BEYOND CONDITIONAL INDEPENDENCE: ROOT CAUSE ANALYSIS WITH DEEP CAUSAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Root cause analysis (RCA) is a critical problem in many real-world scenarios. RCA enables the identification of faulty or failing mechanisms in a system by comparing anomalous observations with corresponding reference (i.e., regular) observations. However, existing approaches rely either on heuristic methods or on conditional independence tests with a strong unconfoundedness assumption, and thus fail to exploit other complicated distributional constraints in the presence of latent variables. To relax these assumptions, we model the underlying system as a causal model and the anomalous system as a change in the structural functions of the same causal model. Specifically, to handle unobserved confounders, we establish an implicit connection between distributional constraint testing and root cause analysis. To adapt our approach to data generated from arbitrary causal models, we employ the deep causal model (DCM) framework, in which we design the causal model using neural networks. Finally, we illustrate how our method, RCA-DCM, can utilize different levels of partial graphical knowledge to perform RCA. We evaluate RCA-DCM against state-of-the-art baselines on simulated datasets, a physics-based causal chamber and two microservice applications. RCA-DCM improves top-1 accuracy over the strongest baseline on both Sock Shop (**0.880** vs. 0.752) and Online Boutique (**0.776** vs. 0.712), and when the true root cause in the causal chamber is unobserved and acts as a latent confounder, it recovers the exact root-cause set more often than any competing method (PRR **0.846** vs. 0.731).

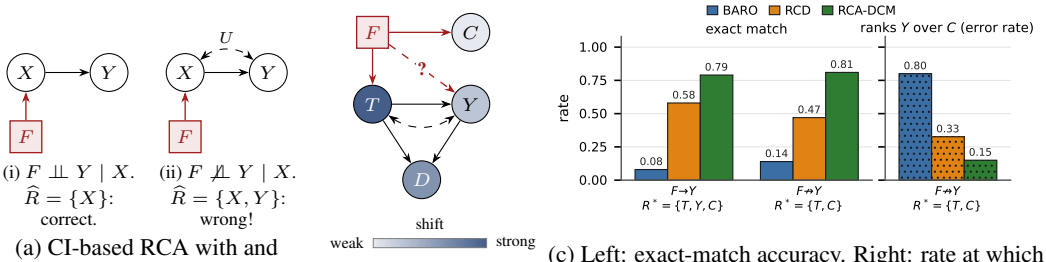
1 INTRODUCTION

Root cause analysis (RCA) aims to identify the components of a complex system that are responsible for anomalies. This is an important problem in many domains, including microservice systems and cloud applications (Wang et al., 2023; Lin et al., 2024; Liu et al., 2021; Ikram et al., 2022a; Shan et al., 2019; Ma et al., 2020), economics (Bai et al., 2024; Inoue & Rossi, 2021), scientific experiments, health monitoring (Strobl & Lasko, 2023; Strobl, 2024), fraud detection (Vanhoeyveld et al., 2020), and credit scoring (Das et al., 2023). Recent critical AI applications, such as failure attribution in LLM-based multi-agent systems (Zhang et al., 2025), can also be modeled as RCA problems.

The RCA problem is challenging because failures or abnormal behavior can propagate through the system's structural dependencies. Many existing RCA methods rely on heuristic rules, statistical measures, or correlation-based analyses (Pham et al., 2024a; Knorr & Ng, 1999; Liu et al., 2017; Micenková et al., 2013; Macha & Akoglu, 2018; Gupta et al., 2019). These approaches may fail to identify the root cause effectively, particularly in systems with complex structures, where a change in a mechanism affects the distributions of downstream variables in nontrivial ways. Thus, causal knowledge is essential for distinguishing the root cause from other variables.

When observations or metrics of the system are available both during normal operation (the normal dataset) and after a failure occurs (the anomalous dataset), causal discovery provides a framework for RCA by modeling anomalies, i.e., shifts in the observed distribution, as the result of a change in the mechanism of the root cause, also known as a soft intervention (Jaber et al., 2020; Okati et al., 2024). These approaches add a binary indicator variable F to the causal graph, with outgoing edges to the intervention targets, to represent their normal mechanisms ($F = 0$) and their post-intervention mechanisms ($F = 1$). Recent works such as Ikram et al. (2022a; 2025b) treat the change from $F = 0$ to $F = 1$ as the occurrence of an anomaly and perform causal discovery to find the children of F,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) CI-based RCA with and without a latent confounder.

(b) Is Y a root cause?

(c) Left: exact-match accuracy. Right: rate at which Y is ranked above the weakly shifted root cause C, which counts as an error when $F \not\to Y$.

Figure 1: Baselines fail when latent confounders are present and non-root-cause variables experience larger distributional shifts than the root causes; Given $F \to T$, deciding whether Y is a root cause ($F \to Y$) is equivalent to deciding whether F is a valid instrument for the effect of T on Y.

which they report as the anomalous nodes. However, these methods rely solely on distributional invariances of the form $p_n(x \mid s) = p_a(x \mid s)$ for subsets s and can fail to correctly rank the root causes in the presence of unobserved confounders, as in the graph in Figure 1(a). Here, the true root cause is $R^* = \{X\}$, indicated by the edge $F \to X$. Since $F \not\perp\!\!\!\perp X$, predicting X as an RC is correct. However, in Fig 1(a.ii), $F \not\perp\!\!\!\perp Y \mid X$ due to the backdoor path $F \to X \leftrightarrow Y$, so using this dependency to declare Y a root cause would be an incorrect prediction. Meanwhile, many non-causal methods based on statistical testing (Pham et al., 2024a; Li et al., 2022b) assume homogeneous anomalies, where the true root causes exhibit the largest distributional shift among all variables, and therefore primarily measure marginal shifts. In the real world, however, a shift originating at a root cause R_1 propagates to its descendants, and a downstream non-root-cause descendant can end up with a larger observed marginal shift than another root cause R_2.

Figure 1(b) shows a scenario, averaged over 100 SCMs, in which both families of approaches fail. We consider a large marginal shift (anomaly) on T and a mild anomaly on C, with Y either a root cause (case 1, $F \to Y$) or not (case 2, $F \not\to Y$). In both cases, the non-root-cause D inherits a larger marginal shift from its parent T than the mildly anomalous C exhibits, so statistical testing approaches such as BARO, which rank variables by marginal shift, rarely recover R^* (Figure 1(c), left). Causal approaches that assume unconfoundedness, such as RCD, fail for a different reason: $F \not\perp\!\!\!\perp Y \mid T$ in both cases, due to the direct edge $F \to Y$ in case 1 and the path $F \to T \leftrightarrow Y$, opened by conditioning on T, in case 2. RCD thus cannot tell the two cases apart and often ranks the non-root-cause Y above the mild root cause C (Figure 1(c), right). Moreover, conditional independence tests such as $D \perp\!\!\!\perp F \mid T, Y$ become unreliable with growing conditioning sets. Hence, neither marginal shifts nor conditional independencies suffice for RCA under unobserved confounding; RCA in this setting requires exploiting other distributional constraints, which no existing work offers. In this paper, we make a key observation: once T has been identified as a root cause ($F \to T$), asking whether Y is also a root cause ($T \leftarrow F \to Y$) in the presence of a confounder is *equivalent to asking whether F is an invalid instrument for the effect of T on Y*. Consequently, if the observed distribution is incompatible with F being a valid instrument, then both T and Y must be root causes. Unlike conditional independence tests, which can be misleading in this setting, the instrument constraint determines when Y should be declared a root cause and ranked above other variables.

It is natural to ask which other distributional constraints Verma & Pearl (1990); Ansanelli et al. (2026) can help distinguish root causes when these fail. However, verifying such constraints, including the instrument constraint, is nontrivial for arbitrary graphs with high-dimensional data. To address these limitations, we propose a new causal framework that models RCA as a form of unknown intervention target detection and leverages deep causal generative models to reason about which mechanism shifts are necessary to explain the observed distributional changes. This DCM-based RCA framework achieves 79% and 81% exact-match accuracy in the setup of Figure 1. Our contributions are:

- We propose `RCA-DCM`, an algorithm for root cause analysis in the presence of an arbitrary number of unobserved confounders, which uses a deep causal model to search for SCMs consistent with both the normal and anomalous datasets and identifies as root causes the variables whose mechanisms must shift to explain the anomaly.
- We characterize the non-identifiable case, in which the root cause cannot be uniquely determined as another variable set can explain the distributional shift. We show how the set of root cause solutions changes when the given graph is mis-specified.

- We also demonstrate that RCA−DCM outperforms state-of-the-art baselines in almost all cases when evaluated on synthetic setups with unobserved confounders, a real-world physics-based testbed (Causal Chamber), and microservice datasets (Sock Shop and Online Boutique).

2 Related Work

Causal inference has become a principled approach to RCA, particularly in microservice systems. Early methods construct a causal graph over service-level metrics, often using constraint-based discovery such as PC, and localize root causes by traversing or ranking nodes with anomaly scores (Chen et al., 2014; Wang et al., 2018; Ma et al., 2019; Meng et al., 2020). More recent works model failures as interventions on causal mechanisms: RCD (Ikram et al., 2022a) and RCG (Ikram et al., 2025b) treat failures as soft interventions and use local causal discovery or partial structural knowledge, CIRCA (Li et al., 2022b) casts RCA as intervention recognition in a causal Bayesian network, and CausalRCA (Xin et al., 2023) learns the graph via gradient-based causal discovery. Other methods relax structural requirements, either with guarantees for restricted graph classes under missing structural knowledge (Orchard et al., 2025) or by exploiting linear-SCM invariances (Li et al., 2024), while statistical approaches such as BARO (Pham et al., 2024a) score distributional shifts without using causal structure. However, these methods either ignore causal structure, rely on conditional independence or invariance tests, or depend on linearity or graphs that may be inaccurate under hidden confounding; none explicitly models nonlinear mechanisms with latent confounders. In contrast, RCA−DCM uses deep causal models to exploit general distributional constraints under an arbitrary number of unobserved confounders. We provide an extended discussion in Appendix B.

3 Background

Definition 3.1 (SCM (Pearl, 2009))**.** An SCM $\mathcal{M}$ is a 4-tuple $\mathcal{M} = (\mathcal{V}, \mathcal{U}, \mathcal{F}, P(.))$, where each observed variable $V_i \in \mathcal{V}$ is realized as an evaluation of a function $f_i^* \in \mathcal{F}$ that looks at a subset of the remaining observed variables $Pa_i \subset \mathcal{V}$, and an unobserved (latent) variable $U_i \in \mathcal{U}$. This refers to the semi-Markovian causal model. P is a product joint distribution over all unobserved variables $\mathcal{U}$.

Definition 3.2 (Graph with Latents, ADMG)**.** Each SCM induces an acyclic directed graph called the *causal graph* over $V \cup U$. A directed edge $V_i \to V_j$ implies that V_i directly appears in the structural equation for V_j. Thus $V_i \to V_j$ iff $V_i \in Pa_j$. The set Pa_j is called the parent set of V_j. We assume this directed graph is acyclic (DAG). The causal structure over observed vertex set $\mathcal{V}$ is represented by the latent projection ADMG $G = (V, E)$. Under the semi-Markovian assumption, each unobserved confounder can appear in the equation of exactly two observed variables. We represent the existence of an unobserved confounder $[U = U_X = U_Y] \in \mathcal{U}$ between X, Y in the SCM with a bidirected edge $X \leftrightarrow Y$ to the causal graph. These graphs are no longer DAGs although still acyclic. V_i is called an ancestor for V_j if there is a directed path from V_i to V_j. Then V_j is said to be a descendant of V_i. The set of ancestors of V_i in graph G is shown by $An_G(V_i)$. We define the partial order induced by G as $\preceq_G$, where $V_i \preceq_G V_j$ if V_i is an ancestor of V_j in G.

Definition 3.3 (Augmented graph)**.** For a set of root causes $R \subseteq \mathbf{V}$, let $G \cup F(R)$ be the mDAG on nodes $\mathbf{V} \cup \{F\}$ obtained from G by adding a visible root F and a directed edge $F \to V_i$ for each $V_i \in R$. We use unrestricted soft-intervention semantics: each V_i is an arbitrary function of its parents in G, of F when $V_i \in R$, and of an independent noise term. In particular a targeted mechanism is allowed to ignore F (the no change intervention).

4 Identifying Root Causes under Unobserved Confounding

Problem Statement: We consider a reference (normal) environment represented by a structural causal model $\mathcal{M}_n = (\mathbf{V}, \mathbf{U}, \mathcal{F}_n, P_n)$ and an anomalous environment represented by another SCM $\mathcal{M}_a = (\mathbf{V}, \mathbf{U}, \mathcal{F}_a, P_a)$, in which some mechanisms $f_i^n \in \mathcal{F}_n$ have shifted to corresponding $f_i^a \in \mathcal{F}_a$. We refer to the set of variables whose mechanisms have shifted as the root causes R^*. We observe two datasets, D_n and D_a, sampled from the normal and anomalous distributions $P_n(\mathbf{V})$ and $P_a(\mathbf{V})$, respectively. Following Ikram et al. (2022a), we introduce a binary indicator variable F as an additional node in the causal graph to represent the normal ($F = 0$) and anomalous ($F = 1$) environments, and add an edge from F to each variable in R^*. Our goal is to detect R^*.

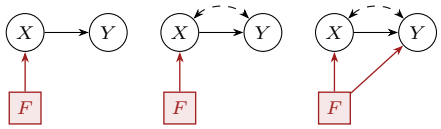
4.1 Root-Cause Solution Sets with a Known Graph

We now characterize root causes under unobserved confounding. Without unobserved confounders, or when they appear only in specific graph structures, the minimal root cause set is unique; in many confounded settings, however, multiple root cause sets are consistent with (P_n, P_a).

Example 4.1 (Intuitive example)**.** First consider Graph 1, which has no unobserved confounders. The conditional distribution $P(Y \mid \mathrm{Pa}(Y))$ represents the causal mechanism of Y. Under the independent causal mechanisms assumption (a shift in one mechanism does not affect another), we can detect a shift in f_V by checking whether $P_n(V \mid \mathrm{Pa}(V)) = P_a(V \mid \mathrm{Pa}(V))$ across the two environments. In this case, we recover the true root causes $\mathbf{R}^* = \{X\}$. Now consider Bow 1 (Figure 2b), where an unobserved variable U causes both X and Y. The ground-truth root cause set is $\mathbf{R}^* = \{X\}$. However, even though Y is not a root cause, a mechanism shift in f_X can change Y's conditional distribution in the anomalous environment, i.e., $P_n(Y \mid X) \neq P_a(Y \mid X)$. **CI test:** Given the normal and anomalous datasets, *conditional independence tests* cannot determine whether Y is a root cause, because the change in Y's distribution can also be produced by a pair of normal and anomalous SCMs in which f_Y shifts as well. Hence $\widehat{\mathcal{R}} = \{X\}$ and $\widehat{\mathcal{R}} = \{X, Y\}$ are equally plausible root cause solutions, both compatible with the given data and graph: $\mathrm{SolutionSet}(\text{CI-tests}) = \{\{X\}, \{X, Y\}\}$.

Instrumental test: We can further check whether the input distribution satisfies additional distributional constraints, such as instrumentality, and use them to shrink the solution set. A variable F is a valid instrument for the causal effect of X on Y if (i) F affects X, (ii) F affects Y only through X, and (iii) F is independent of the common causes U of X and Y. We can use Pearl's instrumental inequality conditions Pearl (1995) to check if a variables is a valid instrument. For binary F, X, and Y: $P(Y{=}y, X{=}x \mid F{=}0) + P(Y{=}1{-}y, X{=}x \mid F{=}1) \leq 1, \forall\, x, y \in \{0, 1\}$. A violation of any of the four inequalities implies that F is not a valid instrument for the effect of X on Y. Thus by contraposition, an edge $F \to Y$ must be present (since i, iii hold by construction, violation must come from ii). Therefore $\{X\}$ alone cannot be a solution, which yields $\mathrm{SolutionSet}(\text{CI-tests} + \text{ IV-fail}) = \{\{X, Y\}\}$.

The next natural question is i) how do we characterize our solution set and ii) how can we enforce constraints to reduce the set. Two rc solutions R_1, R_2 represent two different augmented graphs $G \cup R_1$ and $G \cup R_2$ which can represent distribution set $\mathcal{M}(G \cup R_1)$ and $\mathcal{M}(G \cup R_2)$. Informally, if the input normal-anomalous distribution (P_n, P_a) belongs to $\mathcal{M}(G \cup R_1) \cap \mathcal{M}(G \cup R_2)$, we can say that it can be expressed by both augmented graph. Thus, both R_1, R_2 are rc solutions for (P_n, P_a). Below, we make the concept of solution set precise.



(a) Graph 1 (b) Bow 1 (c) Bow 2

Figure 2: (a) No confounding. (b)–(c) Bow graphs with an unobserved confounder between X and Y (dashed $\leftrightarrow$).

4.1.1 The root-cause solution set

Definition 4.2 (Observational equivalence class $\mathcal{M}(G)$)**.** The set of distributions over the visible variables realizable by G under classical unrestricted semantics with latents of arbitrary cardinality.

Definition 4.3 (Realizable tuple)**.** A pair (P_n, P_a) of distributions over $\mathbf{V}$ is *realized by* (G, R) if there is a joint $P \in \mathcal{M}\big(G \cup F(R)\big)$ over $(\mathbf{V}, F)$ with $P(\mathbf{V} \mid F = 0) = P_n(\mathbf{V})$ and $P(\mathbf{V} \mid F = 1) = P_a(\mathbf{V})$. Since F is a root, its marginal is a free parameter; we only require it to give both regimes positive probability so the conditionals are defined.

Definition 4.4 (Structural dominance, $G_1 \subseteq G_2$)**.** G_2 structurally dominates G_1 if every node, every directed edge and every latent facet of G_1 is present in G_2.

Definition 4.5 (Observational dominance)**.** For two mDAGs G_1, G_2 on the same nodes, G_2 *observationally dominates* G_1 when $\mathcal{M}(G_1) \subseteq \mathcal{M}(G_2)$ at every cardinality of the visible variables

Lemma 4.6 (Structural dominance $\Rightarrow$ Observational dominance Ansanelli et al. (2026))**.** *Let G_1 and G_2 be two mDAGs with the same sets of nodes. If G_2 structurally dominates G_1, then it also observationally dominates it, i.e., $\mathcal{M}(G_1) \subseteq \mathcal{M}(G_2)$*

Figure 2 shows two graphs: $G_1 : G \cup F(\{X\})$ and $G_2 : G \cup F(\{X, Y\})$. Since G_2 structurally dominates G_1 due to the extra $F \to Y$ edge, G_2 realizes a larger set of distributions than G_1 :

$\mathcal{M}(G \cup F(\{X\})) \subseteq \mathcal{M}(G \cup F(\{X, Y\}))$. This implies that if a specific normal and anomalous distribution pair (P_n, P_a) is realized by G_1, it will be realized by G_2 as well. Thus, if $\{X\}$ is a root cause solution, both $\{X, Y\}$ is also valid root cause solution. We formally define the solution set as:

Definition 4.7 (Root cause solution set)**.** For the observed tuple (P_n, P_a) and a graph G, define

$$\begin{aligned} \texttt{SolRC}(G, P_n, P_a) &\coloneqq \big\{\, R \subseteq \mathbf{V} : (P_n, P_a) \text{ realized by } (G, R) \,\big\}, \\ \texttt{minrc}(G, P_n, P_a) &\coloneqq \min_{R \in \texttt{SolRC}(G, P_n, P_a)} |R| \;\in\; \{0, \ldots, n\} \cup \{\infty\}, \end{aligned} \tag{1}$$

with $\texttt{minrc}(G, P_n, P_a) = \infty$ when $\texttt{SolRC}(G, P_n, P_a) = \emptyset$. If the context is clear, we remove the distribution tuple from the notation.

For a graph G, $\texttt{SolRC}(G)$ denotes the family of root-cause sets that explain the data, and the *root-cause number* $\texttt{minrc}(G) = \min_{R \in \texttt{SolRC}(G)} |R|$ is the size of the smallest such set; the inclusion-minimal members of $\texttt{SolRC}(G)$ are the candidate reported sets. Since the data are generated by the true graph G^* and root-cause set R^*, we have $R^* \in \texttt{SolRC}(G^*)$ and hence $\texttt{minrc}(G^*) \leq |R^*|$.

Lemma 4.8. *$\texttt{SolRC}(G)$ is upward closed: if $R \in \texttt{SolRC}(G)$ and $R \subseteq R' \subseteq \mathbf{V}$, then $R' \in \texttt{SolRC}(G)$.*

4.1.2 Exploiting Distributional Constraints for Solution-set Reduction

Having defined the solution set, we determine its members by testing, for each candidate R, the distributional constraints implied by the augmented graph $G \cup F(R)$ against the input distributions: as in Example 4.1, any violated constraint rejects $G \cup F(R)$ and removes R from SolRC. In general, the observational distributions compatible with a latent-variable causal model are restricted by two families of constraints (Evans, 2023; Ansanelli et al., 2026): equality constraints, such as conditional independences (Pearl, 2009) and nested Markov (Verma) constraints (Verma & Pearl, 1990; Richardson et al., 2023), and inequality constraints, such as the instrumental inequality (Pearl, 1995), e-separation inequalities (Evans, 2012), and Bell inequalities (Bell, 1964). Enumerating and testing these constraints individually is non-trivial, particularly since inequality constraints are difficult to derive in general. We bypass this by directly searching for an SCM that is consistent with $G \cup F(R)$ and reproduces both the normal and anomalous distributions; the existence of such an SCM implies that all constraints are satisfied. We establish a connection between SolRC and normal anomalous SCMs as follows:

Definition 4.9 (SCM Pairset, $\widehat{M}_{na}(G^*, \mathtt{R})$)**.** For any $\mathtt{R} \subset \mathbf{V}$, there exists a normal and anomalous SCM pair $\widehat{\mathcal{M}}_n = \big(\{\widehat{f}_X^n\}_{X \in \mathbf{V}},\, \widehat{P}(U)\big)$ and $\widehat{\mathcal{M}}_a = \big(\{\widehat{f}_X^a\}_{X \in \mathbf{V}},\, \widehat{P}(U)\big)$ over $\mathbf{V}$ sharing the exogenous distribution $\widehat{P}(U)$ such that (i) $\widehat{\mathcal{M}}_n$ and $\widehat{\mathcal{M}}_a$ have the true ADMG graph G^*; and (ii) For all variables $V \in \mathtt{R}$, mechanisms change arbitrarily (including no shift) across environments, $\widehat{f}_V^n(\mathrm{pa}(V), U_V) \to \widehat{f}_V^a(\mathrm{pa}(V), U_V)$ while invariant mechanisms $\widehat{f}_V^n(\mathrm{pa}(V), U_V) = \widehat{f}_V^a(\mathrm{pa}(V), U_V)$, for all $V \in \mathbf{V} \setminus \mathtt{R}$. We define the set of all such SCM pairs as SCM Pairset $\widehat{M}_{na}(G^*, \mathtt{R})$.

Definition 4.10 (Population root-cause loss, ℓ^*)**.** For $\mathtt{R} \subseteq \mathbf{V}$, the *population root-cause loss* is the least distributional mismatch attainable by a candidate pair,

$$\ell^*(G^*, \mathtt{R}) \;=\; \inf_{(\widehat{\mathcal{M}}_n, \widehat{\mathcal{M}}_a) \in \widehat{M}_{na}(G^*, \mathtt{R})} \Big[\, d\big(\widehat{P}_n,\, P_n^*\big) + d\big(\widehat{P}_a,\, P_a^*\big) \,\Big] \tag{2}$$

where $\widehat{P}_n$ and $\widehat{P}_a$ are the distributions over $\mathbf{V}$ induced by $\widehat{\mathcal{M}}_n$ and $\widehat{\mathcal{M}}_a$. $\widehat{M}_{na}(G^*, \mathtt{R})$ is the SCM pairs, and d is a discrepancy on distributions over $\mathbf{V}$ with $d(P, Q) = 0 \iff P = Q$.

4.2 RCA-DCM: Root-Cause Detection via SCM Search

Brute Force algorithm: To obtain the members of SolRC, we can: 1. Iterate over all possible $\mathtt{R} \subseteq \mathbf{V}$ as a candidate solution 2. For each R, search for SCM Pairset $\widehat{M}_{na}(G^*, \mathtt{R})$ such that $\ell^*(G^*, \mathtt{R}) = 0$ 3. If there exists at least one SCM pair $(\widehat{\mathcal{M}}_n, \widehat{\mathcal{M}}_a) \in \widehat{M}_{na}$ found, we add R to SolRC as a candidate solution. Nevertheless, we have two challenges: i) this approach requires iteration over $2^{|\mathbf{V}|}$ number of sets and ii) how to find such an SCM pair - is non-trivial. To efficiently find such pair with stochastic gradient descent, we learn a proxy of the SCMs with deep causal models (Definition 4.11) and optimize following our constraints.

Definition 4.11 (Deep causal generative models (DCM) (Kocaoglu et al., 2018; Xia et al., 2021; Rahman & Kocaoglu, 2024)). A neural net architecture $\mathbb{G}$ is called a deep causal generative model (DCM) for an ADMG $G = (\mathcal{V}, \mathcal{E})$ if it is composed of a collection of neural nets, one f_i (or interchangeably f_{V_i}) for each $V_i \in \mathcal{V}$ such that i) *each f_i accepts a sufficiently high-dimensional noise vector N_i,* ii) *the output of f_j is input to f_i iff $V_j \in Pa_G(V_i)$,* iii) $N_i = N_j$ *iff* $V_i \leftrightarrow V_j$. A DCM is trained to learn a proxy of the true SCM. DCM generators are represented as $\mathbb{G} = \{f_1, ..., f_n\}$ parameterized by $\Theta = \{\theta_1, ..., \theta_{|\mathbf{V}|}\}$ where $n = |\mathbf{V}|$. Similar to the original data distribution, $P(\mathbf{V})$, we define $\widehat{P}(\mathbf{V})$ to be the distribution induced by the θ parameterized DCM. Noise vectors N_i replace both the exogenous noises and the unobserved confounders in the true SCM. They are of sufficiently high dimension to induce the observed distribution. We say that a DCM is *representative enough for an ADMG* if the neural networks have sufficiently many parameters to induce any observed distribution induced by any SCM that entails the ADMG. Let $v = [v_1, v_2, ..., v_n]$ st $V_i \in \mathbf{V}$. $v \sim P(\mathbf{V})$ is real samples and $\widehat{v} \sim \widehat{P}(\mathbf{V})$ is DCM generated samples: $\widehat{v}_i = f_i(\widehat{pa}(V_i), N_i)$; $f_i \in \{f_V : V \in \mathbf{V}\}$. A discriminator compares v and $\widehat{v}$ to train $\forall f_i$.

Our aim is to now construct the `SolRC` from a pool of exponentially many candidate subsets by searching for SCMs that are consistent with the causal graph G and satisfy the distributional constraints. We circumvent this problem by iterating over variables in $\mathbf{V}$ and searching for SCMs assuming Y is not a root cause, i.e., the absence of the edge $F \to R$. We test whether there exist two SCMs consistent with the causal graph G^*, allowing changes in all mechanisms except Y, can shift from normal P_n to anomalous P_a distribution. If no such SCM exists, the input distribution must violate at least one equality or inequality constraint implied by the absence of $F \to R$. We therefore conclude that R must be a root cause, without iterating many candidates and identifying the violated constraint explicitly. We claim the following lemma.

Lemma 4.12. *Suppose $\mathcal{D}_n, \mathcal{D}_a$ are generated by $(\mathcal{M}_n^*, \mathcal{M}_a^*)$ with common ADMG G^*, and let latent invariance hold. For any $Y \in \mathbf{V}$, if $\ell^*(G^*, \mathbf{V} \setminus \{Y\}) > 0$ then $\Rightarrow$ $\mathbf{Z} \setminus \{Y\} \notin$ `SolRC`(G^*, P_n^*, P_a^*) for any $\mathbf{Z} \subseteq \mathbf{V}$.*

Theorem 4.13. *For any $Y \in \mathbf{V}$, if $\ell^*(G^*, \mathbf{V} \setminus \{Y\}) > 0$ then $Y \in R^*$.*

Intuitively, if there exists no SCM pairs that achieves $\ell^*(G^*, Y) = 0$, the no solution set without Y exists. Y must belong to all solution set $\mathtt{R} \in$ `SolRC` and thus a true root cause. This reduces $2^{|\mathbf{V}|}$ bruteforce iterations to $|\mathbf{V}|$ iterations.

Proposition 4.14 (Soundness). $\widehat{R} = \bigcap$ `SolRC`$(G^*) \subseteq R^*$

Now, the optimization in Equation 2: SCM search is over a set with infinitely many members where each member is a SCM pair consistent with the causal graph G^*, allowing changes in all mechanisms except Y, can shift from normal P_n to anomalous P_a distribution. We provide our algorithm workflow for an example graph in Figure 5 and our pseudo-code in Algorithm 1. For each variable, we can run the algorithm independently in parallel. For any variable $Y \in \mathbf{V}$, we first initialize two deep causal models $(\widehat{\mathcal{M}}_n, \widehat{\mathcal{M}}_a)$ to represent the normal and anomaly SCM (Line 2). We use causal normalizing flows (Javaloy et al., 2023) as the DCM backbone. In Line 4, $\widehat{\mathcal{M}}_n$ generates samples from the normal joint distribution $\widehat{P}_n(\mathbf{V})$. Next, we copy the model $f_Y(.)$ from $\widehat{\mathcal{M}}_n$ to $\widehat{\mathcal{M}}_a$ (Line 5). In Line 6, we employ $\widehat{\mathcal{M}}_a$ to generate samples from anomaly joint distribution $\widehat{P}_a(\mathbf{V})$. We calculate L by comparing the generated samples with normal and anomaly dataset and backpropogate on L. The losses L_1 and L_2 are estimated empirically using real samples from D_n, D_a and generated samples from $\widehat{P}_n, \widehat{P}_a$. We store L in *score* and rank them in Line 9 to obtain the nodes with highest mechanism shift at the top. We learn proxy to the structural function in the design deep causal models and arrange them according to the causal graph. We execute Algorithm 2 to generate samples from $\mathcal{M}$. Finally, we use distance metric to compare the dissimilarity between learned and true distribution.

Algorithm 1 (Input: G, $D_n \sim P_n(\mathbf{V})$, $D_a \sim P_a(\mathbf{V})$)

1: **for** each $Y \in \mathbf{V}$ in parallel **do**
2: Initialize $\mathcal{M}_n$ and $\mathcal{M}_a$.
3: **for** Run for N epochs **do**
4: Use $\mathcal{M}_n$ to generate samples from $\widehat{P}_n(\mathbf{V})$.
5: Copy f_Y from $\mathcal{M}_n$ to $\mathcal{M}_a$ and freeze it.
6: Use $\mathcal{M}_a$ to generate samples from $\widehat{P}_a(\mathbf{V})$.
7: Backprop: $L = L_1(\widehat{P}_n, P_n) + L_2(\widehat{P}_a, P_a)$
8: $score[Y] = L$
9: Sort *score* and return as *Rank*.

Error analysis of the RCA-DCM ranking: Theorem 4.13 states that if $\ell(G^*, \mathbf{V} \setminus \{Y\}) > 0$, then Y is a root cause. In practice, however, even with perfect model training, the finite sample size leads us to observe $\ell(\cdot) > 0$ for every variable. In Appendix E, we show that under a mild condition, every root cause is scored strictly above every non-root cause, so the $k = |R^*|$ root causes occupy the top k positions of the ranking returned by RCA-DCM.

4.3 Theoretical Guarantees under Graph Mis-specification

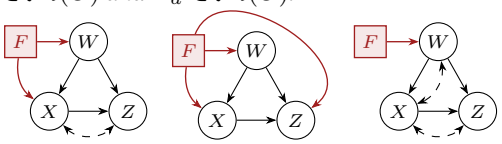
In the previous section, we provided RCA-DCM to find root causes in a system containing unobserved confounders given the true causal graph (ADMG) as input. However, it assumes access to the true ADMG. In most real-world scenarios, it is difficult to find such a graph without domain knowledge. Thus, in this section, we analyze what RCA-DCM outputs when this assumption is violated and we have a sparser or denser graph compared to the true ADMG. The proofs are provided in Appendix F.

First, we answer an important question: can we obtain the true root cause set from an rca algorithm for any arbitrary mis-specified graph? We formally prove that it is only possible when the normal and anomalous distributions generated from the SCM with true grahp G^* are realized by the mis-specified graph G' as well. Note that exploring arbitrary different graph G' is outside the scope of this paper and we enlist this as a limitation of the current work. Lemma 4.15 formalizes this.

Lemma 4.15. *Let G^*, G' be the true and mis-specified graphs, either by adding or by removing edges. Given (P_n, P_a) realized by (G^*, R^*), we have $\texttt{SolRC}(G') \neq \emptyset$ iff $P_n, P_a \in \mathcal{M}(G')$.*

Corollary 4.16. *$\texttt{minrc}(G) < \infty$ if and only if $P_n \in \mathcal{M}(G)$ and $P_a \in \mathcal{M}(G)$.*

Consider the sparse and dense misspecifications G_S and G_D of the 3-node true graph G^* in Figure 3, and suppose both can realize the input pair (P_n, P_a), so that each yields a non-empty SolRC. Since the true root cause set is $R^* = \{W, X\}$, we observe shifts from P_n to P_a in $P(W)$, $P(X \mid W)$, and $P(Z \mid X, W)$, and $\texttt{SolRC}(G^*) = \{\{W, X\}, \{W, X, Z\}\}$. Given the sparse graph G_S (Figure 3b), reproducing all three shifts requires shifting every mechanism f_W, f_X, f_Z, i.e., $F \to \{W, X, Z\}$, so $\texttt{SolRC}(G_S) = \{\{W, X, Z\}\}$. Any proper subset, such as $\{W, X\}$, would imply $Z \perp\!\!\!\perp F \mid \{W, X\}$ in G_S and hence $P_n(Z \mid W, X) = P_a(Z \mid W, X)$, contradicting the observed shift. Given the dense graph G_D (Figure 3c), Lemma 4.17 gives $\texttt{SolRC}(G^*) \subseteq \texttt{SolRC}(G_D)$. Moreover, for some distributions, shifting f_W alone reproduces all three shifts, so $\texttt{SolRC}(G_D) = \{\{W\}, \{W, X\}, \{W, X, Z\}\}$. Hence, $\texttt{SolRC}(G_S) \subset \texttt{SolRC}(G^*) \subset \texttt{SolRC}(G_D)$: a denser graph can admit smaller root cause sets that are infeasible under the true graph, whereas a sparser graph can exclude the true root cause set and require a larger one.



(a) G^* (true) (b) G_S (sparse) (c) G_D (dense)

$\texttt{SolRC}(G_S) \subseteq \texttt{SolRC}(G^*) \subseteq \texttt{SolRC}(G_D)$

Figure 3: Augmented graphs of two misspecifications of the true graph, with $G_S \subset G^* \subset G_D$. All F-edges are in red. While $R(G^*) = \{W, X\}$ is ground truth, we obtain the superset $\widehat{R}(G_S) = \{W, X, Z\}$ for G_S, and the subset $\widehat{R}(G_D) = \{W\}$ for G_D.

Given a sparse graph G_1, we can add directed (causal relations) or bi-directed edges (latent confounders) to obtain a denser graph G_2 such that G_1 is structurally dominated by G_2, written $G_1 \subseteq G_2$. By Lemma 4.6, structural dominance implies that G_2 can realize any pair (P_n, P_a) generated by G_1. Hence, rather than comparing the root causes obtained under a misspecified graph against those obtained under the true graph, we compare how they change between a sparser vs. a denser graph.

Lemma 4.17. *If $G_1 \subseteq G_2$, then $\texttt{SolRC}(G_1) \subseteq \texttt{SolRC}(G_2)$.*

Lemma 4.18. *If $G_1 \subseteq G_2$, then $\texttt{minrc}(G_2, P_n, P_a) \leq \texttt{minrc}(G_1, P_n, P_a)$.*

5 Experimental Results

We evaluate RCA-DCM on synthetic, semi-synthetic, and real-world datasets. We compare it against five representative non-causal and causal RCA approaches based on statistical hypothesis testing (BARO (Pham et al., 2024a)), z-score-based statistical analysis (NSigma (Li et al., 2022a)), regression-based hypothesis testing (CIRCA (Li et al., 2022b)), causal discovery (RCD (Ikram et al., 2022b)), and intervention-aware causal inference (RCG (Ikram et al., 2025a)). We adopt their implementation

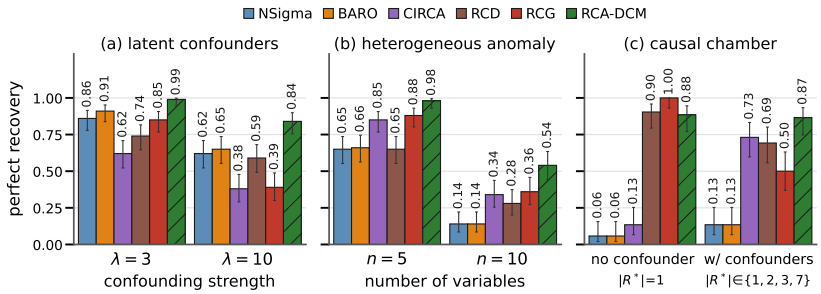


Figure 4: Perfect recovery rate (PRR) with 95% confidence intervals. (a) increasing latent confounding, (b) heterogeneous anomalies where a non-cause can out-shift a true cause, (c) real causal-chamber data with the root cause observed vs. hidden. RCA−DCM remains strong across all three settings; comparisons use a paired test, so overlapping intervals do not imply the absence of a difference.

details and hyperparameters from the RCAEval repository (Pham et al., 2025). **Metrics:** To evaluate the output ranking, we use $T@k = \frac{1}{|\mathbf{E}|}\sum_{e\in\mathbf{E}} \frac{\left|\widehat{R}_e[1,..,k]\cap R_e^*\right|}{\min(k,|R_e^*|)}$, i.e., how many of the true root causes lie in the top k, and $\mathrm{PRR} = \frac{1}{|\mathbf{E}|}\sum_{e\in\mathbf{E}} \mathbf{1}[T@|R_e^*| = 1]$, i.e., in how many cases all true root causes are at the top (i.e., the perfect recovery rate). We provide more experimental details, and discuss the availability of the causal graph, sample-size requirements, and runtime, in Appendices G.

5.1 Simulated Datasets

We follow Yang et al. (2024) to generate synthetic normal, anomalous data from randomly generated graphs, varying the total variable count, latent proportion, and edge density; full details in Appendix G.

Exp 1: Nonlinear model with unobserved confounders: We use graphs with $p = 6$ observed and $m = 4$ latent variables (total $n = 10$) and vary the confounding strength $\lambda \in \{3, 10\}$, which uniformly scales all latent-to-observed coefficients. We take $\lfloor p/2 \rfloor = 3$ observed variables as ground-truth root causes. **Observation:** Figure 4(a) reports PRR for $\lambda \in \{3, 10\}$. RCA−DCM attains 99% and 84%, ahead of every baseline in both settings (per-baseline values are listed in the Appendix). All methods degrade as confounding strengthens: RCA−DCM, starting at 99%, and RCD, starting at 74%, both drop by 15 percentage points, while BARO, NSigma, CIRCA, and RCG drop by 26, 24, 24, and 46 points, respectively. This suggests that RCA−DCM is more robust to strong confounding and better preserves root-cause recovery than the baselines.

Exp 2: Nonlinear model with heterogeneous anomalies: As discussed in Section 1, baselines that rank variables by marginal shift fail when a non-root-cause descendant out-shifts a root cause (heterogeneous anomalies). We construct such heterogeneous anomalies by scaling the injected shift magnitude linearly with topological depth, so that shifts accumulate along directed paths and a downstream descendant out-shifts a genuine upstream root cause. In the generated data, this trap occurs (i.e., some non-root cause exhibits a larger marginal shift than some true root cause) in roughly 60% of trials at $p = 10$. We vary the number of observed variables $n \in \{5, 10\}$ ($n = p$ as $m = 0$). **Observation:** Figure 4(b) shows PRR as the system grows. RCA−DCM achieves the highest rate at both sizes, 98% at $n = 5$ and 54% at $n = 10$, outperforming all five baselines in both settings. The marginal-shift methods (BARO and NSigma) degrade the most, falling to 14% at $n = 10$; this is the failure mode the construction targets, since it makes ranking by shift magnitude wrong in most trials. The graph-based methods (RCG 36%, CIRCA 34%, RCD 28%) also degrade as the system grows. These results suggest that RCA−DCM is more robust to increasing system complexity, where the effects of multiple root causes can compound.

Exp 3: Robustness to graph misspecification. Perturbing *only* the graph supplied to RCA−DCM on the 100 datasets with $\lambda = 10$ from **Exp 1** (true-graph control: PRR of 88% and 84% in two independent runs), adding 50% spurious directed and bidirected edges (PRR = 84%, $p = 0.39$) or deleting 50% of the bidirected edges (PRR = 87%, $p = 1.00$) does not change accuracy significantly under a paired McNemar test; deleting 50% of both directed and bidirected edges degrades it (76%, $p = 0.019$), which is still above every baseline at $\lambda = 10$ (best: BARO, 65%). These results are consistent with Section 4.3 (see Appendix G).

5.2 Causal Chamber (Real-World Physical Testbed)

Exp 4: We evaluate RCA−DCM on the Causal Chambers benchmark (Gamella et al., 2025), a real-world physical testbed. We use the light-tunnel chamber in its standard configuration, which comes with a ground-truth DAG over 38 variables and 57 edges. This yields 52 cases, which we evaluate in two conditions: with the true root cause *observed*, and with it *hidden*, so that it acts as a latent confounder among its former children. In the observed condition, $|R^*| = 1$; in the hidden condition, the hidden cause has between 1 and 7 children, all of which must be recovered as root causes. **Observation:** Figure 4(c) reports PRR in both conditions. When the root cause is observed, there is no confounding and every case has a single root cause; here, RCA−DCM is competitive but does not lead (88%, against 100% for RCG and 90% for RCD). Under this harder setting, the ordering reverses: RCA−DCM attains 87%, against 73% for the best baseline, and it is the only method whose performance barely changes between conditions (−2 points, against −50 for RCG and −21 for RCD). The gap is largest on the cases with multiple root causes: on these cases alone, RCA−DCM attains a PRR of 63%, while CIRCA reaches 25% and RCG 13%.

5.3 Microservice Datasets

Exp 5: We evaluate RCA−DCM on SockShop, a cloud computing dataset recorded from a microservice-based replica of a web application specifically designed for evaluating root cause analysis methods (Pham et al., 2024b). It contains 125 anomaly datasets covering five fault types injected into five services. We assume that *no* causal structure is available to RCA−DCM: instead of a call graph, we supply a *confounded sink*, in which every recorded service points to the service under test and a single shared latent confounds all of them (the justification for this choice is given in the Appendix. This places RCA−DCM at a disadvantage relative to two of the baselines: CIRCA and RCG are given the true edge-inverted call graph, while NSigma, BARO, and RCD do not use a graph at all.

Table 1: Average Top-k accuracy over the five fault types (best in bold).

	Sock Shop			Online Boutique		
Method	T1	T3	T5	T1	T3	T5
DCM	**0.88**	**0.99**	**0.99**	**0.78**	0.90	**0.98**
NSigma	0.75	0.97	**0.99**	0.70	**0.91**	0.96
BARO	0.74	0.98	**0.99**	0.62	0.90	0.94
CIRCA	0.65	0.96	**0.99**	0.52	0.89	**0.98**
RCD	0.38	0.54	0.58	0.49	0.59	0.62
RCG	0.54	0.67	0.70	0.71	0.83	0.85

Exp 6: We repeat the evaluation on Online Boutique (Pham et al., 2024b), a second microservice benchmark with the same five fault types, each injected five times, again yielding 125 anomaly datasets, but over 11 services and with substantially longer traces (roughly 2100 normal and 2100 anomalous observations per case). The graph treatment is identical to Exp 5: RCA−DCM receives only the confounded sink, while CIRCA and RCG receive the true call graph.

Results: Table 1 reports the average top-k accuracy (per-fault results are in Table 3). On SockShop, RCA−DCM attains the highest average top-1 accuracy, 0.88, ahead of NSigma (0.75), BARO (0.74), CIRCA (0.65), RCG (0.54), and RCD (0.38); every gap is significant under a paired test ($p < 0.001$). It is best or tied for best on all five fault types, and its errors are near-misses rather than failures: its average top-3 accuracy reaches 0.99. On Online Boutique, RCA−DCM again ranks first, at 0.78, ahead of RCG (0.71), NSigma (0.70), BARO (0.62), CIRCA (0.52), and RCD (0.49), with an average top-3 accuracy of 0.90; the margins over BARO, CIRCA, and RCD are significant, while those over NSigma and RCG are not at this sample size. Further per-benchmark observations are discussed in Appendix. Notably, RCA−DCM attains these results without any causal structure, while the two graph-based baselines are given the true call graph. Together with the results in Section 5.1, this supports the view that the advantage of RCA−DCM does not depend on access to an accurate structure.

6 Conclusion

In this paper, we address the problem of root cause analysis in presence of any number of unobserved confounders. For that purpose, we propose a sound approach that can detect the root causes in parallel by searching for feasible structural causal models. Finally, we demonstrate our performance on synthetic dataset and real-world physical testbed. In our future work, we aim to relax the assumption on partial order and make it suitable for high-dimensional variables.

Reproducibility statement

The full set of assumptions required for our soundness, completeness, and graph misspecification results is stated in Appendix B, and the corresponding proof details are provided in the appendix. Section 5 describes the datasets, causal graphs, anomaly settings, evaluation metrics, baselines, and experimental protocols needed to reproduce the main results. Training details, implementation choices, and dataset-specific settings are given in Appendix G, with further details for nonlinear models with unobserved confounders in Appendix G.2 and computational complexity in Appendix G.3. Since DCM-RCA trains one DCM for the full graph and one candidate-specific DCM for each node, the candidate-specific models can be trained in parallel, reducing practical wall-clock time. We report variability information where applicable, including error bars for the simulated experiments and accuracy across multiple anomaly instances for the real-world datasets; the reported perfect recovery and top-1 accuracy results are computed over repeated runs or multiple injected anomaly datasets. The datasets are either synthetically generated using the procedures described in the paper or publicly available. We will provide the code and try our best to make it usable, documented, and easy to run for reproducing the main experimental results, and the released code will include instructions for reproducing the reported experiments.

Ethics statement

The research conducted in this paper conforms, in every respect, with the ICLR Code of Ethics. The broader impacts are discussed in Appendix A.2: the paper describes potential benefits for diagnosing failures in complex systems, and also discusses risks from incorrect causal assumptions, misspecified graphs, or use in high-stakes settings without sufficient validation. This work involves no crowdsourcing experiments and no research with human subjects. The datasets used are either synthetically generated using the procedures described in the paper or publicly available, and their creators are credited in the text.

AI use statement

We used generative AI tools only for writing, editing, and formatting purposes. Such tools are not an important, original, or non-standard component of the core methods of this research, and did not affect the methodology, scientific rigor, or originality of the work. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Marina Maciel Ansanelli, Elie Wolfe, and Robert W. Spekkens. The observational partial order of causal structures with latent variables. *Journal of Causal Inference*, 14(1):20250009, 2026. doi: 10.1515/jci-2025-0009.

Xiwen Bai, Jesús Fernández-Villaverde, Yiliang Li, and Francesco Zanetti. The causal effects of global supply chain disruptions on macroeconomic outcomes: evidence and theory. 2024.

John S. Bell. On the Einstein Podolsky Rosen paradox. *Physics Physique Fizika*, 1(3):195–200, 1964.

Kailash Budhathoki, Lenon Minorics, Patrick Blöbaum, and Dominik Janzing. Causal structure-based root cause analysis of outliers. In *International conference on machine learning*, pp. 2357–2369. PMLR, 2022.

Pengfei Chen, Yong Qi, Pengfei Zheng, and Di Hou. Causeinfer: Automatic and distributed performance diagnosis with hierarchical causality graph in large distributed systems. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pp. 1887–1895. IEEE, 2014.

Sanjiv Das, Richard Stanton, and Nancy Wallace. Algorithmic fairness. *Annual Review of Financial Economics*, 15:565–593, 2023.

Robin J Evans. Graphical methods for inequality constraints in marginalized dags. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6. IEEE, 2012.

Robin J. Evans. Latent-free equivalent mDAGs. *Algebraic Statistics*, 14(1):3–16, 2023.

Juan L Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for ai methodology. *Nature Machine Intelligence*, 7(1):107–118, 2025.

Nikhil Gupta, Dhivya Eswaran, Neil Shah, Leman Akoglu, and Christos Faloutsos. Beyond outlier detection: Lookout for pictorial explanation. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I*, pp. 122–138. Springer, 2019.

Azam Ikram, Sarthak Chakraborty, Subrata Mitra, Shiv Saini, Saurabh Bagchi, and Murat Kocaoglu. Root cause analysis of failures in microservices through causal discovery. *Advances in Neural Information Processing Systems*, 35:31158–31170, 2022a.

Azam Ikram, Sarthak Chakraborty, Subrata Mitra, Shiv Kumar Saini, Saurabh Bagchi, and Murat Kocaoglu. Root cause analysis of failures in microservices through causal discovery. In *Advances in Neural Information Processing Systems*, volume 35, 2022b.

Azam Ikram, Kenneth Lee, Shubham Agarwal, Shiv Kumar Saini, Saurabh Bagchi, and Murat Kocaoglu. Root cause analysis of failures from partial causal structures. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 1794–1818. PMLR, 2025a.

Azam Ikram, Kenneth Lee, Shubham Agarwal, Shiv Kumar Saini, Saurabh Bagchi, and Murat Kocaoglu. Root cause analysis of failures from partial causal structures. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025b.

Atsushi Inoue and Barbara Rossi. A new approach to measuring economic policy shocks, with an application to conventional and unconventional monetary policy. *Quantitative Economics*, 12(4): 1085–1138, 2021.

Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems*, 33:9551–9561, 2020.

Adrián Javaloy, Pablo Sánchez-Martín, and Isabel Valera. Causal normalizing flows: from theory to practice. *Advances in Neural Information Processing Systems*, 36:58833–58864, 2023.

Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. Efficient causal graph discovery using large language models. In *ICLR 2024 Workshop on How Far Are We From AGI*, 2024. arXiv:2402.01207.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2024.

Edwin M. Knorr and Raymond T. Ng. Finding intensional knowledge of distance-based outliers. In *VLDB*, volume 99, pp. 211–222, 1999.

Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018.

Jinzhou Li, Benjamin B. Chu, Ines F. Scheller, Julien Gagneur, and Marloes H. Maathuis. Root cause discovery via permutations and cholesky decomposition, 2024.

Mingjie Li, Zeyan Li, Kanglin Yin, Xiaohui Nie, Wenchi Zhang, Kaixin Sui, and Dan 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*, pp. 3230–3240, 2022a.

Mingjie Li, Zeyan Li, Kanglin Yin, Xiaohui Nie, Wenchi Zhang, Kaixin Sui, and Dan 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*, pp. 3230–3240, 2022b. doi: 10.1145/3534678.3539041.

Cheng-Ming Lin, Ching Chang, Wei-Yao Wang, Kuang-Da Wang, and Wen-Chih Peng. Root cause analysis in microservice using neural granger causal discovery. *arXiv preprint arXiv:2402.01140*, 2024.

Dewei Liu, Chuan He, Xin Peng, Fan Lin, Chenxi Zhang, Shengfang Gong, Ziang Li, Jiayu Ou, and Zheshun Wu. Microhecl: High-efficient root cause localization in large-scale microservice systems. In *Proceedings of the 43rd International Conference on Software Engineering: Software Engineering in Practice*, ICSE-SEIP '21, pp. 338–347. IEEE Press, 2021.

Ninghao Liu, Donghwa Shin, and Xia Hu. Contextual outlier interpretation. *arXiv preprint arXiv:1711.10589*, 2017.

Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. Causal discovery with language models as imperfect experts. *arXiv preprint arXiv:2307.02390*, 2023.

Meng Ma, Weilan Lin, Disheng Pan, and Ping Wang. Ms-rank: Multi-metric and self-adaptive root cause diagnosis for microservice applications. In *2019 IEEE International Conference on Web Services (ICWS)*, pp. 60–67. IEEE, 2019.

Meng Ma, Jingmin Xu, Yuan Wang, Pengfei Chen, Zonghua Zhang, and Ping Wang. Automap: Diagnose your microservice-based web applications automatically. In *Proceedings of The Web Conference 2020*, pp. 246–258, 2020.

Meghanath Macha and Leman Akoglu. Explaining anomalies in groups with characterizing subspace rules. *Data Mining and Knowledge Discovery*, 32:1444–1480, 2018.

Yuan Meng, Shenglin Zhang, Yongqian Sun, Ruru Zhang, Zhilong Hu, Yiyin Zhang, Chenyang Jia, Zhaogang Wang, and Dan Pei. Localizing failure root causes in a microservice through causality inference. In *2020 IEEE/ACM 28th International Symposium on Quality of Service (IWQoS)*, pp. 1–10. IEEE, 2020.

Barbora Micenková, Raymond T. Ng, Xuan-Hong Dang, and Ira Assent. Explaining outliers by subspace separability. In *2013 IEEE 13th International Conference on Data Mining*, pp. 518–527. IEEE, 2013.

Nastaran Okati, Sergio Hernan Garrido Mejia, William Roy Orchard, Patrick Blöbaum, and Dominik Janzing. Root cause analysis of outliers with missing structural knowledge. *arXiv e-prints*, pp. arXiv–2406, 2024.

William Roy Orchard, Nastaran Okati, Sergio Hernan Garrido Mejia, Patrick Blöbaum, and Dominik Janzing. Root cause analysis of outliers with missing structural knowledge. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

J. Pearl. *Causality*. Cambridge University Press, 2009. ISBN 9781139643986. URL `https://books.google.com/books?id=LLkhAwAAQBAJ`.

Judea Pearl. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 435–443, 1995.

Luan Pham, Huong Ha, and Hongyu 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, 2024a.

Luan Pham, Huong Ha, and Hongyu Zhang. Root cause analysis for microservice system based on causal inference: How far are we? In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pp. 706–715, 2024b.

Luan Pham, Hongyu Zhang, Huong Ha, Flora Salim, and Xiuzhen Zhang. Rcaeval: A benchmark for root cause analysis of microservice systems with telemetry data. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 777–780, 2025.

Md Musfiqur Rahman and Murat Kocaoglu. Modular learning of deep causal generative models for high-dimensional causal inference. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=bOhzU7NpTB.

Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1):334–361, 2023.

Huasong Shan, Yuan Chen, Haifeng Liu, Yunpeng Zhang, Xiao Xiao, Xiaofeng He, Min Li, and Wei Ding. ϵ-diagnosis: Unsupervised and real-time diagnosis of small-window long-tail latency in large-scale microservice platforms. In *The World Wide Web Conference*, pp. 3215–3222, 2019.

Eric V. Strobl. Counterfactual formulation of patient-specific root causes of disease. *Journal of Biomedical Informatics*, pp. 104585, 2024.

Eric V. Strobl and Thomas A. Lasko. Identifying patient-specific root causes with the heteroscedastic noise model. *Journal of Computational Science*, 72:102099, 2023.

Jellis Vanhoeyveld, David Martens, and Bruno Peeters. Value-added tax fraud detection with scalable anomaly detection techniques. *Applied Soft Computing*, 86:105895, 2020.

Thomas S. Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 255–270, 1990.

Dongjie Wang, Zhengzhang Chen, Jingchao Ni, Liang Tong, Zheng Wang, Yanjie Fu, and Haifeng Chen. Hierarchical graph neural networks for causal discovery and root cause localization. *arXiv preprint arXiv:2302.01987*, 2023.

Ping Wang, Jingmin Xu, Meng Ma, Weilan Lin, Disheng Pan, Yuan Wang, and Pengfei Chen. Cloudranger: Root cause identification for cloud native systems. In *2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, pp. 492–502. IEEE, 2018.

Li Wu, Johan Tordsson, Jasmin Bogatinovski, Erik Elmroth, and Odej Kao. MicroDiag: Fine-grained performance diagnosis for microservice systems. In *2021 IEEE/ACM International Workshop on Cloud Intelligence*, pp. 31–36, 2021.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, 34:10823–10836, 2021.

Ruyue Xin, Peng Chen, and Zhiming Zhao. Causalrca: Causal inference based precise fine-grained root cause localization for microservice applications. *Journal of Systems and Software*, 203:111724, 2023.

Yuqin Yang, Saber Salehkaleybar, and Negar Kiyavash. Learning unknown intervention targets in structural causal models from heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 3187–3195. PMLR, 2024.

Zhenhe Yao, Changhua Pei, Wenxiao Chen, Hanzhang Wang, Liangfei Su, Huai Jiang, Zhe Xie, Xiaohui Nie, and Dan Pei. Chain-of-event: Interpretable root cause analysis for microservices through automatically learning weighted event causal graph. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, pp. 50–61, 2024.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, et al. Which agent causes task failures and when? on automated failure attribution of llm multi-agent systems. *arXiv preprint arXiv:2505.00212*, 2025.

Lecheng Zheng, Zhengzhang Chen, Jingrui He, and Haifeng Chen. Mulan: multi-modal causal structure learning and root cause analysis for microservice systems. In *Proceedings of the ACM Web Conference 2024*, pp. 4107–4116, 2024.

A Additional Discussion

A.1 Limitations and Future works

We assume we have access to a causal structure or its total order. Training neural networks might be resource consuming and time consuming for some applications. We assume that normal and anomalous dataset is clearly separated where anomaly point detect might be challenging for some tasks. Due to convergence error in neural networks, we might always have non-zero value of the loss function even if there is not mechanism shift. Although we showed analysis for denser and sparser graph, in this work, we did not explore with the mis-specified graph is arbitrarily different. We aim to address these limitations in our future work.

A.2 Broader Impact

Root cause analysis in the presence of unobserved confounders can improve the reliability and safety of complex systems by helping practitioners identify failure sources even when some relevant variables are hidden or unmeasured. By using deep causal models, our approach can support more accurate diagnosis in domains such as cloud computing, healthcare, manufacturing, and critical infrastructure, where failures may have significant economic or societal consequences. At the same time, incorrect causal assumptions or misspecified graphs may lead to misleading root cause conclusions, especially in high-stakes applications. Therefore, our method should be used with domain knowledge, uncertainty assessment, and careful validation before deployment in real-world decision-making systems.

B Extended Related Work

Causal graph-based RCA. RCA is an important problem in microservice systems, which have increasingly adopted causal inference as a principled approach. Early causal RCA methods construct a causal graph over service-level metrics and identify the root cause by traversing or ranking nodes with an anomaly score (Chen et al., 2014). Many approaches estimate the causal structure from observational data using constraint-based discovery algorithms such as PC, and then apply scoring methods to traverse the graph and localize the root cause (Wang et al., 2018; Ma et al., 2019; Meng et al., 2020). MicroDiag (Wu et al., 2021) performs fine-grained diagnosis using metric-level dependency graphs, but it remains limited to the observed metrics and does not address unobserved confounding. A large-scale evaluation of causal discovery-based RCA methods further shows that no existing method performs robustly across datasets and failure types (Pham et al., 2024b).

RCA as intervention recognition. Several recent works model system failures as interventions on causal mechanisms, building on the soft-intervention view of distribution shifts (Jaber et al., 2020). RCD (Ikram et al., 2022a) treats a failure as a soft intervention and performs local causal discovery to avoid learning the full graph, and RCG (Ikram et al., 2025b) extends this idea using partial causal structures. CIRCA (Li et al., 2022b) formulates RCA as an intervention recognition problem, based on a sufficient condition for a variable in a causal Bayesian network to be a root cause. CausalRCA (Xin et al., 2023) learns the causal graph with a gradient-based causal discovery method. These methods depend either on expert-provided or learned graphs, which may be inaccurate under noise and hidden confounding, or on conditional independence and invariance tests, whose reliability degrades with continuous variables and growing conditioning sets.

RCA with limited structural knowledge. Orchard et al. (2025) formalize RCA as identifying the mechanism change under a single-target soft intervention, provide theoretical guarantees for restricted graph classes such as polytrees, and propose the efficient SO and ST methods for settings with missing structural knowledge. Cholesky-based RCA (Li et al., 2024) exploits invariances specific to linear SCMs. Neither approach explicitly models nonlinear mechanisms with latent confounding.

Counterfactual and score-based formulations. Budhathoki et al. (2022) formalize RCA as a quantitative contribution analysis based on counterfactuals, attributing an anomaly to the mechanisms whose deviation from normal behavior explains it. Their approach uses a graph learned from normal operation and allows anomalous samples from multiple distributions, but it requires knowing the full

SCM, including invertible functional relations, which are hard to estimate in practice. Okati et al. (2024) propose a score function for RCA that requires no causal knowledge.

Statistical and multi-modal methods. BARO (Pham et al., 2024a) uses multivariate change point detection to model the dependency structure of multivariate time series and provides robust statistical scoring, but it does not use causal structure and, like other statistical testing methods, implicitly assumes that root causes exhibit the largest distributional shifts. Microservice data are also inherently multi-modal, comprising metrics, logs, and traces. Zheng et al. (2024) propose a unified multi-modal causal structure learning framework that combines language-model log representations with contrastive learning to align modality-invariant and modality-specific information, and Yao et al. (2024) construct weighted event causal graphs from multi-modal observations to improve interpretability and alignment with domain knowledge.

Positioning of our work. In contrast to the above, `RCA-DCM` assumes neither causal sufficiency nor linearity and does not rely solely on conditional independence or invariance tests. It requires only a causal graph consistent with the true partial order and uses deep causal models to search for SCMs consistent with both the normal and anomalous datasets, thereby exploiting general distributional constraints in the presence of an arbitrary number of unobserved confounders.

C Assumptions

In this section, we enlist all assumptions considered in our paper. We also repeat them at appropriate places when needed.

Assumption C.1 (Semi-Markovian SCM)**.** The data are generated by an acyclic semi-Markovian SCM; each unobserved confounder affects two observed variables, shown as $V_i \leftrightarrow V_j$ in the ADMG.

Assumption C.2 (Mechanism shift)**.** $\mathcal{M}_n$ and $\mathcal{M}_a$ share the ADMG G^* and $P(\mathbf{U})$, and $f_V^n = f_V^a$ for all $V \notin R^*$. Samples are labeled by regime, $\mathcal{D}_n \sim P_n(\mathbf{V})$ and $\mathcal{D}_a \sim P_a(\mathbf{V})$.

An anomaly changes only the mechanisms of R^*, not the graph or the latent distribution.

Assumption C.3 (Graph access)**.** We are given the true ADMG G^* or partial knowledge about it.

Section 4.3 relaxes this to any $G' \supseteq G^*$, which can be built from the true partial order alone.

Assumption C.4 (Identifiability)**.** $\texttt{SolRC}(G^*) = \{R : R^* \subseteq R \subseteq \mathbf{V}\}$.

Assumption C.5 (Latent invariance)**.** The normal and anomalous SCMs share the same exogenous distribution $P(\mathbf{U})$; only the structural functions of the root causes R^* differ between $\mathcal{M}_n^*$ and $\mathcal{M}_a^*$.

No smaller set of mechanism changes explains the shift; this is used only for completeness.

Assumption C.6 (Root-cause identifiability on the true graph)**.** $\texttt{SolRC}(G^*)$ has a unique inclusion-minimal element.necessarily R^*.

D Possible Root Cause set

Lemma 4.8. *$\texttt{SolRC}(G)$ is upward closed: if $R \in \texttt{SolRC}(G)$ and $R \subseteq R' \subseteq \mathbf{V}$, then $R' \in \texttt{SolRC}(G)$.*

Proof. Let P realize (P_n, P_a) on $G \cup F(R)$. Build P' on $G \cup F(R')$ with the same mechanisms, except that each newly targeted node $V_i \in R' \setminus R$ is given the F-augmented mechanism that ignores F and equals its mechanism in P. The conditional distributions in both regimes are unchanged, so P' realizes (P_n, P_a) and lies in $\mathcal{M}\big(G \cup F(R')\big)$. □

Lemma D.1 (When is the root-cause number finite)**.** *Let G^* and G' be the true and mis-specified graphs, either by adding or by removing edges. Given, (P_n, P_a) realized by (G^*, R^*), we have $\texttt{SolRC}(G') \neq \emptyset$ if and only if $P_n \in \mathcal{M}(G')$ and $P_a \in \mathcal{M}(G')$.*

Proof. (If) $\texttt{SolRC}(G') \neq \emptyset$ implies that there exists some mechanism set R that can be shifted to change from normal distribution P_n to anomalous distribution P_a. By Lemma 4.8, $R \subseteq$

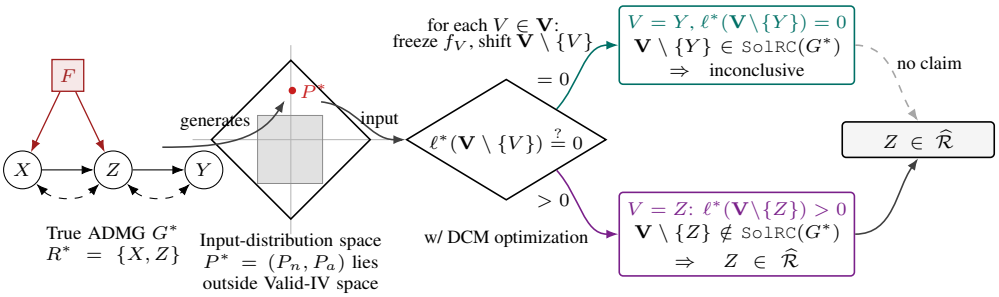


Figure 5: **Workflow example.** In the true augmented graph the fault shifts X and Z, so $R^* = \{X, Z\}$, and latent confounders act on both $X \leftrightarrow Z$ and $Z \leftrightarrow Y$. The induced pair $P^* = (P_n^*, P_a^*)$ falls outside the shaded Valid-IV region of the input space: it violates the instrumental inequalities for the effect of X on Z, which rules out F being excluded from Z (exclusion condition) and hence forces the edge $F \to Z$. Each candidate V is tested by freezing f_V while the mechanisms on $\mathbf{V} \setminus \{V\}$ shift freely. The violation makes $\mathbf{V} \setminus \{Z\}$ infeasible, so $\ell^*(\mathbf{V} \setminus \{Z\}) > 0$ certifies that Z lies in every $\mathtt{R} \in \texttt{SolRC}(G^*)$ and hence in $\widehat{\mathcal{R}}$. For $V = Y$ the score attains zero, which is inconclusive: it is consistent with $\mathbf{V} \setminus \{Y\} \in \texttt{SolRC}(G^*)$, and we draw no conclusion from it.

$\mathbf{V}$(the full set) $\in \texttt{SolRC}(G')$. Thus, $F = 0$ with $\{f_V^n\}_{\forall V \in \mathbf{V}}$, gives a distribution P_n generated by G', hence $P_n \in \mathcal{M}(G')$. Similarly, $F = 1$ with $\{f_V^a\}_{\forall V \in \mathbf{V}}$, gives a distribution P_a generated by G', hence $P_a \in \mathcal{M}(G')$. Precisely, whether the mis-specified graph G' is sparser or denser compared to G^*, G' can realize P_n, and by shifting all mechanisms: $\{f_V^n \to f_V^a\}_{\forall V \in \mathbf{V}}$, can change P_n to P_a.

(Only If) Conversely, suppose $P_n, P_a \in \mathcal{M}(G')$: the normal and anomalous distribution generated by the true graph G^* is realizable by the mis-specified graph G'. We need to prove $\texttt{SolRC}(G') \neq \emptyset$. Realize P_n by G with some latent distribution ν_n and mechanisms f^n, and P_a with latent distribution ν_a and mechanisms f^a.

We can construct a causal model by taking the shared latent to be the independent product $L = (L^n, L^a) \sim \nu_n \times \nu_a$. Let use consider every node as a root cause ($R = \mathbf{V}$).

When $F = 0$, let each visible mechanism read the L^n-component and apply f^n for all $\{f_{V \in \mathbf{V}}\}$, and when $F = 1$, read the L^a-component and apply f^a, for all $\{f_{V \in \mathbf{V}}\}$.

This is a valid element of $\mathcal{M}\big(G \cup F(\mathbf{V})\big)$ and realizes (P_n, P_a), so $\mathbf{V} \in \texttt{SolRC}(G)$. □

Corollary D.2. *$\texttt{minrc}(G) < \infty$ if and only if $P_n \in \mathcal{M}(G)$ and $P_a \in \mathcal{M}(G)$.*

Lemma 4.12. *Suppose $\mathcal{D}_n, \mathcal{D}_a$ are generated by $(\mathcal{M}_n^*, \mathcal{M}_a^*)$ with common ADMG G^*, and let latent invariance hold. For any $Y \in \mathbf{V}$, if $\ell^*(G^*, \mathbf{V} \setminus \{Y\}) > 0$ then $\Rightarrow$ $\mathbf{Z} \setminus \{Y\} \notin \texttt{SolRC}(G^*, P_n^*, P_a^*)$ for any $\mathbf{Z} \subseteq \mathbf{V}$.*

Proof. We first show that $\mathbf{V} \setminus \{Y\} \in \texttt{SolRC}(G^*)$ implies $\ell^*(G^*, \mathbf{V} \setminus \{Y\}) = 0$. Let $P \in \mathcal{M}\big(G^* \cup F(\mathbf{V} \setminus \{Y\})\big)$ realize (P_n^*, P_a^*), with structural equations $V_i = g_i(\mathrm{pa}_{G^*}(V_i), F, U_i)$, where g_i depends on F only if $V_i \neq Y$, and the latents $\mathbf{U}$ are independent of F because F is a root. Setting $\widehat{f}_i^n(\cdot) = g_i(\cdot, 0, \cdot)$ and $\widehat{f}_i^a(\cdot) = g_i(\cdot, 1, \cdot)$ with the common latent distribution $P(\mathbf{U})$ gives an SCM pair on G^* with $\widehat{f}_Y^n = \widehat{f}_Y^a$, i.e., an element of $\widehat{M}_{na}(G^*, \mathbf{V} \setminus \{Y\})$. Since $\mathbf{U} \perp\!\!\!\perp F$, it induces $P(\mathbf{V} \mid F{=}0) = P_n^*$ and $P(\mathbf{V} \mid F{=}1) = P_a^*$, so both discrepancies vanish and $\ell^*(G^*, \mathbf{V} \setminus \{Y\}) = 0$.

By contraposition, $\ell^*(G^*, \mathbf{V} \setminus \{Y\}) > 0$ implies $\mathbf{V} \setminus \{Y\} \notin \texttt{SolRC}(G^*)$. Finally, for any $\mathbf{Z} \subseteq \mathbf{V}$, if $\mathbf{Z} \setminus \{Y\} \in \texttt{SolRC}(G^*)$, then, since $\mathbf{Z} \setminus \{Y\} \subseteq \mathbf{V} \setminus \{Y\}$, Lemma 4.8 would give $\mathbf{V} \setminus \{Y\} \in \texttt{SolRC}(G^*)$, a contradiction. Hence $\mathbf{Z} \setminus \{Y\} \notin \texttt{SolRC}(G^*)$. □

Theorem 4.13. *For any $Y \in \mathbf{V}$, if $\ell^*(G^*, \mathbf{V} \setminus \{Y\}) > 0$ then $Y \in R^*$.*

Proof. By latent invariance, the true pair $(\mathcal{M}_n^*, \mathcal{M}_a^*)$ shares $P(\mathbf{U})$ and differs only in the mechanisms of R^*, so it realizes (P_n^*, P_a^*) on $G^* \cup F(R^*)$ (each $V_i \in R^*$ uses its normal or anomalous mechanism according to F, and every other mechanism ignores F). Hence $R^* \in \texttt{SolRC}(G^*)$. Applying

Lemma 4.12 with $\mathbf{Z} = R^*$ gives $R^* \setminus \{Y\} \notin \texttt{SolRC}(G^*)$. Therefore $R^* \setminus \{Y\} \neq R^*$, i.e., $Y \in R^*$. □

Proposition 4.14 (Soundness)**.** $\widehat{R} = \bigcap \mathit{SolRC}(G^*) \subseteq R^*$

Proof. The true pair $(\mathcal{M}_n^*, \mathcal{M}_a^*)$ realizes (P_n, P_a) while holding every mechanism outside R^* invariant, so $R^* \in \texttt{SolRC}(G^*)$. The intersection of a family is contained in each of its members. □

Assumption D.3. $\texttt{SolRC}(G^*)$ has a unique inclusion-minimal element.

Corollary D.4. *With Assumption D.3,* $\widehat{R} = \bigcap \mathit{SolRC}(G^*) = \arg\min_{R \in \mathit{SolRC}(G^*)} |R| \subseteq R^*$.

Proof. Let m be the unique inclusion-minimal element of $\texttt{SolRC}(G^*)$ (Assumption D.3). Since $\texttt{SolRC}(G^*) \subseteq 2^{\mathbf{V}}$ is finite, every $R \in \texttt{SolRC}(G^*)$ contains an inclusion-minimal element, which must be m; hence $m \subseteq R$ for all $R \in \texttt{SolRC}(G^*)$. Therefore $\bigcap \texttt{SolRC}(G^*) = m$, and $|R| \geq |m|$ for every member, with equality only if $R = m$, so m is the unique minimizer of $|R|$. Finally, $R^* \in \texttt{SolRC}(G^*)$ (Proposition 4.14), so $m \subseteq R^*$. □

E Error analysis in RCA-DCM ranking

Note, Theorem 4.13 says that if $\ell(G^*, \mathbf{V} \setminus \{Y\}) > 0$ then Y is a root cause. However, in practice due to low sample size (with perfect model training), in all cases we find $\ell(.) > 0$. Here, we show that under a condition, in such cases, `RCA-DCM` ranks the true root cause at the top. For any candidate $Y \in \mathbf{V}$ and let $s(Y) = \ell_m(Y) - \ell^*(Y)$ denote the *statistical error* originated due to a low sample size since the loss calculated might not represent the population value. This term is two-sided: the finite set of samples might produce a lower loss, $\ell_m(Y) \in [\ell(Y) - \varepsilon_{\text{stat}}, \ell^*(Y)]$ or a higher loss $\ell_m(Y) \in [\ell^*(Y), \ell(Y) + \varepsilon_{\text{stat}}]$. We assume the empirical loss concentrates around its population value uniformly over candidates: for every $\delta \in (0, 1)$ there is an $\varepsilon_{\text{stat}} > 0$, decreasing in the sample size, such that the event $E = \{\, |s(Y)| \leq \varepsilon_{\text{stat}} \;\; \forall Y \in \mathbf{V} \,\}$ satisfies $\Pr(E) \geq 1 - \delta$.

Proposition E.1 (Correctness of Ranking)**.** *Suppose our model training is perfect, if*

$$\min_{Y \in R^*} \ell^*(Y) > 2\varepsilon_{\text{stat}}$$

then, with probability at least $1 - \delta$*,*

$$\min_{Y \in R^*} \widehat{\ell}(Y) > \max_{Z \notin R^*} \widehat{\ell}(Z),$$

Intuitively, every root cause is scored strictly above every non-root cause, placing the $k = |R^*|$ root causes in its top k positions in the ranking returned by `RCA-DCM`.

Proof. We can construct a bound for the empirical loss as below.

$$\ell^*(Y) - \varepsilon_{\text{stat}} \leq \widehat{\ell}(Y) \leq \ell^*(Y) + \varepsilon_{\text{stat}} \tag{3}$$

Consider a non-root cause variable Z and its loss such that

$$\begin{aligned} Z &= argmax_{V \notin R^*} \widehat{\ell}(V) \\ \widehat{\ell}(Z) &\leq \ell^*(Z) + \varepsilon_{\text{stat}} \\ \widehat{\ell}(Z) &\leq \varepsilon_{\text{stat}} \end{aligned} \tag{4}$$

Here, $\ell^*(Z) = 0$ for non-root cause variables.

Similarly, we consider a root cause variable Y and its loss such that

$$\begin{aligned} Y &= min_{Y \in R^*} \widehat{\ell}(Y) \\ \widehat{\ell}(Y) &\geq \ell^*(Y) - \varepsilon_{\text{stat}} \end{aligned} \tag{5}$$

We subtract $\widehat{\ell}(Z)$ from both sides of Equation 5.

$$\begin{aligned}\widehat{\ell}(Y) - \widehat{\ell}(Z) &\geq \ell^*(Y) - \varepsilon_{\text{stat}} - \widehat{\ell}(Z) \\ \widehat{\ell}(Y) - \widehat{\ell}(Z) &\geq \ell^*(Y) - \varepsilon_{\text{stat}} - \varepsilon_{\text{stat}} \\ \widehat{\ell}(Y) - \widehat{\ell}(Z) &\geq \ell^*(Y) - 2\varepsilon_{\text{stat}}\end{aligned} \tag{6}$$

Thus, if $\ell^*(Y) - 2\varepsilon_{\text{stat}} > 0 \implies \ell^*(Y) > 2\varepsilon_{\text{stat}}$ then $\widehat{\ell}(Y) - \widehat{\ell}(Z) > 0$ which implies that we will obtain the k root causes in the top k of the output ranking. □

F GRAPH MIS-SPECIFICATION

Lemma 4.17. *If $G_1 \subseteq G_2$, then $\texttt{SolRC}(G_1) \subseteq \texttt{SolRC}(G_2)$.*

Proof. Fix any $R \in \texttt{SolRC}(G_1)$. According to Lemma D.1, $R \in \texttt{SolRC}(G_1)$ implies that $P_n = P(\mathbf{V} \mid F = 0) \in \mathcal{M}(G_1)$ and $P_a = P(\mathbf{V} \mid F = 1) \in \mathcal{M}(G_1)$. Since we can shift $P_n \to P_a$ with root causes R, we have $P(\mathbf{V}, F) \in \mathcal{M}\big(G_1 \cup F(R)\big)$.

The augmented graphs $G_1 \cup F(R)$ and $G_2 \cup F(R)$ have identical F-edges, and $G_1 \subseteq G_2$ gives directed-edge and simplicial-complex inclusion, so $G_2 \cup F(R)$ structurally dominates $G_1 \cup F(R)$ in the sense of Definition 4.4. By Lemma 4.6 (structural dominance implies observational dominance),

$$\mathcal{M}\big(G_1 \cup F(R)\big) \subseteq \mathcal{M}\big(G_2 \cup F(R)\big)$$

at every cardinality, in particular at the cardinalities of the data.

Thus $P(\mathbf{V}, F)$ belongs to $\mathcal{M}\big(G_2 \cup F(R)\big)$ (right side) as well. This implies that $P_n = P(\mathbf{V} \mid F = 0) \in \mathcal{M}(G_2)$ and $P_a = P(\mathbf{V} \mid F = 1) \in \mathcal{M}(G_2)$, and the shift $P_n \to P_a$ is possible with root causes R in the causal model constructed with the denser graph G_2. Therefore, according to Lemma D.1, $R \in \texttt{SolRC}(G_2)$.

□

Lemma 4.18. *If $G_1 \subseteq G_2$, then $\texttt{minrc}(G_2, P_n, P_a) \leq \texttt{minrc}(G_1, P_n, P_a)$.*

Proof. By Lemma 4.17, $\texttt{SolRC}(G_1) \subseteq \texttt{SolRC}(G_2)$. Minimizing $|R|$ over the larger family $\texttt{SolRC}(G_2)$ can only decrease the minimum: $\texttt{minrc}(G_2, P_n, P_a) = \min_{R \in \texttt{SolRC}(G_2)} |R| \leq \min_{R \in \texttt{SolRC}(G_1)} |R| = \texttt{minrc}(G_1, P_n, P_a)$. (The inequality holds in $\{0, \ldots, n\} \cup \{\infty\}$, with the convention $\min \emptyset = \infty$.) □

Theorem F.1 (root-cause number under graph misspecification)**.** *Let the observed tuple (P_n, P_a) be generated by the true model (G^*, R^*), so $\texttt{minrc}(G^*) \leq |R^*| < \infty$.*

- *(i)* ***(Over-specification is parsimony-safe.)*** *For any mis-specified graph $G' \supseteq G^*$, $\texttt{minrc}(G') \leq \texttt{minrc}(G^*)$.*
- *(ii)* ***(Under-specification is conservative.)*** *For any mis-specified graph $G' \subseteq G^*$, $\texttt{minrc}(G') \geq \texttt{minrc}(G^*)$. Moreover, if $\texttt{minrc}(G') < \infty$, then $\forall R \in \texttt{SolRC}(G')$, we have $|R| \geq \texttt{minrc}(G^*)$.*

Proof. Both parts are instances of Lemma 4.18. For (i), take $G_D = G' \supseteq G_S = G^*$ to get $\texttt{minrc}(G') \leq \texttt{minrc}(G^*)$.

For (ii), take $G_S = G' \subseteq G_D = G^*$ to get $\texttt{minrc}(G') \geq \texttt{minrc}(G^*)$. Therefore, $\min_{\texttt{SolRC}(G')} |R| \geq \texttt{minrc}(G^*) \implies \forall R \in \texttt{SolRC}(G'), |R| \geq \texttt{minrc}(G^*)$ □

Proposition F.2 (Indispensable root causes, monotone)**.** *If $G_S \subseteq G_D$ then $\bigcap \texttt{SolRC}(G_D) \subseteq \bigcap \texttt{SolRC}(G_S)$.*

Proof. By Lemma 4.17, $\texttt{SolRC}(G_S) \subseteq \texttt{SolRC}(G_D)$. The intersection of a smaller family is larger: $\bigcap \texttt{SolRC}(G_D) \subseteq \bigcap \texttt{SolRC}(G_S)$. □

Corollary F.3 (Faithful inclusion)**.** *Under Assumption C.6, for any $G_S \subseteq G^*$ with $\{P_n, P_a\} \subseteq \mathcal{M}(G_S)$, every inclusion-minimal element of $\texttt{SolRC}(G_S)$ contains R^*. That is, the root-cause set reported on the sparser graph is a superset of the true root causes:*

$$\widehat{R} \supseteq R^* \quad \textit{for every minimal } \widehat{R} \in \texttt{SolRC}(G_S).$$

Proof. A finite upward-closed family with a unique minimal element m consists of all supersets of m, so its intersection is m. By Assumption C.6, $\bigcap \texttt{SolRC}(G^*) = R^*$. By Proposition F.2, $R^* = \bigcap \texttt{SolRC}(G^*) \subseteq \bigcap \texttt{SolRC}(G_S)$. The intersection is contained in every member of $\texttt{SolRC}(G_S)$, in particular in every minimal one. □

Corollary F.4. *Under Assumption C.6, with $\{P_n, P_a\} \subseteq \mathcal{M}(G')$:*

(i) for any $G' \supseteq G^*$, there exists an inclusion-minimal $\widehat{R} \in \texttt{SolRC}(G')$ with $\widehat{R} \subseteq R^*$;

(ii) for any $G' \subseteq G^*$, every inclusion-minimal $\widehat{R} \in \texttt{SolRC}(G')$ satisfies $\widehat{R} \supseteq R^*$.

Proof. i) By Assumption C.6, $\bigcap \texttt{SolRC}(G^*) = R^*$. Since $G_D \supseteq G^*$, by Proposition F.2, $\bigcap \texttt{SolRC}(G_D) \subseteq \bigcap \texttt{SolRC}(G^*) = R^*$.

ii) By Lemma 4.17, $R^* \in \texttt{SolRC}(G^*) \subseteq \texttt{SolRC}(G_D)$, so $R^* \in \texttt{SolRC}(G_D)$. The family $\mathcal{F}_{R^*} := \{S \in \texttt{SolRC}(G_D) : S \subseteq R^*\}$ is nonempty and finite, hence has a $\subseteq$-minimal element $\widehat{R}$. If some $S \in \texttt{SolRC}(G_D)$ satisfied $S \subsetneq \widehat{R}$, then $S \subseteq R^*$ as well, contradicting minimality of $\widehat{R}$ in $\mathcal{F}_{R^*}$. So $\widehat{R}$ is inclusion-minimal in all of $\texttt{SolRC}(G_D)$. □

By Lemma 4.17 and 4.18, if the mis-specified graph G' is denser compared to the true graph G^*, i.e. $G_1 = G^*, G_2 = G'$ then the predicted root cause set will be a subset of the original root cause set. Whereas, if G' is sparse compared to the true graph G^*, the predicted root cause set will be a superset of the original root cause set. We formalize this below.

G Experiment Details

G.1 Training Details

For all continuous experiments we use a flow-based DCM (`dcm_flow`), in which each node is modeled as a conditional normalizing flow given its parents (and its latents, when present) and trained by maximum likelihood. For each candidate we fit two DCMs, one on the normal and one on the anomalous data. Under the shared-update mode used in all our runs we optimize

$$L = L_{\text{anomalous}} + L_{\text{normal}},$$

with equal (unweighted) terms, matching Algorithm 1. The candidate mechanism may be shared across the two models (`trn.upd=shared`), so that only the non-shared components absorb regime-specific change. All models are trained with Adam at a learning rate of 5×10^{-4}, batch size 256, for 50 epochs, using a hidden dimension of 32 and a latent noise dimension of 10. Normal and anomalous features are jointly normalized before training, and candidate node tests are executed in parallel with up to 10 workers (`thr_num`). After training, we rank candidates by a discrepancy between model samples and observed data; primarily the Wasserstein distance for continuous experiments, with MMD and median shift additionally logged as diagnostics;and take the top-k set for PRR. Full defaults are given in `dcm/conf.yaml` and are overridable via Hydra; the supplementary archive contains the training code (`dcm/train.py`) along with configs, training scripts, and experiment runners.

We use the following metrics for evaluating output rankings.

$$T@k \;=\; \frac{1}{|\mathbf{E}|} \sum_{e \in \mathbf{E}} \frac{\left|\widehat{R}_e[1,..,k] \cap R_e^*\right|}{\min(k,\, |R_e^*|)}, \qquad \text{PRR} \;=\; \frac{1}{|\mathbf{E}|} \sum_{e \in \mathbf{E}} \mathbf{1}[\,T@|R_e^*| = 1\,] \tag{7}$$

Causal graph: The causal structural knowledge can be obtained from domain experts or, as recent work shows, through *inexpensive* LLM-based causal discovery Kıcıman et al. (2024); Jiralerspong et al. (2024); Long et al. (2023). In real-world setups such as microservice and agentic frameworks, the call graph or system architecture can also be extracted directly. Thus, we do not consider this requirement a major limitation. Nonetheless, `RCA-DCM` outperforms the baselines in synthetic experiments with both the true graph and a dense graph, and in microservice experiments where no graph is known and we use a sink graph (all variables point to the target node).

Sample size: We evaluate `RCA-DCM` across a range of sample sizes, and it performs consistently well: in synthetic experiments ($N = 5$k), in the Causal Chamber ($N = 1$k), and in microservice setups ($N \sim 700$). More extensive experiments are in Appendix G.3.

Runtime: Our proposed method is computationally more expensive than some baselines (e.g., a single root-cause test takes 0.4 s for RCD vs. 41 s for ours). However, it outperforms them in almost all cases. Thus, our approach targets applications that prioritize reliability under latent confounding over faster runtime.

G.2 Simulated Datasets

We follow (Yang et al., 2024) to generate synthetic normal and anomalous data from randomly generated graphs, varying the total variable count, latent proportion, and edge density. We sample a random DAG over p observed variables plus m latent confounders, observed-to-observed edges follow a random ancestral density, and each latent connects to a small set of observed children. Only observed variables are written to the dataset and latents are never revealed to the algorithms. The causal mechanism of each variable combines linear structural mixing with a per-variable invertible non-linearity. Given exogenous noises, observations are produced by the nonlinear SEM whose graph and mixing structure are shared across both regimes.

In these experiments, we average results over 100 trials and for each trial we draw 5000 *normal* samples with baseline noise scales drawn uniformly from $[a_{\min}, a_{\max}]$ and no interventions, and 5000 *anomalous* samples in which a subset of *observed* variables is selected as root causes; for those targets only, the noise scale is redrawn from a stronger range $[b_{\min}, b_{\max}]$, while non-targets retain their baseline scales. Sample sizes vary across experiments and are stated per experiment. Root causes are therefore noise-scale mechanism shifts, and the intervened observed nodes are the ground-truth root causes. Intervention strengths can optionally be made heterogeneous via a per-variable weight schedule. Methods see only the observed tables and, where applicable, the graph.

Independent exogenous noise $\varepsilon \in \mathbb{R}^{n_L+n_X}$ concatenates latent exogenous shocks $\ell \in \mathbb{R}^{n_L}$ and observed exogenous disturbances $\varepsilon^{\mathsf{X}} \in \mathbb{R}^{n_X}$. A sparse mask G fixes which latent/observed parents may influence each observed sensor; we draw a random column-ℓ_2-normalized coefficient matrix, apply G, and split the observed rows as $\bar{A} = [B_{\text{raw}} \mid C]$ with $B_{\text{raw}} \in \mathbb{R}^{n_X \times n_L}$ and $C \in \mathbb{R}^{n_X \times n_X}$. We scale all latent-to-observed edges by a nonnegative confounding strength λ, $B = \lambda B_{\text{raw}}$, leaving the observed-to-observed support unchanged. Observations are

$$x \;=\; (I_{n_X} - C)^{-1}\big(B\ell + \varepsilon^{\mathsf{X}}\big), \tag{8}$$

i.e. latents and local observed shocks propagate through the directed subgraph on sensors; larger λ increases hidden common-cause leakage into measurements. The default nonlinear generator applies two mixing layers with leaky-ReLU or tanh nonlinearities and randomly drawn mixing weights of controlled condition number.

Regimes and root causes. Normal data uses baseline noise scales drawn uniformly from $[a_{\min}, a_{\max}]$, e.g. $[0, 8]$. For anomalous data we select $\lfloor p/2 \rfloor$ of the p observed nodes as root causes and redraw their noise scales from a stronger range $[b_{\min}, b_{\max}]$, e.g. $[2, 12]$, leaving non-targets at their baseline scales. With $p = 6$ this gives 3 root causes. We draw $N_{\text{normal}} = N_{\text{anomalous}} = 5000$ i.i.d. samples from the same SCM family; other experiments use different sample sizes, reported in their respective sections. Exact hyperparameters are given in the supplementary scripts.

Exp 1: Nonlinear model with unobserved confounders: We use graphs with $p = 6$ observed and $m = 4$ latent variables (total $n = 10$) and vary the confounding strength $\lambda \in \{3, 10\}$, which uniformly scales all latent-to-observed coefficients; a larger λ makes the hidden confounders contribute more strongly to the observed variables. We take $\lfloor p/2 \rfloor = 3$ observed variables as ground-truth root causes.

Observation: Figure 4(a) reports PRR for $\lambda \in \{3, 10\}$. RCA−DCM attains 99% and 84%, ahead of every baseline in both settings: BARO (91%, 65%), NSigma (86%, 62%), RCG (85%, 39%), RCD (74%, 59%), and CIRCA (62%, 38%). All methods degrade as confounding strengthens: RCA−DCM starting at 99% and RCD at 74%, both drop by 15 percentage points, while BARO, NSigma, CIRCA, and RCG drop by 26, 24, 24, and 46 points, respectively. This suggests that RCA−DCM is more robust to strong confounding and better preserves root-cause recovery than the baselines. The baselines fail because a non-root-cause descendant that shares a latent confounder with a root cause inherits a shift that persists even after conditioning on its observed parents, making it indistinguishable from a true root cause unless the latents are explicitly accounted for.

Exp 2: Nonlinear model with heterogeneous anomalies: Baselines that perform RCA mainly by measuring marginal shifts implicitly assume a homogeneous anomaly, in which the true root causes exhibit the largest distributional shifts among all variables. This assumption often fails with multiple root causes: a shift originating at a root cause R_1 propagates to its descendants, so a downstream non-root-cause descendant can end up with a larger observed shift than another root cause R_2, which creates a failure case for these baselines.

We construct such heterogeneous anomalies by scaling the injected shift magnitude linearly with topological depth, so that shifts accumulate along directed paths and a downstream descendant out-shifts a genuine upstream root cause. In the generated data, this trap occurs (i.e., some non-root cause exhibits a larger marginal shift than some true root cause) in roughly 60% of trials at $p = 10$. We follow the same random graph generation process as in the previous experiment, but with no latent confounders and with $\lfloor p/2 \rfloor$ of the observed variables as root causes. We fix 5000 normal and 5000 anomalous samples and vary the number of observed variables $n \in \{5, 10\}$ ($n = p$ as $m = 0$).

Observation: Figure 4(b) shows PRR as the system grows. RCA−DCM achieves the highest rate at both sizes, 98% at $n = 5$ and 54% at $n = 10$, outperforming all five baselines in both settings. The marginal-shift methods (BARO and NSigma) degrade the most, falling to 14% at $n = 10$; this is the failure mode the construction targets, since it makes ranking by shift magnitude wrong in most trials. The graph-based methods perform better (RCG 36%, CIRCA 34%, RCD 28%) because they propagate evidence along edges rather than scoring each node in isolation, so a descendant's large shift can be partly explained by its parents. However, they also degrade as the system grows, since their conditional tests involve larger conditioning sets and become less reliable at a fixed sample size. These results suggest that RCA−DCM is more robust to increasing system complexity, where the effects of multiple root causes can compound.

Exp 3: Robustness to graph misspecification. We next examine the sensitivity of RCA−DCM to errors in the supplied graph. On the same 100 datasets used for $\lambda = 10$ in **Exp 1**, we perturb *only* the graph provided to RCA−DCM, so that any change in performance is attributable solely to the graph error. We compare three perturbations against the true-graph control (PRR of 88% and 84% in two independent runs): adding 50% spurious directed and bidirected edges (PRR $= 84\%$), deleting 50% of the bidirected edges (PRR $= 87\%$), and deleting 50% of both directed and bidirected edges (PRR $= 76\%$). Under a paired McNemar test, neither edge addition ($p = 0.39$) nor bidirected-edge deletion ($p = 1.00$) changes accuracy significantly; only the perturbation that also deletes directed edges degrades it ($p = 0.019$), and even then RCA−DCM retains a PRR of 76%, above every baseline at $\lambda = 10$ (best: BARO, 65%). These results are consistent with our analysis in Section 4.3, which covers misspecified graphs that can still represent the observed normal and anomalous distributions, i.e., $\{P_n, P_a\} \subseteq \mathcal{M}(G')$. Adding edges always preserves this condition, and in our instances deleting bidirected edges largely preserves it as well; in both cases, RCA−DCM performs on par with the control. Deleting directed edges, in contrast, typically takes G' outside this class: each conditional $P(V \mid \mathrm{pa}(V))$ must be fit against an incomplete parent set, so the shift propagated along a missing edge cannot be explained away, and non-root-cause variables can appear as root causes.

G.3 Sensitivity analysis of unequal sample sizes

Algorithm 1 trains two DCMs, one on the normal data and one on the anomalous data, each fit solely to its own sample. An implicit unequal weighting between the two loss terms is therefore not the main risk that sample-size imbalance introduces; what matters is whether each DCM individually has enough data to learn its own SCM.

Suppose r samples are sufficient to learn an SCM with neural networks. Let N_s and N_a denote the normal and anomalous sample sizes, and suppose we are testing whether variable Y is a root cause. With equal weighting, the DCM objective searches for a mechanism f_Y consistent with both the N_s normal and the N_a anomalous samples; if the loss does not reach zero, no such mechanism exists, so Y must be a root cause. However, if $N_a \leq r$, the anomalous samples are insufficient to learn the SCM reliably, which violates the universal-approximation assumption underlying our test, and the algorithm may give incorrect results. We study this empirically below.

Exp 6. Setup. We generate synthetic causal systems with 8 observed variables and 4 unobserved confounders. Graphs are moderately dense ($\sim 80\%$ of possible ancestral edges), with interventions on $\sim 40\%$ of variables (3 root causes) at uniformly drawn strengths. Latent-to-observed edges are moderately strong (scale 2), and the maximum normal-regime noise scale is 8. Models are trained for 50 epochs with equal weights on the normal and anomalous losses, as in Algorithm 1, ranking candidates by Wasserstein discrepancy. We report the *perfect recovery rate* (PRR): an instance counts as a success only when the top-k predictions exactly recover the k true root causes (the same top-k rule is applied to BARO and RCD).

Table 2: Perfect recovery rate (PRR) under varying normal (N_s) and anomalous (N_a) sample sizes.

N_s	N_a	DCM	BARO	RCD	#Instances
1000	1000	**0.63**	0.39	0.32	78
1000	500	**0.58**	0.48	0.48	95
500	500	0.47	0.47	0.31	51
1000	100	0.27	0.35	0.24	131
100	100	0.23	0.24	0.02	62

Findings. (i) Under moderate imbalance ($N_s = 1000$, $N_a = 500$), equal weighting remains effective: DCM PRR stays close to the balanced setting (0.58 vs. 0.63) and still matches or exceeds the baselines. (ii) When the anomalous sample is much smaller ($N_a = 100$, $N_s = 1000$, a $10:1$ ratio), DCM PRR drops sharply ($0.63 \rightarrow 0.27$) and BARO becomes competitive or better. (iii) Balanced small samples ($100/100$) are hard for all methods; scaling N_s and N_a together helps more than inflating N_s alone.

G.4 Causal Chamber (Real-World Physical Testbed)

Exp 4: We evaluate `RCA-DCM` on the Causal Chambers benchmark (Gamella et al., 2025), a real-world physical testbed. We use the light-tunnel chamber in its standard configuration, which comes with a ground-truth DAG over 38 variables and 57 edges. Each experiment applies a single-target intervention of varying strength (strong, mid, or weak) to one of the tunnel's manipulable variables, shifting the target's distribution relative to the `uniform_reference` experiment. We use the reference experiment as the normal dataset (10k samples) and each intervention experiment as an anomalous dataset (1k samples), and we drop variables that are constant within an experiment. This yields 52 cases, which we evaluate in two conditions: with the true root cause *observed*, and with it *hidden*, so that it acts as a latent confounder among its former children. Its children then remain dependent through it, but it is no longer available to condition on. In the observed condition, $|R^*| = 1$; in the hidden condition, the hidden cause has between 1 and 7 children, all of which must be recovered as root causes.

Observation: Figure 4(c) reports PRR in both conditions. When the root cause is observed, there is no confounding and every case has a single root cause; here, `RCA-DCM` is competitive but does not lead (88%, against 100% for RCG and 90% for RCD). Hiding the root cause changes the problem in two ways: it introduces a latent confounder, and it turns each of the hidden variable's children into a root cause, so a case may now have several root causes that must all be recovered. Under this harder setting, the ordering reverses: `RCA-DCM` attains 87%, against 73% for the best baseline, and it is the only method whose performance barely changes between conditions (-2 points, against -50 for RCG and -21 for RCD). The gap is largest on the cases with multiple root causes: on these cases alone, `RCA-DCM` attains a PRR of 63%, while CIRCA reaches 25%, RCG 13%, and RCD 0%. RCD

fails here by design: its elimination procedure keeps only one surviving candidate, so it can recover a single root cause but never a set of several. In contrast, RCA−DCM tests each variable separately for whether its own mechanism changed, so all children of the hidden cause are flagged and appear together at the top of the ranking.

G.5 MICROSERVICE DATASETS

Exp 5: We evaluate RCA−DCM on SockShop, a cloud computing dataset recorded from a microservice-based replica of a web application specifically designed for evaluating root cause analysis methods (Pham et al., 2024b). The repository contains performance issues corresponding to five fault types: CPU hog ("cpu"), memory leak ("mem"), disk I/O stress ("disk"), network delay ("delay"), and packet loss ("loss"). Each fault type is injected five times into each of five services (carts, catalogue, orders, payment, and users), yielding 125 anomaly datasets. Each dataset provides roughly 360 normal and 360 anomalous observations over 14 recorded services, which we split into normal and anomalous segments based on the injection time. We assume that *no* causal structure is available to RCA−DCM: instead of a call graph, we supply a *confounded sink*, in which every recorded service points to the service under test and a single shared latent confounds all of them. The confounder is essential. A plain sink graph, with directed edges into the target but no edges among the remaining services, asserts that those services are mutually independent, which is false in these systems, where co-located services share load and infrastructure (in the normal regime, we measure a mean absolute pairwise correlation of 0.11 on SockShop and 0.18 on Online Boutique). Adding the shared latent removes this false independence claim while encoding no structural knowledge beyond "any service may be responsible, and the services may share unobserved common causes." This places RCA−DCM at a disadvantage relative to two of the baselines: CIRCA and RCG are given the true edge-inverted call graph, while NSigma, BARO, and RCD do not use a graph at all.

Exp 6: We repeat the evaluation on Online Boutique (Pham et al., 2024b), a second microservice benchmark with the same five fault types, each injected five times, again yielding 125 anomaly datasets, but over 11 services and with substantially longer traces (roughly 2100 normal and 2100 anomalous observations per case). The graph treatment is identical to Exp 5: RCA−DCM receives only the confounded sink, while CIRCA and RCG receive the true call graph. Because the two benchmarks differ in topology, service count, and trace length, together they test whether a method's behavior transfers across deployments rather than being tuned to one.

Results: Table 3 reports per-fault top-k accuracy. On SockShop, RCA−DCM attains the highest average top-1 accuracy, 0.88, ahead of NSigma (0.75), BARO (0.74), CIRCA (0.65), RCG (0.54), and RCD (0.38); every gap is significant under a paired test ($p < 0.001$). It is best or tied for best on all five fault types, and its errors are near-misses rather than failures: its average top-3 accuracy reaches 0.99. On Online Boutique, RCA−DCM again ranks first, at 0.78, ahead of RCG (0.71), NSigma (0.70), BARO (0.62), CIRCA (0.52), and RCD (0.49), with an average top-3 accuracy of 0.90; the margins over BARO, CIRCA, and RCD are significant, while those over NSigma and RCG are not at this sample size.

Two observations are worth highlighting. First, no baseline is consistently second: NSigma is the strongest competitor on SockShop, whereas RCG rises from fifth place on SockShop to second on Online Boutique and BARO drops from third to fourth. This is consistent with Ikram et al. (2025b), who suggest that BARO may be tuned to SockShop, and with our synthetic and Causal Chamber results, where the ordering among baselines likewise changes with the setting. RCA−DCM is the only method that ranks first on both benchmarks. Second, packet loss is among the hardest fault types for RCA−DCM on both benchmarks (0.76 on SockShop, tied with disk, and 0.44 on Online Boutique), whereas CPU and memory faults are recovered almost perfectly (1.00 and 0.96–1.00). Loss perturbs a service only intermittently, so the induced mechanism shift is small relative to normal traffic variation. On Online Boutique, most baselines also score lowest on this category, which suggests that the difficulty lies largely in the anomaly signal itself rather than in the attribution method.

Notably, RCA−DCM attains these results without any causal structure, while the two graph-based baselines are given the true call graph. Together with the graph misspecification results in Section 5.1, this supports the view that the advantage of RCA−DCM does not depend on access to an accurate structure.

Table 3: Per-fault Top-1/Top-3/Top-5 accuracy. Best value per column within each dataset is in bold.

		CPU			MEM			DISK			DELAY			LOSS			AVERAGE		
Dataset	Method	T1	T3	T5	T1	T3	T5	T1	T3	T5	T1	T3	T5	T1	T3	T5	T1	T3	T5
Sock Shop	**DCM**	**1.00**	**1.00**	**1.00**	**0.96**	**0.96**	**0.96**	**0.76**	**1.00**	**1.00**	**0.92**	**1.00**	**1.00**	**0.76**	**1.00**	**1.00**	**0.88**	**0.99**	**0.99**
	NSigma	**1.00**	**1.00**	**1.00**	**0.96**	**0.96**	**0.96**	0.64	0.96	**1.00**	0.60	0.92	**1.00**	0.56	**1.00**	**1.00**	0.75	0.97	**0.99**
	BARO	**1.00**	**1.00**	**1.00**	0.48	**0.96**	**0.96**	0.68	0.96	**1.00**	**0.92**	**1.00**	**1.00**	0.64	**1.00**	**1.00**	0.74	0.98	**0.99**
	CIRCA	0.96	**1.00**	**1.00**	0.56	**0.96**	**0.96**	0.68	0.84	**1.00**	0.48	**1.00**	**1.00**	0.56	**1.00**	**1.00**	0.65	0.96	**0.99**
	RCD	0.76	0.92	0.96	0.28	0.48	0.48	0.32	0.52	0.52	0.28	0.48	0.52	0.28	0.32	0.44	0.38	0.54	0.58
	RCG	0.32	0.68	0.80	0.24	0.32	0.32	0.68	0.76	0.80	0.72	0.80	0.80	0.72	0.80	0.80	0.54	0.67	0.70
Online Boutique	**DCM**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.64	0.88	**1.00**	**0.80**	**1.00**	**1.00**	**0.44**	0.60	0.88	**0.78**	0.90	**0.98**
	NSigma	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.56	**1.00**	**1.00**	0.52	0.96	**1.00**	**0.44**	0.60	0.80	0.70	**0.91**	0.96
	BARO	**1.00**	**1.00**	**1.00**	0.36	**1.00**	**1.00**	0.64	0.88	**1.00**	0.64	**1.00**	**1.00**	**0.44**	0.64	0.72	0.62	0.90	0.94
	CIRCA	0.88	**1.00**	**1.00**	0.72	**1.00**	**1.00**	0.20	0.84	**1.00**	0.44	0.92	**1.00**	0.36	0.68	**0.92**	0.52	0.89	**0.98**
	RCD	**1.00**	**1.00**	**1.00**	0.96	**1.00**	**1.00**	0.20	0.36	0.40	0.20	0.28	0.28	0.08	0.32	0.40	0.49	0.59	0.62
	RCG	0.96	**1.00**	**1.00**	0.76	0.80	0.88	**0.72**	0.80	0.80	**0.80**	0.80	0.80	0.32	**0.76**	0.76	0.71	0.83	0.85

G.6 COMPUTATIONAL COMPLEXITY

Asymptotic complexity. DCM-RCA trains one DCM over the full causal graph to model the normal distribution and then trains one candidate-specific DCM for each node in the graph to test whether that node is a root cause. Let $|\mathbf{V}|$ be the number of variables, N the number of samples, B the batch size, T the number of training epochs, and c the cost of one training update for a DCM over the full graph. Each epoch performs $\lceil N/B \rceil$ updates, so training one DCM costs $O(T\lceil N/B \rceil c)$, and evaluating all candidate root causes sequentially requires training $|\mathbf{V}|$ additional DCMs, giving total complexity

$$O(|\mathbf{V}|\, T\lceil N/B \rceil c)\,.$$

However, the candidate-specific DCMs are independent once the normal DCM is trained, so they can be executed in parallel. With P parallel workers, the wall-clock complexity becomes approximately

$$O\left(T\lceil N/B \rceil c\left(1+\left\lceil \frac{|\mathbf{V}|}{P} \right\rceil\right)\right),$$

up to parallelization overhead and memory constraints. Thus, although the total computational cost grows linearly with the number of candidate variables, the practical runtime can be substantially reduced through parallel execution: for $T = 50$, $N = 5000$, $B = 256$, $|\mathbf{V}| = 10$ and $P = 15$, the sequential cost is $11{,}000c$ versus $2000c$ in parallel, a $5.5\times$ speedup.

Empirical wall-clock time. Table 4 reports per-dataset runtime for DCM (`dcm_flow`, 50 epochs) against BARO and RCD in the same synthetic setting (12 variables, 4 latents, $N_s = N_a = 1000$).

Table 4: Per-dataset wall-clock time in the synthetic setting (12 variables, 4 latents, $N_s = N_a = 1000$).

Method	Typical time / dataset
BARO	~ 0.01 s
RCD	~ 0.4 s (median)
DCM (ours)	~ 41 s (median)

DCM is slower than RCD by $10^2\times$ and than BARO by 10^3–$10^4\times$, as expected: BARO is a lightweight distributional score, RCD a localized CI-test search, whereas DCM trains a full multi-node, multi-epoch causal model (here with up to 10 parallel workers). DCM runtime scales with sample size (~ 9 s at $N = 100$, ~ 41 s at $N = 1000$, and $\sim$ 2–3 min at $N_s = 2000$, $N_a = 1000$), while BARO and RCD remain sub-second throughout. These baselines, however, cannot reliably handle systems with unobserved confounders, which our algorithm can.

We do not give a matching closed-form complexity for BARO, RCD, and CIRCA: BARO performs a single pass per variable; RCD's own analysis notes that CI-test-based discovery is in general

Algorithm 2 RunDCM$(G, \mathbb{G})$

Input: Causal graph $G = (\mathcal{V}, \mathcal{E})$, DCM $\mathbb{G}$.
$\widehat{\mathbf{v}} \leftarrow \emptyset$ {Initialize generated samples}
$conf \leftarrow \emptyset$ {Initialize confounding noise storage}
for $U \in latent_confounders(G)$ **do**
  $v_1, v_2 = U.ch1, U.ch2$ {Get the two observed children of latent confounder U}
  $z_U \sim p(z)$ {Sample one shared latent noise}
  $conf[v_1] \leftarrow Append(conf[v_1], z_U)$ {Assign shared noise to v_1}
  $conf[v_2] \leftarrow Append(conf[v_2], z_U)$ {Assign shared noise to v_2}
for $V_i \in \mathcal{V}$ in causal graph G topological order **do**
  $par = get_parents(V_i, G)$ {Get observed parents of V_i}
  $exos \sim p(z)$ {Sample exogenous noise for V_i}
  $conf_i = conf[V_i]$ {Collect latent confounding noise for V_i}
  $v_i = \mathbb{G}_{\theta_i}(exos, conf_i, \widehat{\mathbf{v}}_{par})$ {Generate V_i using its Causal Normalizing Flow module}
  $\widehat{\mathbf{v}} \leftarrow Append(\widehat{\mathbf{v}}, v_i)$ {Store generated value}
Return Samples $\widehat{\mathbf{v}}$ or Fail

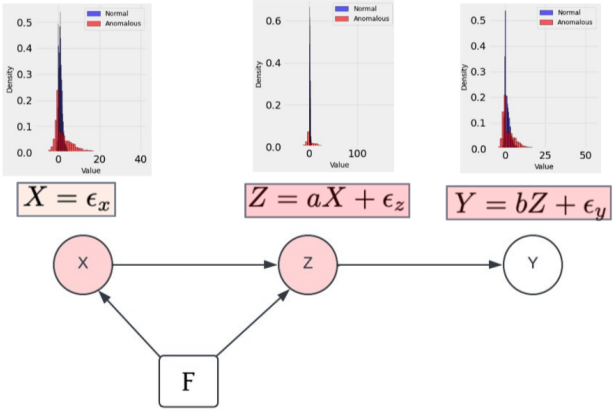


Figure 6: Non-rc Y has higher shift in median than RC X.

exponential in the number of nodes for non-sparse graphs (its localized search targets avoiding this in practice, without a published closed form); and CIRCA's complexity is dominated by its causal-discovery step. We therefore view this as an *accuracy–compute trade-off*: DCM-RCA substantially improves accuracy under latent confounding at the cost of runtime, while the baselines remain preferable when raw speed is the priority. DCM-RCA's cost is justified whenever reliability under unobserved confounding matters.