O}^{mech}, \mathbf{O}^{meas})$ to be independent of the clean SCM variables and noises, and do not represent them explicitly in the DAG. We rather interpret anomalies as hard interventions [35] on the DAG: a mechanistic anomaly $Z_j = 1$ replaces the clean latent mechanism $f_j(\mathrm{Pa}_j^*, U_j)$ by an exogenous draw from O_j^{mech}, while a measurement

anomaly $W_j = 1$ replaces the clean measurement mechanism $X_j = X_j^*$ by an exogenous draw from O_j^{meas}. We assume that anomalies are rare, meaning that the clean assignment occurs for the majority of samples, i.e., $P((\mathbf{W}, \mathbf{Z}) = \mathbf{0}) \gg 0.5$.

Each sample $\mathbf{x}^{(i)}$ can therefore be associated with an assignment-specific graph $\mathcal{G}^{(i)}$. For non-anomalous samples this graph is simply $\mathcal{G}$. For anomalous samples, the edges $\mathrm{Pa}_j^* \to X_j^*$ or $X_j^* \to X_j$ are removed according to the anomaly assignment; see Fig. 2b and Fig. 2c.

An alternative would be to model anomalies as soft interventions [35]. In contrast to the hard interventions used here, they replace the clean mechanism by a different mechanism that may still depend on those parents. For example, a mechanistic anomaly could be modeled as $X_j^* := (1 - Z_j) f_j(\mathrm{Pa}_j^*, U_j) + Z_j g_j(\mathrm{Pa}_j^*, U_j)$. We do not model such soft interventions because anomalies are rare: each failure type typically provides too few samples to reliably estimate a separate parent-dependent mechanism g_j in practice. Without prior knowledge of the failure mechanism, or enough repeated samples from the same failure mode as in causal mixture models [16, 24], soft interventions would add flexibility (variance) without reliable statistical support. We therefore use hard interventions as a minimal model of violations of the usual parent-conditional mechanism.

The key distinction between the two anomaly types lies in propagation. Because observed variables X_j have no children, a measurement anomaly $w_j = 1$ affects only the observed coordinate X_j: the underlying latent state X_j^* and all downstream variables remain unchanged. By contrast, a mechanistic anomaly $z_j = 1$ replaces the mechanism generating X_j^* itself, and therefore can affect the marginal distributions of its descendants. For instance, in Fig. 2b, X_2 is replaced at the measurement level, while the latent causal relations $X_1^* \to X_2^* \to X_3^*$ remain intact. In Fig. 2c, the mechanism generating X_2^* is replaced, thereby perturbing the downstream variables as well.

Finally, we instantiate the sparse mechanism shift principle [SMS; 49] through an independent Bernoulli prior over the latent anomaly indicators, i.e. we model $(\mathbf{W}, \mathbf{Z})$ as jointly independent. This acts as an inductive bias toward simple explanations: we prefer explanations that require fewer root causes, unless the evidence strongly suggests otherwise.

In the next section, we study inference under this model. In particular, we first analyze which intervention probabilities are identifiable once the assignment-conditional densities are fixed, and derive a tractable approximation for inferring root causes and anomaly types from observed data.

4 Inference of Root Causes and Anomaly Types

We formalize inference under the proposed causal anomaly model. We first show how assignment-conditional densities are evaluated, including the marginalization required by measurement anomalies. We then characterize natural equivalences between assignments and analyze when, given fixed assignment-conditional densities, the reduced intervention probabilities are identifiable. Finally, since exact marginalization over assignments is computationally infeasible, we introduce a tractable hard-assignment approximation.

In finite samples, anomalies are often too rare to estimate separate replacement distributions for all failure types. We therefore specify simple symmetric outlier distributions, $\tilde{P}_{O_j^{mech}} = \tilde{P}_{O_j^{meas}}$, so that the model does not favor one anomaly type a priori. For the theoretical analysis, we condition on a given DAG and anomaly-free SCM density. Together with the fixed outlier distributions, these determine the assignment-conditional densities used below. We defer the practical estimation of the anomaly-free density from contaminated data to Sec. 5.

Assignment-conditional densities We now derive the assignment-conditional density $p(\mathbf{x} \mid \mathbf{a})$, the density of a sample $\mathbf{x}$ under a given assignment $\mathbf{a} = (\mathbf{z}, \mathbf{w})$. For outlier indicators $\mathbf{a} = (\mathbf{z}, \mathbf{w})$ of a sample $\mathbf{x}$, define $\mathcal{Z}_\mathbf{a} = \{j : z_j = 1\}$ and $\mathcal{W}_\mathbf{a} = \{j : w_j = 1\}$. Let $\bar{x}_j$ be the unobserved value before measurement replacement if $j \in \mathcal{W}_\mathbf{a}$, and set $\bar{x}_j = x_j$ otherwise. Let $\bar{\mathrm{pa}}_j$ denote the parent values of node j in $\bar{\mathbf{x}}$. Marginalizing the latent values of measurement outliers yields

$$p(\mathbf{x} \mid \mathbf{a}) = \prod_{j \in \mathcal{W}_\mathbf{a}} \tilde{p}_{O_j^{meas}}(x_j) \int \left[\prod_{j \notin \mathcal{Z}_\mathbf{a}} p(\bar{x}_j \mid \bar{\mathrm{pa}}_j) \prod_{j \in \mathcal{Z}_\mathbf{a}} \tilde{p}_{O_j^{mech}}(\bar{x}_j) \right] \prod_{j \in \mathcal{W}_\mathbf{a}} d\bar{x}_j. \tag{1}$$

We provide the detailed derivation of this expression in Appx. A.1. In general, the integral cannot be computed in closed form but can be approximated via Monte Carlo sampling (Sec. 5). The joint density of $\mathbf{x}$ and $\mathbf{a}$ is obtained by multiplying the assignment-conditional density by the prior probability of the assignment:

$$p(\mathbf{x}, \mathbf{a} \mid \boldsymbol{\pi}) = p(\mathbf{x} \mid \mathbf{a}) P(\mathbf{a} \mid \boldsymbol{\pi}) = p(\mathbf{x} \mid \mathbf{a}) \prod_{j=1}^{d} \mathrm{Bern}(z_j; \pi_j^{mech}) \, \mathrm{Bern}(w_j; \pi_j^{meas}).$$

Since assignments can induce identical observable densities, we define an equivalence relation.

Definition 4.1 (Equivalent assignments)**.** Two assignments $\mathbf{a}, \mathbf{a}'$ are called equivalent, written $\mathbf{a} \equiv \mathbf{a}'$, if they induce the same assignment-conditional density, i.e., $p(\mathbf{x} \mid \mathbf{a}) = p(\mathbf{x} \mid \mathbf{a}')$ for all $\mathbf{x}$.

The most common equivalences arise at sink nodes. If j is a sink and $P_{O_j^{mech}} = P_{O_j^{meas}}$, then a mechanistic intervention on X_j^* and a measurement intervention on X_j induce the same observable density: neither can propagate downstream, and both replace only the observed coordinate X_j. Thus, for any fixed intervention pattern on the remaining nodes, $(z_j, w_j) \in \{(1,0), (0,1), (1,1)\}$ are equivalent, and we represent them by $s_j := z_j \vee w_j$. Further equivalences can arise for multi-intervention assignments, since one intervention may remove dependencies through which another intervention would otherwise be distinguishable.

Inference of root causes and anomaly types Under fixed intervention probabilities $\boldsymbol{\pi}$, the sample-conditional assignment posterior is given by

$$q_{\boldsymbol{\pi}}(\mathbf{a} \mid \mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{a} \mid \boldsymbol{\pi})}{\sum_{\mathbf{a}'} p(\mathbf{x}, \mathbf{a}' \mid \boldsymbol{\pi})}.$$

The Bayes-optimal assignment classifier is the maximum-a-posteriori assignment $\arg\max_{\mathbf{a}} q_{\boldsymbol{\pi}}(\mathbf{a} \mid \mathbf{x}) = \arg\max_{\mathbf{a}} \log p(\mathbf{x}, \mathbf{a} \mid \boldsymbol{\pi})$. Importantly, if assignment-conditional distributions overlap on a set of positive probability even the Bayes-optimal classifier has nonzero error. For example, a very unlikely statistical outlier that was generated by the anomaly-free SCM may be misclassified as a mechanistic anomaly. In practice, intervention probabilities are typically unknown. A natural reference objective is the marginalized likelihood

$$\max_{\boldsymbol{\pi}} \sum_{i=1}^{n} \log \left(\sum_{\mathbf{a}} p(\mathbf{x}^{(i)}, \mathbf{a} \mid \boldsymbol{\pi}) \right), \tag{2}$$

which corresponds to a mixture model over fixed assignment-conditional densities. Since different assignments may induce the same observable density, we can group assignments into structural equivalence classes and index mixture components by these classes. This rewrites (2) as a standard mixture model with aggregate weights, where each aggregate weight is the total prior mass of all assignments in the corresponding equivalence class. Because of sink-node equivalence, we use the reduced indicator $s_j := z_j \vee w_j$ for $j \in \mathcal{S}$, where $\mathcal{S}$ denotes the sink nodes of $G[X^*]$, and define $\rho_j := P(s_j = 1)$. Thus, $\boldsymbol{\pi}_{\mathrm{red}}$ collects $(\pi_j^{mech}, \pi_j^{meas})_{j \notin \mathcal{S}}$ and $(\rho_j)_{j \in \mathcal{S}}$. In Appx. A.2, we show that, given a suitable subset of uniquely identifiable aggregate mixture weights, the reduced Bernoulli probabilities $\boldsymbol{\pi}_{\mathrm{red}}$ can be recovered. In particular, π_j^{mech} and π_j^{meas} are separately identifiable only for non-sink nodes, while at sink nodes only $\rho_j = P(z_j = 1 \vee w_j = 1)$ is identifiable.

We show generic identifiability of $\boldsymbol{\pi}_{\mathrm{red}}$ for linear Gaussian SCMs (Appx. A.3) and for fixed-degree polynomial Gaussian SCMs (Appx. A.4). For conciseness, we here state the resulting generic identifiability guarantee in simplified form; the full statement and proof are given in Appx. A.4.

Theorem 4.2 (Generic identifiability under polynomial Gaussian SCMs (simplified))**.** *Under the setup established above, suppose the clean mechanisms are additive-noise polynomial functions of fixed degree $r \geq 1$, the exogenous noises are Gaussian with positive variances, and the mechanistic and measurement outlier distributions are nodewise identical Gaussians. Then, outside a Lebesgue-null set of model parameters, the reduced intervention probabilities $\boldsymbol{\pi}_{\mathrm{red}}$ are identifiable.*

Although identifiable, optimizing Eq. (2) is generally infeasible due to the exponential number of assignments, many of which require marginalizing latent pre-replacement values. Moreover, in the sparse finite-sample regime, many low-probability assignments are observed rarely or not at

all, making their aggregate mixture weights difficult to estimate reliably. We therefore use Eq. (2) as a reference objective, but adopt a hard-assignment approximation for finite-sample inference. Specifically, we approximate the sum over assignments by a maximum:

$$\log \sum_{\mathbf{a}} p(\mathbf{x}, \mathbf{a} \mid \boldsymbol{\pi}) \approx \max_{\mathbf{a}} \log p(\mathbf{x}, \mathbf{a} \mid \boldsymbol{\pi}) \,.$$

The gap is controlled by posterior uncertainty. We have

$$0 \leq \log \sum_{\mathbf{a}} p(\mathbf{x}, \mathbf{a} \mid \boldsymbol{\pi}) - \max_{\mathbf{a}} \log p(\mathbf{x}, \mathbf{a} \mid \boldsymbol{\pi}) \leq H\left(q_{\boldsymbol{\pi}}(\cdot \mid \mathbf{x})\right)$$

Here, $H\left(q_{\boldsymbol{\pi}}(\cdot \mid \mathbf{x})\right) = -\sum_{\mathbf{a}} q_{\boldsymbol{\pi}}(\mathbf{a} \mid \mathbf{x}) \log q_{\boldsymbol{\pi}}(\mathbf{a} \mid \mathbf{x})$ denotes the Shannon entropy of the assignment posterior $q_{\boldsymbol{\pi}}(\cdot \mid \mathbf{x})$. Thus, the approximation is tight when the posterior over assignments is concentrated, which we expect when outliers are strong enough to induce well-separated assignment-conditional densities. The resulting hard-assignment objective remains combinatorial and non-convex. For fixed assignments, the optimal intervention probabilities are their empirical frequencies; our algorithm therefore profiles them out and uses the induced Bernoulli term as an adaptive sparsity cost, as described in Sec. 5. In Sec. 6, we show that the resulting approximation closely matches the optimal assignment classifier with oracle $\boldsymbol{\pi}$ where exact inference is feasible, while scaling to real-world problems.

5 Practical Inference with CALI

We now turn the theoretical formulation into a practical algorithm that takes the observed data $\mathbf{X}$ and a learned or user-specified DAG $\hat{\mathcal{G}}$ as inputs. We first specify how the anomaly-free SCM density and outlier distributions are estimated. We then describe likelihood evaluation under a fixed assignment, including Monte Carlo marginalization for measurement anomalies. Finally, we present CALI,[1] a practical algorithm that jointly updates assignments and intervention probabilities. We present the high-level ideas and algorithms here, while further details, pseudocode, and complexity analysis are given in Appx. A.5. The source code, including the implementation of the proposed algorithm and all experimental evaluations, is publicly available on GitHub[2].

Anomaly-free density. Although the anomaly-free density could be instantiated with arbitrary causal models, we use an additive noise model (ANM) for tractable conditional-density estimation under contamination. Given $\hat{\mathcal{G}}$, each node is modeled as $X_j = f_j(\mathrm{Pa}_j) + \varepsilon_j$. We estimate $\hat{f}_j : \mathbb{R}^k \to \mathbb{R}$ using generalized additive models with cubic spline bases [GAMs; 14]. To reduce sensitivity to anomalies, we fit the mechanisms by median regression using the quantile-regression approach of Fasiolo et al. [9]. Residuals are then defined as $r_j = x_j - \hat{f}_j(\mathrm{pa}_j)$. We estimate their densities using trimmed kernel density estimation: a fraction α of observations farthest from the median is discarded before fitting a Gaussian KDE [34]. Together, the fitted mechanisms $\{\hat{f}_j\}_j$ and residual densities define the estimated anomaly-free conditionals $\hat{P}(X_j \mid \mathrm{Pa}_j)$.

Assignment likelihoods. For the replacement distributions, we use uniform distributions over the observed feature ranges. Given an assignment, likelihood evaluation follows Eq. (1). When measurement anomalies are present, the pre-replacement values are latent; we approximate the corresponding integral by Monte Carlo. For each $j \in \mathcal{W}$, we sample candidate clean values from $\hat{P}(X_j \mid \mathrm{Pa}_j = \mathrm{pa}_j)$, evaluate the remaining factors in Eq. (1), and average over samples.

Assignment optimization. For exact assignment classification under fixed $\boldsymbol{\pi}_{\text{red}}$, used only in small oracle comparisons, we enumerate all 4^d assignments for each sample and choose the maximizer of $p(\mathbf{x}, \mathbf{a} \mid \boldsymbol{\pi}_{\text{red}})$. In contrast, CALI approximates the profiled hard-assignment objective heuristically. We initialize all samples with the all-clean assignment and profile $\boldsymbol{\pi}_{\text{red}}$ by empirical assignment frequencies. Candidate interventions are prioritized by low residual likelihood under the clean density. A candidate is accepted only when the resulting gain in assignment-conditional likelihood outweighs the induced increase in Bernoulli sparsity cost. This procedure iterates until no candidate

[1] **C**ausal **A**nomaly **L**ocalization and **I**nspection

[2] `https://github.com/hsuhr99/measurement_mechanistic_rca`

update improves the objective. It reduces the search from exhaustive enumeration over assignments to a greedy scan over feature-level candidates. In Sec. 6, we show that CALI closely matches exact classification with oracle $\boldsymbol{\pi}_{\text{red}}$, where enumeration is feasible, while scaling to real-world problems.

Root-cause scores and type confidence. CALI returns a discrete assignment $\mathbf{a}$ for every sample, but it is often useful to rank candidate root causes by plausibility. Since exact posteriors are intractable, we use a likelihood-gain surrogate. For an assignment $\mathbf{a}$, let $\mathbf{a}[j \leftarrow 0]$ denote the assignment obtained by setting $(z_j, w_j) = (0, 0)$, and let $\mathbf{a}[j \leftarrow \text{mech}]$ and $\mathbf{a}[j \leftarrow \text{meas}]$ denote the assignments obtained by setting $(z_j, w_j) = (1, 0)$ and $(0, 1)$, respectively, leaving all other indicators unchanged. Let $\ell_j^0 := \log p(\mathbf{x}, \mathbf{a}[j \leftarrow 0] \mid \boldsymbol{\pi})$, $\ell_j^{\text{mech}} := \log p(\mathbf{x}, \mathbf{a}[j \leftarrow \text{mech}] \mid \boldsymbol{\pi})$, and $\ell_j^{\text{meas}} := \log p(\mathbf{x}, \mathbf{a}[j \leftarrow \text{meas}] \mid \boldsymbol{\pi})$. We define the root-cause score as $\Delta_j := \max\{\ell_j^{\text{mech}}, \ell_j^{\text{meas}}\} - \ell_j^0$, where Δ_j measures the best likelihood gain obtained by treating j as a root cause. This score is especially useful because the uniform replacement distribution is only a simple approximation, which is likely misspecified in practice. Relative likelihood gains can still rank plausible root causes highly, even when conservative replacement likelihoods make the absolute gains small or negative. We also define the type confidence as $\tilde{C}_j := \left|\ell_j^{\text{mech}} - \ell_j^{\text{meas}}\right|$. It measures approximately how strongly the model favors one anomaly type over the other, conditional on j being a root cause.

6 Experiments

We empirically evaluate CALI across various settings. In addition to the main experiments presented here, we evaluate detection performance, sensitivity to sample size and dimensionality, runtime, and robustness to violations of the modeling assumptions, including correlated root causes, soft interventions, controlled DAG misspecification, confounding, heteroscedastic noise, extreme outliers, global measurement noise, and endogenous intervention probabilities in Appx. A.9.

Data To be able to evaluate on data with known ground truth, we generate 2000 samples $\mathbf{x}$ from an SCM defined over a random Erdös–Rényi DAG with 15 nodes. Each variable X_j is specified as $X_j = f_j\big(\text{Pa}(X_j)\big) + U_j$, where f_j is a quadratic or cubic polynomial and $U_j \sim \mathcal{N}(0, \sigma^2)$. For 10% of samples, outliers are injected by shifting the marginal distribution of X_j by l empirical standard deviations and replacing the actual value with a random sample from the shifted marginal distribution, and recalculating downstream values for mechanistic outliers. We refer to l as *outlier strength*.

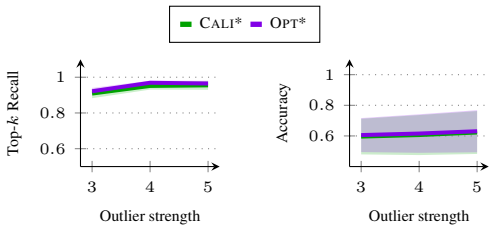


Figure 3: CALI closely matches exhaustive model-optimal assignment classification. Root-cause localization (Top-k recall, left) and anomaly-type classification accuracy (right) on synthetic data, compared against the model-optimal classifier obtained by exhaustive enumeration over assignments.

In addition, we evaluate on two real-world datasets with known consensus DAGs: the Sachs dataset [44] and the *Causal Chambers* benchmark [11]. Both contain observational and interventional data. In Causal Chambers, interventions are applied at different strengths (*mid* and *strong*), whereas interventions in the Sachs dataset induce weaker marginal shifts than the *mid* interventions. We simulate mechanistic outliers by injecting a small fraction of interventional samples, and measurement outliers by replacing individual features with values drawn from the post-intervention distribution. We consider three settings with respect to outlier strength: low (Sachs), medium (Causal Chambers, *mid*), and high (Causal Chambers, *strong*). Further details on data-generation and processing are provided in Appx. A.7.

Instantiating CALI To evaluate the best possible performance, we consider CALI with access to the true causal graph $\mathcal{G}^*$. Since the ground truth is often unknown, we also consider the case where CALI is given a learned causal DAG. To this end, we apply CALI to the outputs of various causal discovery algorithms on independently generated synthetic datasets and assess downstream performance (see Appx. A.9.8). The best performance occurs with NOGAM [28], with AE-based prefiltering, which we thus use for subsequent experiments. We write CALI* for when it has access to the true graph, and CALI when the causal graph is learned.

Evaluation Metrics For root cause localization, we report Top-k Recall, where k is set dynamically to the true number of root causes of the respective sample, which calculates what percentage of the true root causes is included in the top k candidates. This metric is detection-agnostic and ensures a fair comparison across methods. However, it does not necessarily reflect performance in deployment, as explaining anomalies that are not detected provides no actionable insight. Therefore, we also report an evaluation that takes detectability into account in Appx. A.9.1, which also includes further metrics. For the classification of outlier types, we report the accuracy over correctly detected root causes, evaluating the classification performance for actionable anomalies. The shaded regions in figures show mean ± 1.96 standard errors across 20 independent runs.

6.1 Comparison to Model-Optimal Assignment Classification

To evaluate whether the approximations used by CALI affect performance in practice, we compare it to exhaustive model-optimal assignment classification under oracle intervention probabilities $\boldsymbol{\pi}_{\text{red}}$, denoted OPT*. OPT* is optimal only with respect to the estimated likelihood, not the true data-generating distribution. Since it enumerates all assignments, it is exponential in d, so we restrict this comparison to the synthetic setting with 500 samples and 3 nodes. Fig. 3 shows that CALI* closely matches OPT* in both root-cause localization and anomaly-type classification accuracy, while being substantially more efficient (47.57s vs. 1190.60s on average). This suggests that the approximations introduce little loss in the settings where exact comparison is possible.

6.2 Root Cause Localization

To investigate how well CALI performs compared to existing root cause analysis methods, we grant all methods access to the ground-truth DAG where applicable and mark such methods with a *. We consider SCM* an adapted version of SCM-based attribution approaches [4, 51, 53], SMTR* and SCOR[3] [32], and AE-SHAP (autoencoder with Shapley values). We refer to Appx. A.6 for detailed implementations and hyperparameter selection. Since all competitors assume a given split between the normal and anomalous regimes, and are thus not directly applicable in the in-sample setting, we combine them with the best-performing sample detector aside from CALI to obtain the split.

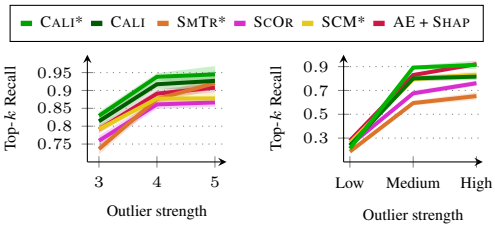


Figure 4: Root Cause Localization (Top-k recall, with k set to the true number of root causes in each sample) for synthetic data (left) and real data (right). * denotes ground-truth DAG access.

We report the results in Fig. 4. On synthetic data, CALI* achieves the best performance, with CALI using a learned graph as a close second. On the real-world datasets, all methods struggle under the weak outliers in the Sachs dataset, resulting in uniformly low Top-k recall. In contrast, on the Causal Chambers dataset, CALI* already attains a Top-k recall of 90% under the medium outlier strength, a level that is only matched by a single competitor under strong outliers. A surprising observation is that AE-SHAP is very competitive for strong interventions on the Causal Chamber dataset. It performs on par with CALI*, and outperforms CALI.

One likely reason why CALI performs worse than CALI* is the causal discovery step. We provide causal discovery metrics for both the original setting and the clean, anomaly-free Causal Chamber dataset in Tab. 4 in Appx. A.9.1. The learned graphs remain substantially different from the reference graph even when discovery is performed on clean observational data, suggesting that the performance gap between CALI and CALI* primarily reflects limitations of causal discovery on this dataset, rather than contamination alone. Another factor for the good performance of AE-SHAP may be the graph structure of the Causal Chamber dataset. The DAG is very shallow, with maximum depth 2, which reduces the penalty for ignoring causal propagation structure and gives association-based methods an unusually high chance of identifying the correct root causes.

Importantly, our evaluation measures root-cause localization under oracle detection. When we take detectability into account (see Appx. A.9.5), CALI* and CALI outperform competing approaches by an even larger margin.

[3] Short for SMOOTHTRAVERSAL, SCOREORDER

Table 1: Top 3 measurement (left) and mechanistic (right) anomalies identified by CALI on the NYC Taxi dataset, ranked by confidence. Pick-up and drop-off coordinates are summarized as air distance, and detected root causes are shown in **bold**. The second measurement example has a *Distance* value that is clearly incompatible with physical constraints, since it would imply a speed exceeding three times the speed of sound. All remaining attributes indicate a very short trip, supporting the measurement-error interpretation. For the second and third mechanistic examples, CALI correctly detects the anomalous pricing mechanism corresponding to the fixed JFK flat fare of $52.

Distance	Air Dist	Time	Fare	Tip	Total
3.90	3.16	**94.65**	10.00	2.00	13.30
9.40	0.01	0.20	2.50	0.00	3.30
0.60	0.98	**33.52**	5.00	0.00	6.30

Distance	Air Dist	Time	Fare	Tip	Total
0.00	0.06	0.75	**220.00**	44.05	264.35
0.00	0.00	0.30	**52.00**	0.00	58.13
1.80	1.39	8.48	**52.00**	11.63	69.76

6.3 Anomaly Type Classification

Next, we evaluate whether CALI can accurately classify the types of detected root causes. As a baseline, we consider MARG, which identifies outlier types using marginal outlierness. A root cause is classified as a measurement outlier if none of its child values exceed a predefined outlierness threshold. A key limitation of this baseline is the difficulty of threshold selection. Since we want to eliminate errors in causal discovery for the comparison, we only consider CALI* and MARG* for this experiment. We provide the results with learned graphs in Appx. A.9.7. To ensure a fair comparison, we provide MARG* with the root causes identified by CALI* and tune its significance threshold on synthetic data. Detailed implementation and hyperparameter selection are described in Appx. A.6.

Fig. 5 shows that CALI* outperforms MARG*. Despite the poor localization performance on the Sachs dataset, those root causes that are correctly identified are classified with high accuracy. Especially under weak outliers CALI beats MARG by a large margin. Another major advantage of CALI is that it performs robustly across different settings without requiring a threshold, as we further illustrate in the "influence of degree" experiment in Appx. A.9.7.

6.4 Case Study on NYC Taxi Data

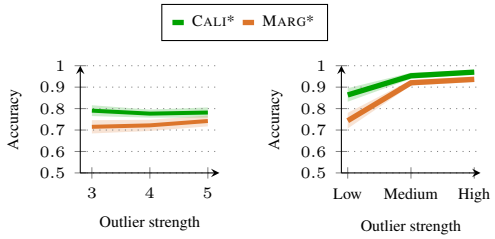


Figure 5: Classifying measurement and mechanistic anomalies. Accuracy on synthetic data (left) and real-world data (right). Random guessing corresponds to 50%.

To qualitatively verify the anomalies that CALI finds in real data, we perform a case study on the NYC Yellow Taxi Trip Dataset [8], which contains 12.7 million taxi trips with pickup and dropoff coordinates, times, fares, tips, extra charges, and the applied rate. We run CALI on a random subset of 5 000 entries using a selected set of continuous features. We illustrate the DAG used in the analysis in Fig. 9 in the Appx. A.7, and the examples identified by CALI in Tab. 1. By filtering the root causes by types, and sorting based on gain and confidence we selected 3 anomalies with high likelihood gain and high classification confidence.

First, we look into the measurement outliers in Tab. 1 (left). The first and third examples have recorded travel times of 94 and 33 minutes and distances of 3.9 and 0.6 miles, respectively, but especially the fares (10 and 5) are suspiciously low compared to what one would expect from accrued time-based charges alone. This inconsistency between distance, time, and fare, makes an incorrectly recorded travel time by far the most plausible explanation. The second example is perhaps even more clearly a measurement error: the trip distance of 9.4 miles is incompatible with the pickup and drop-off coordinates, time, and fare, all of which instead indicate a very short trip.

Next, we consider the mechanistic outliers in Tab. 1 (right). The first example has a fare that is unusually high for a trip of less than a minute and zero-distance, yet consistent with the total amount and tip, reflecting a possibly fraudulent transaction. The second and third samples are textbook definitions of a mechanistic outlier as metadata indicates that these are rides under the special JFK flat fare, which sets a fixed fare of $52 independent of distance or time.

Overall, we find that CALI successfully identifies measurement and mechanistic outliers in real-world data, with explanations that are plausible and actionable.

7 Conclusion

We studied root cause analysis from a new perspective, distinguishing measurement errors from genuine mechanistic anomalies by modeling both as interventions on latent and observed variables. We cast the problem as latent-variable maximum likelihood estimation and introduced a practical algorithm that achieves competitive root-cause localization while additionally enabling classification of outlier types. Finally, a case study on real-world data demonstrated that our approach produces plausible and actionable anomaly characterizations.

Limitations Our formulation reveals that mechanistic and measurement anomalies are not always distinguishable. In particular, at sink nodes, the two types of anomalies have identical observable effects and cannot be distinguished even with infinite data. For non-sink nodes, the right classification can also not be guaranteed since even under optimal classification there is a nonzero probability of misclassification when assignment-conditional distributions overlap. However, our experiments and especially the case study show that in practice, our method can provide useful and actionable insights into the nature of anomalies.

CALI requires a DAG, and performance depends on how well the true mechanisms, noise model, and outlier types fit our modelling assumptions. In particular, the formal model treats anomalies as exogenous hard interventions and does not model ordinary non-anomalous measurement error. Empirically, CALI remains robust to several violations of these assumptions, but performance can degrade when the model is severely misspecified. Our model also assumes a clean separation between measurements and the underlying causal system. If a recorded sensor value is fed into a downstream causal mechanism, this assumption is violated: the measurement itself becomes part of the causal system, rather than a pure leaf readout of an underlying latent state.

To enable computational tractability, we have to approximate the marginalized distribution with a hard approximation and can not provide exact posteriors. Furthermore, if the DAG is unknown, it must be estimated using a causal discovery method, whose limitations may be exacerbated by contamination. CALI is currently restricted to tabular data, although its main ideas could be extended to other data types.

Future Work A promising direction is causal discovery under contamination. Rather than treating all anomalies as nuisance samples, future work could exploit their structure: measurement outliers may distort causal signals, whereas mechanistic outliers can act as informative perturbations. Distinguishing these cases may therefore benefit structure learning from contaminated data and help close the gap between CALI and CALI*.

Another interesting direction is to amortize anomaly assignment inference using prior-fitted networks (PFNs) [30]. Rather than relying on the current hard-assignment approximation, one could train a transformer on synthetic causal models and anomaly-generation tasks to predict a distribution over plausible anomaly assignments. Such a model could incorporate assignment uncertainty directly, avoid explicit enumeration, and reduce dependence on a single learned graph by conditioning on the causal information that can be inferred from the observed data. This would allow the model to use causal structure when it is identifiable, while reflecting both structural and assignment uncertainty when it is not.

Acknowledgments and Disclosure of Funding

H.S. is a PhD student at the Saarbrücken Graduate School of Computer Science and was supported by the Liberate AI project, funded by the Initiative and Networking Fund of the Helmholtz Association.

References

[1] F. Angiulli, F. Fassetti, S. Nisticò, and L. Palopoli. Explaining outliers and anomalous groups via subspace density contrastive loss. *Machine Learning*, 113:7565–7589, 2024.

[2] C. Bachechi, F. Rollo, and L. Po. Detection and classification of sensor anomalies for simulating urban traffic scenarios. *Cluster Computing*, 25(4):2793–2817, 2022.

[3] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: identifying density-based local outliers. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, page 93–104. ACM, 2000.

[4] K. Budhathoki, L. Minorics, P. Blöbaum, and D. Janzing. Causal structure-based root cause analysis of outliers. In *International conference on machine learning*, pages 2357–2369. PMLR, 2022.

[5] P. Bühlmann, J. Peters, and J. Ernest. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42, 2014.

[6] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[7] J. F. Crenshaw, J. B. Kalmbach, A. Gagliano, Z. Yan, A. J. Connolly, A. I. Malz, S. J. Schmidt, L. D. E. S. Collaboration, et al. Probabilistic forward modeling of galaxy catalogs with normalizing flows. *The Astronomical Journal*, 168(2):80, 2024.

[8] Elemento. NYC Yellow Taxi Trip Data. `https://www.kaggle.com/datasets/elemento/nyc-yellow-taxi-trip-data`, n.d. Kaggle dataset (accessed 2025-11-11).

[9] M. Fasiolo, S. N. Wood, M. Zaffran, R. Nedellec, and Y. Goude. qgam: Bayesian nonparametric quantile regression modeling in r. *Journal of Statistical Software*, 100:1–31, 2021.

[10] L. Friedman Antwarg, R. Miller, B. Shapira, and L. Rokach. Explaining anomalies detected by autoencoders using Shapley Additive Explanations. *Expert Systems with Applications*, 186: 115736, 2021.

[11] J. L. Gamella, J. Peters, and P. Bühlmann. Causal chambers as a real-world physical testbed for AI methodology. *Nature Machine Intelligence*, 7(1):107–118, 2025.

[12] X. Han, L. Zhang, Y. Wu, and S. Yuan. On Root Cause Localization and Anomaly Mitigation through Causal Inference. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 699–708, New York, NY, USA, 2023. Association for Computing Machinery.

[13] S. A. Haque, M. Rahman, and S. M. Aziz. Sensor anomaly detection in wireless sensor networks for healthcare. *Sensors*, 15(4):8764–8786, 2015.

[14] T. J. Hastie. Generalized additive models. *Statistical models in S*, pages 249–307, 2017.

[15] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

[16] S. Hu, Z. Chen, V. Partovi Nia, L. Chan, and Y. Geng. Causal inference and mechanism clustering of a mixture of additive noise models. *Advances in neural information processing systems*, 31, 2018.

[17] A. Ikram, K. Lee, S. Agarwal, S. K. Saini, S. Bagchi, and M. Kocaoglu. Root cause analysis of failures from partial causal structures. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

[18] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. *ArXiv*, abs/1312.6114, 2013.

[19] N. LaRosa, J. Farber, P. Venkitasubramaniam, R. Blum, and A. Al Rashdan. Separating sensor anomalies from process anomalies in data-driven anomaly detection. *IEEE Signal Processing Letters*, 29:1704–1708, 2022.

[20] J. Li, B. B. Chu, I. F. Scheller, J. Gagneur, and M. H. Maathuis. Root cause discovery via permutations and cholesky decomposition. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2025.

[21] M. Li, Z. Li, K. Yin, X. Nie, W. Zhang, K. Sui, and D. Pei. Causal inference-based root cause analysis for online service systems with intervention recognition. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3230–3240, 2022.

[22] Z. Li, Y. Zhao, N. Botta, C. Ionescu, and X. Hu. COPOD: Copula-Based Outlier Detection . In *2020 IEEE international conference on data mining (ICDM)*, pages 1118–1123. IEEE, 2020.

[23] Z. Li, Y. Zhu, and M. Van Leeuwen. A Survey on Explainable Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data*, 18(1):1–54, 2023.

[24] F. Liu and L. Chan. Causal discovery on discrete data with extensions to mixture model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–19, 2015.

[25] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 413–422. IEEE, 2008.

[26] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[27] P. C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55, 1936.

[28] F. Montagna, N. Noceti, L. Rosasco, K. Zhang, and F. Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *Conference on Causal Learning and Reasoning*, pages 726–751. PMLR, 2023.

[29] R. K. Mothilal, A. Sharma, and C. Tan. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 607–617, New York, NY, USA, 2020. Association for Computing Machinery.

[30] S. Müller, N. Hollmann, S. P. Arango, J. Grabocka, and F. Hutter. Transformers can do bayesian-inference by meta-learning on prior-data. In *Fifth Workshop on Meta-Learning at the Conference on Neural Information Processing Systems*, 2021.

[31] P. Nguyen, T. Tran, S. Gupta, T. Nguyen, and S. Venkatesh. Root Cause Explanation of Outliers under Noisy Mechanisms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38: 20508–20515, 2024.

[32] W. R. Orchard, N. Okati, S. H. G. Mejia, P. Blöbaum, and D. Janzing. Root Cause Analysis of Outliers with Missing Structural Knowledge. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[33] G. Pang, C. Shen, L. Cao, and A. Hengel. Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, 54:1–38, 2021.

[34] E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

[35] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

[36] R. Perry, J. Von Kügelgen, and B. Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. *Advances in Neural Information Processing Systems*, 35:10904–10917, 2022.

[37] J. Peters and P. Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

[38] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT press, 2017.

[39] L. Pham, H. Ha, and H. Zhang. Baro: Robust root cause analysis for microservices via multivariate bayesian online change point detection. *Proceedings of the ACM on Software Engineering*, 1(FSE):2214–2237, 2024.

[40] D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. *ArXiv*, abs/1505.05770, 2015.

[41] M. T. Ribeiro, S. Singh, and C. Guestrin. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[42] P. Rolland, V. Cevher, M. Kleindessner, C. Russell, D. Janzing, B. Schölkopf, and F. Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pages 18741–18753. PMLR, 2022.

[43] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft. Deep One-Class Classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

[44] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[45] M. Sakurada and T. Yairi. Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction. In *Proceedings of the 2nd Workshop on Machine Learning for Sensory Data Analysis (MLSDA)*, page 4–11. ACM, 2014.

[46] O. Salem, A. Guerassimov, A. Mehaoua, A. Marcus, and B. Furht. Sensor fault and patient anomaly detection and classification in medical wireless sensor networks. In *2013 IEEE international conference on communications (ICC)*, pages 4373–4378. IEEE, 2013.

[47] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13:1443–1471, 2001.

[48] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

[49] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Towards Causal Representation Learning. *ArXiv*, abs/2102.11107, 2021.

[50] B. W. Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

[51] E. Strobl and T. A. Lasko. Sample-specific root causal inference with latent variables. In *Conference on Causal Learning and Reasoning*, pages 895–915. PMLR, 2023.

[52] E. V. Strobl. Counterfactual formulation of patient-specific root causes of disease. *Journal of Biomedical Informatics*, 150:104585, 2024.

[53] E. V. Strobl and T. A. Lasko. Identifying patient-specific root causes with the heteroscedastic noise model. *Journal of Computational Science*, 72:102099, 2023.

[54] Y. Sun, R. S. Blum, and P. Venkitasubramaniam. Unifying Explainable Anomaly Detection and Root Cause Analysis in Dynamical Systems. *arXiv preprint arXiv:2502.12086*, 2025.

[55] S. Wachter, B. D. Mittelstadt, and C. Russell. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 2018.

[56] S. Xu, S. Mameche, and J. Vreeken. Information-theoretic causal discovery in topological order. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

[57] S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *The annals of Mathematical statistics*, 39(1):209–214, 1968.

[58] Y. Zhao, Z. Nasrullah, and Z. Li. Pyod: A python toolbox for scalable outlier detection. *Journal of machine learning research*, 20:1–7, 2019.

[59] X. Zheng, C. Dan, B. Aragam, P. Ravikumar, and E. Xing. Learning sparse nonparametric dags. In *International conference on artificial intelligence and statistics*, pages 3414–3425. Pmlr, 2020.

A Appendix

A.1 Derivation of assignment-conditional densities

In the main text, the anomaly-free density is represented by the fixed clean conditionals $p(x_j^* | \, \mathrm{pa}_j^*)$. We use the same notation here. The density under the normal reference distribution is induced by the causal factorization of the anomaly-free SCM over latent states:

$$p_{\text{normal}}(\mathbf{x}^*) = \prod_{j=1}^{d} p(x_j^* \mid \mathrm{pa}_j^*).$$

Since we assume perfect measurement in the anomaly-free setting, $X_j = X_j^*$, the induced observed density is

$$p_{\text{normal}}(\mathbf{x}) = \prod_{j=1}^{d} p(x_j \mid \mathrm{pa}_j).$$

If we now consider the density conditional on a certain intervention pattern $\mathbf{a} = (\mathbf{z}, \mathbf{w})$, we need to distinguish between mechanistic and measurement interventions. For simplicity, we first consider these cases separately. In case of a mechanistic intervention at node j, the latent value is replaced by a draw from an exogenous distribution, which is perfectly observed. Hence the conditional density is given by the typical causal factorization, except that the conditional for node j is replaced by the exogenous distribution. We denote $\mathcal{Z}_\mathbf{a} = \{j : z_j = 1\}, \mathcal{W}_\mathbf{a} = \{j : w_j = 1\}$ and obtain

$$p(\mathbf{x} \mid \boldsymbol{z}) = \prod_{j \in \mathcal{Z}_\mathbf{a}} \tilde{p}_{O_j^{mech}}(x_j) \prod_{j \notin \mathcal{Z}_\mathbf{a}} p(x_j \mid \mathrm{pa}_j).$$

For a measurement intervention, the observed value is replaced by a draw from an exogenous distribution, but the latent value that nonetheless propagated through the causal system is unobserved. Hence, we need to marginalize over this latent value. To separate the latent values that we integrate over from the observed values, we define $\bar{x}_j$ to be the unobserved value before measurement replacement for $j \in \mathcal{W}_\mathbf{a}$, and set $\bar{x}_j = x_j$ otherwise. Let $\bar{\mathrm{pa}}_j$ denote the parent vector evaluated at $\bar{\mathbf{x}}$. This means that the notation is somewhat overloaded since $\bar{x}_j$ can both denote an integration variable, and an observed value, depending on whether $j \in \mathcal{W}_\mathbf{a}$ or not. We obtain

$$p(\mathbf{x} \mid \boldsymbol{w}) = \prod_{j \in \mathcal{W}_\mathbf{a}} \tilde{p}_{O_j^{meas}}(x_j) \int \left[\prod_{j=1}^{d} p(\bar{x}_j \mid \bar{\mathrm{pa}}_j) \right] \prod_{j \in \mathcal{W}_\mathbf{a}} d\bar{x}_j.$$

By combining the two cases in the density under arbitrary assignments $\mathbf{a}$, we obtain

$$p(\mathbf{x} \mid \mathbf{a}) = \prod_{j \in \mathcal{W}_\mathbf{a}} \tilde{p}_{O_j^{meas}}(x_j) \int \left[\prod_{j \notin \mathcal{Z}_\mathbf{a}} p(\bar{x}_j \mid \bar{\mathrm{pa}}_j) \prod_{j \in \mathcal{Z}_\mathbf{a}} \tilde{p}_{O_j^{mech}}(\bar{x}_j) \right] \prod_{j \in \mathcal{W}_\mathbf{a}} d\bar{x}_j,$$

which is Eq. (1) in the main text.

A.2 Identifiability of intervention probabilities

We reduce identifiability of the intervention probabilities $\boldsymbol{\pi}_{\mathrm{red}}$ to identifiability of the weights of fixed assignment-conditional densities.

Theorem A.1 (Identifiability of reduced intervention probabilities)**.** *Let $\{\mathbf{a}_k\}_{k=1}^m$ be a subset of assignments whose densities $p(\mathbf{x} \mid \mathbf{a}_k)$ correspond to singleton equivalence classes under the reduced sink-node parameterization. Let m_k denote their mixture weights. Write the reduced Bernoulli coordinates as*

$$\boldsymbol{q} = \big((\pi_j^{mech})_{j \in [d] \setminus \mathcal{S}}, (\pi_j^{meas})_{j \in [d] \setminus \mathcal{S}}, (\rho_j)_{j \in \mathcal{S}}\big), \qquad \rho_j := P(z_j = 1 \vee w_j = 1).$$

Assume:

(i) $q_\ell \in (0,1)$ *for all reduced Bernoulli coordinates* q_ℓ *of* $\boldsymbol{q}$*;*
(ii) the weights m_k *are uniquely identifiable from the full mixture;*
(iii) the matrix $A \in \mathbb{R}^{m\times(1+2d-|\mathcal{S}|)}$ *with rows*

$$A_k = \big(1, z^{(k)}_{[d]\setminus\mathcal{S}}, w^{(k)}_{[d]\setminus\mathcal{S}}, s^{(k)}_{\mathcal{S}}\big)$$

has full column rank.

Then the reduced intervention probabilities $\boldsymbol{\pi}_{\mathrm{red}}$ *are identifiable. In particular, at sink nodes only* $\rho_j = P(z_j = 1 \vee w_j = 1)$ *is identifiable, not* π_j^{mech} *and* π_j^{meas} *separately.*

Proof. Let $\mathcal{N} := [d] \setminus \mathcal{S}$ denote the non-sink nodes and define $s_j := z_j \vee w_j$ for $j \in \mathcal{S}$. For an assignment $\mathbf{a}_k$, let

$$y_k = \big(z^{(k)}_{\mathcal{N}}, w^{(k)}_{\mathcal{N}}, s^{(k)}_{\mathcal{S}}\big) \in \{0,1\}^{2d-|\mathcal{S}|}.$$

The reduced prior mass of $\mathbf{a}_k$ can be written as

$$m_k = \prod_\ell q_\ell^{y_{k\ell}}(1-q_\ell)^{1-y_{k\ell}}.$$

By assumption (ii), these weights m_k are uniquely determined from the observed mixture.

By assumption (i), all reduced Bernoulli coordinates satisfy $q_\ell \in (0,1)$. Hence all selected weights are positive and logarithms are well-defined. Taking logarithms and rewriting in log-odds form gives

$$\log m_k = c + \sum_\ell y_{k\ell}\theta_\ell,$$

where

$$\theta_\ell := \log\frac{q_\ell}{1-q_\ell}, \qquad c := \sum_\ell \log(1-q_\ell).$$

Equivalently, in the original notation,

$$\theta = \big(c, (u_j)_{j\in\mathcal{N}}, (v_j)_{j\in\mathcal{N}}, (r_j)_{j\in\mathcal{S}}\big)^\top,$$

with

$$u_j := \log\frac{\pi_j^{mech}}{1-\pi_j^{mech}}, \qquad v_j := \log\frac{\pi_j^{meas}}{1-\pi_j^{meas}}, \qquad r_j := \log\frac{\rho_j}{1-\rho_j}.$$

Stacking the equations gives

$$A\theta = b, \qquad b = (\log m_1, \dots, \log m_m)^\top.$$

By assumption (iii), A has full column rank, so θ is uniquely determined. Applying the inverse logit map to the non-intercept coordinates recovers

$$\pi_j^{mech}, \pi_j^{meas} \quad (j \in \mathcal{N}), \qquad \rho_j \quad (j \in \mathcal{S}).$$

Thus $\boldsymbol{\pi}_{\mathrm{red}}$ is identifiable on the open parameter space.

Finally, the reduction at sink nodes identifies only

$$\rho_j = P(z_j = 1 \vee w_j = 1),$$

because mechanistic and measurement interventions at a sink induce the same reduced assignment-conditional distribution. Thus π_j^{mech} and π_j^{meas} are not separately identifiable at sink nodes. □

Remark A.2 (Boundary cases). The assumption $q_\ell \in (0,1)$ is used only to make the log-odds parameterization well-defined. Some boundary cases can also be handled.

For a reduced Bernoulli coordinate q_ℓ, the product form of the assignment weights implies

$$q_\ell = 0 \quad \Rightarrow \quad m_k = 0 \text{ for every selected assignment with } y_{k\ell} = 1,$$

and

$$q_\ell = 1 \quad \Rightarrow \quad m_k = 0 \text{ for every selected assignment with } y_{k\ell} = 0.$$

Converses hold when the selected assignments contain the relevant opposing indicator values and all other coordinates are compatible with positive mass. Under such coverage conditions, boundary coordinates can be identified from the zero pattern of the identifiable weights.

After fixing the identified boundary coordinates, one discards assignments whose weights are forced to be zero and restricts the matrix A to the columns corresponding to the remaining free Bernoulli coordinates, together with the intercept. On this restricted problem the remaining coordinates lie in $(0, 1)$, so the log-odds argument above applies. If the restricted assignment matrix has full column rank, the remaining coordinates are uniquely identified.

In the anomaly setting of this paper, we assume that the all-clean assignment has non-negligible, indeed dominant, mass. This excludes boundary cases with $q_\ell = 1$, since any such coordinate would assign zero probability to the all-clean assignment.

Lemma A.3 (Sufficient condition for identifiable selected weights)**.** *For a set of fixed densities $\{p_k\}_{1\le k\le K}$, and a subset $\{p_j\}_{j\in\mathcal{J}}$, assume that*

(a) the selected densities $\{p_j\}_{j\in\mathcal{J}}$ are linearly independent, and

(b) their span intersects the span of the remaining densities only trivially:

$$\operatorname{span}\{p_j : j \in \mathcal{J}\} \cap \operatorname{span}\{p_j : j \notin \mathcal{J}\} = \{0\}.$$

Then the mixture weights m_j of the selected components, $j \in \mathcal{J}$, are uniquely identifiable from the full mixture.

A stronger but simpler sufficient condition is global linear independence of the full density family $\{p_k\}_{1\le k\le K}$.

Proof. Assume that two sets of mixture weights induce the same mixture density:

$$\sum_{k=1}^{K} m_k p_k = \sum_{k=1}^{K} m'_k p_k.$$

Then

$$\sum_{k=1}^{K} (m_k - m'_k) p_k = 0.$$

Splitting the sum into selected and remaining components gives

$$\sum_{j\in\mathcal{J}} (m_j - m'_j) p_j = -\sum_{k\notin\mathcal{J}} (m_k - m'_k) p_k.$$

The left-hand side lies in $\operatorname{span}\{p_j : j \in \mathcal{J}\}$, while the right-hand side lies in $\operatorname{span}\{p_k : k \notin \mathcal{J}\}$. By assumption (b), their common value must lie in the trivial intersection of these two spans, and hence must be zero:

$$\sum_{j\in\mathcal{J}} (m_j - m'_j) p_j = 0.$$

By assumption (a), the selected densities are linearly independent. Therefore

$$m_j - m'_j = 0 \qquad \text{for all } j \in \mathcal{J}.$$

Thus the selected mixture weights are uniquely identifiable.

If the full family $\{p_k\}_{k=1}^{K}$ is linearly independent, then the same argument applied to the full sum immediately yields $m_k = m'_k$ for every $k = 1, \ldots, K$. In particular, the selected weights are identifiable. □

A.3 Generic identifiability for linear Gaussian SCMs

We show in the following that for linear Gaussian SCMs, and Gaussian outlier distributions, the reduced intervention probabilities are generically identifiable.

Theorem A.4 (Generic identifiability in linear-Gaussian SCMs)**.** *Fix a known DAG $G = (V, E)$ over d variables with a known topological order. Consider the linear-Gaussian SCM*

$$X_j^* = \sum_{k \in \mathrm{Pa}(j)} b_{kj} X_k^* + \varepsilon_j, \qquad \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2), \qquad \sigma_j^2 > 0,$$

with mutually independent noises. Observations satisfy $X_j = X_j^*$ unless a measurement intervention occurs.

For each node j, assume mechanistic and measurement replacement variables share the same Gaussian distribution

$$O_j^{\mathrm{mech}}, O_j^{\mathrm{meas}} \sim \mathcal{N}(\alpha_j, \tau_j^2), \qquad \tau_j^2 > 0,$$

independent of all other variables.

For each assignment $a = (z, w)$, let $p_a(x; \theta)$ denote the corresponding assignment-conditional density, where

$$\theta = (b_{kj}, \sigma_j, \alpha_j, \tau_j) \in \Theta$$

and the admissible parameter space is

$$\Theta = (\mathbb{R})^{|E|} \times (0, \infty)^d \times \mathbb{R}^d \times (0, \infty)^d.$$

Thus the coefficients b_{kj} and replacement means α_j are unrestricted real parameters, while standard deviations are strictly positive.

Define structural equivalence by

$$a \sim a' \quad \Longleftrightarrow \quad p_a(x; \theta) = p_{a'}(x; \theta) \quad \textit{for all } \theta \in \Theta.$$

Let S denote the set of sink nodes of G.

Then, outside a Lebesgue-null subset of Θ, the assignment densities indexed by structural equivalence classes are linearly independent. Consequently, the corresponding aggregate mixture weights are identifiable. Moreover, the clean assignment together with singleton interventions yields a full-rank system, and therefore the reduced intervention probabilities

$$\boldsymbol{q} = \big((\pi_j^{mech})_{j \in [d] \setminus \mathcal{S}}, (\pi_j^{meas})_{j \in [d] \setminus \mathcal{S}}, (\rho_j)_{j \in \mathcal{S}}\big), \qquad \rho_j := P(z_j = 1 \vee w_j = 1)$$

are generically identifiable for reduced Bernoulli coordinates satisfying $q_\ell \in [0, 1)$.

We first show in the following lemma, that for an arbitrary DAG, that the clean assignment and all singleton assignments, are pairwise not structurally equivalent, and also not structurally equivalent to any multi-intervention assignment. It is sufficient to show that we can construct one admissible parameter setting such that the assignment-conditional densities differ. Together with the standard linear independence of finite families of distinct Gaussian densities, this will imply that the selected densities are linearly independent and have trivial intersection with the span of the remaining components.

Lemma A.5 (Clean and singleton assignments are not merged with multi-intervention assignments)**.** *Fix a known DAG G with topological order $1, \ldots, d$, and let S denote the set of sink nodes. Consider the linear-Gaussian SCM*

$$X_j^* = \sum_{k \in \mathrm{Pa}(j)} b_{kj} X_k^* + \varepsilon_j, \qquad \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2), \qquad \sigma_j^2 > 0,$$

with clean measurements $X_j = X_j^*$ unless a measurement intervention occurs. For each node j, let mechanistic and measurement replacement variables have the same distribution

$$O_j^{\mathrm{mech}}, O_j^{\mathrm{meas}} \sim \mathcal{N}(\alpha_j, \tau_j^2), \qquad \tau_j^2 > 0.$$

Let a_0 denote the clean assignment. For $j \notin S$, let a_j^{mech} denote the singleton mechanistic intervention at j, and a_j^{meas} the singleton measurement intervention at j. For $j \in S$, let a_j^{sink} denote the reduced singleton intervention $s_j = z_j \vee w_j = 1$.

There exists an admissible parameter setting θ_0 such that the assignment-conditional densities corresponding to

$$a_0, \qquad a_j^{\mathrm{mech}}\ (j \notin S), \qquad a_j^{\mathrm{meas}}\ (j \notin S), \qquad a_j^{\mathrm{sink}}\ (j \in S)$$

are distinct, except for the intended sink-node collapse between z_j and w_j when $j \in S$, and are also distinct from any assignment involving two or more reduced interventions.

Hence, the selected assignments form distinct structural equivalence classes.

Proof. We exhibit one admissible parameter setting at which the relevant assignment-conditional distributions are distinct. This is sufficient to prove non-structural-equivalence, since structural equivalence means equality of the assignment-conditional densities for all admissible parameter values.

Choose

$$\sigma_j^2 = \tau_j^2 = 1 \qquad \text{for all } j.$$

Choose every structural coefficient on an edge to be

$$b_{kj} = 1.$$

Let

$$\alpha_1, \ldots, \alpha_d$$

be algebraically independent over $\mathbb{Q}$. Such a choice is possible because $\mathbb{R}$ contains finite sets of arbitrary size that are algebraically independent over $\mathbb{Q}$. In particular, the α_j are linearly independent over $\mathbb{Q}$, and each α_j is nonzero.

For any assignment $a = (z, w)$, write

$$R(a) := \{j : z_j \vee w_j = 1\}$$

for the set of nodes with a reduced intervention.

We first record a consequence of the linear SCM under the above parameter choice. For every assignment a and every node r, the mean $\mathbb{E}_a[X_r]$ is an integer linear combination of replacement means

$$\{\alpha_\ell : \ell \preceq r\},$$

where $\ell \preceq r$ means that $\ell = r$ or ℓ is an ancestor of r. Indeed, this follows by induction along the topological order. If $w_r = 1$, then $X_r = O_r^{\text{meas}}$, so

$$\mathbb{E}_a[X_r] = \alpha_r.$$

If $w_r = 0$ and $z_r = 1$, then $X_r = X_r^* = O_r^{\text{mech}}$, so again

$$\mathbb{E}_a[X_r] = \alpha_r.$$

If $w_r = z_r = 0$, then $X_r = X_r^*$ and

$$\mathbb{E}_a[X_r] = \sum_{k \in \text{Pa}(r)} \mathbb{E}_a[X_k^*],$$

so the claim follows from the induction hypothesis for the parents of r.

Consequently, let a and a' be two assignments, and let r be the earliest node in topological order at which their reduced intervention indicators differ:

$$z_r \vee w_r \neq z'_r \vee w'_r.$$

Exactly one of the two assignments has a reduced intervention at r. The assignment with the reduced intervention contributes the direct term α_r to the mean of X_r, while the other assignment does not. Any contribution to the mean of X_r coming from strict ancestors of r is an integer linear combination of replacement means α_ℓ with $\ell \neq r$. Therefore

$$\mathbb{E}_a[X_r] - \mathbb{E}_{a'}[X_r] = \pm\alpha_r + \sum_{\ell \neq r} c_\ell \alpha_\ell$$

for some integers c_ℓ. Since the α_ℓ are linearly independent over $\mathbb{Q}$, this difference is nonzero.

We now compare the selected assignments with arbitrary assignments.

Case 1: The clean assignment. Under the clean assignment a_0, all structural noises have mean zero and no replacement variables are used. Hence

$$\mathbb{E}_{a_0}[X_j] = 0 \qquad \forall j.$$

Let a be any non-clean assignment. Then $R(a) \neq \emptyset$. Let r be the earliest node in $R(a)$. Since no strict predecessor of r has a reduced intervention, no replacement mean from an ancestor can propagate into X_r. If $w_r = 1$, then

$$\mathbb{E}_a[X_r] = \alpha_r.$$

If $w_r = 0$ and $z_r = 1$, then also

$$\mathbb{E}_a[X_r] = \alpha_r.$$

In both cases $\mathbb{E}_a[X_r] = \alpha_r \neq 0$, whereas $\mathbb{E}_{a_0}[X_r] = 0$. Therefore

$$p_a(\cdot; \theta_0) \neq p_{a_0}(\cdot; \theta_0),$$

and no non-clean assignment is structurally equivalent to the clean assignment.

Case 2: A selected singleton versus an assignment with a different reduced intervention set. Fix a selected singleton reduced assignment at node j. If $j \notin S$, this selected singleton is either a_j^{mech} or a_j^{meas}. If $j \in S$, it is the reduced sink singleton a_j^{sink}, represented by one of the assignments with $s_j = z_j \vee w_j = 1$ and no other reduced interventions.

Let a be any assignment such that

$$R(a) \neq \{j\}.$$

Let

$$r := \min_{\text{topological order}} \big(R(a) \triangle \{j\}\big)$$

be the earliest node at which the reduced intervention pattern of a differs from that of the selected singleton. By the preliminary argument,

$$\mathbb{E}_a[X_r] - \mathbb{E}_{\mathrm{single}}[X_r] = \pm\alpha_r + \sum_{\ell \neq r} c_\ell \alpha_\ell$$

for some integers c_ℓ. The coefficient of α_r is nonzero. Hence, by linear independence of $\alpha_1, \ldots, \alpha_d$ over $\mathbb{Q}$,

$$\mathbb{E}_a[X_r] \neq \mathbb{E}_{\mathrm{single}}[X_r].$$

Thus the observed mean vectors differ, and therefore

$$p_a(\cdot; \theta_0) \neq p_{\mathrm{single}}(\cdot; \theta_0).$$

So a is not structurally equivalent to the selected singleton.

This covers all assignments whose reduced intervention set is not exactly $\{j\}$, including assignments that remove the singleton intervention at j, assignments that add interventions at other nodes, and assignments that do both.

Case 3: A selected singleton versus assignments with the same reduced intervention set. It remains to compare assignments whose reduced intervention set is exactly

$$R(a) = \{j\},$$

but whose intervention type at j may differ from the selected singleton.

Case 3a: Sink nodes. First suppose $j \in S$. Since j is a sink node, it has no descendants in $G[X^*]$. A mechanistic intervention at j replaces X_j^* by O_j^{mech}, and the observation satisfies $X_j = X_j^*$ unless a measurement intervention occurs. A measurement intervention at j replaces X_j by O_j^{meas}. The two replacement variables have the same distribution by assumption, and no downstream observed coordinates are affected because j has no descendants. Thus the assignments with

$$(z_j, w_j) = (1, 0), \quad (0, 1), \quad (1, 1)$$

all induce the same observed distribution. This is exactly the intended sink-node collapse represented by

$$s_j = z_j \vee w_j.$$

Case 3b: Non-sink nodes. Now suppose $j \notin S$. Then j has at least one child in $G[X^*]$. Choose

$$k \in \mathrm{Ch}(j).$$

Because all edge coefficients have been set to 1, the coefficient of X_j^* in the linear expansion of X_k^* is the number of directed paths from j to k. Since k is a child of j, this number is at least one. Denote this strictly positive integer by

$$\beta_{jk} > 0.$$

Since no assignment considered in this case intervenes on k, we have $X_k = X_k^*$.

We compare the three possible assignments with reduced intervention set $\{j\}$.

Mechanistic singleton versus measurement singleton. Under the measurement singleton a_j^{meas}, the latent variable X_j^* is not intervened on. Since all structural noises have mean zero and there are no other interventions,

$$\mathbb{E}_{a_j^{\mathrm{meas}}}[X_j^*] = 0.$$

Hence no mean shift is propagated from j to k, and

$$\mathbb{E}_{a_j^{\mathrm{meas}}}[X_k] = 0.$$

Under the mechanistic singleton a_j^{mech}, the latent variable X_j^* is replaced by O_j^{mech} with mean α_j. Therefore the mean shift propagated to k is

$$\mathbb{E}_{a_j^{\mathrm{mech}}}[X_k] = \beta_{jk}\alpha_j.$$

Since $\beta_{jk} > 0$ and $\alpha_j \neq 0$, this mean is nonzero. Hence a_j^{mech} and a_j^{meas} induce different observed distributions and are not structurally equivalent.

Both interventions at j versus the measurement singleton. Let a_j^{both} denote the assignment with

$$z_j = w_j = 1$$

and no other reduced interventions. Under a_j^{both}, the latent variable X_j^* is replaced by O_j^{mech}, so the same mean shift is propagated to k:

$$\mathbb{E}_{a_j^{\mathrm{both}}}[X_k] = \beta_{jk}\alpha_j.$$

Under a_j^{meas},

$$\mathbb{E}_{a_j^{\mathrm{meas}}}[X_k] = 0.$$

Thus a_j^{both} and a_j^{meas} induce different observed distributions and are not structurally equivalent.

Both interventions at j versus the mechanistic singleton. Under the mechanistic singleton a_j^{mech}, the observed coordinate satisfies

$$X_j = X_j^* = O_j^{\mathrm{mech}}.$$

In the linear expansion of X_k, the coefficient multiplying X_j^* is β_{jk}, while all remaining terms are independent of O_j^{mech}. Hence

$$\mathrm{Cov}_{a_j^{\mathrm{mech}}}(X_j, X_k) = \mathrm{Cov}(O_j^{\mathrm{mech}}, \beta_{jk}O_j^{\mathrm{mech}}) = \beta_{jk}\tau_j^2 \neq 0.$$

Under a_j^{both}, the observed coordinate is

$$X_j = O_j^{\mathrm{meas}},$$

whereas the latent variable propagating to descendants is

$$X_j^* = O_j^{\mathrm{mech}}.$$

The variables O_j^{meas} and O_j^{mech} are independent, and all other terms in X_k are independent of O_j^{meas}. Hence

$$\mathrm{Cov}_{a_j^{\mathrm{both}}}(X_j, X_k) = 0.$$

Thus a_j^{both} and a_j^{mech} induce different covariance matrices and are not structurally equivalent.

Therefore, for every non-sink node j, the mechanistic singleton, the measurement singleton, and the assignment with both interventions at j are pairwise not structurally equivalent.

Combining Cases 1–3, the clean assignment and the selected singleton reduced assignments are not structurally equivalent to any other assignment, except for the intended sink-node collapse at original sink nodes. This collapse is exactly represented by the reduced intervention

$$s_j = z_j \vee w_j.$$

□

Now we show that one such parameter setting θ_0 is sufficient to establish that the linear independence conditions hold generically, in the sense that the set of parameters for which they do not hold is a Lebesgue-null subset of the parameter space.

Proof of Theorem A.4. We proceed in steps.

Step 1: Gaussian form. Under any assignment a, the SCM remains linear Gaussian. Since both marginalization, and hard replacement preserve the Gaussian property, we have

$$p_a(x;\theta) = \mathcal{N}(x;\mu_a(\theta),\Sigma_a(\theta)).$$

Both $\mu_a(\theta)$ and $\Sigma_a(\theta)$ are polynomial functions of θ.

Step 2: Generic separation. For a, a', define

$$F_{a,a'}(\theta) = (\mu_a(\theta) - \mu_{a'}(\theta), \mathrm{vec}(\Sigma_a(\theta) - \Sigma_{a'}(\theta))).$$

Each coordinate is polynomial. If $a \not\sim a'$, then there exists θ_0 such that $F_{a,a'}(\theta_0) \neq 0$, hence at least one coordinate polynomial is nonzero. Its zero set has measure zero.

Taking a finite union over all assignment pairs gives a measure-zero set E such that, for all $\theta \notin E$, distinct structural equivalence classes yield distinct Gaussian distributions. Classical finite-mixture identifiability results imply that distinct non-degenerate Gaussian component densities are linearly independent [57]. Hence, generically, the assignment-conditional densities indexed by structural equivalence classes are linearly independent outside a Lebesgue-null subset of parameter space.

Step 3: Full-rank construction. By Lemma A.5, the clean and the reduced singleton assignments form distinct equivalence classes. Their reduced assignment rows are

$$(1, 0, \ldots, 0)$$

and standard basis perturbations in the coordinates

$$(z_{[d]\setminus S}, w_{[d]\setminus S}, s_S).$$

Subtracting the clean row yields the standard basis of $\mathbb{R}^{2d-|S|}$, so the matrix has rank $1 + 2d - |S|$.

Step 4: Recovery of probabilities. If $q_\ell \in (0,1)$ for all reduced Bernoulli coordinates, all conditions of Theorem A.1 are generically satisfied. We exclude the case $q_\ell = 1$, since it would assign zero probability to the all-clean assignment, contrary to the rare-anomaly setting. Zero-boundary cases $q_\ell = 0$ can nevertheless be handled in the generic construction: the selected assignments include the all-clean assignment and all singleton reduced interventions. Let m_0 be the all-clean weight and m_ℓ the singleton weight for reduced coordinate ℓ. Since

$$m_0 = \prod_r (1 - q_r), \qquad m_\ell = q_\ell \prod_{r \neq \ell} (1 - q_r),$$

we have, whenever $m_0 > 0$,

$$q_\ell = \frac{m_\ell}{m_0 + m_\ell}.$$

Thus $q_\ell = 0$ is identified exactly when the corresponding singleton assignment has zero weight.

□

A.4 Generic linear independence beyond linear-Gaussian SCMs

We now extend the generic linear-independence result from linear-Gaussian SCMs to fixed-degree polynomial Gaussian SCMs. The argument uses two facts. First, the polynomial Gaussian model and its linear-Gaussian submodel induce the same structural equivalence classes over assignments. Second, for polynomial Gaussian SCMs, moments of polynomial test functions are polynomial, hence real analytic, functions of the model parameters.

Throughout this subsection, fix a DAG $G = (V, E)$ and a maximum polynomial degree $r \geq 1$. For each node j, write

$$f_j(x_{\mathrm{Pa}(j)}; \beta_j) = \sum_{\gamma \in \Gamma_j(r)} \beta_{j,\gamma} x_{\mathrm{Pa}(j)}^{\gamma},$$

where $\Gamma_j(r)$ is the finite set of multi-indices of total degree at most r over the parent variables of j. Let

$$\beta = (\beta_{j,\gamma})_{j,\gamma}$$

collect all polynomial coefficients, and let

$$\theta = (\beta, \sigma, \alpha, \tau),$$

where

$$\sigma = (\sigma_1, \ldots, \sigma_d) \in (0, \infty)^d, \qquad \alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{R}^d, \qquad \tau = (\tau_1, \ldots, \tau_d) \in (0, \infty)^d.$$

Thus the natural parameter space is

$$\Theta = \mathbb{R}^{|\Gamma|} \times (0, \infty)^d \times \mathbb{R}^d \times (0, \infty)^d, \qquad |\Gamma| := \sum_{j=1}^{d} |\Gamma_j(r)|.$$

The clean structural equations are

$$X_j^* = f_j(X_{\mathrm{Pa}(j)}^*; \beta_j) + \sigma_j \eta_j, \qquad \eta_j \sim \mathcal{N}(0, 1),$$

and the clean measurement equations are deterministic identities,

$$X_j = X_j^*.$$

A mechanistic intervention at j replaces the structural equation for X_j^* by

$$X_j^* = O_j^{\mathrm{mech}} = \alpha_j + \tau_j \eta_j^{\mathrm{mech}},$$

whereas a measurement intervention at j replaces the measurement equation by

$$X_j = O_j^{\mathrm{meas}} = \alpha_j + \tau_j \eta_j^{\mathrm{meas}}.$$

All clean noises and replacement noises are mutually independent, and

$$\eta_j^{\mathrm{mech}}, \eta_j^{\mathrm{meas}} \sim \mathcal{N}(0, 1)$$

are independent copies. Thus, for each node j, O_j^{mech} and O_j^{meas} have the same node-specific Gaussian distribution but are independent draws. Replacement distributions are node-specific: replacement sources at different nodes are not identified with one another, even if some parameter values make their Gaussian distributions accidentally coincide.

For the polynomial Gaussian class, write

$$a \sim_{\mathrm{poly}} a'$$

if the assignment-conditional observed laws under a and a' are equal for every admissible polynomial-Gaussian parameter value $\theta \in \Theta$. Similarly, write

$$a \sim_{\mathrm{lin}} a'$$

if the assignment-conditional observed laws are equal for every admissible linear-Gaussian parameter value. These are structural equivalence relations: they require equality for all admissible parameter values in the corresponding model class, not merely equality at a particular parameter value.

Definition A.6 (Observed active graph)**.** Fix an assignment $a = (z, w)$. The active graph associated with a is constructed from the clean causal graph as follows.

The graph contains latent vertices $X_1^*, \ldots, X_d^*$, observed vertices $X_1, \ldots, X_d$, and active source vertices

$$C_j \quad (z_j = 0), \qquad R_j^{\mathrm{mech}} \quad (z_j = 1), \qquad R_j^{\mathrm{meas}} \quad (w_j = 1).$$

The source labels are

$$\lambda(C_j) = (\mathrm{clean}, j), \qquad \lambda(R_j^{\mathrm{mech}}) = \lambda(R_j^{\mathrm{meas}}) = (\mathrm{repl}, j).$$

The equality of the two replacement labels reflects that O_j^{mech} and O_j^{meas} are independent copies from the same node-specific replacement distribution.

The directed edges are:

- $C_j \to X_j^*$ if $z_j = 0$, and $R_j^{\mathrm{mech}} \to X_j^*$ if $z_j = 1$;
- $X_k^* \to X_j^*$ for every $k \in \mathrm{Pa}(j)$, but only if $z_j = 0$;
- $X_j^* \to X_j$ if $w_j = 0$, and $R_j^{\mathrm{meas}} \to X_j$ if $w_j = 1$.

Thus, a mechanistic intervention removes the incoming latent causal edges into X_j^*, whereas a measurement intervention removes only the measurement edge $X_j^* \to X_j$. In particular, when $z_j = w_j = 1$, the mechanistic replacement source may still propagate through X_j^* to latent descendants, while the measurement replacement source affects only the observed coordinate X_j.

We then reduce the graph in two steps.

1. Remove every non-observed vertex that has no directed path to any observed vertex, together with its incident edges.
2. If R_j^{mech} reaches observations only through the deterministic identity measurement path

   $$R_j^{\mathrm{mech}} \to X_j^* \to X_j,$$

   contract this path into a single edge

   $$R_j^{\mathrm{mech}} \to X_j.$$

The second reduction can occur only when $w_j = 0$, since otherwise the measurement edge $X_j^* \to X_j$ is absent. If $w_j = 1$, the mechanistic replacement source is either pruned, if it has no observed descendants, or retained, if it propagates to downstream observed variables.

The resulting reduced graph is called the observed active graph and is denoted $\bar{H}(a)$.

Two assignments a and a' are graphically equivalent, written $a \equiv_{\mathrm{gr}} a'$, if $\bar{H}(a)$ and $\bar{H}(a')$ are isomorphic by a directed graph isomorphism that fixes every latent vertex X_j^*, fixes every observed vertex X_j, and preserves source labels λ. Source vertices with the same label may be permuted.

Lemma A.7 (Polynomial and linear models have the same structural equivalences)**.** *Assume the setup above, and assume $r \geq 1$, so that the polynomial Gaussian SCM class contains the linear-Gaussian submodel. Then, for any two assignments a, a',*

$$a \sim_{\mathrm{poly}} a' \quad \Longleftrightarrow \quad a \equiv_{\mathrm{gr}} a' \quad \Longleftrightarrow \quad a \sim_{\mathrm{lin}} a'.$$

Hence the polynomial Gaussian class and the linear-Gaussian submodel induce the same structural equivalence classes over assignments.

Proof. We first prove that graphical equivalence implies structural equivalence. Suppose $a \equiv_{\mathrm{gr}} a'$. The observed active graph contains exactly the exogenous source variables, active structural mechanisms, and active measurement maps that can affect the observed vector. The isomorphism between $\bar{H}(a)$ and $\bar{H}(a')$ fixes every latent vertex X_j^* and every observed vertex X_j. Hence it preserves which named structural mechanism f_j is evaluated at each latent node and which observed coordinate is produced at each measurement node. The isomorphism may only permute source vertices with the same label, and such vertices are independent copies with the same distribution.

The pruning step is distribution-preserving because pruned vertices have no directed path to any observed variable. The contraction step is distribution-preserving because it contracts only the deterministic identity measurement path

$$R_j^{\mathrm{mech}} \to X_j^* \to X_j, \qquad X_j = X_j^*,$$

and only when R_j^{mech} has no other observed descendants. A contracted mechanistic replacement source at node j has the same label as a measurement replacement source at node j because these are independent copies from the same node-specific replacement distribution.

Therefore a and a' induce the same recursive computation of the observed vector in distribution for every admissible polynomial-Gaussian parameter value. Hence

$$a \sim_{\mathrm{poly}} a'.$$

Since the linear-Gaussian model is a submodel of the polynomial Gaussian model, the same argument gives

$$a \sim_{\mathrm{lin}} a'.$$

It remains to prove the converse. Suppose

$$a \not\equiv_{\mathrm{gr}} a'.$$

We show that a and a' are already separated inside the linear-Gaussian submodel.

Assign a distinct formal indeterminate b_e to every latent edge $e \in E(G[X^*])$. Let

$$\mathcal{R} = \mathbb{R}[b_e : e \in E(G[X^*])]$$

be the corresponding polynomial ring, and let $\mathcal{K}$ be its fraction field.

For each active source vertex s in $\bar{H}(a)$, define its transfer vector

$$T_a(s) \in \mathcal{R}^d$$

by

$$[T_a(s)]_i = \sum_{\pi : s \rightsquigarrow X_i} \prod_{e \in \pi \cap E(G[X^*])} b_e,$$

where the sum is over all directed paths from s to the observed vertex X_i in $\bar{H}(a)$. Source edges and clean measurement edges have coefficient one. Define the labeled transfer multiset

$$\mathcal{T}(a) = \big\{(\lambda(s), T_a(s)) : s \text{ is an active source of } \bar{H}(a)\big\},$$

with multiplicity. Define $\mathcal{T}(a')$ analogously.

We claim that $\mathcal{T}(a)$ determines $\bar{H}(a)$ up to label-preserving isomorphism. Indeed, each monomial in the i-th coordinate of $T_a(s)$ corresponds to a directed source-to-observation path from s to X_i. Since the latent graph is acyclic and all latent vertices are named, a directed path is uniquely determined by its latent edge set together with its source and observed target. Direct source-to-observation paths correspond to the empty monomial in the appropriate coordinate. All coefficients are nonnegative integers, so path contributions cannot cancel. Hence the support of the transfer polynomials recovers exactly the labeled source-to-observation path set. By construction of $\bar{H}(a)$, every retained non-observed vertex and every retained edge lies on at least one directed path from an active source to an observed vertex. Therefore the union of the recovered paths is precisely the reduced observed active graph. Thus

$$\mathcal{T}(a) = \mathcal{T}(a') \quad \Longrightarrow \quad a \equiv_{\mathrm{gr}} a'.$$

By contraposition,

$$a \not\equiv_{\mathrm{gr}} a' \quad \Longrightarrow \quad \mathcal{T}(a) \neq \mathcal{T}(a')$$

as labeled multisets.

Now work symbolically in the linear-Gaussian submodel with zero intercepts. Let the clean source C_j have mean 0 and variance σ_j^2, and let every replacement source with label (repl, j) have mean α_j and variance τ_j^2. All sources are mutually independent. Treat

$$\{b_e\}_{e \in E(G[X^*])}, \qquad \{\alpha_j, \sigma_j^2, \tau_j^2\}_{j=1}^d$$

as formal indeterminates.

Under assignment a, the symbolic observed mean and covariance are

$$\mu_a = \sum_{j=1}^{d} \alpha_j \sum_{s : \lambda(s) = (\mathrm{repl}, j)} T_a(s),$$

and

$$\Sigma_a = \sum_{j=1}^{d} \sigma_j^2 \sum_{s:\lambda(s)=(\mathrm{clean},j)} T_a(s)T_a(s)^\top + \sum_{j=1}^{d} \tau_j^2 \sum_{s:\lambda(s)=(\mathrm{repl},j)} T_a(s)T_a(s)^\top.$$

The same formulas hold for a'.

Assume, for contradiction, that

$$\mu_a = \mu_{a'}, \qquad \Sigma_a = \Sigma_{a'}.$$

Since $\alpha_j, \sigma_j^2, \tau_j^2$ are formal indeterminates, equality of these polynomial expressions implies equality of the coefficient of each such indeterminate.

For a clean label (clean, j), there is at most one active clean source. Let its transfer vector under a be $t_j(a)$, and set $t_j(a) = 0$ if the source is absent after pruning. Define $t_j(a')$ analogously. Equality of the coefficient of σ_j^2 gives

$$t_j(a)t_j(a)^\top = t_j(a')t_j(a')^\top.$$

Over the fraction field $\mathcal{K}$, equality of two rank-one outer products implies equality of the vectors up to sign:

$$t_j(a') = \pm t_j(a).$$

Because transfer vectors have entries in $\mathcal{R}$ with nonnegative coefficients, the negative sign is possible only when both vectors are zero. Thus

$$t_j(a) = t_j(a').$$

For a replacement label (repl, j), there are at most two active replacement sources, corresponding to the mechanistic and measurement replacement at node j. Let their transfer vectors under a be $u_j(a)$ and $v_j(a)$, using the zero vector for an absent or pruned source. Define $u_j(a')$ and $v_j(a')$ analogously. Equality of the coefficient of α_j in the mean gives

$$u_j(a) + v_j(a) = u_j(a') + v_j(a'),$$

and equality of the coefficient of τ_j^2 in the covariance gives

$$u_j(a)u_j(a)^\top + v_j(a)v_j(a)^\top = u_j(a')u_j(a')^\top + v_j(a')v_j(a')^\top.$$

Combining the two identities yields

$$\big(u_j(a) - v_j(a)\big)\big(u_j(a) - v_j(a)\big)^\top = \big(u_j(a') - v_j(a')\big)\big(u_j(a') - v_j(a')\big)^\top.$$

Again over $\mathcal{K}$, this implies

$$u_j(a') - v_j(a') = \pm\big(u_j(a) - v_j(a)\big).$$

Together with equality of the sums, this gives

$$\{u_j(a), v_j(a)\} = \{u_j(a'), v_j(a')\}$$

as unordered pairs.

Therefore equality of the symbolic means and covariances would force equality of every clean-source transfer vector and every replacement-source transfer multiset. Equivalently,

$$\mathcal{T}(a) = \mathcal{T}(a'),$$

contradicting $a \not\equiv_{\mathrm{gr}} a'$. Hence at least one entry of

$$\mu_a - \mu_{a'} \quad \text{or} \quad \Sigma_a - \Sigma_{a'}$$

is a nonzero polynomial in the symbolic linear-Gaussian parameters.

The admissible linear-Gaussian parameter space contains a nonempty open set: edge coefficients $b_e \in \mathbb{R}$, replacement means $\alpha_j \in \mathbb{R}$, and positive variances $\sigma_j^2, \tau_j^2 > 0$. A nonzero polynomial cannot vanish on this whole open set. Therefore there exists an admissible linear-Gaussian parameter value with positive variances such that

$$(\mu_a, \Sigma_a) \neq (\mu_{a'}, \Sigma_{a'}).$$

At this parameter value, the observed Gaussian laws differ:

$$p_a(\cdot;\theta) \neq p_{a'}(\cdot;\theta).$$

Thus

$$a \not\sim_{\mathrm{lin}} a'.$$

Since the linear-Gaussian model is contained in the polynomial Gaussian model, also

$$a \not\sim_{\mathrm{poly}} a'.$$

Combining the two directions proves

$$a \sim_{\mathrm{poly}} a' \quad \Longleftrightarrow \quad a \equiv_{\mathrm{gr}} a' \quad \Longleftrightarrow \quad a \sim_{\mathrm{lin}} a',$$

as claimed. □

Now, we can use this to extend the result of linear independence of equivalence classes under the linear model to the polynomial model

Lemma A.8 (Finite witness extension)**.** *Let $\Theta \subseteq \mathbb{R}^q$ be connected and open, and let $\mathcal{A} = \{a_1, \ldots, a_m\}$ be finite. For each $a_i \in \mathcal{A}$ and $\theta \in \Theta$, let μ_i^θ be a probability measure on $\mathbb{R}^d$.*

Suppose there exist measurable test functions $h_1, \ldots, h_m : \mathbb{R}^d \to \mathbb{R}$ such that

$$M_{i\ell}(\theta) := \int h_\ell(x)\, d\mu_i^\theta(x)$$

is well-defined and real analytic in θ on Θ. If there exists $\theta_0 \in \Theta$ such that

$$D(\theta_0) := \det\left[M_{i\ell}(\theta_0)\right]_{i,\ell=1}^m \neq 0,$$

then

$$\mu_1^\theta, \ldots, \mu_m^\theta$$

are linearly independent for Lebesgue-almost-every $\theta \in \Theta$.

Proof. Define

$$D(\theta) := \det\left[M_{i\ell}(\theta)\right]_{i,\ell=1}^m.$$

By assumption, D is real analytic on Θ. Since $D(\theta_0) \neq 0$, D is not identically zero. The zero set of a nontrivial real-analytic function on a connected open set has Lebesgue measure zero. Hence $D(\theta) \neq 0$ for Lebesgue-almost-every θ.

For such θ, suppose that the measures were linearly dependent. Then there would exist $c = (c_1, \ldots, c_m) \neq 0$ such that

$$\sum_{i=1}^m c_i \mu_i^\theta = 0$$

as a finite signed measure. Integrating against h_ℓ, for $\ell = 1, \ldots, m$, gives

$$\sum_{i=1}^m c_i M_{i\ell}(\theta) = 0.$$

Equivalently, $c^\top M(\theta) = 0$. Since $D(\theta) \neq 0$, the matrix $M(\theta)$ is invertible, and therefore $c = 0$, a contradiction. □

Lemma A.9 (Polynomial moments under interventions)**.** *Consider a known-DAG polynomial Gaussian additive-noise SCM of fixed degree at most r. For each node,*

$$X_j^* = f_j(X_{\mathrm{Pa}(j)}^*; \beta_j) + \sigma_j \eta_j, \qquad \eta_j \sim \mathcal{N}(0,1),$$

where f_j is polynomial and $\sigma_j > 0$. Let mechanistic and measurement replacement variables be independent Gaussian variables with the same node-specific distribution

$$O_j^{\mathrm{mech}} = \alpha_j + \tau_j \eta_j^{\mathrm{mech}}, \qquad O_j^{\mathrm{meas}} = \alpha_j + \tau_j \eta_j^{\mathrm{meas}}, \qquad \tau_j > 0,$$

where

$$\eta_j^{\mathrm{mech}}, \eta_j^{\mathrm{meas}} \sim \mathcal{N}(0,1)$$

are independent standard Gaussian variables.

Let θ collect the structural coefficients, noise standard deviations, replacement means, and replacement standard deviations:

$$\theta = (\beta, \sigma, \alpha, \tau).$$

For a fixed assignment a, let

$$\mu_a^\theta = \mathcal{L}_\theta^a(X)$$

denote the observed assignment-conditional distribution. If $h : \mathbb{R}^d \to \mathbb{R}$ is polynomial, then

$$\theta \mapsto \int h(x)\, d\mu_a^\theta(x)$$

is polynomial in θ.

Proof. Collect all independent standard Gaussian variables into

$$E = (\eta_1, \dots, \eta_d, \eta_1^{\mathrm{mech}}, \dots, \eta_d^{\mathrm{mech}}, \eta_1^{\mathrm{meas}}, \dots, \eta_d^{\mathrm{meas}}).$$

For an assignment $a = (z, w)$, the variables are generated recursively in a topological order by

$$X_j^* = \begin{cases} f_j(X^*_{\mathrm{Pa}(j)}; \beta_j) + \sigma_j \eta_j, & z_j = 0, \\ \alpha_j + \tau_j \eta_j^{\mathrm{mech}}, & z_j = 1, \end{cases}$$

and

$$X_j = \begin{cases} X_j^*, & w_j = 0, \\ \alpha_j + \tau_j \eta_j^{\mathrm{meas}}, & w_j = 1. \end{cases}$$

Here α_j and τ_j are shared replacement parameters for the mechanistic and measurement replacement distributions at node j, while η_j^{mech} and η_j^{meas} are independent standard Gaussian variables.

Since the graph is acyclic and all f_j are polynomial, recursive substitution yields

$$X = T_a(E; \theta),$$

where every coordinate of T_a is polynomial in E and in the model parameters

$$\theta = (\beta, \sigma, \alpha, \tau).$$

Thus, for polynomial h, the composition

$$h(T_a(E; \theta))$$

is polynomial in E and θ. Taking expectation with respect to the standard Gaussian vector E replaces monomials in E by finite Gaussian moments, so

$$\mathbb{E}[h(T_a(E; \theta))]$$

is polynomial in θ. □

Lemma A.10 (Polynomial tests separate finite Gaussian mixtures)**.** *Let*

$$\nu = \sum_{i=1}^m c_i\, \mathcal{N}(\mu_i, \Sigma_i)$$

be a finite signed combination of non-degenerate Gaussian probability measures on $\mathbb{R}^d$. If

$$\int h(x)\, d\nu(x) = 0$$

for every polynomial h, then $\nu = 0$.

Proof. The moment-generating function of ν is

$$M_\nu(t) = \int e^{t^\top x}\, d\nu(x) = \sum_{i=1}^{m} c_i \exp\left(t^\top \mu_i + \frac{1}{2} t^\top \Sigma_i t\right).$$

This is an entire function of $t \in \mathbb{C}^d$. Its derivatives at $t = 0$ are the polynomial moments of ν, which vanish by assumption. Hence $M_\nu \equiv 0$. Therefore the characteristic function

$$\widehat{\nu}(s) = M_\nu(is)$$

vanishes for all $s \in \mathbb{R}^d$. By uniqueness of the Fourier transform for finite signed measures, $\nu = 0$. □

Lemma A.11 (Finite polynomial witnesses at a Gaussian point)**.** *Let*

$$\mu_1, \dots, \mu_m$$

be linearly independent non-degenerate Gaussian probability measures on $\mathbb{R}^d$. Then there exist polynomial test functions

$$h_1, \dots, h_m$$

such that

$$\det\left[\int h_\ell(x)\, d\mu_i(x)\right]_{i,\ell=1}^{m} \neq 0.$$

Proof. Suppose no such finite collection of polynomial tests existed. Then the polynomial moment functionals would have rank strictly smaller than m on $\operatorname{span}\{\mu_1, \dots, \mu_m\}$. Hence there would exist coefficients $c_1, \dots, c_m$, not all zero, such that

$$\nu = \sum_{i=1}^{m} c_i \mu_i$$

annihilates every polynomial test function. By lemma A.10 $\nu = 0$, contradicting the assumed linear independence of $\mu_1, \dots, \mu_m$. Therefore a finite collection of polynomial tests with full-rank moment matrix exists. □

Theorem A.12 (Generic identifiability for polynomial Gaussian SCMs)**.** *Fix a known DAG G, and consider the polynomial Gaussian additive-noise SCM class of fixed degree at most $r \geq 1$ with parameter space Θ defined above. For each $\theta \in \Theta$, let $\{\mu_a^\theta : a \in \mathcal{A}\}$ denote the assignment-conditional observed probability measures indexed by the full finite set of reduced assignment classes, where assignments are quotiented by structural equivalences, including the sink-node collapse $s_j = z_j \vee w_j$ for $j \in S$.*

Then, outside a Lebesgue-null subset of Θ, the assignment-conditional observed measures $\{\mu_a^\theta : a \in \mathcal{A}\}$ are linearly independent as finite signed measures. Moreover, the corresponding assignment-conditional densities are linearly independent, the aggregate mixture weights of all reduced assignment classes are identifiable, and the reduced intervention probabilities

$$\{\pi_j^{\mathrm{mech}}, \pi_j^{\mathrm{meas}} : j \notin S\}, \qquad \{\rho_j = P(z_j \vee w_j) : j \in S\}$$

are identifiable from the fixed-component mixture, for all reduced Bernoulli coordinates $q_\ell \in [0, 1)$.

Proof. **Generic linear independence of measures.**

By Lemma A.7, the reduced assignment classes in the polynomial Gaussian class coincide with those in the linear-Gaussian submodel. By Theorem A.4, we can therefore choose a witness point θ_0 in the linear-Gaussian submodel, with all higher-order polynomial coefficients set to zero, outside the corresponding Lebesgue-null exceptional set. At this point, the assignment-conditional densities indexed by the full reduced structural equivalence classes are linearly independent. Since these densities are densities with respect to Lebesgue measure, the corresponding observed probability measures

$$\mu_{a_1}^{\theta_0}, \dots, \mu_{a_m}^{\theta_0}$$

are linearly independent as finite signed measures. Moreover, at θ_0 the model is linear Gaussian, so these measures are non-degenerate Gaussian measures. Lemma A.11 therefore implies that there exist polynomial test functions $h_1, \dots, h_m$ such that

$$D(\theta_0) := \det \left[\int h_\ell(x)\, d\mu_{a_i}^{\theta_0}(x) \right]_{i,\ell=1}^{m} \neq 0.$$

For general $\theta \in \Theta$, define

$$D(\theta) := \det \left[\int h_\ell(x)\, d\mu_{a_i}^{\theta}(x) \right]_{i,\ell=1}^{m}.$$

By Lemma A.9, each matrix entry is polynomial in $\theta = (\beta, \sigma, \alpha, \tau)$. Hence D is a polynomial, and therefore real analytic, function on Θ. Since $D(\theta_0) \neq 0$, D is not identically zero. Because Θ is connected and open, the zero set of D has Lebesgue measure zero. For every θ outside this zero set, Lemma A.8 implies that

$$\mu_{a_1}^{\theta}, \dots, \mu_{a_m}^{\theta}$$

are linearly independent as finite signed measures.

From measures to densities and mixture weights. For every $\theta \in \Theta$ and every assignment a, the observed law μ_a^θ is absolutely continuous with respect to Lebesgue measure. Indeed, all structural noises and replacement variables have strictly positive Gaussian densities and enter additively with positive standard deviations. For each observed coordinate X_j, choose an independent Gaussian source that enters X_j directly: η_j if $z_j = w_j = 0$, η_j^{mech} if $z_j = 1, w_j = 0$, and η_j^{meas} if $w_j = 1$. Conditional on all remaining exogenous sources, the map from these d chosen Gaussian variables to $(X_1, \dots, X_d)$ is triangular in topological order with nonzero diagonal entries σ_j or τ_j. Therefore the conditional law has a Lebesgue density, and integrating over the remaining exogenous variables preserves absolute continuity. Thus we may write

$$d\mu_a^\theta(x) = p_a^\theta(x)\, dx.$$

Any linear relation among the measures has Radon–Nikodym derivative equal to the corresponding linear relation among the densities. Hence linear independence of the measures and of the densities coincide. Therefore, outside the same null set, the reduced assignment-conditional densities are linearly independent. By Lemma A.3, the aggregate mixture weights of all reduced assignment classes are identifiable in the associated fixed-component mixture.

Recovery of reduced intervention probabilities.

The full assignment set contains the clean assignment and all singleton interventions. By Lemma A.5, each of these assignments is not structurally equivalent to any other assignment, except for the intended sink-node collapse: the lemma exhibits a linear-Gaussian parameter setting at which the corresponding assignment-conditional densities differ. Because this linear-Gaussian setting is contained in the polynomial model class, the same non-equivalence holds for the polynomial class. Thus the full reduced assignment set contains distinct classes for the clean assignment and all reduced singleton interventions, whose aggregate weights are identifiable.

For these selected classes, the reduced assignment matrix A has one row

$$(1, 0, \dots, 0)$$

for the clean assignment and one row

$$(1, e_\ell)$$

for each reduced singleton coordinate

$$(z_{[d]\setminus S}, w_{[d]\setminus S}, s_S).$$

Subtracting the clean row from each singleton row gives the standard basis of $\mathbb{R}^{2d-|S|}$. Hence A has full column rank.

If all reduced Bernoulli coordinates satisfy $q_\ell \in (0, 1)$, the log-odds argument of Theorem A.1 recovers the reduced intervention probabilities. The case $q_\ell = 1$ is excluded by the rare-anomaly setting, since it would assign zero probability to the all-clean assignment. Zero boundary cases

$q_\ell = 0$ are also identifiable from the clean and singleton weights. Let m_0 be the all-clean weight and m_ℓ the singleton weight for coordinate ℓ. Since

$$m_0 = \prod_r (1 - q_r), \qquad m_\ell = q_\ell \prod_{r \neq \ell} (1 - q_r),$$

we have, whenever $m_0 > 0$,

$$q_\ell = \frac{m_\ell}{m_0 + m_\ell}.$$

Thus $q_\ell = 0$ is identified exactly when the corresponding singleton weight is zero. This recovers

$$\pi_j^{\mathrm{mech}}, \quad \pi_j^{\mathrm{meas}} \quad (j \notin S), \qquad \rho_j \quad (j \in S).$$

□

Remark A.13 (Scope). The results above cover two analytically tractable model classes: linear Gaussian SCMs and fixed-degree polynomial Gaussian additive-noise SCMs containing the linear-Gaussian submodel. The finite-witness strategy may extend to broader finite-dimensional nonlinear SCM classes whenever the relevant test moments are real analytic in the model parameters, but establishing such moment analyticity requires additional assumptions, for example on domains and integrability.

Our identifiability statement is deliberately limited. We do not claim to recover the structural parameters of the clean SCM or the parameters of the replacement distributions. Rather, we show that, for generic fixed choices of these parameters, the reduced intervention probabilities are identifiable from the corresponding fixed-component mixture.

This is separate from the identifiability of the clean SCM itself. In nonlinear additive-noise models, the causal direction and, more generally, the DAG can be identifiable under standard ANM assumptions, except for known degenerate cases such as the linear-Gaussian setting [15]. In linear Gaussian SCMs, the DAG is generally identifiable only up to Markov equivalence without additional assumptions; full DAG identifiability can be obtained under equal error variances [37].

A.5 Algorithmic Details

Algorithm 1 CALI

Input: DAG $\mathcal{G}$, dataset $\mathbf{X}$
Output: anomaly assignment $\mathbf{A} = (\mathbf{Z}, \mathbf{W})$, feature scores $\mathcal{S}$, confidences $\mathcal{C}$
for all $X_k \in \mathcal{G}$ **do**
  $\hat{f}_k \leftarrow \textsc{RobustlyRegress}(\mathrm{Pa}_k, X_k)$
  $R_k \leftarrow X_k - \hat{f}_k(\mathrm{Pa}_k)$
  $\mathcal{L}(R_k) \leftarrow \textsc{EstimateNoise}(R_k)$
end for
$\mathbf{A} \leftarrow \text{MLE-}\textsc{Assign}(\mathbf{X}, \mathcal{G}, \{\hat{f}_k\}, \{\mathcal{L}(R_k)\})$
$\mathcal{S} \leftarrow \textsc{Score}(\mathbf{A}, \mathbf{X}, \mathcal{G}, \{\hat{f}_k\}, \{\mathcal{L}(R_k)\})$
$\mathcal{C} \leftarrow \textsc{Conf}(\mathbf{A}, \mathbf{X}, \mathcal{G}, \{\hat{f}_k\}, \{\mathcal{L}(R_k)\})$
return $\mathbf{A}, \mathcal{S}, \mathcal{C}$

This section provides additional details on the proposed algorithm. Algorithm 1 in Sec. 5 illustrates the high-level procedure. Here, we elaborate on the key components: the fitting of the causal mechanisms, the estimation of the noise distributions, and the heuristic assignment and gain calculations.

One practical approximation is that we do not compute the full scores in Equation (5) for all samples and features. For samples that are unassigned and thus likely clean, we only compute the $O(1)$ mechanistic gains, avoiding the costly marginalization over all features for measurement gains. The rationale is that unassigned samples are expected to have low anomaly scores, so considering both outlier types provides minimal additional benefit.

A.5.1 Causal Mechanisms

To estimate the causal mechanisms, we use `qgam` [9] with $q = 0.5$. We select the number of knots per parent as a simple rule of thumb:

$$m_{j,k} = \min\left\{\max\left(5, \lfloor N^{1/3}/2 \rfloor\right), 30\right\}.$$

Knots are placed at empirical quantile breakpoints of the parent covariates, so that dense regions receive more knots (allowing flexibility) while sparse regions receive fewer (reducing variance).

A.5.2 Marginalization

Here, we provide further details on how to approximate the integral

$$\int \underbrace{\prod_{j \notin \mathcal{Z}_\mathbf{a}} p(x_j \mid \mathrm{pa}_j)}_{\text{'clean' likelihood}} \underbrace{\prod_{j \in \mathcal{Z}_\mathbf{a}} \tilde{p}_{O^{mech}{}_j}(x_j)}_{\text{mechanistic outlier density}} \underbrace{\prod_{j \in \mathcal{W}_\mathbf{a}} \mathrm{d}x_j}_{\text{latent clean values}} . \tag{3}$$

We approximate (3) via Monte Carlo integration over the latent clean values. Let $\mathbf{a} = (\mathbf{z}, \mathbf{w})$ be a fixed assignment and define $\mathcal{Z}_\mathbf{a} = \{j : z_j = 1\}$ and $\mathcal{W}_\mathbf{a} = \{j : w_j = 1\}$. For each $j \in \mathcal{W}_\mathbf{a}$, let U_j denote the structural noise variable of node j. Since we explicitly learn the distributions P_{U_j}, which capture the clean conditional noise, from the residuals of the causal mechanism fits, we can sample from these distributions. Latent clean values are obtained by sampling M times from the corresponding conditional distributions P_{U_j} for nodes in $\mathcal{W}_\mathbf{a}$ and propagating samples in topological order, and the resulting likelihoods are then averaged:

$$\int \underbrace{\prod_{j \notin \mathcal{Z}_\mathbf{a}} p(x_j \mid \mathrm{pa}_j)}_{\text{'clean' likelihood}} \underbrace{\prod_{j \in \mathcal{Z}_\mathbf{a}} \tilde{p}_{O^{mech}{}_j}(x_j)}_{\text{mechanistic outlier density}} \underbrace{\prod_{j \in \mathcal{W}_\mathbf{a}} \mathrm{d}x_j}_{\text{latent clean values}}$$

$$\approx \frac{1}{M} \sum_{m=1}^{M} \underbrace{\prod_{j \notin \mathcal{Z}_\mathbf{a} \cup \mathcal{W}_\mathbf{a}} p(x_j^{(m)} \mid \mathrm{pa}_j^{(m)})}_{\text{'clean' likelihood}} \underbrace{\prod_{j \in \mathcal{Z}_\mathbf{a} \setminus \mathcal{W}_\mathbf{a}} \tilde{p}_{O^{mech}{}_j}(x_j^{(m)})}_{\text{mechanistic outlier density}},$$

$$x_j^{(m)} = \begin{cases} \hat{f}_j(\mathrm{pa}_j^{(m)}) + \tilde{u}_j^{(m)}, & j \in \mathcal{W}_\mathbf{a} \setminus \mathcal{Z}_\mathbf{a}, \\ \tilde{o}_j^{(m)}, & j \in \mathcal{W}_\mathbf{a} \cap \mathcal{Z}_\mathbf{a}, \\ x_j, & j \notin \mathcal{W}_\mathbf{a}, \end{cases} \qquad \tilde{u}_j^{(m)} \sim P_{U_j}, \tilde{o}_j^{(m)} \sim P_{O_j^{mech}}.$$

By the law of large numbers, the Monte Carlo approximation converges to the true marginalization for $M \to \infty$.

A.5.3 Heuristic MLE Assignment

The assignment procedure is summarized in Algorithm 2. The algorithm iteratively and greedily accepts candidate outlier assignments based on *optimistic gains*, which can be estimated without performing full marginalization. The procedure is decomposed into two main steps: assigning measurement outliers and then attributing residuals to mechanistic outliers.

Measurement outliers are first considered by sorting candidates according to their optimistic gains. Each candidate's gain is compared against the optimal alternative gain obtained from mechanistic attribution. Only candidates with positive optimistic gains that exceed the mechanistic alternative are accepted. The process continues until no remaining measurement candidates have positive optimistic gains.

For a measurement outlier candidate at node j and sample i, the optimistic gain $g_{i,j}^{opt}$ assumes that all affected residual likelihoods improve maximally, and then subtracts the cost of encoding the outlier and updating the Bernoulli probability. Specifically, the Bernoulli probability update can be

approximated using the log of the binomial coefficient: encoding a sequence of k successes and $n-k$ failures with prefix codes based on empirical probabilities yields

$$-k\log\frac{k}{n}-(n-k)\log\frac{n-k}{n}\approx\log\binom{n}{k},$$

i.e., the sequence-level Bernoulli code length is asymptotically equivalent to the binomial code length. For node j with $\sum_{i=1}^{N}W_j$ current measurement outliers, adding a new sample incurs a prior cost

$$-\log\tilde{p}_{O_j^{meas}}+\Big(\log\binom{N}{\sum_{i=1}^{N}W_j+1}-\log\binom{N}{\sum_{i=1}^{N}W_j}\Big).$$

The full optimistic measurement gain is then

$$\begin{aligned}g_{i,j}^{meas-opt}=\sum_{k\in\{j\}\cup\mathrm{Ch}(j)}&\Big[-\log\mathcal{L}(r_{ik})-\min_r\{-\log\mathcal{L}(r)\}\Big]+\log\tilde{p}_{O_j^{meas}}(x_{i,j})\\&-\Big[\log\binom{N}{\sum_{i=1}^{N}W_j+1}-\log\binom{N}{\sum_{i=1}^{N}W_j}\Big],\end{aligned}$$

where $\mathrm{Ch}(j)$ denotes the children of node j, $\mathcal{L}(r_{ik})$ is the likelihood of the residual for node k and sample i, and $\tilde{p}_{O_j^{meas}}$ is the current estimated prior of a measurement outlier at node j.

Intuitively,

1. The first term captures the maximal improvement in likelihood for affected nodes.
2. The second term penalizes the prior probability of adding a new outlier.
3. The third term accounts for the combinatorial cost of updating the number of outliers.

Mechanistic gains are defined analogously, but depend only on the residual of the node itself, i.e.

$$\begin{aligned}g_{i,j}^{mech-opt}=&-\log\mathcal{L}(r_{ij})-\min_r\{-\log\mathcal{L}(r)\}\\&+\log\tilde{p}_{O_j^{mech}}(x_{i,j})\\&-\Big[\log\binom{N}{\sum_{i=1}^{N}Z_j+1}-\log\binom{N}{\sum_{i=1}^{N}Z_j}\Big].\end{aligned}$$

If a node is not in the MB of a measurement outlier, this estimate is exact. Otherwise, exact gains are evaluated by calculating the likelihood after adding the outlier to the assignment, approximating marginalization as needed.

The overall procedure is as follows: MLE-ASSIGN performs the greedy measurement and mechanistic assignment passes. Candidates are processed in descending order of optimistic gain for measurements, and descending residual cost for mechanistic outliers. The final assignment $\mathbf{A}=(\mathbf{Z},\mathbf{W})$ is then used to compute scores and classes.

A.5.4 Complexity.

Fitting one `qgam` per non-root node k has complexity $O(Nm_k^2)$, where m_k is the number of spline basis functions used for node k. With

$$m_k=O\Big(N^{1/3}|\,\mathrm{Pa}_k\,|\Big),$$

the total fitting cost is

$$\sum_{k=1}^{d}O(Nm_k^2)=O\left(N^{5/3}\sum_{k=1}^{d}|\,\mathrm{Pa}_k\,|^2\right).$$

In the dense worst case this is $O(N^{5/3}d^3)$, while for bounded in-degree Δ it is $O(N^{5/3}d\Delta^2)$.

Mechanistic gains are inexpensive once the clean residual likelihoods are cached, giving $O(1)$ per candidate. For measurement gains, we perform a constant number of model evaluations in the

Algorithm 2 MLE-ASSIGN

Input: dataset D of size N, DAG $\mathcal{G}$, regressors $\{\hat{f}_k\}$, likelihoods $\{\mathcal{L}(R_k)\}$
Output: assignment $\mathbf{A} = (\mathbf{Z}, \mathbf{W})$
Notation: $g_{i,j}^{\text{meas-opt}}/g_{i,j}^{\text{mech-opt}}$ optimistic gains; $g_{i,j}^{\text{meas}}/g_{i,j}^{\text{mech}}$ actual gains
Initialization:
$\mathbf{A} \leftarrow (\mathbf{0}, \mathbf{0})$
Compute $g_{i,j}^{\text{opt}}$ for all (i, j)
$C \leftarrow \text{sorted}(\{(i, j)\}, g_{i,j}^{\text{meas-opt}})$ (descending)
Measurement pass:
for all $(i, j) \in C$ **with** $g_{i,j}^{\text{opt}} > 0$ **do**
  Compute $g_{i,j}^{\text{meas}}$
  $g_{i,j}^{\text{comp}} \leftarrow \sum_{k \in \{j\} \cup \text{Ch}(j),\ g_{i,k}^{\text{mech}} > 0} g_{i,k}^{\text{mech}}$
  if $g_{i,j}^{\text{meas}} > g_{i,j}^{\text{comp}}$ **then**
    $\mathbf{W}[i, j] \leftarrow 1$
    Update $g_{i,\ell}^{\text{opt}}$ for affected (i, ℓ)
  end if
end for
Mechanistic pass:
$C \leftarrow \text{sorted}(\{(i, j)\}, g_{i,j}^{\text{mech-opt}})$ (descending)
for all $(i, j) \in C$ **do**
  if $\mathbf{W}[i, j] = 1$ **then**
    continue
  end if
  Compute $g_{i,j}^{\text{mech}}$
  if $g_{i,j}^{\text{mech}} > 0$ **then**
    $\mathbf{Z}[i, j] \leftarrow 1$
  else
    continue
  end if
end for
return $\mathbf{A} = (\mathbf{Z}, \mathbf{W})$

Markov-blanket space MB_j. Thus computing measurement gains for all samples and features costs

$$O\left(N \sum_{j=1}^{d} |\,\text{MB}_j\,|\right),$$

which is $O(Nd^2)$ in the dense worst case and $O(Nd\Delta_{\text{MB}})$ when Markov blankets have size at most Δ_{MB}.

The greedy search evaluates only a linear number of candidate gains per sample, rather than enumerating the exponential assignment space. Therefore the total worst-case complexity is

$$O\left(N^{5/3} \sum_{k=1}^{d} |\,\text{Pa}_k\,|^2 + N \sum_{j=1}^{d} |\,\text{MB}_j\,|\right).$$

In the dense worst case this becomes

$$O(N^{5/3}d^3 + Nd^2) = O(N^{5/3}d^3),$$

whereas for bounded in-degree and bounded Markov-blanket size it is

$$O(N^{5/3}d + Nd).$$

In practice, parent sets and Markov blankets are typically much smaller than d, and only a subset of measurement candidates is evaluated.

A.6 Implementation Details

In this section, we provide both the implementation details of all algorithms, we compared to, and the hyperparameter selection that we performed.

A.6.1 Implementation

CALI is implemented in Python, and uses the R package `qgam` [9]. Most of our competing detection methods, namely AE [45], DEEPSVDD [43], IFOREST [25], COPOD [22], LOF [3], are implemented with the Python library `pyod` [58]. We implemented the NF-based [40] anomaly detection using the Python library `pzflow` [7].

The competing RCA methods are combined with the best sample detector (aside from CALI) of the respective experiment. For the synthetic data, this is AE and for the real data NF. After applying the sample detectors, we threshold such that the highest $p\%$ of samples with the highest scores are removed from training, where p corresponds to the true percentage of anomalies.

For the calculation of Shapley values, we use the kernel explainer implemented in the `shap` library [26] with 100 background samples. For SMTR and SCOR, we use the published implementations. For DICE [29], and M2OE [1], we adapted the implementations published by the authors to our setting. We combined DICE with a thresholded autoencoder to simulate a classification problem.

As mentioned in the main paper, SCM* is an adaptation of existing SCM-based attribution approaches [4, 51, 53]. The key distinction in our setting is that we aim to attribute anomalousness at the level of an entire sample, rather than the outlierness of an individual feature [4] or a binary target [51]. In our formulation, sample anomalousness can be quantified with the joint density, which factorizes into independent noise distributions. Consequently, the conditional likelihoods already capture the individual feature contributions, making an explicit attribution step unnecessary. For the SCM, we use the exact same implementation as CALI to ensure a fair comparison.

A.6.2 Hyperparameters

Since all of our tasks are unsupervised, we tuned all hyperparameters, by selecting the ones with the best performance in our synthetic setup with different seeds than the data used for the actual evaluation.

CALI The only hyperparameter of CALI are the number of Monte Carlo samples, which we fixed to $K = 100$ across all experiments, and the KDE parameters. We fix the trimmed fraction $\alpha = 0.01$, and use a Gaussian kernel with bandwidth selected via Silverman's rule of thumb [50] across all experiments.

Detection methods For competing detection methods, we conducted grid searches over: $bs \in \{32, 64, 128\}$, $e \in \{10, 100, 500\}$, $est \in \{100, 200, 300\}$, $neigh \in \{10, 20, 40\}$. The best configurations are summarized in Tab. 2.

Table 2: Best hyperparameters for detection methods.

Method	Batch size (bs)	Epochs (e)	Estimators (est)	Neighbors ($neigh$)
AE	32	10	–	–
NF	32	10	–	–
DEEPSVDD	128	100	–	–
IFOREST	–	–	300	–
LOF	–	–	–	10

RCA methods For the RCA methods, we searched over the same ranges of epochs and batch sizes. For M2OE, we additionally tuned the number of neighbors in the KNN step, $K \in \{10, 50, 100\}$, and for SMTR and SCOR we tested the anomaly detectors {MEDIANCDFQUANTILE,INVERSEDENSITY}. The selected configurations are shown in Tab. 3.

MARG For MARG, we also compare two marginal outlier detectors based on estimated density and median distance, as well as significance levels $\alpha \in \{0.001, 0.01, 0.05, 0.1\}$ with and without Bonferroni correction. The best configuration uses the density-based detector with $\alpha = 0.01$ with Bonferroni correction.

Table 3: Best hyperparameters for feature-level baselines.

Method	Batch size (bs)	Epochs (e)	KNN neighbors (K)	Detector
M2OE	32	500	100	-
SCOR	-	-	-	MEDIANCDFQUANTILE
SMTR	-	-	-	MEDIANCDFQUANTILE

A.6.3 Error bars

Unless stated otherwise, shaded regions in all plots denote mean ± 1.96 standard errors across 20 independent runs.

A.6.4 Hardware

All experiments were run on a CPU cluster. Each individual run used a single CPU core, and no GPU acceleration was required. Peak memory usage was below 8 GB per run. The reported experiments required approximately 200 CPU-hours in total, dominated by repeated synthetic and robustness runs. Representative individual runtimes for the main methods are reported in Appx. A.9.3.

A.7 Datasets

This section provides further detail about the synthetic and real-world datasets.

A.7.1 Synthetic data

We generate synthetic observations $X \in \mathbb{R}^{N=2000 \times D=15}$ from a structural causal model (SCM) over a random DAG $\mathcal{G}$ with 15 nodes. Each edge (i, j), $i > j$, is included independently with probability $p_e = 0.3$, yielding an acyclic graph.

Clean variables follow

$$X_j^* = f_j(\mathrm{Pa}_j) + U_j,$$

where $U_j \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$, and f_j is an additive polynomial mechanism normalized to $[0, 1]$ on the clean data.

Polynomial generation: Each function f_j is constructed as a sum of univariate polynomial components over its parent nodes. For a single parent input A, we generate a polynomial of length l as

$$f(A) = \sum_{i=1}^{l} a_i (A - b_i)^{e_i},$$

where

- a_i are random coefficients sampled uniformly from $[-1, -0.1] \cup [0.1, 1]$,
- b_i are random shifts sampled uniformly from $[-3, 3]$,
- $e_i \in \{2, 3\}$ are randomly chosen exponents.

For multivariate parents, the node function is the sum of the univariate polynomials applied independently to each parent:

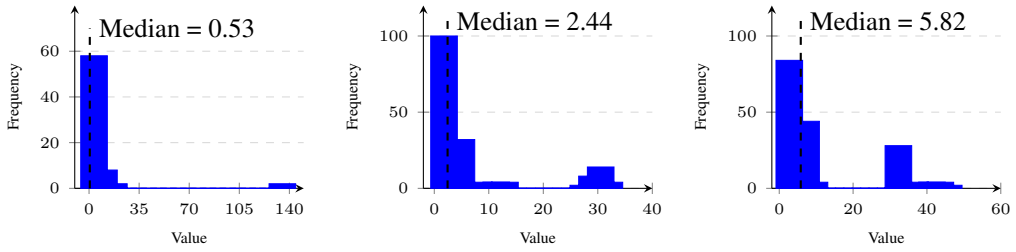
$$f_j(\mathrm{Pa}_j) = \sum_{k \in \mathrm{Pa}_j} f_{jk}(X_k),$$

where each f_{jk} is generated as above. After evaluating the raw polynomials on the parent data, the outputs are normalized to $[0, 1]$ using min/max scaling to ensure numerical stability. We set $l = 2$ for all experiments.

Asymmetric outliers are introduced independently by sampling from the true distribution of X_j^* and shifting by k empirical standard deviations. For mechanistic outliers, downstream nodes are recalculated with resampled noise. This procedure is repeated across five seeds, and averages are reported.

We choose outlier ratios to obtain 10% sample contamination, with measurement and mechanistic outliers equally probable. Specifically,

$$\pi_j^{meas} = \pi_j^{mech} = 1 - (1 - 0.1)^{\frac{1}{2 \cdot 15}} \approx 0.0035.$$



(a) Sachs (b) Causal Chamber, mid (c) Causal Chamber, strong

Figure 6: Distribution of absolute z-scores of interventional samples in the Sachs dataset (left), and Causal Chamber dataset with mid (center) and strong (right) interventions.

In every synthetic experiment, we average results over 20 random seeds, generating new graphs, mechanisms, and outlier injections each time.

A.7.2 Sachs dataset

The Sachs dataset consists of simultaneous measurements of 11 phosphorylated proteins and phospholipids in human immune cells [44]. The data comprise both observational and interventional samples collected under multiple experimental conditions, and are commonly used as a benchmark for causal discovery. A consensus causal graph, derived from prior biological knowledge, is typically treated as ground truth. We present this graph in Fig. 7. Although the underlying biological signaling network may contain feedback loops, the benchmark graph used in most evaluations is represented as a directed acyclic graph (DAG). In our experiments, we follow this convention and remove the edge from `PIP3` to `plcg` to ensure acyclicity. The dataset contains 853 observational samples, and we introduce outliers as follows.

We simulate two types of outliers: (i) mechanistic, created by injecting a small fraction of interventional samples into the observational set, and (ii) measurement, created by selecting a column of an observational sample and replacing its value with the one of a value from a sample of the respective interventional distribution. For each variable, we introduce five mechanistic and five measurement outliers. We repeat this procedure across 20 random seeds.

A.7.3 Causal Chambers dataset

In addition, we use the *Causal Chamber* benchmark [11]. It provides observational and interventional data from a controlled physical system with known ground-truth DAG. The graph is provided in Fig. 8. In each run we sample 2 000 observational samples and introduce outliers as follows.

As for the Sachs dataset, we simulate two types of outliers: (i) mechanistic, created by injecting a small fraction of interventional samples into the observational set, and (ii) measurement, created by selecting a column of an observational sample and replacing its value with the one of a value from a sample of the respective interventional distribution. Precisely, we randomly add $\frac{N}{20}\frac{0.05}{0.95} \approx 26$ random samples from each interventional dataset to contaminate approximately 5% of samples with mechanistic outliers. Additionally, we create the same amount of measurement outliers for each variable, by randomly selecting a column, and replacing it with the interventional column of an interventional sample. Again, we repeat this procedure across 20 random seeds.

The benchmark includes interventions labels as *mid* and *strong* for all variables but $L_{32}, L_{31}, L_{22}, L_{21}, L_{12}, L_{11}$. We evaluate the performance on the different strengths separately. Since the Sachs dataset contains much weaker interventions, we can separate the datasets by weak, medium and strong interventions. We provide the distribution of absolute z-scores of the interventional samples in Fig. 6.

A.7.4 NYC Yellow Taxi Trip Dataset

For the case study, we use the New York City Taxi & Limousine Commission (NYC TLC) Yellow Taxi trip dataset obtained from Kaggle [8]. The analysis is restricted to data from January 2015 only. The dataset contains trip-level records for yellow taxis in New York City, including pickup and

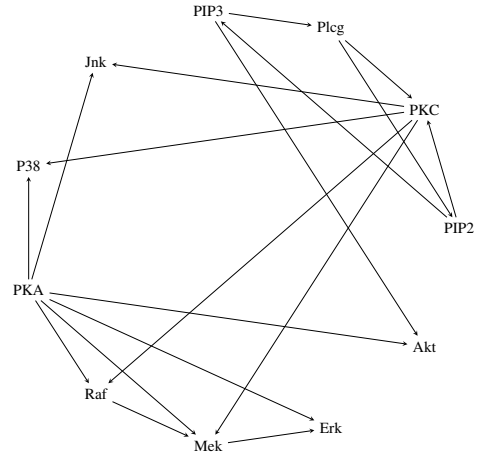


Figure 7: Consensus DAG of the Sachs dataset.

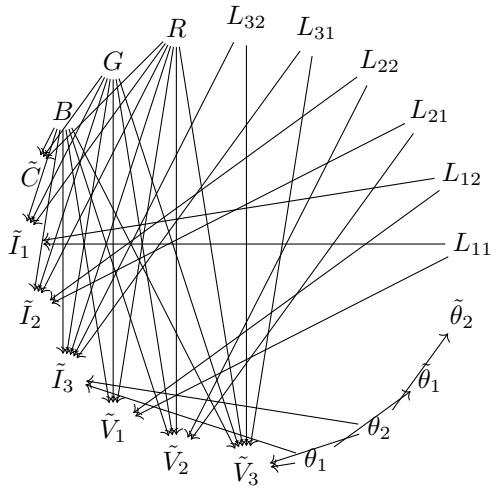


Figure 8: DAG of the light tunnel dataset of the *Causal Chamber* benchmark.

drop-off timestamps, trip distance, fare-related variables, passenger count, and geographic pickup and drop-off information recorded as coordinates. As an administrative dataset collected by the NYC TLC, it provides comprehensive coverage of yellow taxi activity during the study period. We preprocess the data by removing records with trivially missing or incorrect values such as zero coordinates, and select a random sample of size of $5\,000$ for our case study.

A.7.5 Dataset licenses and terms.

The Sachs protein-signaling dataset is publicly available; we use the version distributed on Zenodo under a CC BY 4.0 license. The Causal Chambers datasets are distributed by the authors under a CC BY 4.0 license. The NYC Yellow Taxi Trip records are publicly available from the NYC Taxi and Limousine Commission (TLC); we use them according to the applicable NYC TLC / NYC Open Data terms of use.

A.8 Additional Related Work

In addition, to the related work in the field of root cause analysis, we briefly discuss the broader field of outlier detection and explainable outlier detection.

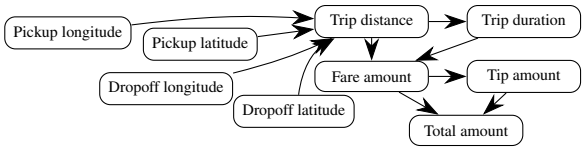


Figure 9: Causal relationships of the NYC Taxi Dataset.

Outlier Detection Outlier detection aims to identify observations that deviate from the bulk of the data. Traditional statistical methods, such as feature-wise Z-scores and Mahalanobis distance [27], are interpretable but limited in capturing complex dependencies. Machine learning-based approaches, including Local Outlier Factor [LOF 3], Isolation Forests [25], and One-Class SVM [47], improve flexibility and robustness to structured deviations. Deep models, such as autoencoders [45], normalizing flows [40], and VAEs [18], further capture highly nonlinear relationships at the cost of interpretability. For a comprehensive survey, see [6, 33].

Explainable Outlier Detection To improve interpretability, feature attribution methods like LIME [41] and SHAP [26] have been adapted to anomaly scores [10], while counterfactual approaches [29, 55] identify minimal changes that would alter predictions. More recent methods [1] use masked transformation-based optimization to localize contributing features. For a survey on the field, see [23]. These methods provide feature-level explanations, but are largely correlation-based and rely on the learned structure of non-interpretable black-box models, which makes correct interpretation and classification of outlier types challenging.

A.9 Additional Experiments

In this section, we present a series of additional experiments designed to further evaluate and analyze our approach.

A.9.1 Additional metrics

In this section, we report additional evaluation metrics for the root cause localization experiments presented in the main paper.

In Tab. 4, we report the causal discovery diagnostics for the Causal Chambers dataset. We report DAG Precision, Recall, F1 and SHD for both the AE-filtered and clean observational settings. The results show that the learned graph is already substantially different from the reference graph in the anomaly-free settings, indicating that causal discovery on real-world data remains a challenging problem.

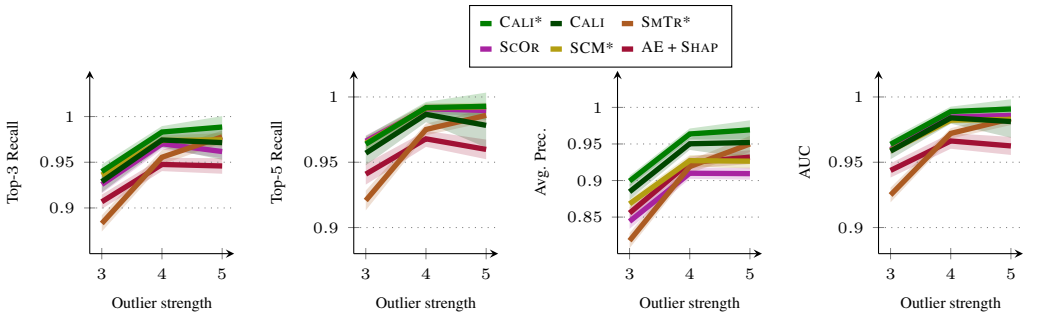
For the main experiments, we provide Top-3 Recall, Top-5 Recall, Average Precision (Avg. Prec.), and AUC in Fig. 10 and Fig. 11. Both Avg. Prec. and AUC are computed individually for each anomalous sample and then averaged across all anomalous samples.

A.9.2 Violated assumptions

In this section, we report additional experiments that evaluate the performance of CALI under violations of its assumptions, including confounding, controlled DAG misspecification, soft interventions, extreme outliers, heteroscedasticity and correlated root causes.

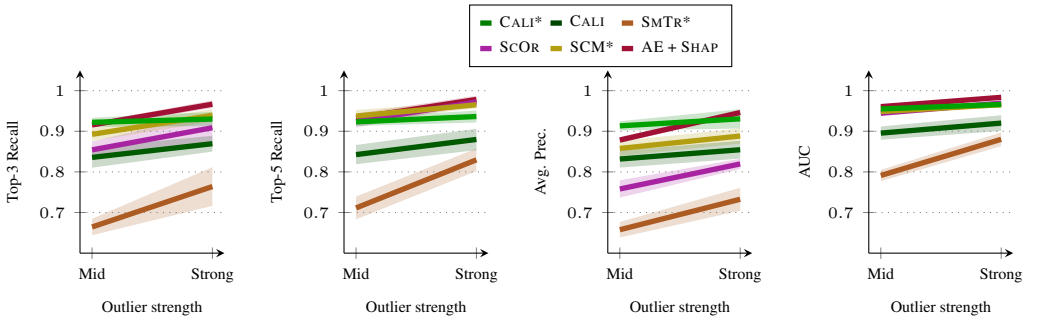
Table 4: Causal Chambers causal-discovery diagnostics. DAG Precision, Recall, F1 and SHD compare the learned graph to the reference graph. The clean-observational rows report a single clean-only discovery run, so no standard deviation is shown.

Strength	Setting	DAG Precision	DAG Recall	DAG F1	SHD
Mid	AE-filtered	0.108 ± 0.017	0.137 ± 0.021	0.121 ± 0.018	58.350 ± 2.007
Mid	Clean obs.	0.143	0.179	0.159	56.0
Strong	AE-filtered	0.136 ± 0.043	0.188 ± 0.065	0.158 ± 0.051	59.450 ± 3.017
Strong	Clean obs.	0.104	0.128	0.115	57.0



(a) Top-3-Recall (b) Top-5-Recall (c) Average Precision (d) Area Under the Curve

Figure 10: Additional metrics for root cause localization on synthetic data.

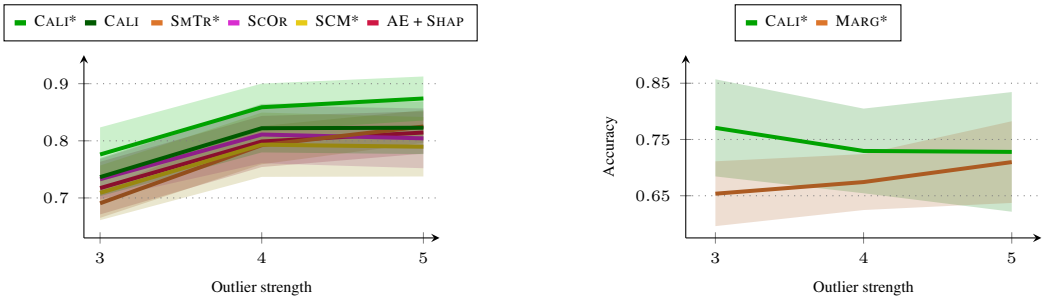


(a) Top-3-Recall (b) Top-5-Recall (c) Average Precision (d) Area Under the Curve

Figure 11: Additional metrics for root cause localization on the Causal Chamber dataset.

Correlated root causes Instead of sampling mechanistic and measurement outliers independently, we introduce correlation by sampling pairs of variables for which we jointly inject mechanistic and measurement outliers. The results are shown in Fig. 12. CALI* maintains strong performance in both localization and classification, indicating that violating the independence assumption with sparse but correlated perturbations does not significantly affect the performance.

Dag misspecification We analyze the performance of CALI and competitors under three types of DAG misspecification: (i) edge additions, (ii) edge removals, and (iii) edge reversals. For edge additions, we randomly add edges between pairs of nodes that are not connected in the true DAG. For edge removals, we randomly remove existing edges from the true DAG. For edge reversals, we randomly select existing edges and reverse their direction. In each case, we introduce a fixed number



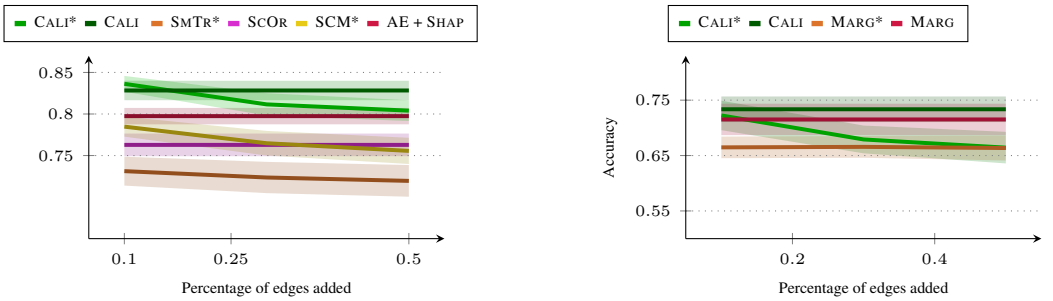
(a) Localization under correlated root causes (Top-k recall, higher is better). We set k to the true number of root causes in each sample.

(b) Classification under correlated root causes (accuracy, higher is better).

Figure 12: Performance of CALI under correlated root causes.

Table 5: Causal Chambers graph-perturbation sensitivity. The reference column reports top-k recall. Add, delete, and reverse columns report paired changes in top-k recall relative to the corresponding reference graph.

Strength	Edge errors	Reference	Add	Delete	Reverse
Mid	4	0.872 ± 0.027	$+0.003 \pm 0.005$	-0.014 ± 0.015	-0.033 ± 0.013
Mid	12	0.872 ± 0.027	$+0.006 \pm 0.003$	-0.093 ± 0.049	-0.102 ± 0.033
Mid	20	0.872 ± 0.027	$+0.006 \pm 0.005$	-0.153 ± 0.052	-0.139 ± 0.029
Strong	4	0.903 ± 0.039	-0.002 ± 0.002	-0.021 ± 0.022	-0.028 ± 0.014
Strong	12	0.903 ± 0.039	$+0.016 \pm 0.019$	-0.070 ± 0.030	-0.067 ± 0.023
Strong	20	0.903 ± 0.039	$+0.018 \pm 0.024$	-0.100 ± 0.018	-0.108 ± 0.033



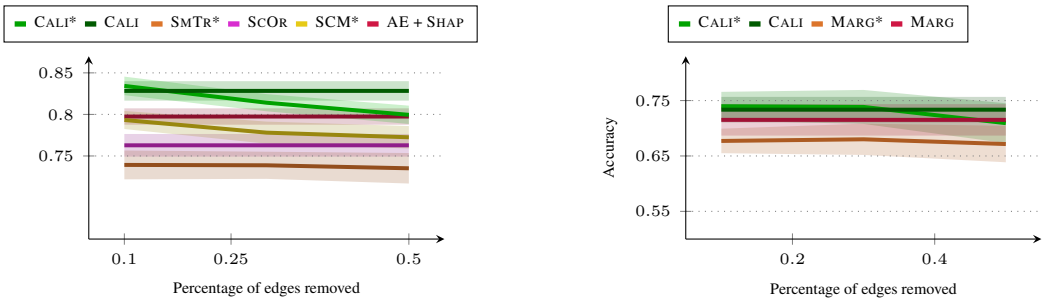
(a) Localization under edge additions.

(b) Classification under edge additions.

Figure 13: Performance of CALI under edge additions.

of modifications to the DAG and evaluate the performance of CALI and competitors on the modified graph. The results are shown in Fig. 13, Fig. 14 and Fig. 15. As expected, DAG misspecification leads to a decrease in performance for all methods that rely on a given DAG structure. In Tab. 5, we report the top-k recall of CALI under different levels of DAG misspecification on the Causal Chambers dataset. The results show that CALI is robust to moderate levels of DAG misspecification, with only a slight decrease in performance as the number of edge errors increases. However, as the number of edge errors becomes large, the performance of CALI degrades more significantly, highlighting the importance of accurate DAG specification for optimal performance.

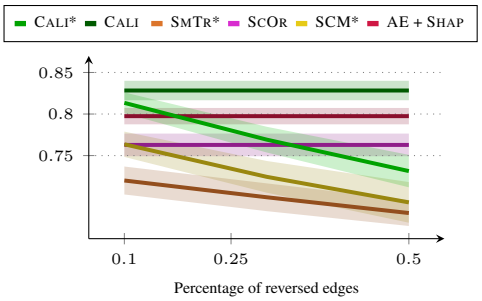
Confounding We evaluate the performance of CALI and its competitors under confounding under varying confounding levels at a shift strength of 3 standard deviations. We add $p\%$ (of the default of 15) new variables to the generating SCM, which are hidden from the algorithms. Methods that typically receive the ground truth DAG are now provided with the true subgraph among the observed variables. We restrict the evaluation of localization and classification to the observed variables, since root causes in unobserved variables can trivially neither be correctly localized nor classified. The results in Fig. 16 show that performance decreases slightly for the variant that relies on causal



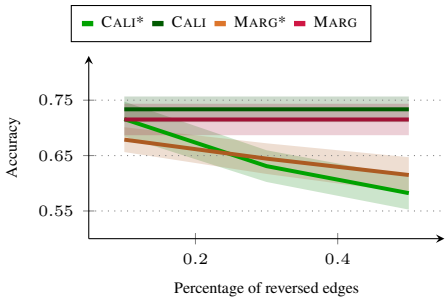
(a) Localization under edge removals.

(b) Classification under edge removals.

Figure 14: Performance of CALI under confounding and DAG misspecification.

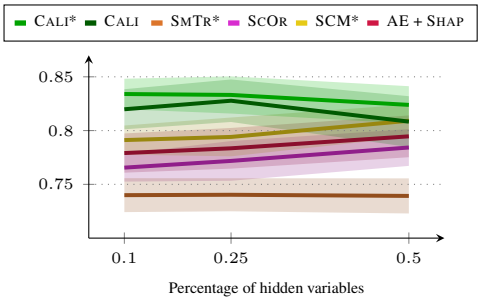


(a) Localization under edge reversals.

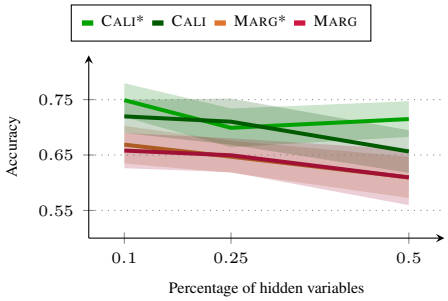


(b) Classification under edge reversals.

Figure 15: Performance of CALI under confounding and DAG misspecification.



(a) Localization under confounding



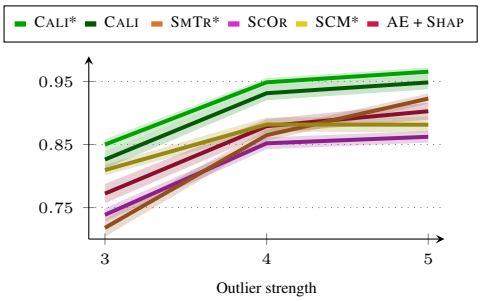
(b) Classification under confounding

Figure 16: Performance of CALI under confounding.

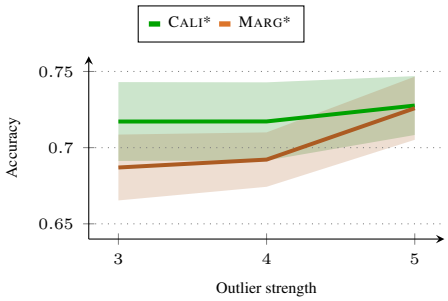
discovery to learn the DAG, but the overall effect remains modest even when up to half of the variables in the generating SCM are hidden.

Soft interventions We evaluate the performance of CALI and its competitors under soft interventions. Instead of resampling from the shifted marginal distribution, we introduce additive perturbations to the original values. We vary the magnitude of the perturbations in standard deviations of the marginal, as we did for the hard interventions. Fig. 17 shows the results. CALI* and CALI outperform the competitors in both localization and classification similarly to the hard intervention setting.

Endogenous intervention probabilities This experiment tests a violation of the assumption that intervention probabilities are independent of the pre-fault state. We first generate the clean latent sample X^0 as in the default synthetic setting. Instead of selecting intervention targets independently of X^0, we sample anomalous observations with probability proportional to $\exp(\lambda(2\,\mathrm{rank}(X^0)-1))$,



(a) Localization under soft interventions.



(b) Classification under soft interventions.

Figure 17: Performance of CALI under soft interventions.

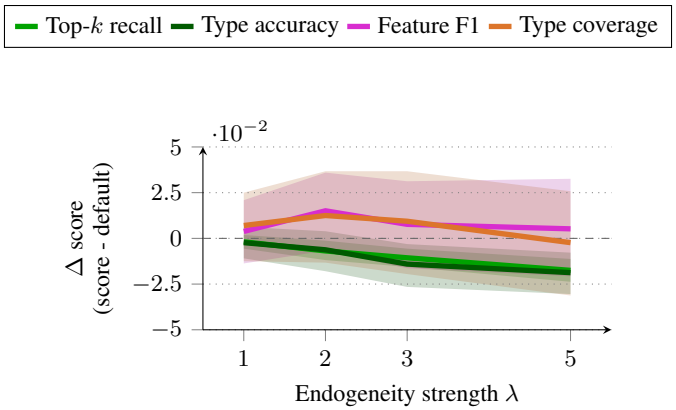


Figure 18: Performance of CALI* when intervention probabilities depend on the pre-fault state. We draw anomalies with sample weights proportional to $\exp(\lambda(2\,\mathrm{rank}(X^0) - 1))$, where $\mathrm{rank}(X^0) \in [0, 1]$ is the percentile rank of the counterfactual pre-fault state and λ controls the strength of the dependence. Changes are measured relative to the default synthetic setting, i.e., each score is the metric under the robustness setting minus the corresponding baseline score.

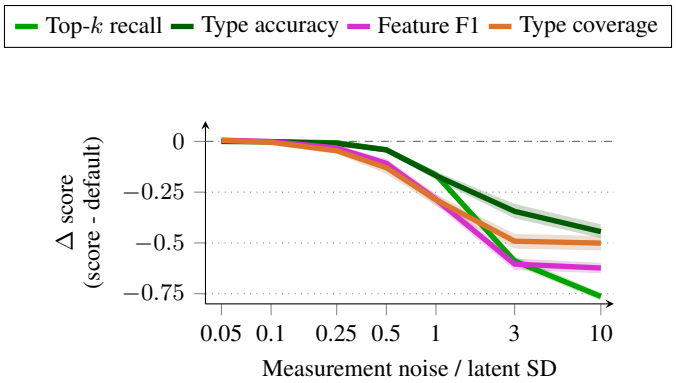
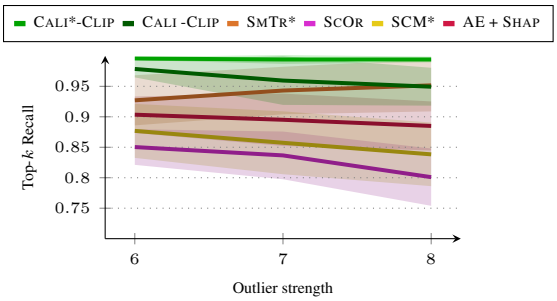


Figure 19: Performance of CALI* under additional non-anomalous measurement noise. Noise strength is measured as the standard deviation of the added measurement noise divided by the standard deviation of the corresponding latent variable, so a value of 1 means equal standard deviations. Changes are measured relative to the default synthetic setting, i.e., each score is the metric under the robustness setting minus the corresponding baseline score.

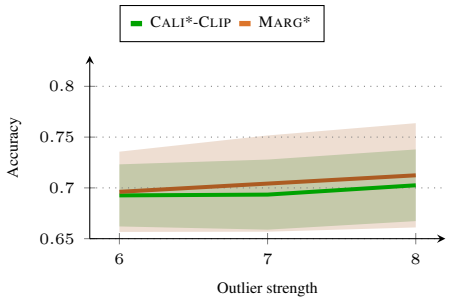
where $\mathrm{rank}(X^0) \in [0, 1]$ is the percentile rank of the pre-fault value. Thus, $\lambda = 0$ recovers state-independent anomaly occurrence, while larger λ makes anomalies increasingly concentrated in high-valued pre-fault states. The results show only a mild degradation as the dependence strength increases, suggesting that CALI* is robust to moderate endogeneity in the anomaly-generation process.

Measurement noise This experiment tests the effect of ordinary, non-anomalous measurement noise, which is not part of our formal model. After generating the latent causal variables and injected anomalies as in the default synthetic setting, we observe $X_j = X_j^* + \eta_j$ instead of $X_j = X_j^*$, with independent Gaussian noise η_j added to every measured variable. The noise strength is the ratio between the standard deviation of η_j and the standard deviation of the corresponding latent variable X_j^*. Small amounts of additional measurement noise have little effect, while large noise levels increasingly obscure the causal signal needed for localization and type classification.

Extreme outliers We evaluate CALI and its competitors under increasingly extreme outliers. This setting challenges an implicit assumption of CALI: the learned conditional models and residual densities are most reliable near the support of the observed data, whereas very large shifts force the spline models to extrapolate far beyond the training range. As a simple remedy, we evaluate

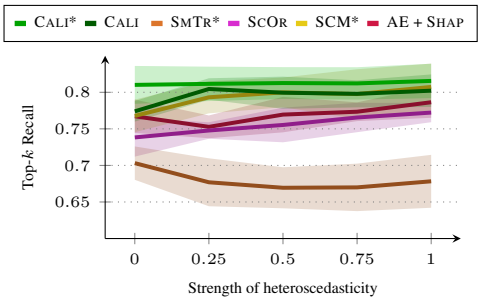


(a) Localization under extreme outliers.

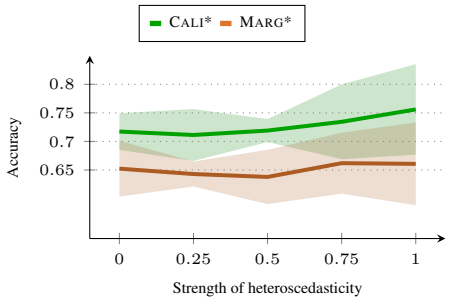


(b) Classification under extreme outliers.

Figure 20: Performance of CALI under extreme outliers.



(a) Localization under heteroscedastic noise.



(b) Classification under heteroscedastic noise.

Figure 21: Performance of CALI under heteroscedastic noise.

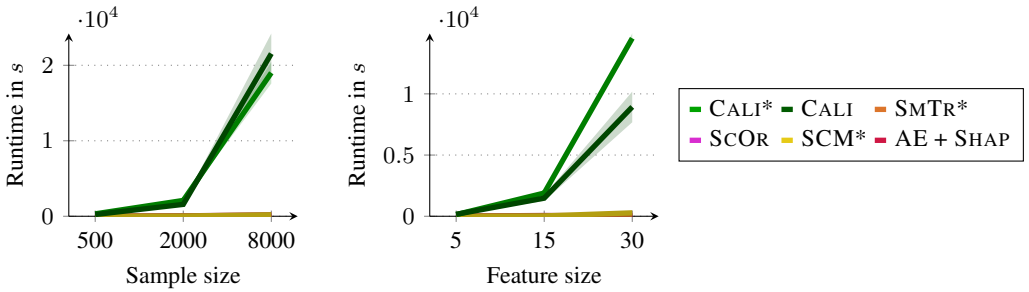
a clipped variant of CALI that clamps each feature marginally to the interval given by the median $\pm 6 \cdot 1.4826$ median absolute deviations, estimated from the observed contaminated data.

Fig. 20 shows that CALI*-CLIP performs consistently best across all extreme outlier strengths. In contrast, the learned-DAG variant of CALI-CLIP gradually degrades as the outlier strength increases and is eventually overtaken by SMTR*. Most remaining methods degrade with increasing outlier strength, indicating that a simple robust marginal clipping step is sufficient to remove much of the extrapolation pathology. In contrast, for the classification task, we observe that for sufficiently extreme outliers CALI* no longer outperforms MARG*, indicating that for sufficiently strong outliers the simple marginal heuristic is no longer dominated by the more sophisticated conditional model.

Heteroscedasticity We evaluate the performance of CALI and its competitors under heteroscedastic noise. We model heteroscedastic noise by letting the variance depend on a normalized signal profile $\phi_{j,i}$ derived from standardized deviations of $s_{j,i} = f_j\big(x_{\mathrm{Pa}(j),i}\big)$. The noise scale is $\sigma_{j,i} = \sigma_0[(1-h) + h\,\phi_{j,i}]$, where $h \in [0,1]$ interpolates between homoscedastic ($h = 0$) and fully heteroscedastic ($h = 1$) noise. We set $\sigma_0 = 0.1$ as for the other experiments and $\rho = 0.25$. We keep the mean shift at 3 standard deviations and vary the heteroscedasticity strength parameter h (0.0, 0.25, 0.5, 0.75, 1.0). The results in Fig. 21 show that CALI remains robust under heteroscedastic noise, and the causal discovery variant even benefits from the increased variance.

A.9.3 Runtime

In Fig. 22, we report the runtime of CALI and the competing methods. We evaluate scaling in two regimes: first fixing the number of features to $d = 15$ and varying the sample size, and then fixing $N = 2000$ and varying the number of features. CALI is slower than the baselines because it evaluates feature-level intervention candidates for all samples, including measurement candidates whose likelihoods require additional Markov-blanket evaluations. For the learned-graph variant, the reported runtime also includes causal discovery.



(a) Sample size (b) Feature size

Figure 22: Runtimes for increasing sample and feature size.

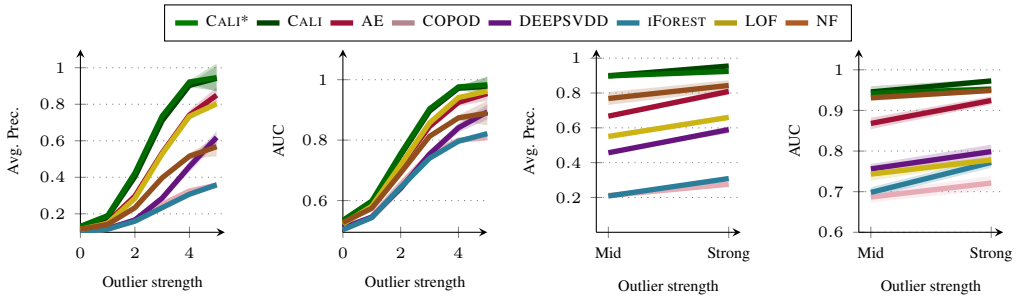


Figure 23: Outlier Detection (↑ is better). Left: Synthetic data. Right: Causal Chamber dataset.

When runtime is a concern, a simple practical acceleration is to restrict assignment updates to samples that are most unlikely under the fitted clean density. In our current implementation, CALI considers candidate root causes for all samples whose updates improve the global objective. One could instead score only the top-k samples with largest negative log-likelihood under the clean model, or use this set as an initial active set, thereby reducing the dominant candidate-evaluation cost while preserving focus on the most anomalous observations.

A.9.4 Anomaly detection

We compare against AE [45], DEEPSVDD [43], IFOREST [25], COPOD [22], LOF [3], and NF [40].

We report AUC and Average Precision (AP) for synthetic data and the Causal Chamber dataset (Fig. 23). Despite primarily being designed for characterization, CALI also performs great at detection.

A.9.5 Additional RCA experiments

In this section, we provide various additional RCA experiments to further analyze the performance of CALI and its competitors.

Varying parameters in synthetic data generation. We repeat the synthetic experiment but vary either sample size, feature size, edge probability, or the expected number of root causes per expected sample. We use the same default settings as previously ($N = 2000, D = 15, p_e = 0.3$) and vary one of the factors. To analyze how the methods work under multiple root causes, we do one additional experiment we no longer use the independent Bernoulli draws, but instead randomly place rc_{count} root causes in 10% of samples, deciding with a coin flip whether we create a measurement or mechanistic outlier. We present the results in Fig. 24.

Individual outlier types We do an experiment under the default conditions, but only add one outlier type to evaluate how well the methods work for either type. We present the results in Fig.26.

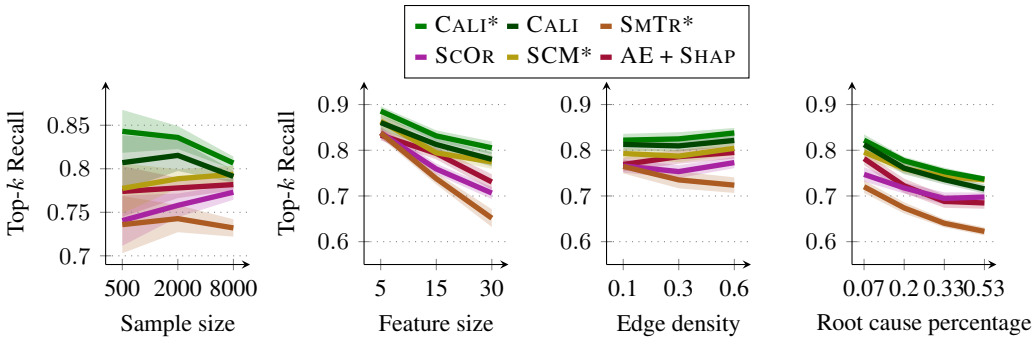


Figure 24: Root cause localization under varying parameters of the synthetic data generation.

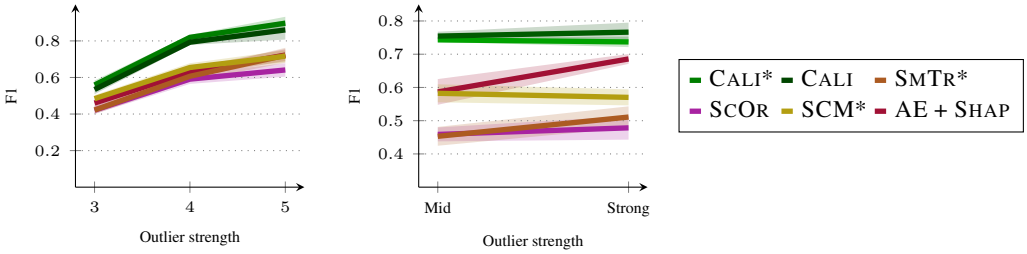


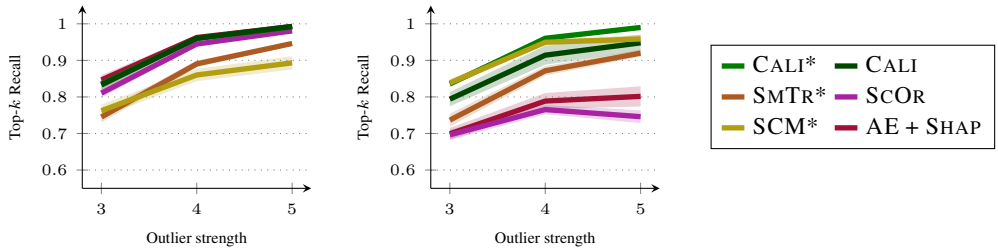
Figure 25: F1 scores of fixed assignments (↑ better). Left: Synthetic data, Right: Causal Chamber dataset.

While many methods work well when the data contains only measurement outliers, CALI*, SCM* and CALI perform best when it contains only mechanistic outliers.

Realistic RCA evaluation. To assess how many root causes can be realistically identified in practice, we conduct an additional experiment where we only evaluate RCA methods on samples that are actually detected as anomalous by a sample-level detector. This evaluation captures the practically relevant question of how many true root causes can be recovered when both anomaly detection and localization are imperfect, rather than assuming oracle knowledge of anomalous samples We evaluate all competing RCA methods in combination with AE, which is the best-performing sample-level detector on the synthetic data, in order to obtain fixed binary predictions. Specifically, we first use the true sample contamination ratio to determine the corresponding score quantile of AE, yielding fixed anomaly predictions at the sample level. Next, for the samples predicted as anomalous, we apply the same procedure at the feature level using the RCA scores: we compute the ratio of root causes among truly anomalous samples and use it to select the corresponding quantile of the RCA scores. This process results in a binary prediction matrix indicating predicted root causes for each detected sample.

For CALI, this post-processing step is unnecessary, as the method directly outputs feature assignments through the learned model. We report Precision, Recall, and F1 score, with results shown in Fig. 25. On synthetic data, CALI* and CALI outperform all competing approaches by a large margin, achieving F1 scores of up to 0.8 for shifts of at least four standard deviations. This is particularly encouraging because, unlike competing methods, CALI does not rely on thresholding based on the true contamination ratio, which is typically unknown in practice. On the Causal Chamber dataset, CALI* and CALI likewise attain the best overall performance.

RCA for individual outlier types. Here we generate data with only one outlier type at a time to evaluate how well the methods can localize either mechanistic or measurement outliers. We use the same default settings as previously ($N = 2000, D = 15, p_e = 0.3$), and either only add measurement or mechanistic outliers. The results are presented in Fig. 26.



(a) Measurement outliers (b) Mechanistic outliers

Figure 26: Root cause localization for the individual outlier types.

Table 6: Root cause localization and classification on the Sachs dataset. The highest performance is bolded.

Metric \ Method	AE-REC	AE-SHAP	MARG*	MARG	CALI	CALI*	SCM*	SCOR	SMTR
Top-3 Recall	0.412 ± 0.050	0.435 ± 0.041			**0.438 ± 0.057**	0.402 ± 0.042	0.412 ± 0.049	0.414 ± 0.043	0.347 ± 0.036
Top-5 Recall	0.548 ± 0.052	0.573 ± 0.039			**0.620 ± 0.075**	0.587 ± 0.068	0.544 ± 0.054	0.550 ± 0.041	0.564 ± 0.048
Top-k Recall	0.269 ± 0.026	**0.279 ± 0.046**			0.236 ± 0.036	0.192 ± 0.036	0.236 ± 0.036	0.240 ± 0.032	0.187 ± 0.032
Avg. Prec.	0.416 ± 0.027	**0.430 ± 0.034**			0.410 ± 0.034	0.375 ± 0.032	0.399 ± 0.030	0.376 ± 0.027	0.337 ± 0.024
AUC	0.586 ± 0.039	0.606 ± 0.032			**0.623 ± 0.042**	0.604 ± 0.043	0.591 ± 0.028	0.584 ± 0.033	0.558 ± 0.024
Accuracy			0.663 ± 0.077	0.847 ± 0.075	**0.898 ± 0.097**	0.860 ± 0.090			

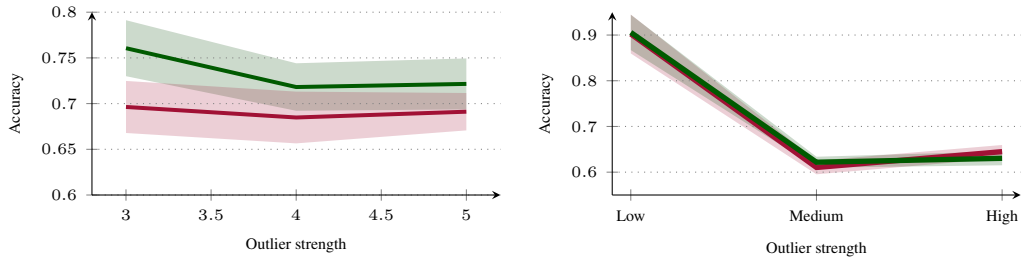
A.9.6 Sachs dataset

The full results of the Sachs experiment are reported in Table 6. Clearly, the dataset is quite challenging, as all methods achieve much lower RCA performance compared to the synthetic data or the Causal Chamber dataset. Somewhat surprisingly, CALI outperforms CALI* on most root-cause localization metrics, while MARG and CALI achieve the best overall classification performance. Notably, for the root causes that can be detected, the classification accuracy of 84% is even better than for the synthetic data. We attribute the fact that CALI outperforms CALI* to the factor that the data may not be entirely faithful to the assumed ground-truth DAG. Since it is known that the system contains feedback loops, the agreed-upon DAG is only an approximation. The fact that MARG* performs significantly worse than MARG further supports this hypothesis.

A.9.7 Classification experiments

In this section we provide additional experiments related to the classification of outlier types.

CALI and MARG variants In Fig. 27, we compare the classification performance of CALI and MARG with a learned causal DAG. The experimental conditions are identical to the main paper. Surprisingly, the performance on the Sachs dataset is much better than on Causal Chamber despite weaker interventions, possibly because the learned DAG is more accurate. Overall, CALI outperforms MARG on synthetic data, whereas no method is consistently better on the real data. Since this is very different from the results with the ground-truth DAG, we attribute this to the learned DAG restricting classifications to trivial cases, which MARG can handle well.



(a) Synthetic Data (b) Real data

Figure 27: Classification performance of CALI and MARG with learned causal DAG.

Sample classification Instead of inspecting classification performance at the level of detected root causes, as done in the main paper, we also consider classification at the sample level. In our causal setting, each sample may contain measurement failure(s), mechanistic failure(s), or both. However, the presence of unidentifiable sink nodes complicates a direct sample-level categorization, as failures at such nodes cannot be causally distinguished without additional assumptions. Representing all possible combinations would therefore require introducing several additional and potentially unintuitive classes.

[19] propose a classification method for a related but simpler problem. Their approach distinguishes between *sensor anomalies*, defined as failures affecting exactly one observed feature, and *process anomalies*, defined as failures affecting multiple features, in addition to the *normal* class. Under this definition, mechanistic failures at sink nodes are interpreted as sensor anomalies, and multiple independent mechanistic failures are classified as process anomalies. Consequently, their method does not aim to identify causal failure mechanisms in the data-generating process, but rather distinguishes whether none, exactly one, or more than one marginal distribution of the observed variables has changed.

Nevertheless, we compare CALI with the approach of LaRosa et al. under their classification scheme to assess whether incorporating causal structure also improves performance on this simpler sample-level task. Sample-level predictions for CALI are obtained by aggregating the inferred root-cause assignments. Specifically, whenever CALI assigns a mechanistic root cause to a node with successors, or assigns multiple root causes, we classify the sample as a *process anomaly*. When exactly one root cause is assigned and it corresponds either to a measurement failure or to a failure at a sink node, we classify the sample as a *sensor anomaly*. Samples without any assigned root cause are classified as *normal*.

The method of LaRosa et al. is originally designed for time-series data and is therefore adapted to the tabular setting. First, we separate the data into clean and anomalous samples using an autoencoder. Next, we generate 100 random groups of features of size three. These groups serve as inputs to gradient boosting prediction models, one for each feature, resulting in a total of $100 \times D$ predictors (1,500 predictors for $D = 15$). Classification is performed by thresholding the residuals of these predictors into normal and abnormal; we evaluate multiple significance levels $\alpha \in \{0.001, 0.01, 0.05, 0.1\}$.

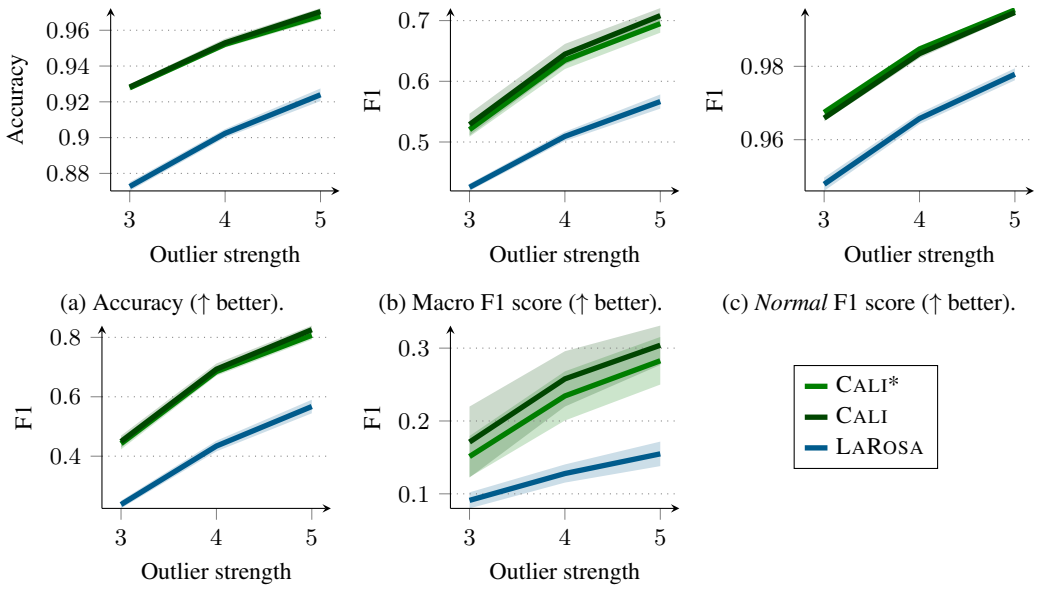
Based on these binary residual predictions, LaRosa et al. define the following classification rule: if no residual is classified as abnormal, the sample is labeled *normal*. If there exists a feature i for which all corresponding groups predict abnormal residuals, and no other feature j (predicted from any group not containing i) is classified as abnormal, the sample is labeled *sensor anomaly*. In all remaining cases, the sample is classified as *process anomaly*.

We present the results in Figure 28.

Influence of degree To analyze classification behavior under different structural conditions, we begin with a three-variable chain $X_1 \to X_2 \to X_3$ and introduce both a measurement and a mechanistic outlier at X_2. We then progressively add parent and child nodes to X_2 and evaluate how classification accuracy changes. For each graph structure, we generate random nonlinear functions and noise following the setup of the previous synthetic experiments. Classification accuracy is averaged over 50 runs, counting only those runs in which both root causes are correctly identified. Fig. 29 summarizes the results.

Surprisingly, MARG* outperforms CALI* when X_2 has two parents and two children; however, its performance degrades once an additional parent and child are added. This is probably because the optimal significance threshold to determine marginal outlierness depends on the degree. This suggests that if the outliers are strong enough that propagation can easily be detected, we need the right threshold, which is impossible to choose in the unsupervised setting. In contrast, the performance of CALI* improves consistently as the degree of the outlier node increases, without requiring the selection of a significance threshold.

Confidence Last but not least, we look at the confidence scores produced by CALI. Specifically, we evaluate how well the confidence scores rank correct classifications, by evaluating the AUC and calculating the Spearman correlation. We present the results in Fig. 30.



(a) Accuracy (↑ better).

(b) Macro F1 score (↑ better).

(c) *Normal* F1 score (↑ better).

(d) *Sensor anomaly* F1 score (↑ better).

(e) *Process anomaly* F1 score (↑ better).

Figure 28: Classification into the classes *Normal*, *Sensor Anomaly*, *Process Anomaly*. Accuracy, macro F1, and per-class F1 scores are reported due to class imbalance.

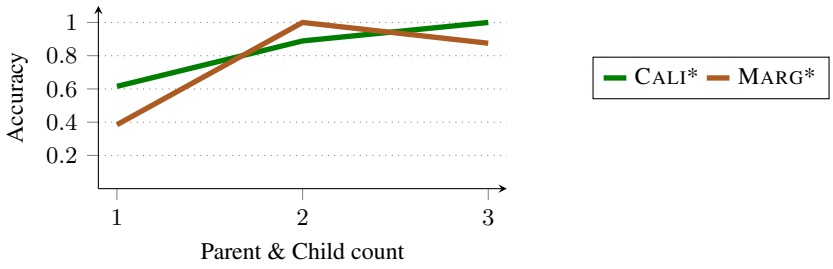
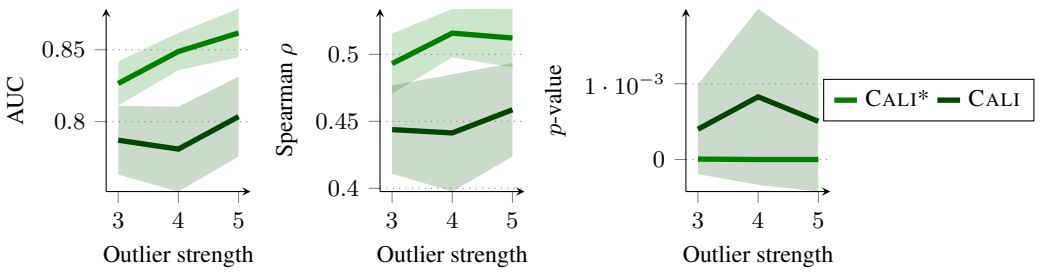


Figure 29: Influence of node degree on classification accuracy.



(a) AUC (↑ better).

(b) Spearman's ρ (↑ better)

(c) Significance of Spearman correlation (↓ better).

Figure 30: We evaluate the confidence scores of CALI by evaluating the ranking of true positive predictions, in terms of AUC and Spearman correlation. Overall the confidence scores produce a clear but imperfect signal, which is likely due to their heuristic nature. The confidence scores are weakly positively correlated with true positive predictions, and clearly significant at $\alpha = 0.05$.

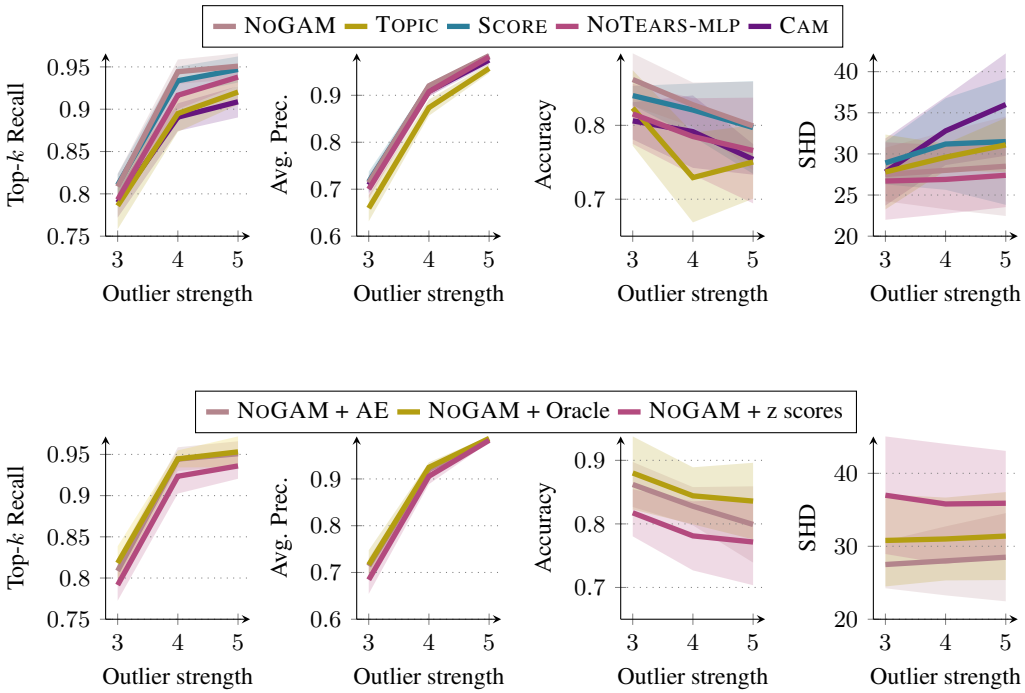


Figure 31: Performance of CALI when combined with different causal discovery algorithms. Top-k Recall, and Avg. Prec. assess the downstream root cause localization, Accuracy assesses outlier type classification and Structural Hamming Distance (SHD) similarity to the true DAG. Top: Different algorithms with AE prefiltering. Bottom: NOGAM with different prefiltering strategies.

A.9.8 Causal Discovery

In this experiment, we evaluate the performance of CALI on DAGs learned using different causal discovery algorithms, namely TOPIC [56], CAM [5], SCORE [42], NOGAM [28], and NOTEARS-MLP [59]. We run the algorithms under three different settings on our synthetic data. First, they have access to the clean data. Second, we apply filtering to the corrupted data using either an autoencoder or robust z-scores based on the median and MAD. In both cases, we remove values outside the 0.95 quantile and then learn the DAG from the filtered data. Overall the AE prefiltering performs better than z-scores, and NOGAM achieves the best downstream performance. We report all methods with AE prefiltering, and the different prefiltering strategies with NOGAM in Fig. 31.

A.9.9 Non-causal RCA

For non-causal RCA methods without a notion of causal structure, we evaluate AE-SHAP, AE-REC, M2OE [1], and DICE [29]. We present the results in Fig. 32. AE-SHAP performs best with M2OE coming in second. DICE clearly shows the worst performance. We want to emphasize that none of these methods are designed for the task that we evaluate them on, and we include them here for completeness. Our main experiments indicate that AE-SHAP still performs surprisingly well, even without access to the causal structure or causal guarantees.

A.10 Broader Impact

Distinguishing measurement errors from genuine mechanism shifts can have positive impact in domains where anomaly responses differ substantially, such as scientific measurement, engineering systems, and operations monitoring. Correctly identifying measurement anomalies may support data correction or sensor maintenance, while mechanistic anomalies may indicate genuine changes in the underlying system.

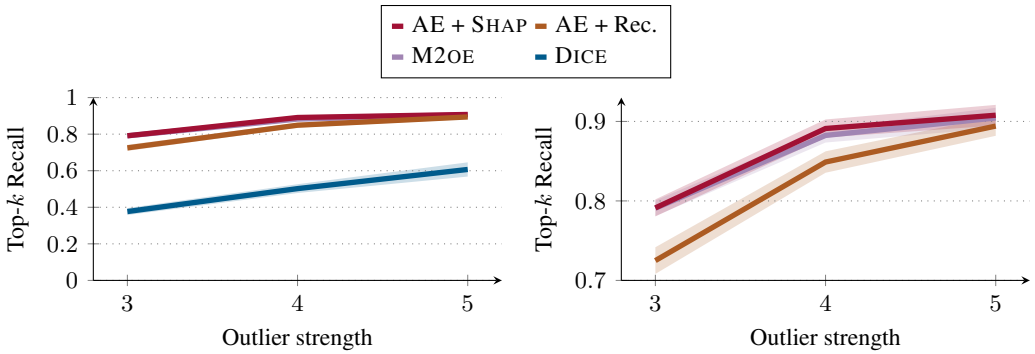


Figure 32: Root cause localization (↑ better) on synthetic data.

At the same time, incorrect classifications could lead to inappropriate interventions, for example correcting away a genuine process change or escalating a mere measurement error. This risk is particularly important in high-stakes applications. The method should therefore be used as decision support rather than as an automated decision-making system, and its conclusions should be interpreted in light of the assumed or estimated causal graph and the modeling assumptions discussed in the paper.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the paper's scope: they claim a causal formalization of measurement and mechanistic anomalies, an identifiability characterization, and an efficient method for localization and type classification. These claims are supported by the theoretical analysis, experiments on synthetic and real datasets, learned-DAG evaluations, and the case study. The assumptions are also reflected in the paper, including the need for a given or estimated causal DAG and the unsupervised in-sample setting.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sec. 7 discusses indistinguishability of anomaly types at sink nodes and classification uncertainty when assignment-conditional distributions overlap, the assumptions of exogenous hard interventions and pure-leaf measurements without ordinary measurement error, dependence on a given or estimated DAG and model specification, the hard-assignment approximation that precludes exact posteriors, and the restriction to tabular data. Appx. A.9.2 evaluates robustness to assumption violations, including endogenous intervention probabilities and ordinary measurement noise.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper states the assumptions for its theoretical results in the corresponding theorem statements, including the fixed DAG/model setting, the reduced sink-node parameterization, mixture-weight identifiability, open-parameter assumptions, and rank conditions. Full proofs are provided in the appendix for the identifiability result and the generic identifiability results for linear Gaussian and fixed-degree polynomial Gaussian SCMs. The main text summarizes the consequences of these results and refers readers to the appendix for the formal statements and proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the information needed to reproduce the main experimental results. It describes the synthetic SCM generation, outlier injection procedure, real datasets, evaluation metrics, graph assumptions, learned-DAG setting, and the baseline comparison protocol in the main text. Additional details on data generation and processing, baseline implementations and hyperparameter selection, the CALI algorithm, complexity, and robustness experiments are provided in the appendix. The datasets used are either synthetic, or publicly available.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via

detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the source code necessary to reproduce all experiments at `https://github.com/hsuhr99/measurement_mechanistic_rca`, including a dockerized environment and a README file with instructions for running the experiments. The repository also includes scripts for generating synthetic data and loading and preprocessing the public real-world datasets used in the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

Justification: The paper specifies the main experimental settings needed to interpret the results, including the synthetic data-generating process, outlier injection procedure, real datasets, evaluation metrics, graph assumptions, learned-DAG setting, and comparison protocol. Additional details on preprocessing, baseline implementations, hyperparameter selection, algorithmic choices, and robustness experiments are provided in the appendix and supplementary code.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification:The main quantitative plots report shaded regions showing mean ± 1.96 standard errors across 20 independent runs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We mention in the appendix, that all experiments were run on a CPU cluster. Each individual run used a single CPU core, and no GPU acceleration was required. Peak memory usage was below 8 GB per run. The reported experiments required approximately 200 CPU-hours in total, dominated by repeated synthetic and robustness runs. Representative individual runtimes for the main methods are reported in Appx. A.9.3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and believe the research conforms to it. The paper uses synthetic data and publicly available benchmark datasets and does not involve human-subject experiments or sensitive personal data collection. Potential limitations and appropriate scope of the method are discussed in the paper.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The appendix includes a broader-impact discussion. It describes potential positive impacts from better distinguishing measurement errors from genuine mechanism shifts, and potential negative impacts from incorrect classifications leading to inappropriate interventions, especially in high-stakes domains. It also emphasizes that the method should be used as decision support and interpreted in light of the causal and modeling assumptions.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a high-risk model, scraped dataset, or dual-use generative system requiring safeguards. The released code implements a methodological framework for anomaly characterization, and the experiments use synthetic data and publicly available benchmark datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the original sources for all external datasets, benchmarks, and methods used in the experiments. Dataset URLs, access information, and available license or terms-of-use information are provided in the paper or supplementary material. The released code environment lists package versions for software dependencies, and we use these assets according to their stated terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We provide the source code for the proposed method and experiments online at `https://github.com/hsuhr99/measurement_mechanistic_rca`. The repository includes a README, dockerized environment, scripts for generating synthetic data and running experiments, and documentation of the main configuration options. Like the paper, the code is licensed under the CC-BY 4.0 license.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The contribution does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.