Who caused Y? Learning causal parents using Independent Risk Minimization

Felix Schur
Seminar for Statistics, ETH Zürich
felix.schur@stat.math.ethz.ch

Sorawit Saengkyongam*
sorawitj@gmail.com

Jonas Peters
Seminar for Statistics, ETH Zürich
jonas.peters@stat.math.ethz.ch

Abstract

We study identifiability and estimation of direct causes, that is, the parents of a designated target Y in a structural causal model, using only observational data. Our main assumption is an additive-noise model for Y: the value of Y equals some function of its parents plus noise that is independent of those parents. We allow for general functional relationships and hidden confounding among all other variables. We prove that under a 'no-backward' additive-noise condition the true parent set, $\mathrm{PA}(Y)$, is identified from the observational distribution by a simple population principle: among all candidate variable sets that make the regression residual of Y independent of the regressors, choose the one with the smallest residual variance, and—if several tie—the smallest set. To justify the 'no-backward' condition, we propose a novel identifiability scheme based on identifiability witnesses: we prove that in finite-dimensional analytic model classes, for which we have an identifiability witness (that is, a single identifiable model), residual independence for sets containing descendants of Y holds only on Lebesgue-null exceptional parameter sets. This scheme is strong enough to recover known identifiability results. For finite data, we propose Independent Risk Minimization (`IndRM`), which offers a simple, local, and non-interventional route to isolating the direct causes of Y in multivariate systems, avoiding global model assumptions and full-graph search.

1 Introduction and Related Work

Causal discovery matters because decisions in genomics, medicine, economics, and climate science hinge on identifying direct drivers of a target Y—for example, reconstructing gene-regulatory interactions from expression data or mapping protein-signaling pathways from single-cell measurements [8, 10, 21, 30]. Yet most existing approaches depend on global assumptions—such as acyclicity, faithfulness (`IC`, `PC`, `FCI`), additive-noise (`RESIT`, `CAM`), equal error variances, or (other) full-graph models that allow for penalized likelihood score optimization [3, 4, 21, 22, 24, 35]—that can be fragile in real-world applications. In contrast to existing work, we pursue a local identifiability principle that targets $\mathrm{PA}(Y)$ from observational data under a local additive-noise condition, avoiding global requirements. More concretely, this paper proposes a simple learning principle called Independent Risk Minimization (`IndRM`) that identifies $\mathrm{PA}(Y)$ from observational data when Y admits an additive-noise mechanism. The principle has three ingredients: (i) *residual independence* $Y - \mathbb{E}[Y \mid S] \perp\!\!\!\perp S$ for valid explanatory sets S, (ii) *variance optimality* among such sets, and

*Most work done while at ETH Zürich. Now at Apple.

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

(iii) *cardinality minimality* to break ties. We show that, under natural graphical and distributional assumptions plus a no-backward-ANM-with-descendants property (which we prove generically holds), the unique variance- and size-optimal residual-independent set is $\mathrm{PA}(Y)$.

Determining the direct causes of a response Y from observational data is a fundamental problem in causal discovery. Constraint-based methods (such as `PC`, `FCI`, `CD-NOD` [15, 35, 36]) recover Markov equivalence classes by exploiting conditional independence implied by the Markov property and faithfulness [26]. Score-based procedures such as `GES` [4] optimize global criteria over DAGs and, when the score is likelihood-equivalent and relies only on Markov and faithfulness information, typically identify a Markov equivalence class rather than a unique DAG. Other score-based methods can go beyond Markov equivalence by adding stronger functional assumptions; for example, `CAM` uses an additive-noise likelihood model to identify a full DAG under suitable conditions [3]. The distinction relevant for our work is not constraint-based versus score-based search per se, but whether the method relies on global graph-wide assumptions or on a target-local condition. The Markov blanket $\mathrm{MB}(Y)$ consists of parents, children, and spouses of Y. We, instead, aim to identify the direct causes of Y, that is, $\mathrm{PA}(Y)$, so identifying the Markov equivalence class is generally insufficient. Constraint-based approaches rely on accurate conditional independence testing. It has been shown that this is possible only under strong assumptions, e.g., on the complexity of the conditional mean functions [33]. Since they are global methods, constraint-based approaches perform many such conditional independence tests; to the best of our knowledge it is unknown whether there are interesting distributions (beyond, for example, the Gaussian distribution) satisfying such assumptions.

A complementary line of work exploits *functional* restrictions, most prominently additive-noise models (ANMs), where a variable is expressed as a function of its causes plus noise that is independent of those causes (and other ancestors). Identifiability in the multivariate case [23] can be deduced from identifiability in a bivariate setting [14], and analogous identifiability results have been established for the more general post-nonlinear model [40]; see also [26]. Linear non-Gaussian models (`LiNGAM`) exploit non-Gaussianity to identify full DAGs in linear settings [34]. `RESIT` [24] uses residual independence of the additive noise in ANMs to estimate the causal graph. The first step of `IndRM` takes advantage of the same residual independence idea as `RESIT`; however, we then use a local procedure to estimate the parent set $\mathrm{PA}(Y)$, while they use global ordering and pruning. Their approach therefore relies on the global ANM assumption, whereas we require local additive noise only.

The potential benefits of local discovery methods have been stressed before. Invariant Causal Prediction (ICP) [13, 17, 25] targets $\mathrm{PA}(Y)$ directly without recovering the full DAG. A key distinction, however, is that ICP requires data from multiple environments to exploit invariance, whereas our approach operates from purely observational data under a single distribution, replacing the multi-environment requirement with the assumption on the additive-noise mechanism of Y. Another line of work estimates the causal neighborhood of a target based on conditional independences (possibly in combination with score-based methods), without recovering the full graph. For example, Y-structure methods [19] can certify an unconfounded edge $X \to Z$ from a specific local conditional-independence pattern, even with latent common causes. More recent local methods such as LOAD [31] use neighborhoods of a treatment-outcome pair to decide effect identifiability and recover adjustment sets. `CMB` [9] and `PCD-by-PCD` [39] learn local parent–child sets and orient edges at the target with local orientation rules, and `HLCD` [18] prunes and orients a constraint-based local skeleton using local scores. These methods may leave some edges at the target unoriented and use conditional independence testing, whereas `IndRM` selects a parent set directly by residual independence, risk, and cardinality. Zheng et al. [41] decide whether a variable is a cause of the target using supplied direct, ancestral, and non-ancestral background knowledge, an input that `IndRM` does not use. There is also work on local causal discovery for a target variable under structural restrictions on the target mechanism, defining an identifiable set as the intersection of all nonempty structurally plausible parent sets, which is a subset of $\mathrm{PA}(Y)$ under their assumptions [1]. They also propose a separate score-based estimator of the full parent set, combining residual independence, covariate significance, and, where applicable, fit to a specified noise distribution. In contrast, we focus on the additive-noise target model and use residual variance, followed by cardinality, to select among residual-independent candidate sets, which allows us to identify the full parent set. Their consistency result assumes a unique structurally plausible parent set [1, Proposition 2], whereas Theorem 3.3 identifies $\mathrm{PA}(Y)$ under our assumptions even when several candidate sets are residual-independent.

The plausibility of ANM-based identifiability has been argued for via algorithmic information theory: a backward additive-noise representation is argued to be non-generic because it requires a finely tuned alignment between the marginal of the putative cause and the conditional of the putative effect given that cause [16]. These results only cover the bivariate case and are not sufficiently strong for our purpose. Our results from Section 3.1 provide a result that is similar in spirit but strong enough for justifying our local discovery procedure.

Contributions. This paper makes three main contributions.

1. We introduce a local identifiability framework called `IndRM` (Definition 2.1) for $\mathrm{PA}(Y)$ from observational data based on an additive noise model for the target node Y and a no-backward condition (Assumption 1); we allow for arbitrary structural equations and hidden confounding elsewhere in the system.
2. We motivate the no-backward condition as a genericity condition. In finite-dimensional analytic SCM families, the parameters with descendant residual independence form a Lebesgue-null set unless this is an identity of the model class.
3. Based on the local identifiability criterion we develop the estimator `IndRM`, prove consistency of its threshold-screen variant, and demonstrate its empirical performance on synthetic and real-world data.

2 SCM with Additive Noise in Y

Consider a data-generating process over variables V whose causal structure is represented by a DAG G. In this work, we assume that there is an underlying SCM [2, 21] inducing both the distribution over V and the DAG G. We focus on a single target variable Y and aim to recover its direct causes, that is, its parents in the graph G. Write $\mathrm{Desc}(Y)$ for the strict descendants of Y in G, that is, the nodes reachable from Y by a directed path involving at least two nodes.

Setting 1 (SCM with additive noise in Y)**.** *Let $d \in \mathbb{N}$. Let $\mathbb{C}$ be an SCM over variables*[2] *$V := (V_1, \dots, V_d, Y)$ with a DAG G and structural assignments $\forall j \in [d], V_j := f_j(\mathrm{PA}(V_j), U_j)$ and*

$$Y := f(\mathrm{PA}(Y)) + \epsilon,$$

where $(U_1, \dots, U_d, \epsilon)$ are jointly independent centered, finite variance noise variables. Assume also $\mathbb{E}[Y^2] < \infty$. We allow a subset $H \subseteq \{V_1, \dots, V_d\} \setminus \mathrm{PA}(Y)$ to be unobserved. We denote the induced distribution by P^V and assume that it is absolutely continuous with respect to Lebesgue measure and faithful to G. The goal is to identify $\mathrm{PA}(Y)$ from $P^{V \setminus H}$ (population case) or from an i.i.d. sample from $P^{V \setminus H}$.

Our setting does not restrict the conditional distributions of $V_j | \mathrm{PA}(V_j)$ since there always exists a measurable function f_j and $U_j \perp\!\!\!\perp \mathrm{PA}(V_j)$ such that $V_j := f_j(\mathrm{PA}(V_j), U_j)$ [see, e.g., 26, Prop. 7.1]. That is, we restrict the model class of only one conditional (namely, the one of $Y | \mathrm{PA}(Y)$), and for the remaining variables, there may be hidden variables and/or non-additive models. This makes Setting 1 considerably more general than the settings considered in previous works: we allow for violations of the assumptions made by ANM-based approaches (in contrast to `CAM` or `RESIT` by [3, 24]) and allow for settings where no non-trivial conditional independence test with uniform asymptotic level guarantees is known (in contrast to settings, in which `PC` or `FCI` [35] are known to be consistent).

Next, we translate the target-local additive-noise assumption into an identifiability criterion for the parent set $\mathrm{PA}(Y)$. The intuition is as follows: under the true parent set, the regression residual of Y equals the structural noise ϵ, and is therefore independent of the regressors. This motivates residual independence as the first requirement for a candidate explanation of Y. Residual independence by itself leaves two natural ambiguities. First, a candidate set may omit a true parent (without violating residual independence). In that case, one expects it to miss part of the parental signal $f(\mathrm{PA}(Y))$, and hence to leave a larger residual variance. This motivates comparing residual-independent sets by their residual variance. Second, a candidate set may contain $\mathrm{PA}(Y)$ together with redundant variables. Once the parents are included, additional uninformative variables may preserve both

[2] By an abuse of notation, we use the same letters $V_1, \dots, V_d$ for both the random variables and their corresponding nodes in the induced graph G and sometimes treat V as a set rather than a vector.

residual independence and residual variance. This motivates using cardinality only as a final tie-breaker, selecting the smallest set among the variance minimizers. We now turn this intuition into a precise population criterion.

Let $\mathcal{S} \subseteq \{V_1, \dots, V_d\} \setminus H$ be a set of candidate variables satisfying $\mathrm{PA}(Y) \subseteq \mathcal{S}$. In particular, $\mathcal{S}$ could simply be the set of all observed, non-target variables $\{V_1, \dots, V_d\} \setminus H$. For $S \subseteq \mathcal{S}$, the same symbol S also denotes the associated random vector, ordered as in V; lowercase s denotes a realization. For any $S \subseteq \mathcal{S}$ let the population regression function be denoted by $m_S(s) \coloneqq \mathbb{E}[Y \mid S = s]$ and define the *population residual* as $r_S \coloneqq Y - m_S(S)$. Define the family of *residual-independent sets* as $\mathcal{R} \coloneqq \{S \subseteq \mathcal{S} \mid r_S \perp\!\!\!\perp S\}$ and among these the variance-optimal *residual-independent risk minimizing* ones $\mathcal{R}_{\min} \coloneqq \arg\min_{S \in \mathcal{R}} \mathbb{E}[r_S^2]$.

Definition 2.1. *We call a subset* $\mathrm{S_{PA}}$ *of* $\mathcal{S}$ *an* `IndRM`-set *if* $\mathrm{S_{PA}} \in \arg\min_{S \in \mathcal{R}_{\min}} |S|$.

3 Identifiability

We now give a population identifiability result for $\mathrm{PA}(Y)$ via `IndRM`, introduced in Definition 2.1. The result relies on the target-local additive-noise model in Setting 1, and on one additional assumption that rules out explaining Y by an additive-noise model built from variables that include descendants of Y (Assumption 1). All quantities defined in Section 2, namely m_S, r_S, $\mathcal{R}$, $\mathcal{R}_{\min}$, and the `IndRM`-sets, involve only Y and candidates $S \subseteq \mathcal{S} \subseteq \{V_1, \dots, V_d\} \setminus H$. They are therefore functionals of the observed distribution $P^{V \setminus H}$, in which the hidden variables are integrated out. In contrast, the graph G, the descendants $\mathrm{Desc}(Y)$, and faithfulness refer to all variables, including those in H. Since H contains no parent of Y, the residual of the parent set can be computed from the observed distribution; Proposition B.1 in Appendix B.5 shows that identification can fail if parents are unobserved. We begin with two observations. First, under Setting 1, the true parent set $\mathrm{PA}(Y)$ is residual-independent (Lemma 3.1).

Lemma 3.1. *Assume Setting 1. Then* $\mathrm{PA}(Y) \in \mathcal{R}$*, and its residual risk equals the noise variance:* $\mathbb{E}\left[r^2_{\mathrm{PA}(Y)}\right] = \mathbb{E}[\epsilon^2]$.

Second, any residual-independent set that is not equal to the true parent set and does not contain descendants of Y must have residual variance at least as large as that of the parent set; moreover, if the residual variances are equal, then the set must have larger cardinality (Lemma 3.2).

Lemma 3.2 (Risk and cardinality comparison)**.** *Assume Setting 1. Let* $S \subseteq \mathcal{S}$ *satisfy* $S \cap \mathrm{Desc}(Y) = \emptyset$*. Then* $\mathbb{E}[r_S^2] \geqslant \mathbb{E}[r^2_{\mathrm{PA}(Y)}]$*. Moreover, if* $\mathbb{E}[r_S^2] = \mathbb{E}[r^2_{\mathrm{PA}(Y)}]$*, then* $\mathrm{PA}(Y) \subseteq S$*; in particular, if* $S \neq \mathrm{PA}(Y)$*, then* $|S| > |\mathrm{PA}(Y)|$.

Thus, there is only a single remaining obstruction for proving that an `IndRM`-set $\mathrm{S_{PA}}$ equals the parents of Y: there could be a set that contains descendants of Y and nevertheless yields a backward additive-noise representation of Y. We now assume that such sets do not exist.

Assumption 1 (No backward ANM through descendants)**.** *Assume Setting 1 and let* $S \subseteq \mathcal{S}$ *with* $S \cap \mathrm{Desc}(Y) \neq \emptyset$*. Then* $r_S \not\perp\!\!\!\perp S$*, that is,* $S \notin \mathcal{R}$.

Assumption 1 does not follow directly from existing literature: [16] rules out a reverse additive-noise model between two variables, but S may be a multivariate set containing descendants of Y together with other variables; standard ANM identifiability theorems [14, 24, 34, 40] do not apply because they assume additive-noise structure throughout the graph, while we assume it only for the target Y and allow arbitrary mechanisms, and possible hidden confounding, elsewhere; finally, even under the additional assumption of a global ANM, Assumption 1 would not follow directly: classical identifiability results[3] only rule out that there are two different full-graph ANM models, whereas the above statement is about the specific variable Y and does not imply the existence of an ANM with a different graph. Thus, Assumption 1 is a genuinely local additional condition.

In Section 3.1, we justify Assumption 1 as a genericity condition in finite-dimensional analytic SCM families: if descendant residual independence is not an identity of the model class, then it holds only on a Lebesgue-null exceptional parameter set. We also connect this condition to standard ANM

[3]Existing identifiability proofs are often 'global' in that, for example, they start with two different graphs and identify specific reversed edges with a specific property – adjusting such proofs for our purpose seems non-trivial.

asymmetry: it fails in the linear Gaussian case but holds generically in the regular bivariate linear non-Gaussian and nonlinear Gaussian settings.

We can now state the main result of this section.

Theorem 3.3 (Identifiability of $\mathrm{PA}(Y)$)**.** *Assume Setting 1 and Assumption 1. Then there exists a unique* `IndRM`*-set* $\mathrm{S_{PA}}$*, and it equals the true parent set of* Y*, that is,* $\mathrm{S_{PA}} = \mathrm{PA}(Y)$.

Proof. By Lemma 3.1, $\mathrm{PA}(Y) \in \mathcal{R}$. By Assumption 1 every $S \in \mathcal{R}$ satisfies $S \cap \mathrm{Desc}(Y) = \emptyset$. Therefore, by Lemma 3.2, every $S \in \mathcal{R}$ satisfies $\mathbb{E}[r_S^2] \geqslant \mathbb{E}[r^2_{\mathrm{PA}(Y)}]$. Hence, $\mathrm{PA}(Y) \in \mathcal{R}_{\min}$. Moreover, if $S \in \mathcal{R}_{\min}$ and $S \neq \mathrm{PA}(Y)$, then equality of risks holds, so Lemma 3.2 gives $|S| > |\mathrm{PA}(Y)|$. Thus, $\mathrm{PA}(Y)$ is the unique cardinality-minimal element of $\mathcal{R}_{\min}$, and hence the unique `IndRM`-set. □

3.1 Identifiability witnesses and validity of Assumption 1

Theorem 3.3 relies on Assumption 1. We now give regularity conditions under which it holds generically. To do so, fix a finite DAG G with target node Y, and consider a *modular* SCM class, in which the structural assignment of each non-target node belongs to a class that depends only on the number of its parents. Formally, for each $m \geqslant 0$, let $\mathcal{A}_m$ be a class of pairs (g, Q), where $g : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ is measurable and Q is a probability law on $\mathbb{R}$, with $\mathcal{A}_0 \neq \emptyset$, and let $\mathcal{A}_m^Y$ be a class of such pairs of additive-noise form $g(z, n) = f(z) + n$. For a finite DAG H containing Y, let $\mathcal{M}_H$ be the set of distributions induced by SCMs over H with jointly independent exogenous noises $N_v \sim Q_v$ and assignments $v := g_v(\mathrm{PA}_H(v), N_v)$, where $(g_v, Q_v) \in \mathcal{A}_{|\mathrm{PA}_H(v)|}$ for $v \neq Y$ and $(g_Y, Q_Y) \in \mathcal{A}^Y_{|\mathrm{PA}_H(Y)|}$, where the pairs can be chosen separately for each node, that is, without any constraints linking the choices at different nodes. We assume that these classes are closed under permuting the parent arguments and under ignoring parents: for $m' \geqslant m$, an assignment in $\mathcal{A}_m$ (resp. $\mathcal{A}_m^Y$), viewed as a function of m' parent arguments that ignores $m' - m$ of them, belongs to $\mathcal{A}_{m'}$ (resp. $\mathcal{A}_{m'}^Y$). Family members need not be faithful to H. We assume that $\mathbb{E}\,|Y| < \infty$ under every distribution in these families, so that the conditional means below are defined. For the genericity result in Proposition 3.4, assume that each class $\mathcal{A}_m$ and $\mathcal{A}_m^Y$ is the image of an open and connected subset of a Euclidean space under a parameterization; ignoring parents then corresponds to particular parameter values, for example, zero coefficients. Taking products over the nodes of G, we obtain $\mathcal{M}_G = \{P_{G,\theta} : \theta \in \Theta_G\}$, where $\Theta_G \subseteq \mathbb{R}^{q_G}$ is open and connected as a finite product of open and connected sets. We write $P_\theta = P_{G,\theta}$, $\Theta = \Theta_G$, and $q = q_G$.

Having specified the model class, we can now formulate the main assumption of Section 3.1. For $S \subseteq \mathcal{S}$, write $m_{S,\theta}(s) = \mathbb{E}_\theta[Y \mid S = s]$, and $r_{S,\theta} = Y - m_{S,\theta}(S)$. If $r_{S,\theta} \perp\!\!\!\perp S$, then for every pair of *test functions* $\varphi : \mathbb{R}^{|S|} \to \mathbb{R}, \psi : \mathbb{R} \to \mathbb{R}$, we have that $\mathrm{Cov}_\theta(\varphi(S), \psi(r_{S,\theta})) = 0$.[4] Thus, residual independence imposes functional equations on the parameters of the SCM. The following assumption says that, for descendant-containing sets, at least one such equation is nontrivial.

Assumption 2 (Analytic residual nonidentity)**.** *For every* $S \subseteq \mathcal{S}$ *satisfying* $S \cap \mathrm{Desc}(Y) \neq \emptyset$*, there exist test functions* φ_S, ψ_S *such that* $F_S(\theta) := \mathrm{Cov}_\theta(\varphi_S(S), \psi_S(r_{S,\theta}))$ *is real analytic in* θ *and is not identically zero on* Θ*.*

Analyticity of F_S must be verified for the chosen parameterized model family. Lemma B.5 shows that this holds under suitable regularity and integrability conditions on the joint density, regression function, and test functions.

The second part of Assumption 2, namely that F_S is not identically zero, follows from the existence of an *identifiability witness*, that is, a single model θ for which we find residual dependence $F_S(\theta) \neq 0$. In the bivariate linear non-Gaussian case (Proposition 3.7), the witness is obtained through characteristic functions: if all residual-dependence functionals $G_t(a) = \mathrm{Cov}(X_a, e^{itr_a})$ vanished identically, their derivatives at $a = 0$ would force the characteristic function of Y to be Gaussian. Non-Gaussianity therefore implies that some G_t, or one of its real or imaginary parts, is not identically zero.

Proposition 3.4 (Generic no-backward residual independence)**.** *Assume Assumption 2. Then, for Lebesgue-a.e.* $\theta \in \Theta$*, we have that for all* $S \subseteq \mathcal{S}$ *with* $S \cap \mathrm{Desc}(Y) \neq \emptyset$

$$r_{S,\theta} \not\perp\!\!\!\perp S.$$

[4]Throughout, test functions are measurable, and both arguments of each covariance, as well as their product, are required to be integrable under the distribution at which it is evaluated.

More precisely, all $\theta \in \Theta$ for which this statement does not hold, are contained in a relatively closed, nowhere-dense, Lebesgue-null set $Z \subseteq \Theta$. Furthermore, for every nonempty compact $K \subseteq \Theta \setminus Z$, there is a constant $c_K > 0$ such that for all $\theta \in K$ and all S with $S \cap \mathrm{Desc}(Y) \neq \emptyset$ the witnesses F_S from Assumption 2 satisfy $|F_S(\theta)| \geqslant c_K$. In particular, if θ is drawn from any distribution on Θ absolutely continuous with respect to Lebesgue measure, then P_θ satisfies the no-backward condition in Assumption 1 almost surely.

Proposition 3.4 shows that, under analytic regularity assumptions, once residual independence for a descendant-containing set is not an identity of the model class (that is, there exists an identifiability witness), the parameter values for which it holds form a Lebesgue-null set. In this sense, Assumption 1 is a genericity assumption analogous to faithfulness: it excludes exact analytic coincidences rather than structural independence implied by the graph. The linear Gaussian model is the canonical case where Assumption 2 fails: If Y and S are jointly Gaussian, then the regression residual $r_S = Y - \mathbb{E}[Y \mid S]$ is uncorrelated with every component of S. Joint Gaussianity turns this orthogonality into independence, so $r_S \perp\!\!\!\perp S$ for every S, including sets containing descendants of Y. Thus, linear Gaussian models are a case in which residual independence cannot distinguish causal direction.

Reduction to the bivariate case. We now provide a convenient property of our approach based on witnesses: Under some regularity conditions, it is sufficient to find a witness in the bivariate case. To formalize this, we impose the following extension property on the collection of SCM families.

Definition 3.5 (Latent extension property)**.** *The modular SCM class satisfies the* latent extension property *if $\mathcal{M}_{Y\to X} \subseteq \mathrm{Marg}_{Y,X}(\mathcal{M}_{Y\to V\to X})$, where $\mathrm{Marg}_{Y,X}$ denotes taking the (Y,X)-marginal of all distributions in the class. Equivalently, every $P_{Y,X} \in \mathcal{M}_{Y\to X}$ admits an extension $P_{Y,V,X} \in \mathcal{M}_{Y\to V\to X}$ whose (Y,X)-marginal is $P_{Y,X}$.*

In particular, consider the modular additive-noise class in which $\mathcal{A}_m$ and $\mathcal{A}_m^Y$ consist of all pairs (g, Q) with $g(z,n) = f(z) + n$, $f \in \mathcal{F}_m$, and $Q \in \mathcal{N}$, where the function classes $\mathcal{F}_m$ are closed under permuting and ignoring arguments and $\mathcal{F}_1$ contains the identity. If every $Q \in \mathcal{N}$ is the law of $E + U$ for independent E and U with non-degenerate laws in $\mathcal{N}$, the latent extension property holds: an assignment $X = f(Y) + N$ can be replaced by $V = f(Y) + E$ and $X = V + U$ with $E + U =_d N$, and marginalizing V recovers the joint distribution of (Y, X).

We now prove that the condition $F_S \not\equiv 0$, see Assumption 2, can be verified in the bivariate case.

Theorem 3.6 (Reduction to bivariate witnesses)**.** *Assume that the modular SCM class satisfies the latent extension property. Suppose that the bivariate class $\mathcal{M}_{Y\to X}$ admits a witness, i.e., there exist $P^\star_{Y,X} \in \mathcal{M}_{Y\to X}$ and test functions φ, ψ such that $\mathrm{Cov}_{P^\star}(\varphi(X), \psi(Y - \mathbb{E}[Y \mid X])) \neq 0$. Then, for every $S \subseteq \mathcal{S}$ with $S \cap \mathrm{Desc}(Y) \neq \emptyset$, $\mathcal{M}_G$ admits a witness: there exist $\theta_S \in \Theta_G$ and test functions $\tilde\varphi, \psi$ such that, under P_{G,θ_S}, $\mathrm{Cov}(\tilde\varphi(S), \psi(Y - \mathbb{E}[Y \mid S])) \neq 0$.*

Examples for bivariate identifiability. Under their stated assumptions, the next two propositions provide bivariate witnesses that Theorem 3.6 can extend to multivariate models. For example, consider linear SCMs where the noise for node j is given by $N_j := \sigma_j \varepsilon_j + \tau_j W_j$, with $\sigma_j > 0$, and $\tau_j \in \mathbb{R}$, where ε_j is standard Gaussian, all W_j have the same fixed centered non-Gaussian distribution with finite variance, and all ε_j, W_j are mutually independent. This class has the latent extension property, as $\sigma_j\varepsilon_j + \tau_j W_j$ has the law of $(\sigma_j\varepsilon_{j,1}/\sqrt{2} + \tau_j W_j) + \sigma_j\varepsilon_{j,2}/\sqrt{2}$ for mutually independent $\varepsilon_{j,1}, \varepsilon_{j,2}, W_j$, and Proposition 3.7(ii) supplies a bivariate witness with Gaussian noise for Y ($\tau_Y = 0$) and non-Gaussian noise for X ($\tau_X \neq 0$, by Cramér's theorem [5]). Hence, if the witness functions F_S are real analytic on all of Θ, then the no-backward condition in Assumption 1 holds for Lebesgue-a.e. $\theta \in \Theta$ by Proposition 3.4.

Proposition 3.7 (Generic linear non-Gaussian descendants)**.** *Let $I \subseteq \mathbb{R}$ be an open interval containing 0. Consider the bivariate family $Y = N_Y, X_a = aY + N_X$, $a \in I \setminus \{0\}$, where $N_Y \perp\!\!\!\perp N_X$, both noises are centered, and $0 < \mathrm{Var}(N_Y), \mathrm{Var}(N_X) < \infty$. Suppose that either*

- *(i) N_Y is non-Gaussian and Assumption 3 with $h(y) = y$ holds; or*
- *(ii) N_Y is Gaussian and N_X is non-Gaussian.*

In case (i), Assumption 2 holds on each connected component of $I \setminus \{0\}$; consequently, by Proposition 3.4, the no-backward condition in Assumption 1 holds for Lebesgue-a.e. $a \in I \setminus \{0\}$. In case (ii),

$r_a \not\perp\!\!\!\perp X_a$ *for every* $a \in I \setminus \{0\}$*, so the no-backward condition in Assumption 1 holds for every such* a*.*

Proposition 3.8 (Generic nonlinear Gaussian descendants)**.** *Let* $I \subseteq \mathbb{R}$ *be an open interval containing* 0*. Consider the bivariate family* $Y = N_Y, X_b = b\,h(Y) + N_X, b \in I \setminus \{0\}$*, where* N_Y *and* N_X *are independent centered Gaussian variables, and* $h \notin \operatorname{span}\{1, Y\}$ *in* $L^2(P_Y)$*. Assume that the induced family of joint laws of* (Y, X_b) *are sufficiently regular (they satisfy Assumption 3). Then Assumption 2 holds for* $S = \{X_b\}$ *on each connected component of* $I \setminus \{0\}$*. Consequently, by Proposition 3.4, for Lebesgue-a.e.* $b \in I \setminus \{0\}$*, the no-backward condition in Assumption 1 holds.*

Residual independence of a descendant-containing set can occur even when the mechanism of Y is nonlinear: in Example 1 (Appendix B.4), $Y = C^3 + \epsilon$ and $D = aY + \eta$ with Gaussian noises, and $\{C, D\}$ violates Assumption 1. Hence a nonlinear mechanism of Y alone does not suffice; in Proposition 3.8, the nonlinearity enters through the mechanism of the descendant.

4 `IndRM`: Independent Risk Minimization

Based on Section 3, we propose `IndRM` for i.i.d. observational data from $P^{V \setminus H}$; it does not require interventional data. Section 4.1 describes `IndRM` and `IndRM-TS` (`IndRM` with threshold screen), a variant for theoretical analysis that replaces the two resampling-based decision rules by fixed thresholds; we propose to use `IndRM` in practice. Section 4.2 proves that `IndRM-TS` recovers $\mathrm{PA}(Y)$, and Section 4.3 discusses screening and the all-data implementation.

4.1 Algorithm

We now describe `IndRM`; pseudocode is given in Algorithms 1 to 3 in Appendix A. We assume that we observe n i.i.d. draws from $P^{V \setminus H}$, that is, of Y and of the observed variables $\{V_1, \dots, V_d\} \setminus H$. We write Y_i and S_i for the values of Y and of a candidate S in the i-th draw, $i \in \{1, \dots, n\}$, and randomly partition the sample into three disjoint folds $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$. The partition is drawn independently of the observations, and we write $n_j := |\mathcal{I}_j|$ for $j \in \{1, 2, 3\}$. The three folds are used for regression, independence screening, and risk comparison, respectively. This sample splitting keeps the fitted regressors independent of the two downstream testing stages. The candidates are the subsets $S \subseteq \mathcal{S}$, including the empty set; Section 4.3 restricts this family when $\mathcal{S}$ is large. For $S = \emptyset$, the regression estimate is the training-fold mean of Y. Its residual is trivially independent of the empty regressor, so $\emptyset \in \mathcal{R}$, and both screening rules below always retain $\emptyset$. Let $\beta_n \downarrow 0$ be the overall familywise error budget. We split the testing budget evenly across the two inferential stages, $\beta_n^{\mathrm{HSIC}} := \beta_n/2$ and $\beta_n^{\mathrm{cmp}} := \beta_n/2$. For each candidate set $S \subseteq \mathcal{S}$, fit a regression estimator $\hat{m}_S^{\mathcal{I}_1} \leftarrow \widehat{\mathbb{E}}_{\mathcal{I}_1}[Y \mid S]$ using the observations in $\mathcal{I}_1$. On the second fold, form residuals $\widehat{r}_{S,i}^{\mathcal{I}_2} := Y_i - \hat{m}_S^{\mathcal{I}_1}(S_i), i \in \mathcal{I}_2$, and test residual–set independence with `HSIC` [12], using a Westfall–Young max-T step-down adjustment with common permutations [38]. For nonempty S, the test statistic $\widehat{H}_n(S)$ is the biased empirical `HSIC` of the pairs $(S_i, \widehat{r}_{S,i}^{\mathcal{I}_2})$, $i \in \mathcal{I}_2$, defined in Appendix A. The sets that are not rejected at level β_n^{HSIC} form $\widehat{\mathcal{R}}_n$. For a single candidate with fixed kernels and oracle residuals, permutation calibration is valid under residual independence. With fitted residuals and many candidates, we do not claim familywise error control and treat β_n as a tuning parameter. Next, recompute residuals on the third fold, $\widehat{r}_{S,i}^{\mathcal{I}_3} := Y_i - \hat{m}_S^{\mathcal{I}_1}(S_i), i \in \mathcal{I}_3$, and let $\widehat{R}_n(S)$, the mean of $(\widehat{r}_{S,i}^{\mathcal{I}_3})^2$ over $i \in \mathcal{I}_3$, denote the held-out risk of S. We then compare all surviving sets pairwise. For each ordered pair of distinct sets $(S, T) \in \widehat{\mathcal{R}}_n \times \widehat{\mathcal{R}}_n$ with $S \neq T$, we test the null hypothesis that the mean squared error of S is no larger than that of T, using a one-sided studentized sign-flip statistic. We apply a Romano–Wolf step-down correction [27–29] to the full family of these ordered-pair comparisons at level β_n^{cmp}, yielding adjusted pairwise p-values $\tilde{p}_{n,\mathrm{pair}}^{\mathrm{cmp}}(S, T)$ for all $S \neq T$. For each $S \in \widehat{\mathcal{R}}_n$, let $\tilde{p}_n^{\mathrm{cmp}}(S)$ be the smallest of the values $\tilde{p}_{n,\mathrm{pair}}^{\mathrm{cmp}}(S, T)$, $T \in \widehat{\mathcal{R}}_n \setminus \{S\}$, that is, the smallest adjusted p-value among all comparisons testing whether S is worse than another surviving set. We then define $\widehat{\mathcal{R}}_n^{\min} := \{S \in \widehat{\mathcal{R}}_n : \tilde{p}_n^{\mathrm{cmp}}(S) > \beta_n^{\mathrm{cmp}}\}$. Thus, $\widehat{\mathcal{R}}_n^{\min}$ consists of those sets that are not found to be significantly worse (in terms of residual variance) than any other set in $\widehat{\mathcal{R}}_n$. The procedure returns the family $\widehat{\mathrm{PA}}_n(Y)$ of sets of smallest cardinality in $\widehat{\mathcal{R}}_n^{\min}$. Here and below, $\arg\min$ denotes the family of all minimizers. If several sets are returned, for example because of tied population risks, any tie-break that does not use the ground truth may be applied; in the experiments, we average every metric over the returned sets.

`IndRM-TS` **(threshold screen).** The resampling-based decision rules are convenient in practice, but their analysis with fitted residuals and tied risks is delicate. For example, if $W \in \mathcal{S} \setminus \mathrm{PA}(Y)$ is a non-descendant of Y, the oracle residuals of $\mathrm{PA}(Y)$ and $\mathrm{PA}(Y) \cup \{W\}$ both equal ϵ, so their squared-residual difference has variance zero and the studentized comparison degenerates. We therefore also consider a variant that keeps the folds, regression fits, fitted residuals, candidate family, and cardinality step of `IndRM`, but replaces both decision rules by direct thresholds. For a screening threshold $h_n \downarrow 0$ and a risk tolerance $\lambda_n \downarrow 0$, let $\widehat{\mathcal{R}}_n^{\mathrm{thr}}$ consist of $\emptyset$ and every nonempty $S \subseteq \mathcal{S}$ with $\widehat{H}_n(S) \leqslant h_n$, let $\widehat{\mathcal{R}}_{n,\lambda}^{\min}$ consist of those $S \in \widehat{\mathcal{R}}_n^{\mathrm{thr}}$ whose held-out risk $\widehat{R}_n(S)$ exceeds the smallest one in $\widehat{\mathcal{R}}_n^{\mathrm{thr}}$ by at most λ_n, and return the family $\widehat{\mathrm{PA}}_n^{\mathrm{thr}}(Y)$ of sets of smallest cardinality in $\widehat{\mathcal{R}}_{n,\lambda}^{\min}$. In the experiments, we use $h_n = 0.05\, n^{-1/4}$ and $\lambda_n = 0.25\, \widehat{\mathrm{Var}}_{\mathcal{I}_3}(Y)\, n_3^{-1/4}$, where $\widehat{\mathrm{Var}}_{\mathcal{I}_3}(Y)$ is the empirical variance of Y on $\mathcal{I}_3$. These constants were fixed in advance and not tuned on the reported benchmarks. In this variant, the kernels of $\widehat{H}_n$ use scales estimated on $\mathcal{I}_1$ only (Appendix A). Pseudocode is given in Algorithm 4.

4.2 Consistency

By Theorem 3.3, in the population problem the parent set $\mathrm{PA}(Y)$ is the unique set that is both variance-optimal among the residual-independent sets and cardinality-minimal among the variance minimizers. We now show that `IndRM-TS` recovers this target. Let $\mathcal{C}$ denote the candidate family, that is, the collection of all subsets of $\mathcal{S}$ or, more generally, a fixed collection of subsets of $\mathcal{S}$ that contains $\emptyset$. Restrict the empirical search to $\mathcal{C}$. Define $\mathcal{R}$ and $\mathcal{R}_{\min}$ as in Section 2, with $\mathcal{C}$ in place of the collection of all subsets of $\mathcal{S}$, and write $R(S) \coloneqq \mathbb{E}[r_S^2]$ and $\mathfrak{T}(P) \coloneqq \arg\min_{S \in \mathcal{R}_{\min}} |S|$ for the population risk and the family of `IndRM`-sets. The proof of Theorem 3.3 uses only that $\mathrm{PA}(Y)$ is a candidate. Hence, under Setting 1 and Assumption 1, $\mathfrak{T}(P) = \{\mathrm{PA}(Y)\}$ whenever $\mathrm{PA}(Y) \in \mathcal{C}$. We write $\widehat{m}_S \coloneqq \hat{m}_S^{\mathcal{I}_1}$.

Assumption 5 in Appendix B.9 states the conditions: a fixed finite candidate family, folds of proportional sizes, $\mathbb{E}[Y^4] < \infty$, bounded characteristic kernels that depend Lipschitz-continuously on their parameters and on the residuals, a regression estimator whose L^2 error vanishes at some rate r_n, and a threshold h_n and a tolerance λ_n that vanish more slowly than the estimation errors of the `HSIC` statistics and the risks. They hold, for example, for kernel ridge regression with a suitable Sobolev RKHS, Gaussian, Laplace, or Matérn ($\nu \geqslant 1/2$) `HSIC` kernels, and the rate $n^{-1/4}$ used in our experiments; the fixed ridge penalty of our experiments is not covered.

Theorem 4.1 (Consistency of `IndRM-TS`)**.** *Under Assumption 5,*

$$\mathbb{P}\Big(\widehat{\mathcal{R}}_n^{\mathrm{thr}} = \mathcal{R},\ \widehat{\mathcal{R}}_{n,\lambda}^{\min} = \mathcal{R}_{\min},\ \widehat{\mathrm{PA}}_n^{\mathrm{thr}}(Y) = \mathfrak{T}(P)\Big) \to 1.$$

If, in addition, Setting 1 and Assumption 1 hold and $\mathrm{PA}(Y) \in \mathcal{C}$*, then* $\mathbb{P}\big(\widehat{\mathrm{PA}}_n^{\mathrm{thr}}(Y) = \{\mathrm{PA}(Y)\}\big) \to 1$.

Theorem 4.1 is a selection-consistency result for a fixed population and a fixed finite candidate family.

4.3 Screening and all-data implementation

Searching all subsets of size at most K of q candidates requires $M(q, K) = \sum_{k=0}^{K} \binom{q}{k} = O(q^K)$ regression fits and residual scores. For larger $\mathcal{S}$, we therefore compute a screened set $\widehat{\mathcal{S}} \subseteq \mathcal{S}$ on $\mathcal{I}_1$ only, freeze it, and search $\mathcal{H}_K = \{S \subseteq \widehat{\mathcal{S}} : |S| \leqslant K\}$. The population target within $\mathcal{H}_K$ is $\{\mathrm{PA}(Y)\}$ whenever the screen retains every parent and $K \geqslant |\mathrm{PA}(Y)|$, and Corollary B.8 in Appendix B.9 shows that Theorem 4.1 carries over if the screen retains every parent with probability tending to one. A parent removed by the screen is an end-to-end false negative.

Splitting reduces the sample size available to each stage, and results vary across random partitions. In practice, we therefore also use an all-data implementation in which regression fitting, residual screening, and risk comparison all use the full sample. It resembles the approximate test in Method II of [25], which ignores the sampling variability of the fitted coefficients, and it is a finite-sample heuristic outside Theorem 4.1; cross-fitting schemes that retain formal guarantees are left for future work.

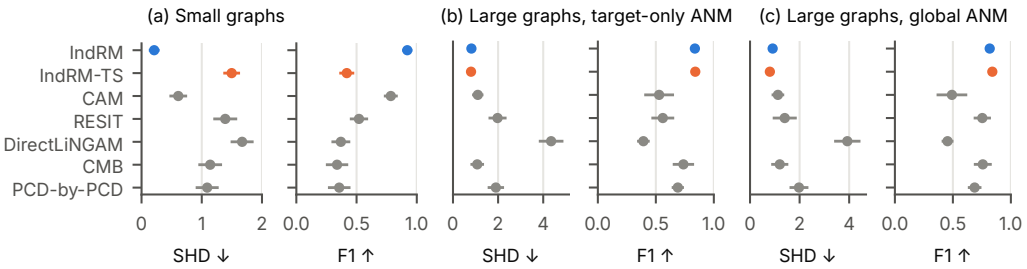
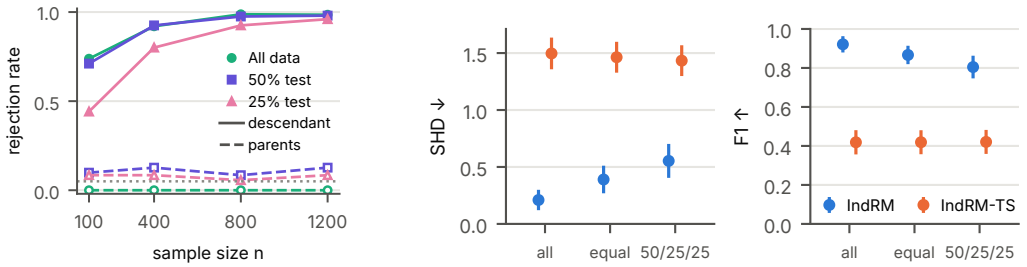


Figure 1: Parent-set recovery of `IndRM`, `IndRM-TS`, and the baselines: SHD and F_1 score on (a) small graphs (100 datasets) and on large graphs in (b) the target-only and (c) the global regime (40 datasets each, pooled over d). False discovery rates are given in Appendix D.2. Error bars are 95% confidence intervals.



(a) `HSIC` rejection rates

(b) Sample splits

Figure 2: Small-graph benchmark. (a) Rejection rates of the `HSIC` test for sets containing a descendant of Y and for nonempty true parent sets, with residuals tested on all data or on a held-out 50% or 25% of the sample; the dotted line is the level 0.05. (b) SHD and F_1 score of `IndRM` and `IndRM-TS` with all data and with equal and $50/25/25$ splits. Error bars are 95% confidence intervals.

5 Empirical Results

We evaluate `IndRM` on two synthetic benchmarks and on the Sachs protein-signaling data. We report `IndRM` (Algorithm 1) and `IndRM-TS` (Algorithm 4) with the constants of Section 4.1. In the main comparisons, both use the all-data implementation of Section 4.3; sample splits are compared in Figure 2b. For an estimate $\widehat{S}$ of the parent set $S^\star$ of a target, we report the local structural Hamming distance $|\widehat{S} \triangle S^\star|$ (SHD) and the F_1 score; false discovery rates are given in Appendix D.2. For baselines that output partially oriented graphs, only edges oriented into the target count as selected parents. Details on data, configurations, and baselines are given in Appendix C, and code is available at `https://github.com/fschur/who-caused-y`.

Small-graph benchmark. Each of 100 datasets consists of $n = 3000$ observations from a random four-node DAG in which only the target follows an additive-noise model; the other variables are non-additive in their noise. We compare with the global methods `DirectLiNGAM`, `RESIT`, and `CAM` and the local methods `CMB` and `PCD-by-PCD`.

Figure 1a shows that `IndRM` recovers the parent set most accurately ($F_1 = .92$, against .78 for `CAM`, the best baseline). `CMB` and `PCD-by-PCD` recover few parents ($F_1 = .34$ and .36). Figure 2a shows that, as the sample size grows, the `HSIC` rejection rate of sets containing a descendant of Y approaches one, while that of the true parent set stays low, in line with Assumption 1. Using all data in every step performs best (Figure 2b), and other `HSIC` kernels give similar results (Appendix D).

`IndRM-TS` performs considerably worse ($F_1 = .42$; Figure 1a). At this sample size, h_n lies far above the `HSIC` statistics of most descendant-containing sets, so the screen removes almost no candidate (Appendix C); because all noise variances are small, a descendant of Y then predicts Y about as well as its parents and is often selected. The screen becomes effective as n grows (Theorem 4.1), but fixed

Table 1: Sachs data: SHD averaged and F_1 pooled over the 11 targets; *Empty* returns no parents. Best values in bold.

	IndRM	IndRM-TS	CAM	RESIT	DirectLiNGAM	CMB	PCD-by-PCD	Empty
SHD ↓	2.67	2.02	**1.36**	1.91	1.64	1.55	1.55	1.55
F_1 ↑	.16	.42	**.44**	.40	.36	.00	.00	.00

constants do not adapt to the scale of the HSIC statistics. The permutation tests of IndRM do, and we recommend IndRM in practice.

Large-graph benchmark. To study larger candidate sets, we generate data with $n = 1000$ and $d \in \{10, 20, 50, 100\}$ candidates in which Y has two parents, two children, and a non-additive mean function. We consider two regimes: in the target-only regime, only Y follows an additive-noise model; in the global regime, every variable does. For each d and regime, we generate 10 datasets. All methods receive the same screened candidates: the union of the 8 candidates with the largest absolute correlation with Y and the 8 with the largest marginal HSIC, computed on half of the sample. The screen keeps $q \in \{8, \dots, 13\}$ candidates, including both parents in every dataset, so that with $K = 3$, IndRM compares between $M(8, 3) = 93$ and $M(13, 3) = 378$ candidate sets.

Both variants of IndRM attain the highest F_1 score in both regimes (Figure 1b, c) and in every dimension (Table 4). Among the baselines, CMB attains the highest F_1 score in both regimes. RESIT, which assumes an additive-noise model for every variable, drops from $F_1 = .75$ in the global regime to .56 in the target-only regime, whereas IndRM attains .82 and .84. The threshold screen again removes almost no candidate set (99.8% of the candidate sets pass; Appendix C). Here, the risk and cardinality steps still recover both parents in most datasets because the children of Y carry substantial noise of their own and improve the prediction of Y only slightly.

Sachs data. We use the 853 observations of the Sachs data [30][5] recorded under CD3/CD28 stimulation without further interventions. Each of the 11 variables is the target in turn, and estimates are compared with its parents in the 17-edge consensus network, which we treat as a reference rather than as ground truth. IndRM searches subsets of size at most $K = 3$ of the $q = 6$ candidates with the largest gradient-boosting importance, that is, $M(6, 3) = 42$ sets; this screen retains 14 of the 17 reference parents. IndRM-TS attains an F_1 score similar to the strongest baselines, CAM and RESIT, but a larger SHD, and IndRM performs worse (Table 1)[6]. CMB and PCD-by-PCD orient no edge into any target and thus coincide with the empty estimate; only CAM attains a lower SHD than this estimate. For 7 of the 9 non-root targets, the residuals of the target given its reference parents are dependent on these parents (Appendix C), so the additive-noise assumption appears to be violated for most targets.

Overall, when only the target follows an additive-noise model, IndRM recovers the parent set more accurately than global and local causal discovery baselines. On Sachs, where this model appears to be violated, it is not superior. IndRM is computationally more expensive than the baselines (Appendix D.5).

6 Discussion and Future Work

This paper introduces a local method for identifying the direct causes of a target variable Y from observational data. Under an additive-noise model for Y and a no-backward additive-noise condition, we show that the true parent set $\mathrm{PA}(Y)$ is characterized by residual independence, minimal residual variance, and minimal cardinality. Concretely, we propose IndRM, a simple local procedure that avoids global causal-model assumptions, and prove consistency of a variant with deterministic thresholds. Our work is limited in that IndRM makes a binary decision for each candidate set and does not quantify uncertainty; returning rankings of parent sets is an interesting direction for future research. The procedure may extend to other globally identifiable classes, such as post-nonlinear ANMs [40]. Finally, we believe that the proposed framework of identifiability witnesses has the potential to discover new classes of globally identifiable models.

[5] We could not identify a license for this dataset; however, the authors provided written permission for its use in this work.

[6] An earlier version of this paper pooled 2000 rows from all experimental conditions and used $K = 2$ and $q = 8$; under that protocol, IndRM had the lowest SHD.

Acknowledgments and Disclosure of Funding

This work received no third-party funding, however, Sorawit Saengkyongam received a salary from Apple during parts of the time he worked on this project. The authors declare no competing interests.

References

[1] J. Bodik and V. Chavez-Demoulin. Structural restrictions in local causal discovery: identifying direct causes of a target variable. *Biometrika*, 113(1):asaf042, 2026.

[2] S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

[3] P. Bühlmann, J. Peters, and J. Ernest. CAM: causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

[4] D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

[5] H. Cramér. Über eine Eigenschaft der normalen Verteilungsfunktion. *Mathematische Zeitschrift*, 41:405–414, 1936.

[6] B. Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

[7] S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020.

[8] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, pages 127–135, 2000.

[9] T. Gao and Q. Ji. Local causal discovery of direct causes and effects. In *Advances in Neural Information Processing Systems*, volume 28, pages 2512–2520, 2015.

[10] C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019.

[11] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert–Schmidt norms. In *Algorithmic Learning Theory (ALT)*, volume 3734 of *Lecture Notes in Artificial Intelligence*, pages 63–77. Springer, 2005.

[12] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20, pages 585–592, 2007.

[13] C. Heinze-Deml, J. Peters, and N. Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

[14] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21, pages 689–696, 2008.

[15] B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

[16] D. Janzing and B. Steudel. Justifying additive noise model-based causal discovery via algorithmic information theory. *Open Systems & Information Dynamics*, 17(2):189–212, 2010.

[17] L. Kook, S. Saengkyongam, A. R. Lundborg, T. Hothorn, and J. Peters. Model-based causal feature selection for general response types. *Journal of the American Statistical Association*, 120(550):1090–1101, 2025.

[18] Z. Ling, H. Peng, Y. Zhang, D. Cheng, X. Wu, P. Zhou, and K. Yu. Hybrid local causal discovery. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5725–5733, 2025. doi: 10.24963/ijcai.2025/637.

[19] S. Mani, P. Spirtes, and G. F. Cooper. A theoretical study of Y structures for causal discovery. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 314–323. AUAI Press, 2006.

[20] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

[21] J. Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, New York, USA, 2nd edition, 2009.

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

[23] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Identifiability of causal graphs using functional models. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 589–598. AUAI Press, 2011.

[24] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014.

[25] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

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

[27] J. P. Romano and M. Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108, 2005.

[28] J. P. Romano and M. Wolf. Control of generalized error rates in multiple testing. *The Annals of Statistics*, 35(4):1378–1408, 2007.

[29] J. P. Romano and M. Wolf. Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters*, 113:38–40, 2016.

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

[31] M. Schubert, T. Claassen, and S. Magliacane. Local causal discovery for statistically efficient causal inference. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 300 of *Proceedings of Machine Learning Research*, pages 1990–1998. PMLR, 2026.

[32] M. Scutari. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010.

[33] R. D. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

[34] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006.

[35] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT Press, 2000.

[36] P. L. Spirtes, C. Meek, and T. S. Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 499–506, 1995.

[37] Z. Szabó and B. K. Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018.

[38] P. H. Westfall and S. S. Young. *Resampling-based multiple testing: examples and methods for p-value adjustment*. Wiley, 1993.

[39] J. Yin, Y. Zhou, C. Wang, P. He, C. Zheng, and Z. Geng. Partial orientation and local structural learning of causal networks for prediction. In *Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008*, volume 3 of *Proceedings of Machine Learning Research*, pages 93–105. PMLR, 2008.

[40] K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 647–655, Corvallis, Oregon, USA, 2009. AUAI Press.

[41] Q. Zheng, Y. Liu, and Y. He. Local causal discovery with background knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(7):7969–7980, 2026. doi: 10.1109/TPAMI.2026.3667409.

A Algorithms

Algorithm 1 summarizes `IndRM`, Algorithms 2 and 3 give its two step-down adjustments, and Algorithm 4 gives `IndRM-TS`.

Algorithm 1 `IndRM`

Input candidate set $\mathcal{S}$, splits $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$, resamples B_n, nominal level β_n
$\beta_n^{\mathrm{HSIC}} \leftarrow \beta_n/2$, $\beta_n^{\mathrm{cmp}} \leftarrow \beta_n/2$
Fit on $\mathcal{I}_1$:
  For all $S \subseteq \mathcal{S}$, fit $\hat{m}_S^{\mathcal{I}_1}$ on $\mathcal{I}_1$
Filter by independence on $\mathcal{I}_2$:
  For all $S \subseteq \mathcal{S}$ and $i \in \mathcal{I}_2$, set $\widehat{r}_{S,i}^{\mathcal{I}_2} \leftarrow Y_i - \hat{m}_S^{\mathcal{I}_1}(S_i)$
  $\{\tilde{p}_n^{\mathrm{HSIC}}(S)\} \leftarrow$ `HSIC-SD`$(\mathcal{I}_2, \{S \subseteq \mathcal{S}\}, B_n)$ ▷ see Algorithm 2
  $\widehat{\mathcal{R}}_n \leftarrow \{S \subseteq \mathcal{S} \mid \tilde{p}_n^{\mathrm{HSIC}}(S) > \beta_n^{\mathrm{HSIC}}\}$
Compare all surviving models pairwise on $\mathcal{I}_3$:
  For all $S \in \widehat{\mathcal{R}}_n$ and $i \in \mathcal{I}_3$, set $\widehat{r}_{S,i}^{\mathcal{I}_3} \leftarrow Y_i - \hat{m}_S^{\mathcal{I}_1}(S_i)$
  $\{\tilde{p}_n^{\mathrm{cmp}}(S)\}_{S \in \widehat{\mathcal{R}}_n} \leftarrow$ `TTest-SD`$\big(\mathcal{I}_3, \widehat{\mathcal{R}}_n, B_n\big)$ ▷ see Algorithm 3
  $\widehat{\mathcal{R}}_n^{\min} \leftarrow \{S \in \widehat{\mathcal{R}}_n \mid \tilde{p}_n^{\mathrm{cmp}}(S) > \beta_n^{\mathrm{cmp}}\}$
Output $\widehat{\mathrm{PA}}_n(Y) \leftarrow \arg\min_{S \in \widehat{\mathcal{R}}_n^{\min}} |S|$

Notation and conventions. For paired observations $\{(a_i, b_i)\}_{i=1}^m$ and kernels k and ℓ,

$$\widehat{\mathrm{HSIC}}_{\mathrm{b}}(a, b) \coloneqq m^{-2} \operatorname{tr}(\mathbf{KCLC}), \qquad \mathbf{C} \coloneqq I_m - m^{-1}\mathbf{1}\mathbf{1}^\top,$$

denotes the biased empirical `HSIC`, a V-statistic [11], where $\mathbf{K}_{ij} = k(a_i, a_j)$ and $\mathbf{L}_{ij} = \ell(b_i, b_j)$. It is symmetric in its two arguments. The empty candidate is always retained: in Algorithm 2, its observed and permuted statistics are set to 0, which gives the adjusted p-value 1, and in Algorithm 4 it is included by construction. Hence the screened families are never empty. If only one candidate survives the screen, Algorithm 3 assigns it the p-value 1. In Algorithm 3, if $s(S,T) = 0$, we set $T_n(S,T)$ to $+\infty$, $-\infty$, or 0 according to whether $\bar{d}(S,T)$ is positive, negative, or zero, and we set $T_b(S,T) = 0$ for all b. If $s(S,T) > 0$ but $s_b(S,T) = 0$, we use the same convention for $T_b(S,T)$ according to the sign of $\bar{d}_b(S,T)$.

Algorithm 2 HSIC-SD (Westfall–Young max-T step-down adjustment with common permutations)

Input index set $\mathcal{I}_2$, hypotheses $\mathcal{H}$, resamples B_n, residuals $\{\widehat{r}_{S,i}^{\mathcal{I}_2}\}_{i \in \mathcal{I}_2, S \in \mathcal{H}}$, regressors $\{S_i\}_{i \in \mathcal{I}_2}$
Output adjusted p-values $\{\tilde{p}_n^{\mathrm{HSIC}}(S)\}_{S \in \mathcal{H}}$
For $S \in \mathcal{H}$, set $T_n(S) \leftarrow \widehat{\mathrm{HSIC}}_{\mathrm{b}}\big(\{\widehat{r}_{S,i}^{\mathcal{I}_2}\}_{i \in \mathcal{I}_2}, \{S_i\}_{i \in \mathcal{I}_2}\big)$
Order $T_n(S)$ decreasing: $T_{(1)} \geqslant \cdots \geqslant T_{(m)}$ with $S_{(1)}, \ldots, S_{(m)}$, $m \leftarrow |\mathcal{H}|$
for $b \leftarrow 1$ **to** B_n **do** ▷ common permutations
  Draw a permutation π_b of $\mathcal{I}_2$
  for $S \in \mathcal{H}$ **do**
    $T_b(S) \leftarrow \widehat{\mathrm{HSIC}}_{\mathrm{b}}\big(\{\widehat{r}_{S,i}^{\mathcal{I}_2}\}_{i \in \mathcal{I}_2}, \{S_{\pi_b(i)}\}_{i \in \mathcal{I}_2}\big)$
  end for
end for
Initialize $\tilde{p}_n^{\mathrm{HSIC}}(S) \leftarrow 0$ for all $S \in \mathcal{H}$
for $j \leftarrow 1$ **to** m **do**
  $M_b^{(j)} \leftarrow \max_{\ell \in \{j, \ldots, m\}} T_b\big(S_{(\ell)}\big)$ for $b \in \{1, \ldots, B_n\}$
  $\tilde{p}_{\mathrm{raw}}^{\mathrm{HSIC}}\big(S_{(j)}\big) \leftarrow \dfrac{1 + \#\{b \mid M_b^{(j)} \geqslant T_{(j)}\}}{1 + B_n}$
  $\tilde{p}_n^{\mathrm{HSIC}}\big(S_{(j)}\big) \leftarrow \max\{\tilde{p}_n^{\mathrm{HSIC}}\big(S_{(j-1)}\big),\ \tilde{p}_{\mathrm{raw}}^{\mathrm{HSIC}}\big(S_{(j)}\big)\}$ with $\tilde{p}_n^{\mathrm{HSIC}}\big(S_{(0)}\big) \leftarrow 0$
end for
return $\{\tilde{p}_n^{\mathrm{HSIC}}(S) \mid S \in \mathcal{H}\}$

Algorithm 3 TTest-SD (Romano–Wolf step-down adjustment with studentized sign flips)

Input index set $\mathcal{I}_3$, hypotheses $\mathcal{H}$, resamples B_n, residuals $\{\widehat{r}_{S,i}^{\mathcal{I}_3}\}_{i\in\mathcal{I}_3, S\in\mathcal{H}}$
Output adjusted p-values $\{\tilde{p}_n^{\mathrm{cmp}}(S)\}_{S\in\mathcal{H}}$
$n_3 \leftarrow |\mathcal{I}_3|, \quad \mathcal{H}^{\neq} \leftarrow \{(S,T)\in\mathcal{H}\times\mathcal{H} \mid S\neq T\}$
for $(S,T)\in\mathcal{H}^{\neq}$ **do**
    For $i\in\mathcal{I}_3$ set $d_i(S,T) \leftarrow (\widehat{r}_{S,i}^{\mathcal{I}_3})^2 - (\widehat{r}_{T,i}^{\mathcal{I}_3})^2$
    $\bar{d}(S,T) \leftarrow n_3^{-1}\sum_{i\in\mathcal{I}_3} d_i(S,T)$
    $s^2(S,T) \leftarrow (n_3-1)^{-1}\sum_{i\in\mathcal{I}_3}(d_i(S,T)-\bar{d}(S,T))^2$
    For $i\in\mathcal{I}_3$ set $d_i^{\circ}(S,T) \leftarrow d_i(S,T) - \bar{d}(S,T)$
    $T_n(S,T) \leftarrow \sqrt{n_3}\,\bar{d}(S,T)/s(S,T)$ ▷ one-sided, larger means S is worse than T
end for
Order $\{T_n(S,T)\}_{(S,T)\in\mathcal{H}^{\neq}}$ decreasing:

$$T_{(1)} \geqslant \cdots \geqslant T_{(m)}$$

with ordered pairs $(S,T)_{(1)},\ldots,(S,T)_{(m)}$, where $m \leftarrow |\mathcal{H}^{\neq}|$
for $b\leftarrow 1$ **to** B_n **do** ▷ common Rademacher sign-flips
    Draw i.i.d. $\{\xi_i^{(b)}\}_{i\in\mathcal{I}_3}$ with $\mathbb{P}(\xi_i^{(b)}=\pm 1)=1/2$
    for $(S,T)\in\mathcal{H}^{\neq}$ **do**
        $\bar{d}_b(S,T) \leftarrow n_3^{-1}\sum_{i\in\mathcal{I}_3}\xi_i^{(b)}\,d_i^{\circ}(S,T)$
        $s_b^2(S,T) \leftarrow (n_3-1)^{-1}\sum_{i\in\mathcal{I}_3}\big(\xi_i^{(b)}d_i^{\circ}(S,T) - \bar{d}_b(S,T)\big)^2$
        $T_b(S,T) \leftarrow \sqrt{n_3}\,\bar{d}_b(S,T)/s_b(S,T)$
    end for
end for
Initialize $\tilde{p}_{n,\mathrm{pair}}^{\mathrm{cmp}}(S,T)\leftarrow 0$ for all $(S,T)\in\mathcal{H}^{\neq}$
for $j\leftarrow 1$ **to** m **do**
    $M_b^{(j)} \leftarrow \max_{\ell\in\{j,\ldots,m\}} T_b\big((S,T)_{(\ell)}\big)$ for $b\in\{1,\ldots,B_n\}$
    $\tilde{p}_{\mathrm{raw,pair}}^{\mathrm{cmp}}\big((S,T)_{(j)}\big) \leftarrow \dfrac{1+\#\{b \mid M_b^{(j)} \geqslant T_{(j)}\}}{1+B_n}$
    $\tilde{p}_{n,\mathrm{pair}}^{\mathrm{cmp}}\big((S,T)_{(j)}\big) \leftarrow \max\Big\{\tilde{p}_{n,\mathrm{pair}}^{\mathrm{cmp}}\big((S,T)_{(j-1)}\big),\ \tilde{p}_{\mathrm{raw,pair}}^{\mathrm{cmp}}\big((S,T)_{(j)}\big)\Big\}$ with $\tilde{p}_{n,\mathrm{pair}}^{\mathrm{cmp}}\big((S,T)_{(0)}\big)\leftarrow 0$
end for
for $S\in\mathcal{H}$ **do**
    if $\mathcal{H}\setminus\{S\}=\varnothing$ **then**
        $\tilde{p}_n^{\mathrm{cmp}}(S)\leftarrow 1$
    else
        $\tilde{p}_n^{\mathrm{cmp}}(S) \leftarrow \min_{T\in\mathcal{H}\setminus\{S\}}\tilde{p}_{n,\mathrm{pair}}^{\mathrm{cmp}}(S,T)$
    end if
end for
return $\{\tilde{p}_n^{\mathrm{cmp}}(S) \mid S\in\mathcal{H}\}$

For `IndRM-TS`, our implementation standardizes each component of S by its standard deviation on $\mathcal{I}_1$, standardizes the residuals by the standard deviation of Y on $\mathcal{I}_1$, and uses Gaussian kernels with fixed bandwidths,

$$k_S(s,s') = \exp\Big\{-\frac{\|\widehat{D}_S^{-1}(s-s')\|^2}{2|S|}\Big\}, \qquad \ell_S(e,e') = \exp\Big\{-\frac{(e-e')^2}{2\widehat{\sigma}_Y^2}\Big\},$$

where $\widehat{D}_S$ is the diagonal matrix of the training-fold standard deviations of the components of S and $\widehat{\sigma}_Y$ is the training-fold standard deviation of Y. Gaussian kernels are bounded, characteristic, and Lipschitz in each argument, and on compact subsets of $(0,\infty)$ they are Lipschitz in their scales, uniformly in their arguments. If all these variances are positive and $\mathbb{E}\,\|S\|^4<\infty$ for every candidate,

the estimated scales converge at rate $n_1^{-1/2}$. Under proportional splitting, Assumption 5(iii) therefore holds with $b_n = n^{-1/2}$.

Algorithm 4 `IndRM-TS`

Input candidate family $\mathcal{C} \ni \emptyset$, splits $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$, threshold h_n, tolerance λ_n
for $S \in \mathcal{C}$ **do**
    Fit $\widehat{m}_S$ on $\mathcal{I}_1$; estimate and freeze the kernel parameters of $\widehat{H}_n(S)$ on $\mathcal{I}_1$
    If $S \neq \emptyset$, compute $\widehat{H}_n(S)$ from $\{(S_i, Y_i - \widehat{m}_S(S_i))\}_{i \in \mathcal{I}_2}$
    Compute $\widehat{R}_n(S) \leftarrow n_3^{-1} \sum_{i \in \mathcal{I}_3} \{Y_i - \widehat{m}_S(S_i)\}^2$
end for
$\widehat{\mathcal{R}}_n^{\mathrm{thr}} \leftarrow \{\emptyset\} \cup \{S \in \mathcal{C} : S \neq \emptyset,\ \widehat{H}_n(S) \leqslant h_n\}$
$\widehat{\mathcal{R}}_{n,\lambda}^{\min} \leftarrow \{S \in \widehat{\mathcal{R}}_n^{\mathrm{thr}} : \widehat{R}_n(S) \leqslant \min_{T \in \widehat{\mathcal{R}}_n^{\mathrm{thr}}} \widehat{R}_n(T) + \lambda_n\}$
Output $\widehat{\mathrm{PA}}_n^{\mathrm{thr}}(Y) \leftarrow \arg\min_{S \in \widehat{\mathcal{R}}_{n,\lambda}^{\min}} |S|$

B Proofs

B.1 Proof of Lemma 3.1

Under Setting 1, every parent of Y is a non-descendant of Y. In an acyclic SCM, each variable is a measurable function of the exogenous noises of itself and its ancestors, and Y is not an ancestor of any non-descendant of Y. Hence $\mathrm{PA}(Y)$ is a function of noises other than ϵ, and joint independence of the noises gives $\epsilon \perp\!\!\!\perp \mathrm{PA}(Y)$. Together with $\mathbb{E}[\epsilon] = 0$, this yields

$$m_{\mathrm{PA}(Y)}(\mathrm{PA}(Y)) = \mathbb{E}[Y \mid \mathrm{PA}(Y)] = f(\mathrm{PA}(Y)).$$

Hence

$$r_{\mathrm{PA}(Y)} = Y - m_{\mathrm{PA}(Y)}(\mathrm{PA}(Y)) = Y - f(\mathrm{PA}(Y)) = \epsilon.$$

Since $\epsilon \perp\!\!\!\perp \mathrm{PA}(Y)$, we have $\mathrm{PA}(Y) \in \mathcal{R}$. Moreover,

$$\mathbb{E}[r_{\mathrm{PA}(Y)}^2] = \mathbb{E}[\epsilon^2].$$

B.2 Proof of Lemma 3.2

Assume Setting 1, and let $S \subseteq \mathcal{S}$ satisfy

$$S \cap \mathrm{Desc}(Y) = \emptyset.$$

Because $S \cap \mathrm{Desc}(Y) = \emptyset$, the structural noise ϵ is independent of $(\mathrm{PA}(Y), S)$. Indeed, in the extended graph that includes the node ϵ and the edge $\epsilon \to Y$, the node ϵ is d-separated from the non-descendants of Y. This follows from the SCM Markov property.

Since $\mathbb{E}[\epsilon] = 0$ and $\epsilon \perp\!\!\!\perp (\mathrm{PA}(Y), S)$, we have

$$m_S(S) = \mathbb{E}[Y \mid S] = \mathbb{E}[f(\mathrm{PA}(Y)) + \epsilon \mid S] = \mathbb{E}[f(\mathrm{PA}(Y)) \mid S].$$

Therefore

$$r_S = Y - m_S(S) = f(\mathrm{PA}(Y)) - \mathbb{E}[f(\mathrm{PA}(Y)) \mid S] + \epsilon.$$

Consequently, the cross term has expectation zero, and

$$\mathbb{E}[r_S^2] = \mathbb{E}\Big[(f(\mathrm{PA}(Y)) - \mathbb{E}[f(\mathrm{PA}(Y)) \mid S])^2\Big] + \mathbb{E}[\epsilon^2].$$

By Lemma 3.1,

$$\mathbb{E}[r_{\mathrm{PA}(Y)}^2] = \mathbb{E}[\epsilon^2].$$

Hence

$$\mathbb{E}[r_S^2] \geqslant \mathbb{E}[r_{\mathrm{PA}(Y)}^2].$$

If $\mathbb{E}[r_S^2] = \mathbb{E}[r_{\mathrm{PA}(Y)}^2]$, the decomposition gives $f(\mathrm{PA}(Y)) = g(S)$ almost surely, where $g(S) := \mathbb{E}[f(\mathrm{PA}(Y)) \mid S]$. Suppose that some $Z \in \mathrm{PA}(Y)$ does not belong to S. Then $Y = g(S) + \epsilon$ almost surely, and $\epsilon \perp\!\!\!\perp (Z, S)$ because $\epsilon \perp\!\!\!\perp (\mathrm{PA}(Y), S)$. Hence $Y \perp\!\!\!\perp Z \mid S$. But $Z \to Y$ is an edge of G and $Z, Y \notin S$, so Z and Y are d-connected given S, and faithfulness of P^V to G gives $Y \not\perp\!\!\!\perp Z \mid S$, a contradiction. Therefore $\mathrm{PA}(Y) \subseteq S$, and $|S| > |\mathrm{PA}(Y)|$ whenever $S \neq \mathrm{PA}(Y)$.

B.3 Proof of Proposition 3.4

Fix a candidate set $S \subseteq \mathcal{S}$ satisfying

$$S \cap \mathrm{Desc}(Y) \neq \emptyset.$$

If

$$r_{S,\theta} \perp\!\!\!\perp S,$$

then for every admissible pair of test functions φ, ψ,

$$\mathrm{Cov}_\theta(\varphi(S), \psi(r_{S,\theta})) = 0.$$

In particular, for the test function pair (φ_S, ψ_S) from Assumption 2,

$$r_{S,\theta} \perp\!\!\!\perp S \quad \Longrightarrow \quad F_S(\theta) = 0.$$

Therefore

$$\{\theta \in \Theta : r_{S,\theta} \perp\!\!\!\perp S\} \subseteq \{\theta \in \Theta : F_S(\theta) = 0\} =: Z_S.$$

By Assumption 2, F_S is real analytic on Θ and is not identically zero. Since Θ is open and connected, the zero set of a nonzero real analytic function has Lebesgue measure zero. Moreover, Z_S is relatively closed in Θ by continuity and has empty interior in Θ. Hence

$$\lambda(\{\theta \in \Theta : F_S(\theta) = 0\}) = 0,$$

where λ denotes Lebesgue measure on $\mathbb{R}^q$. Consequently,

$$\lambda(\{\theta \in \Theta : r_{S,\theta} \perp\!\!\!\perp S\}) = 0.$$

There are only finitely many candidate sets $S \subseteq \mathcal{S}$. Let Z be the union of Z_S over all descendant-containing candidates. This finite union is relatively closed in Θ, nowhere dense, and Lebesgue-null, and it contains every parameter at which the no-backward condition fails. Thus $\Theta \setminus Z$ is open and dense in Θ and has full Lebesgue measure in Θ. Thus, if θ is drawn from any distribution on Θ that is absolutely continuous with respect to Lebesgue measure, then almost surely

$$r_{S,\theta} \not\perp\!\!\!\perp S \qquad \text{for all } S \subseteq \mathcal{S} \text{ with } S \cap \mathrm{Desc}(Y) \neq \emptyset.$$

This is the no-backward condition in Assumption 1. Finally, suppose that descendant-containing candidates exist, and let $K \subseteq \Theta \setminus Z$ be nonempty and compact. For every such S, continuity and nonvanishing of F_S on K give $\min_{\theta \in K} |F_S(\theta)| > 0$. Taking the minimum of these finitely many positive numbers gives the constant $c_K > 0$.

B.4 A nonlinear backward ANM

Example 1 (Nonlinearity does not exclude backward ANMs)**.** *Let $C \sim \mathcal{N}(0,1)$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, and $\eta \sim \mathcal{N}(0, \tau^2)$ be mutually independent, with $a \neq 0$ and $\sigma, \tau > 0$. The SCM*

$$Y = C^3 + \epsilon, \qquad D = aY + \eta$$

is faithful to $C \to Y \to D$, but $r_{\{C,D\}} \perp\!\!\!\perp (C, D)$ and

$$\mathbb{E}[r^2_{\{C,D\}}] = \frac{\sigma^2\tau^2}{a^2\sigma^2 + \tau^2} < \sigma^2 = \mathbb{E}[r^2_{\{C\}}].$$

Thus a set containing a descendant passes residual independence and has lower risk than the true parent set $\{C\}$.

Proof. Set $W = D - aC^3 = a\epsilon + \eta$ and $\lambda = a\sigma^2/(a^2\sigma^2 + \tau^2)$. Gaussian conditioning gives

$$\mathbb{E}[Y \mid C, D] = (1 - a\lambda)C^3 + \lambda D, \qquad r_{\{C,D\}} = (1 - a\lambda)\epsilon - \lambda\eta.$$

The pair $(r_{\{C,D\}}, W)$ is jointly Gaussian with zero covariance, and C is independent of this pair. Hence $r_{\{C,D\}} \perp\!\!\!\perp (C, W)$, and therefore $r_{\{C,D\}} \perp\!\!\!\perp (C, D)$. Its variance is $\sigma^2\tau^2/(a^2\sigma^2 + \tau^2) < \sigma^2$.

For faithfulness, the chain implies $C \perp\!\!\!\perp D \mid Y$. The covariances

$$\mathrm{Cov}(C, Y) = 3, \qquad \mathrm{Cov}(C, D) = 3a, \qquad \mathrm{Cov}(Y, D) = a\,\mathrm{Var}(Y)$$

are nonzero, so no pair is marginally independent. Also $\mathrm{Cov}(Y, D \mid C) = a\sigma^2 \neq 0$. Finally, $1 - a\lambda = \tau^2/(a^2\sigma^2 + \tau^2) > 0$, so $\mathbb{E}[Y \mid C, D]$ depends on C. Since (C, D) has a positive joint density, this excludes $C \perp\!\!\!\perp Y \mid D$. Thus there are no additional conditional independences. □

B.5 Impossibility with unobserved parents

Proposition B.1 (Impossibility with unobserved parents)**.** *If parents of Y may be unobserved, the observed distribution need not determine whether an observed variable is a direct parent of Y, even within faithful linear-Gaussian SCMs satisfying the no-backward condition in Assumption 1.*

Proof. Fix $\rho \in (0,1)$. Consider two SCMs with observed variables (Z, Y):

$$\begin{aligned} &\text{A:} \quad Z = N_Z, \quad Y = \rho Z + \sqrt{1-\rho^2}\, N_Y, \\ &\text{B:} \quad L = N_L, \quad Z = L + U_Z, \qquad Y = L + U_Y, \end{aligned}$$

where the exogenous noises are independent, with $N_Z, N_Y \sim \mathcal{N}(0,1)$, $N_L \sim \mathcal{N}(0,\rho)$, and $U_Z, U_Y \sim \mathcal{N}(0, 1-\rho)$. Both observed vectors are centered Gaussian with unit variances and covariance ρ, so their laws agree. In A, Z is a parent of Y; in B, the unobserved L is the only parent of Y, and Z is a proxy.

Model A is faithful to $Z \to Y$ because $\rho \neq 0$. In B, all three marginal covariances equal $\rho > 0$, while

$$\mathrm{Cov}(Z, Y \mid L) = 0, \qquad \mathrm{Cov}(L, Z \mid Y) = \mathrm{Cov}(L, Y \mid Z) = \rho(1-\rho) > 0.$$

Joint Gaussianity therefore implies faithfulness to $Z \leftarrow L \to Y$. The observed candidate Z is not a descendant of Y in either model, so the no-backward condition holds vacuously. The identical observed laws and different direct-parent status prove the claim. □

For the common observed law, Gaussian regression gives $r_{\{Z\}} = Y - \rho Z \perp\!\!\!\perp Z$, whereas $r_\emptyset = Y$. Their risks satisfy $\mathbb{E}[r^2_{\{Z\}}] = 1 - \rho^2 < 1 = \mathbb{E}[r^2_\emptyset]$. Thus, with $\mathcal{S} = \{Z\}$, `IndRM` selects $\{Z\}$ in both models: a parent in A and a proxy in B.

B.6 Proofs of Proposition 3.7 and Proposition 3.8

Both propositions concern the bivariate family

$$Y = N_Y, \qquad X_b = b\, h(Y) + N_X, \qquad b \in I,$$

with $N_Y \perp\!\!\!\perp N_X$, centered N_X with finite variance, and $h \in L^2(P_Y)$; Proposition 3.7 uses $h(y) = y$. Write $\mu := \mathbb{E}[Y]$, $m_b(x) := \mathbb{E}[Y \mid X_b = x]$, $r_b := Y - m_b(X_b)$, and, for $t \in \mathbb{R}$,

$$G_t(b) := \mathrm{Cov}\big(X_b, e^{itr_b}\big).$$

If $r_b \perp\!\!\!\perp X_b$, then $G_t(b) = 0$ for every t.

Assumption 3 (Regularity of the bivariate witnesses)**.** *N_X has a strictly positive, continuously differentiable density p with $\int |x|\,|p'(x)|\,dx < \infty$ and $xp(x) \to 0$ as $|x| \to \infty$. Moreover, for every $t \in \mathbb{R}$,*

- *(a) the real and imaginary parts of G_t are well defined on I and real analytic on each connected component of $I \setminus \{0\}$;*
- *(b) the map $(b,x) \mapsto m_b(x)$ is continuously differentiable on a neighborhood of $\{0\} \times \mathbb{R}$, and $\partial_b m_b(x)|_{b=0}$ may be computed by differentiating the numerator and the denominator of*

$$m_b(x) = \frac{\mathbb{E}[Y\, p(x - bh(Y))]}{\mathbb{E}[p(x - bh(Y))]}$$

  under the expectation sign;
- *(c) G_t is differentiable at $b = 0$, and $G_t'(0)$ may be computed by differentiating under the expectation sign, viewing X_b and r_b as functions of (Y, N_X).*

Lemma B.5 gives sufficient conditions for (a). Analyticity is required only away from the baseline $b = 0$; at $b = 0$, only differentiability is used.

Lemma B.2 (First-order witness)**.** *Under Assumption 3, for every $t \in \mathbb{R}$, we have $G_t(0) = 0$ and*

$$G_t'(0) = \mathrm{Cov}\big(h(Y), e^{it(Y-\mu)}\big) - it\,\mathrm{Cov}\big(h(Y), Y\big)\ \mathbb{E}\big[e^{it(Y-\mu)}\big].$$

Proof. At $b = 0$, $X_0 = N_X$ is independent of Y, so $m_0 \equiv \mu$, $r_0 = Y - \mu$, and $G_t(0) = \mathrm{Cov}(N_X, e^{it(Y-\mu)}) = 0$. Write $s := p'/p$. At $b = 0$, the numerator and the denominator of $m_b(x)$ equal $\mu p(x)$ and $p(x)$, and their derivatives in b equal $-\mathbb{E}[Y h(Y)]\, p'(x)$ and $-\mathbb{E}[h(Y)]\, p'(x)$. The quotient rule gives $\partial_b m_b(x)|_{b=0} = -\mathrm{Cov}(Y, h(Y))\, s(x)$. Since m_0 is constant, the chain rule gives $\partial_b r_b|_{b=0} = \mathrm{Cov}(Y, h(Y))\, s(N_X)$. Differentiating $G_t(b) = \mathbb{E}[X_b e^{itr_b}] - \mathbb{E}[X_b]\, \mathbb{E}[e^{itr_b}]$ at $b = 0$, and using $\partial_b X_b = h(Y)$, $X_0 = N_X$, $\mathbb{E}[N_X] = 0$, and the independence of Y and N_X, we obtain

$$G_t'(0) = \mathrm{Cov}\big(h(Y), e^{it(Y-\mu)}\big) + it\, \mathrm{Cov}\big(Y, h(Y)\big)\, \mathbb{E}\big[e^{it(Y-\mu)}\big]\, \mathbb{E}[N_X s(N_X)].$$

Integrating by parts over $[-R, R]$ and letting $R \to \infty$ gives $\mathbb{E}[N_X s(N_X)] = \int x p'(x)\, dx = -1$, because $xp(x) \to 0$ and $\int |x|\, |p'(x)|\, dx < \infty$. □

Proof of Proposition 3.7. Case (i). Here $h(y) = y$ and $\mu = 0$. With $v := \mathrm{Var}(Y)$ and $\phi_Y(t) := \mathbb{E}[e^{itY}]$, Lemma B.2 gives $G_t'(0) = \mathbb{E}[Y e^{itY}] - it\, v\, \phi_Y(t)$. If $G_t'(0) = 0$ for every t, then, since $\phi_Y'(t) = i\, \mathbb{E}[Y e^{itY}]$, the characteristic function would solve $\phi_Y'(t) = -vt\, \phi_Y(t)$ with $\phi_Y(0) = 1$, so $\phi_Y(t) = \exp(-vt^2/2)$ and Y would be Gaussian. Hence $G_{t_0}'(0) \neq 0$ for some t_0, and one of

$$F(a) = \mathrm{Cov}\big(X_a, \cos(t_0 r_a)\big), \qquad F(a) = \mathrm{Cov}\big(X_a, \sin(t_0 r_a)\big),$$

the real and imaginary parts of G_{t_0}, satisfies $F(0) = 0$ and $F'(0) \neq 0$. Therefore $F(a) \neq 0$ for all sufficiently small $a \neq 0$, so F is not identically zero on either connected component J of $I \setminus \{0\}$, and it is real analytic on J by Assumption 3(a). Since $r_a \perp\!\!\!\perp X_a$ implies $F(a) = 0$, Assumption 2 holds on J for $S = \{X_a\}$, and Proposition 3.4 gives the conclusion.

Case (ii). Here $Y \sim \mathcal{N}(0, v)$ and N_X is non-Gaussian; no regularity condition is needed. Fix $a \neq 0$ and write $X_a = G + N_X$ with $G := aY \sim \mathcal{N}(0, \tau^2)$, $\tau^2 := a^2 v$, independent of N_X. With φ_τ the $\mathcal{N}(0, \tau^2)$ density, the density $p_{X_a}(x) = \mathbb{E}[\varphi_\tau(x - N_X)]$ of X_a is smooth and strictly positive, and differentiating under the expectation gives Tweedie's formula [6]

$$\mathrm{Var}(G \mid X_a = x) = \tau^2 + \tau^4 (\log p_{X_a})''(x).$$

Hence $\mathbb{E}[r_a^2 \mid X_a = x] = \mathrm{Var}(Y \mid X_a = x) = v + a^2 v^2 (\log p_{X_a})''(x)$. If $r_a \perp\!\!\!\perp X_a$, the left-hand side is constant for almost every x, and hence, by continuity, for every x. Thus $\log p_{X_a}$ is a quadratic polynomial, and integrability of p_{X_a} forces X_a to be Gaussian. By Cramér's decomposition theorem [5], the independent summands G and N_X would then both be Gaussian, a contradiction. Hence $r_a \not\perp\!\!\!\perp X_a$ for every $a \neq 0$. □

Proof of Proposition 3.8. Here $Y \sim \mathcal{N}(0, v)$ with $v > 0$, and N_X is centered Gaussian, so the density condition in Assumption 3 holds. Suppose that $G_t'(0) = 0$ for every t. Gaussian integration by parts gives $\mathbb{E}[Y e^{itY}] = it\, v\, \mathbb{E}[e^{itY}]$, so Lemma B.2 yields $\mathbb{E}\big[\{h(Y) - \mathbb{E}\, h(Y) - cY\} e^{itY}\big] = 0$ for every t, where $c := \mathrm{Cov}(h(Y), Y)/v$. The finite signed measure with density $h(y) - \mathbb{E}\, h(Y) - cy$ with respect to P_Y therefore has vanishing Fourier transform, so $h(Y) = \mathbb{E}\, h(Y) + cY$ almost surely, which contradicts $h \notin \mathrm{span}\{1, Y\}$. Hence $G_{t_0}'(0) \neq 0$ for some t_0, and the argument for case (i) of Proposition 3.7 applies verbatim. The proof uses Gaussianity only for N_Y; for N_X, the density condition in Assumption 3 suffices. □

B.7 Proof of Theorem 3.6

We first record two consequences of modularity. A pair $(g, Q) \in \mathcal{A}_1$ induces the transition kernel $K(u, \cdot) = \mathrm{law}(g(u, N))$, $N \sim Q$, and a pair in $\mathcal{A}_0^Y$ induces a law Q_0 of Y. Hence, for a chain $H = (Y \to v_1 \to \cdots \to v_k)$, the family $\mathcal{M}_H$ consists exactly of the laws $Q_0 \otimes K_1 \otimes \cdots \otimes K_k$ of Markov chains whose initial law is induced by $\mathcal{A}_0^Y$ and whose transition kernels are induced by $\mathcal{A}_1$, chosen independently.

Lemma B.3 (Iterated latent insertion)**.** *Assume that the modular SCM class satisfies the latent extension property. Then, for every $k \geqslant 1$ and every chain $H = (Y \to v_1 \to \cdots \to v_k)$,*

$$\mathcal{M}_{Y \to v_k} \subseteq \mathrm{Marg}_{Y, v_k}(\mathcal{M}_H).$$

Proof. Fix $P \in \mathcal{M}_{Y \to X}$. We show by induction on k that some $P_k \in \mathcal{M}_{H_k}$, where $H_k = (Y \to v_1 \to \cdots \to v_k)$ with $v_k = X$, has (Y, X)-marginal P. The case $k = 1$ is trivial. Let $P_k = Q_0 \otimes K_1 \otimes \cdots \otimes K_k \in \mathcal{M}_{H_k}$ have (Y, X)-marginal P. Its (Y, v_1)-marginal $Q_0 \otimes K_1$ lies

in $\mathcal{M}_{Y\to v_1}$. By the latent extension property, some $Q_0' \otimes K' \otimes K'' \in \mathcal{M}_{Y\to w\to v_1}$ has the same (Y, v_1)-marginal. Hence $Q_0' = Q_0$, and $\int K'(y, dw)\, K''(w, \cdot) = K_1(y, \cdot)$ for Q_0-almost every y, by uniqueness of regular conditional distributions. By modularity,

$$P_{k+1} \coloneqq Q_0 \otimes K' \otimes K'' \otimes K_2 \otimes \cdots \otimes K_k$$

belongs to $\mathcal{M}_{Y\to w\to v_1\to\cdots\to v_k}$, and marginalizing over w yields P_k. In particular, the (Y, X)-marginal of P_{k+1} is P. Relabelling the nodes completes the induction step. □

Lemma B.4 (Extension from a path to G)**.** *Let π be a directed path in G starting at Y, let H_π be the DAG consisting of the nodes and edges of π, and let W_π be its node set. Then every $P \in \mathcal{M}_{H_\pi}$ is the W_π-marginal of some $P_{G,\theta} \in \mathcal{M}_G$ under which the variables outside W_π are mutually independent and independent of the variables in W_π.*

Proof. Let P be induced by assignments (g_v, Q_v), $v \in W_\pi$, over H_π. Define an SCM over G with jointly independent noises as follows. Each $v \in W_\pi$ keeps its noise law and applies g_v to its parent in H_π, if any, ignoring all other parents in G. Each $Z \notin W_\pi$ is assigned $Z \coloneqq g_0(N_Z)$, $N_Z \sim Q$, for a fixed $(g_0, Q) \in \mathcal{A}_0$, ignoring all its parents in G. By closure under permuting and ignoring parents, all assignments are admissible, so the induced law lies in $\mathcal{M}_G$. The variables in W_π are computed from $(N_v)_{v\in W_\pi}$ exactly as in H_π, so their joint law is P, and each $Z \notin W_\pi$ is a function of N_Z alone. □

Fix $S \subseteq \mathcal{S}$ with $S \cap \mathrm{Desc}(Y) \neq \emptyset$. Choose $X \in S \cap \mathrm{Desc}(Y)$ with minimal directed distance from Y, and let

$$\pi : \quad Y = V_0 \to V_1 \to \cdots \to V_k = X$$

be a shortest directed path from Y to X. By minimality of X, none of the intermediate variables $V_1, \ldots, V_{k-1}$ belongs to S.

Let $P^\star_{Y,X} \in \mathcal{M}_{Y\to X}$ be the bivariate witness, so that for some test functions φ, ψ,

$$\mathrm{Cov}_{P^\star}(\varphi(X), \psi(Y - \mathbb{E}_{P^\star}[Y \mid X])) \neq 0.$$

By Lemma B.3, $P^\star_{Y,X}$ can be realized as the (Y, X)-marginal of a distribution $P_\pi \in \mathcal{M}_{H_\pi}$, where H_π is the DAG consisting of the nodes and edges of π and W_π is its node set.

By Lemma B.4, P_π extends to $P_{G,\theta_S} \in \mathcal{M}_G$, with the off-path variables mutually independent and independent of the path. All remaining expectations and covariances without a subscript are under P_{G,θ_S}. Since $S \cap W_\pi = \{X\}$, the vector $S \setminus \{X\}$ is independent of (Y, X). Hence

$$\mathbb{E}[Y \mid S] = \mathbb{E}[Y \mid X].$$

Consequently,

$$Y - \mathbb{E}[Y \mid S] = Y - \mathbb{E}[Y \mid X].$$

Define

$$\tilde{\varphi}(S) \coloneqq \varphi(X),$$

where X denotes the coordinate of S corresponding to the chosen descendant. Then

$$\begin{aligned}\mathrm{Cov}(\tilde{\varphi}(S), \psi(Y - \mathbb{E}[Y \mid S])) &= \mathrm{Cov}(\varphi(X), \psi(Y - \mathbb{E}[Y \mid X])) \\ &= \mathrm{Cov}_{P^\star}(\varphi(X), \psi(Y - \mathbb{E}_{P^\star}[Y \mid X])) \neq 0.\end{aligned}$$

Thus, $\mathcal{M}_G$ admits a witness for S. Since S was arbitrary, a bivariate witness yields a witness for every descendant-containing set.

B.8 Analyticity of residual-dependence of $F(\theta)$

The genericity argument in Section 3.1 assumes that the residual-dependence function

$$F_S(\theta) = \mathrm{Cov}_\theta(\varphi(S), \psi(r_{S,\theta}))\,, \qquad r_{S,\theta} = Y - \mathbb{E}_\theta[Y \mid S],$$

is analytic in the SCM parameter θ. This subsection records a sufficient condition: the integrands extend holomorphically in the parameter and have an integrable envelope. This condition must be checked for the chosen model family and test functions.

Assumption 4 (Holomorphic integrands)**.** *Let $\Theta \subseteq \mathbb{R}^q$ be open, fix $S \subseteq \mathcal{S}$ and test functions φ, ψ, and suppose that, for every $\theta \in \Theta$, the joint law of (Y, S) has a density p_θ with respect to a fixed σ-finite product measure $\mu = \mu_Y \otimes \mu_S$. Define*

$$g_{1,\theta}(y,s) = \varphi(s)p_\theta(y,s), \qquad g_{2,\theta}(y,s) = \psi\big(y - m_{S,\theta}(s)\big)p_\theta(y,s),$$
$$g_{3,\theta}(y,s) = \varphi(s)g_{2,\theta}(y,s).$$

For every $\theta_0 \in \Theta$, there exist an open neighborhood $U \subseteq \mathbb{C}^q$ of θ_0 with $U \cap \mathbb{R}^q \subseteq \Theta$ and nonnegative $M_1, M_2, M_3 \in L^1(\mu)$ such that, for μ-almost every (y, s) and each j, the map $\theta \mapsto g_{j,\theta}(y,s)$ on $U \cap \mathbb{R}^q$ extends to a holomorphic function on U, the extension is jointly measurable in (θ, y, s), and $|g_{j,\theta}(y,s)| \leqslant M_j(y,s)$ for all $\theta \in U$.

Lemma B.5 (Analyticity of residual-dependence function)**.** *Assume Assumption 4 for test functions φ and ψ. Then*

$$\theta \mapsto F_S(\theta) = \mathrm{Cov}_\theta(\varphi(S), \psi(r_{S,\theta}))$$

is real analytic.

Proof. Fix θ_0 and take U and M_j from Assumption 4; shrinking U, we may assume that it is a polydisc. Put $G_j(\theta) \coloneqq \int g_{j,\theta}\, d\mu$ for $\theta \in U$. By dominated convergence, G_j is continuous on U. For each coordinate θ_ℓ and each closed triangle Δ in the corresponding coordinate disc, Fubini's theorem and Goursat's theorem give $\oint_{\partial\Delta} G_j\, d\theta_\ell = \int \oint_{\partial\Delta} g_{j,\theta}\, d\theta_\ell\, d\mu = 0$, so G_j is holomorphic in each variable separately by Morera's theorem. A continuous function that is holomorphic in each variable separately is holomorphic (Osgood's lemma), and the restriction of a holomorphic function to $U \cap \mathbb{R}^q$ is real analytic. Hence G_1, G_2, G_3, and $F_S = G_3 - G_1G_2$ are real analytic near θ_0. □

B.9 Proofs for Section 4.2

Assumption 5 (Conditions for consistency)**.** *Assume the following.*

- *(i) The candidate family $\mathcal{C}$ is fixed, finite, and contains $\emptyset$. The partition is drawn independently of the observations, and $n_j/n \to \pi_j \in (0,1)$ for $j \in \{1,2,3\}$.*
- *(ii) $\mathbb{E}[Y^4] < \infty$.*
- *(iii) For each nonempty $S \in \mathcal{C}$, the statistic $\widehat{H}_n(S)$ uses kernels $k_{S,\eta}$ on $\mathbb{R}^{|S|}$ and $\ell_{S,\eta}$ on $\mathbb{R}$ with a finite-dimensional parameter η, such as bandwidths or scales. The parameter is prespecified or estimated from $\mathcal{I}_1$ only, and the estimate $\widehat{\eta}_S$ satisfies $\|\widehat{\eta}_S - \eta_S\| = O_p(b_n)$ for a limit η_S and a deterministic $b_n \downarrow 0$, where $b_n = 0$ for prespecified parameters. On a neighborhood of η_S, the kernels take values in $[0,1]$ and are Lipschitz in η uniformly in their arguments, and $\ell_{S,\eta}$ is Lipschitz in each argument with a common constant. The limiting kernels $k_S \coloneqq k_{S,\eta_S}$ and $\ell_S \coloneqq \ell_{S,\eta_S}$ are characteristic.*
- *(iv) For a deterministic $r_n \downarrow 0$, $\rho_n \coloneqq \max_{S \in \mathcal{C}} \mathbb{E}[\{\widehat{m}_S(S) - m_S(S)\}^2 \mid \mathcal{I}_1] = O_p(r_n)$, where the expectation is over an observation independent of $\mathcal{I}_1$.*
- *(v) $h_n \downarrow 0$ and $n_2^{-1/2} + \sqrt{r_n} + b_n = o(h_n)$.*
- *(vi) $\lambda_n \downarrow 0$ and $n_3^{-1/2} + \sqrt{r_n} = o(\lambda_n)$. A data-dependent tolerance $\widehat{\lambda}_n$ may be used instead if $\widehat{\lambda}_n/\lambda_n \to 1$ in probability for such a deterministic λ_n.*

For nonempty S, let $H(S) \coloneqq \mathrm{HSIC}_{k_S,\ell_S}(S, r_S)$ denote the population `HSIC` with the limiting kernels; it vanishes if and only if $r_S \perp\!\!\!\perp S$ [37, Theorem 3]. Lemma B.6 in Appendix B.9 shows that the regression error enters both $\widehat{H}_n$ and $\widehat{R}_n$ only through $\sqrt{r_n}$. Consistency of the regression therefore suffices, and proportional splits are allowed; the faster rate $n_2\rho_n \to 0$, which generally excludes them, is not needed. Thresholds and tolerances that vanish more slowly than these errors recover the population screen and its risk minimizers.

Assumption 5(iii) holds, for example, for Gaussian, Laplace, and Matérn kernels with smoothness $\nu \geqslant 1/2$ whose length scales are fixed or proportional to training-fold standard deviations: these kernels take values in $[0,1]$, are characteristic, and are Lipschitz in their arguments and in the length scale on compact subsets of $(0,\infty)$. With estimated scales, $b_n = n^{-1/2}$ if the variances are positive

and the fourth moments finite (Appendix A). Corollary B.7 in Appendix B.9 verifies Assumption 5(iv) for kernel ridge regression whose RKHS is a Sobolev space of smoothness $s > d_{\max}/2$, for example, that of a Matérn kernel, where $d_{\max}$ is the largest candidate size. Then $r_n = n^{-2s/(2s+d_{\max})}$, so equal or other proportional splits suffice, and thresholds of order $n^{-\gamma}$ with $\gamma < s/(2s + d_{\max})$ are admissible, including the rate $n^{-1/4}$ used in our experiments. The required regularization grows with n_1, so the fixed ridge penalty used in our experiments is not covered.

Throughout, $\mathbb{P}_{n_j}$ denotes the empirical average over $\mathcal{I}_j$, and $e_S \coloneqq \widehat{m}_S - m_S$, which is a function determined by $\mathcal{I}_1$. Maxima over S range over $\mathcal{C}$, and over its nonempty members whenever `HSIC` is involved.

We first state and prove the error bounds used in Section 4.2.

Lemma B.6 (Uniform score errors)**.** *Under Assumption 5(i)–(iv),*

$$\max_{\emptyset \neq S \in \mathcal{C}} |\widehat{H}_n(S) - H(S)| = O_p\big(n_2^{-1/2} + \sqrt{r_n} + b_n\big), \qquad \max_{S \in \mathcal{C}} |\widehat{R}_n(S) - R(S)| = O_p\big(n_3^{-1/2} + \sqrt{r_n}\big).$$

The first maximum is defined as zero if $\mathcal{C} = \{\emptyset\}$.

Proof of Lemma B.6. Step 0: a conditional Markov bound. The folds $\mathcal{I}_2$ and $\mathcal{I}_3$ are independent of $\mathcal{I}_1$ and consist of i.i.d. observations. Hence $\mathbb{E}[\mathbb{P}_{n_j} e_S(S)^2 \mid \mathcal{I}_1] = \mathbb{E}[e_S(S)^2 \mid \mathcal{I}_1] \leqslant \rho_n$ for $j \in \{2, 3\}$, and, for every $M > 0$, the conditional Markov inequality and a union bound give

$$\mathbb{P}\Big(\max_S \mathbb{P}_{n_j} e_S^2 > M r_n\Big) \leqslant \mathbb{P}\big(\rho_n > \sqrt{M}\, r_n\big) + \frac{|\mathcal{C}|}{\sqrt{M}}.$$

Since $\rho_n = O_p(r_n)$, it follows that $\max_S \mathbb{P}_{n_j} e_S^2 = O_p(r_n)$.

Step 1: `HSIC`. We use two results of Mooij et al. [20] for the empirical `HSIC` of m pairs: by their Lemma 14, which restates Gretton et al. [11, Theorem 3], it differs from the population `HSIC` by $O_p(m^{-1/2})$ for i.i.d. pairs and kernels with values in $[0, 1]$; by their Lemma 16, for a first kernel bounded by one and a second kernel that is Lipschitz with constant λ in each argument, replacing the second sample $y \in \mathbb{R}^m$ by y' changes it by at most $32\lambda m^{-1/2}\|y - y'\|$. They normalize by $(m-1)^{-2}$ instead of m^{-2}; since the two statistics differ by the factor $((m-1)/m)^2 = 1 - O(m^{-1})$ and are bounded, both results hold for our normalization, the first up to an additional $O(m^{-1})$ term.

For symmetric $A, B \in \mathbb{R}^{m \times m}$, let $V(A, B) \coloneqq m^{-2} \operatorname{tr}(A\mathbf{C}B\mathbf{C})$. Expanding $\mathbf{C} = I_m - m^{-1}\mathbf{1}\mathbf{1}^\top$ shows that V is bilinear with $|V(A, B)| \leqslant 4 \max_{ij} |A_{ij}| \max_{ij} |B_{ij}|$.

Fix a nonempty $S \in \mathcal{C}$, put $m = n_2$, and for $i \in \mathcal{I}_2$ write $r_i \coloneqq Y_i - m_S(S_i)$ and $\widehat{r}_i \coloneqq Y_i - \widehat{m}_S(S_i) = r_i - e_S(S_i)$. Let $\widehat{\mathbf{K}}$ and $\widehat{\mathbf{L}}$ be the Gram matrices of $k_{S,\widehat{\eta}_S}$ on (S_i) and of $\ell_{S,\widehat{\eta}_S}$ on $(\widehat{r}_i)$, so that $\widehat{H}_n(S) = V(\widehat{\mathbf{K}}, \widehat{\mathbf{L}})$. Let $\mathbf{K}$ and $\mathbf{L}^\circ$ be the corresponding matrices with the limiting parameter η_S, let $\mathbf{L}$ be the Gram matrix of ℓ_S on (r_i), and put $\widetilde{H}_n(S) \coloneqq V(\mathbf{K}, \mathbf{L})$. Let c be the Lipschitz constant in the parameter and L_ℓ the Lipschitz constant of ℓ_S in each argument.

(a) *Kernel parameters.* With probability tending to one, $\widehat{\eta}_S$ lies in the neighborhood of Assumption 5(iii). There, bilinearity gives

$$|V(\widehat{\mathbf{K}}, \widehat{\mathbf{L}}) - V(\mathbf{K}, \mathbf{L}^\circ)| \leqslant |V(\widehat{\mathbf{K}} - \mathbf{K}, \widehat{\mathbf{L}})| + |V(\mathbf{K}, \widehat{\mathbf{L}} - \mathbf{L}^\circ)| \leqslant 8c\, \|\widehat{\eta}_S - \eta_S\| = O_p(b_n).$$

(b) *Fitted residuals.* By Lemma 16 of Mooij et al. [20],

$$|V(\mathbf{K}, \mathbf{L}^\circ) - \widetilde{H}_n(S)| \leqslant 32 L_\ell\, m^{-1/2} \big\| \big(e_S(S_i)\big)_{i \in \mathcal{I}_2} \big\| = 32 L_\ell \big(\mathbb{P}_{n_2} e_S^2\big)^{1/2},$$

which is $O_p(\sqrt{r_n})$ uniformly in S by Step 0.

(c) *Oracle statistic.* $\widetilde{H}_n(S)$ is the empirical `HSIC` of the i.i.d. pairs (S_i, r_i), $i \in \mathcal{I}_2$, with the fixed kernels k_S and ℓ_S, so Lemma 14 of Mooij et al. [20] gives $\widetilde{H}_n(S) - H(S) = O_p(n_2^{-1/2})$.

Combining (a)–(c) and taking the maximum over the finitely many candidates proves the first bound.

Step 2: held-out risk. Fix $S \in \mathcal{C}$. On $\mathcal{I}_3$, the fitted residuals are $r_{S,i} - e_S(S_i)$, so

$$\widehat{R}_n(S) - R(S) = (\mathbb{P}_{n_3} - P) r_S^2 - 2\, \mathbb{P}_{n_3}(r_S e_S) + \mathbb{P}_{n_3} e_S^2.$$

By conditional Jensen, $\mathbb{E}[m_S(S)^4] \leqslant \mathbb{E}[Y^4]$, so $\mathbb{E}[r_S^4] \leqslant 8(\mathbb{E}[Y^4]+\mathbb{E}[m_S(S)^4]) < \infty$. Chebyshev's inequality yields $(\mathbb{P}_{n_3} - P)r_S^2 = O_p(n_3^{-1/2})$ and $\mathbb{P}_{n_3} r_S^2 = O_p(1)$. By the Cauchy–Schwarz inequality and Step 0, $|\mathbb{P}_{n_3}(r_S e_S)| \leqslant (\mathbb{P}_{n_3} r_S^2)^{1/2}(\mathbb{P}_{n_3} e_S^2)^{1/2} = O_p(\sqrt{r_n})$, and $\mathbb{P}_{n_3} e_S^2 = O_p(r_n) = O_p(\sqrt{r_n})$ because $r_n \to 0$. Taking the maximum over the finitely many candidates proves the second bound. □

Proof of Theorem 4.1. Let $\varepsilon_n \coloneqq \max_{S\neq\emptyset} |\widehat{H}_n(S) - H(S)|$ and $a_n \coloneqq \max_S |\widehat{R}_n(S) - R(S)|$. By Lemma B.6 and Assumption 5(v)–(vi), $\varepsilon_n = o_p(h_n)$ and $a_n = o_p(\lambda_n)$. If a data-dependent tolerance $\widehat{\lambda}_n$ is used, replace λ_n by $\widehat{\lambda}_n$ below; since $\widehat{\lambda}_n/\lambda_n \to 1$ in probability, all events below still have probability tending to one.

Screen. The residual of $\emptyset$ is independent of the empty regressor, so $\emptyset \in \mathcal{R}$, and $\emptyset \in \widehat{\mathcal{R}}_n^{\mathrm{thr}}$ by construction. Since the limiting kernels are bounded and characteristic, $H(S) = 0$ for every nonempty $S \in \mathcal{R}$ and $H(S) > 0$ for every $S \in \mathcal{C} \setminus \mathcal{R}$ [37, Theorem 3]. Let $\Delta_H \coloneqq \min_{S\in\mathcal{C}\setminus\mathcal{R}} H(S)$, with $\Delta_H \coloneqq \infty$ if $\mathcal{C} = \mathcal{R}$; finiteness of $\mathcal{C}$ gives $\Delta_H > 0$. The event $E_{1,n} \coloneqq \{\varepsilon_n < h_n,\ h_n + \varepsilon_n < \Delta_H\}$ has probability tending to one because $\varepsilon_n = o_p(h_n)$ and $h_n \to 0$. On $E_{1,n}$, every nonempty $S \in \mathcal{R}$ satisfies $\widehat{H}_n(S) \leqslant \varepsilon_n < h_n$, and every $S \in \mathcal{C} \setminus \mathcal{R}$ satisfies $\widehat{H}_n(S) \geqslant \Delta_H - \varepsilon_n > h_n$. Hence $\widehat{\mathcal{R}}_n^{\mathrm{thr}} = \mathcal{R}$ on $E_{1,n}$.

Risk. Let $R_\star \coloneqq \min_{S\in\mathcal{R}} R(S)$ and $\Delta_R \coloneqq \min_{S\in\mathcal{R}\setminus\mathcal{R}_{\min}}\{R(S) - R_\star\}$, with $\Delta_R \coloneqq \infty$ if $\mathcal{R} = \mathcal{R}_{\min}$; again $\Delta_R > 0$. The event $E_{2,n} \coloneqq \{2a_n < \lambda_n,\ \lambda_n + 2a_n < \Delta_R\}$ has probability tending to one. On $E_{1,n} \cap E_{2,n}$, the minimal empirical risk $\widehat{R}_\star \coloneqq \min_{T\in\mathcal{R}} \widehat{R}_n(T)$ satisfies $|\widehat{R}_\star - R_\star| \leqslant a_n$. Hence every $S \in \mathcal{R}_{\min}$ satisfies $\widehat{R}_n(S) - \widehat{R}_\star \leqslant 2a_n < \lambda_n$, and every $S \in \mathcal{R} \setminus \mathcal{R}_{\min}$ satisfies $\widehat{R}_n(S) - \widehat{R}_\star \geqslant \Delta_R - 2a_n > \lambda_n$. Therefore $\widehat{\mathcal{R}}_{n,\lambda}^{\min} = \mathcal{R}_{\min}$ and $\widehat{\mathrm{PA}}_n^{\mathrm{thr}}(Y) = \arg\min_{S\in\mathcal{R}_{\min}} |S| = \mathfrak{T}(P)$ on $E_{1,n} \cap E_{2,n}$, which proves the first claim.

Under Setting 1 and Assumption 1, the proof of Theorem 3.3, which uses only that $\mathrm{PA}(Y)$ is a candidate, gives $\mathfrak{T}(P) = \{\mathrm{PA}(Y)\}$ whenever $\mathrm{PA}(Y) \in \mathcal{C}$. This proves the second claim. □

The following corollary verifies Assumption 5(iv) for kernel ridge regression with Sobolev kernels and shows that equal or other proportional splits suffice.

Corollary B.7 (Sobolev kernel ridge regression with proportional splits)**.** *Assume Assumption 5(i)–(iii) with $b_n = O(n^{-1/2})$, let $d_{\max} \coloneqq \max_{S\in\mathcal{C}} |S| \geqslant 1$, and let $s > d_{\max}/2$. Suppose that, for every nonempty $S \in \mathcal{C}$,*

- *(a) S takes values in a bounded, open, connected set $\mathcal{X}_S \subseteq \mathbb{R}^{|S|}$ with C^∞ boundary and has a Lebesgue density that is bounded away from zero and infinity on $\mathcal{X}_S$;*
- *(b) $\widehat{m}_S$ is kernel ridge regression on $\mathcal{I}_1$, that is, it minimizes $n_1^{-1}\sum_{i\in\mathcal{I}_1}\{Y_i - g(S_i)\}^2 + \mu_{n_1}\|g\|_{\mathcal{H}_S}^2$ over an RKHS $\mathcal{H}_S$ that is norm-equivalent to the Sobolev space $H^s(\mathcal{X}_S)$, with $\mu_{n_1} \asymp n_1^{-2s/(2s+d_{\max})}$; for example, $\mathcal{H}_S$ can be the RKHS of a Matérn kernel with smoothness parameter $s - |S|/2$;*
- *(c) $\|m_S\|_{\mathcal{H}_S} \leqslant C$, and $\mathbb{E}[|Y - m_S(S)|^m \mid S] \leqslant \frac{1}{2}m!\,\sigma^2 M^{m-2}$ almost surely for all integers $m \geqslant 2$, for constants $C, \sigma, M < \infty$.*

Let $\widehat{m}_\emptyset$ be the training-fold mean of Y. Then Assumption 5(iv) holds with $r_n = n^{-2s/(2s+d_{\max})}$. Consequently, the conclusions of Theorem 4.1 hold for every split with $n_j/n \to \pi_j \in (0,1)$, including equal splits, whenever

$$h_n \asymp n^{-\gamma}, \qquad \lambda_n \asymp n^{-\zeta}, \qquad 0 < \gamma, \zeta < \frac{s}{2s + d_{\max}}.$$

In the dual form $(\mathbf{K} + \lambda_{\mathrm{mat}} I)\alpha = \mathbf{Y}$ of kernel ridge regression, this regularization corresponds to the matrix ridge $\lambda_{\mathrm{mat}} = n_1\mu_{n_1} \asymp n_1^{d_{\max}/(2s+d_{\max})}$, which grows with n_1. Since $s > d_{\max}/2$ implies $s/(2s + d_{\max}) > 1/4$, the rate $n^{-1/4}$ of the thresholds used in our experiments is admissible; by Assumption 5(vi), so is the variance-scaled tolerance $0.25\,\widehat{\mathrm{Var}}_{\mathcal{I}_3}(Y)\,n_3^{-1/4}$ when $\mathrm{Var}(Y) > 0$. The fixed matrix ridge used in our experiments, however, is not covered by the corollary.

Proof of Corollary B.7. Fix a nonempty $S \in \mathcal{C}$ and put $p \coloneqq d_{\max}/(2s) < 1$. Because the density of S is bounded away from zero and infinity on $\mathcal{X}_S$ and $\mathcal{H}_S$ is norm-equivalent to $H^s(\mathcal{X}_S)$, the eigenvalues of the integral operator of the regression kernel on $L^2(P_S)$ decay like $i^{-2s/|S|}$ [7, Section 4], and hence are bounded by a constant times $i^{-1/p}$ since $|S| \leqslant d_{\max}$. The kernel is bounded, which gives the embedding condition of Fischer and Steinwart [7] with exponent one and shows that m_S is bounded. Condition (c) gives their source condition with exponent one and their noise-moment condition. Their L^2 learning rate (Theorem 1(ii) there, with $\gamma = 0$) for regularization $\mu_{n_1} \asymp n_1^{-1/(1+p)} = n_1^{-2s/(2s+d_{\max})}$ gives

$$\mathbb{E}\big[\{\widehat{m}_S(S) - m_S(S)\}^2 \mid \mathcal{I}_1\big] = \|\widehat{m}_S - m_S\|^2_{L^2(P_S)} = O_p\big(n_1^{-2s/(2s+d_{\max})}\big).$$

For $S = \emptyset$, the conditional mean squared error is $(\overline{Y}_{\mathcal{I}_1} - \mathbb{E}\,Y)^2 = O_p(n_1^{-1})$. Taking the maximum over the finite family and using $n_1 \asymp n$ gives $\rho_n = O_p(r_n)$ with $r_n = n^{-2s/(2s+d_{\max})}$.

Since $s/(2s + d_{\max}) < 1/2$, the terms $n_2^{-1/2}$, $n_3^{-1/2}$, and $b_n = O(n^{-1/2})$ are of smaller order than $\sqrt{r_n} = n^{-s/(2s+d_{\max})}$. Both error terms in Assumption 5(v)–(vi) are therefore $O(n^{-s/(2s+d_{\max})})$, which is $o(h_n)$ and $o(\lambda_n)$ for $h_n \asymp n^{-\gamma}$ and $\lambda_n \asymp n^{-\zeta}$ with $\gamma, \zeta < s/(2s + d_{\max})$. Theorem 4.1 applies. □

Search after screening. The following corollary justifies the screened search in Section 4.3. It is conditional on parent retention by the screen.

Corollary B.8 (Search after screening)**.** *Let $K \geqslant 1$ and $\mathcal{C}_K \coloneqq \{S \subseteq \mathcal{S} : |S| \leqslant K\}$, and suppose that Assumption 5 holds with $\mathcal{C} = \mathcal{C}_K$. Let $\widetilde{\mathcal{S}}_n \subseteq \mathcal{S}$ be a data-dependent screened set, let $\mathcal{H}_n \coloneqq \{S \subseteq \widetilde{\mathcal{S}}_n : |S| \leqslant K\}$, and let $\widehat{\mathrm{PA}}_n^{\mathrm{thr}}(Y)$ be the output of Algorithm 4 with candidate family $\mathcal{H}_n$, using the same regression fits and kernel parameters for each candidate as in the search over $\mathcal{C}_K$. If Setting 1 and Assumption 1 hold, $|\,\mathrm{PA}(Y)| \leqslant K$, and $\mathbb{P}(\mathrm{PA}(Y) \subseteq \widetilde{\mathcal{S}}_n) \to 1$, then $\mathbb{P}(\widehat{\mathrm{PA}}_n^{\mathrm{thr}}(Y) = \{\mathrm{PA}(Y)\}) \to 1$.*

Proof. For a subfamily $\mathcal{C}' \subseteq \mathcal{C}_K$ containing $\emptyset$, write $\mathfrak{T}(P; \mathcal{C}')$ for the population target defined with $\mathcal{C}'$ as the candidate family, and let $\Delta_R(\mathcal{C}')$ be the corresponding risk gap from the proof of Theorem 4.1. Define ε_n, a_n, and Δ_H as in that proof with $\mathcal{C} = \mathcal{C}_K$, and let Δ_R^* be the minimum of $\Delta_R(\mathcal{C}')$ over the finitely many such subfamilies, so that $\Delta_R^* > 0$. The regression fit and kernel parameters of a candidate do not depend on which other candidates are searched, so the scores of the members of $\mathcal{C}'$ are the same as in the search over $\mathcal{C}_K$, and the gaps $\Delta_H(\mathcal{C}')$ and $\Delta_R(\mathcal{C}')$ are at least Δ_H and Δ_R^*. The proof of Theorem 4.1, with Δ_R replaced by Δ_R^*, therefore shows that, on an event whose probability tends to one, the output of Algorithm 4 with candidate family $\mathcal{C}'$ equals $\mathfrak{T}(P; \mathcal{C}')$ simultaneously for all such $\mathcal{C}'$. Since $\mathcal{H}_n$ is one of these subfamilies, $\widehat{\mathrm{PA}}_n^{\mathrm{thr}}(Y) = \mathfrak{T}(P; \mathcal{H}_n)$ on this event. On the event $\{\mathrm{PA}(Y) \subseteq \widetilde{\mathcal{S}}_n\}$, we have $\mathrm{PA}(Y) \in \mathcal{H}_n$ because $|\,\mathrm{PA}(Y)| \leqslant K$, and the proof of Theorem 3.3 gives $\mathfrak{T}(P; \mathcal{H}_n) = \{\mathrm{PA}(Y)\}$. Both events have probability tending to one. □

The proof does not use that $\widetilde{\mathcal{S}}_n$ depends only on $\mathcal{I}_1$. We nevertheless compute the screen on the training fold so that $\mathcal{I}_2$ and $\mathcal{I}_3$ remain untouched by it, which matters for the calibration of the resampling-based rules.

C Experimental details

Metrics. Let TP, FP, and FN be the numbers of true positives, false positives, and false negatives of an estimate $\widehat{S}$ of $S^\star$. Then SHD is $\mathrm{FP} + \mathrm{FN}$ and FDR is $\mathrm{FP}/(\mathrm{TP} + \mathrm{FP})$, set to 0 if $\widehat{S} = \emptyset$. F_1 is the harmonic mean of precision $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, set to 1 if $\widehat{S} = S^\star = \emptyset$ and to 0 if only $\widehat{S}$ is empty, and recall $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, set to 1 if $S^\star = \emptyset$. Synthetic results are means over datasets with normal-approximation 95% intervals. On Sachs, SHD is averaged over the 11 targets, and FDR and F_1 are computed from TP, FP, and FN summed over targets.

Small-graph benchmark. Each synthetic dataset is generated from a random four-node DAG. The nodes are ordered topologically as $1, \dots, 4$, and for each ordered pair $i < j$ we draw an edge $i \to j$ independently with probability 0.85. The target index $T \in \{1, \dots, 4\}$ is drawn uniformly at random, and the observed response is $Y = X_T$. The remaining three variables are used as candidate predictors.

For each node j, we draw an independent noise variable

$$\epsilon_j \sim \mathrm{Unif}[-\sqrt{3}\sigma_j, \sqrt{3}\sigma_j], \qquad \sigma_j = 0.1.$$

The first node is generated as $X_1 \sim \mathrm{Unif}[-1, 1]$. For $j > 1$, let

$$P_j = (A_{1j}X_1, \dots, A_{j-1,j}X_{j-1})$$

denote the vector of previous variables, with non-parent coordinates masked to zero. If $j = T$, the target node is generated by an additive-noise mechanism

$$X_j = g_j(P_j) + \epsilon_j.$$

The function g_j is a random smooth function constructed from $m = 40$ RBF features. Specifically, we sample m anchor points $Z_1, \dots, Z_m$ from the rows of P_j, use the median heuristic for the RBF bandwidth, draw coefficients

$$a_{j\ell} \sim N(0, 1.5^2/m),$$

and add a linear signal $P_j^\top \beta_j$, where each nonzero coefficient has a random sign and magnitude drawn uniformly from $[0.35, 0.80]$. Thus

$$g_j(P_j) = \sum_{\ell=1}^{m} a_{j\ell} k(P_j, Z_\ell) + P_j^\top \beta_j,$$

where k is the RBF kernel with bandwidth chosen by the median heuristic. For non-target nodes $j \neq T$ with $j > 1$, the structural equations are deliberately non-additive in the noise. We form an augmented input

$$A_j = (P_j, \epsilon_j, P_j\epsilon_j),$$

construct another random RBF feature expansion on A_j, add the same type of masked linear signal $P_j^\top \beta_j$, and set

$$X_j = \tanh\left(\sum_{\ell=1}^{m} a_{j\ell} k(A_j, Z_\ell) + P_j^\top \beta_j + 0.1\epsilon_j \right).$$

Consequently, only the target node is guaranteed to satisfy an additive-noise model, while the remaining structural equations generally violate the global ANM assumption.

The true parent set used for scoring is

$$\mathrm{PA}(Y) = \{X_i : i < T,\ A_{iT} = 1\}.$$

We generate 100 datasets of $n = 3000$ observations each.

Large-graph benchmark. The parents of Y are two independent root variables with unit-variance uniform distributions, and

$$Y = f(X_{P_1}, X_{P_2}) + 0.65\,\epsilon_Y,$$

where ϵ_Y is unit-variance uniform noise and f is the sum of two RBF main effects and an RBF interaction, which contribute $1/4$, $1/4$, and $1/2$ of $\mathrm{Var}\, f = 1$. Y has two children, each with parents Y and one parent of Y, and two to four descendants in total; the remaining candidates descend from the parents of Y, and every variable has at most two parents. Every other non-root variable X_j has a conditional mean $g_j(P_j)$ given by a random RBF function of its parents P_j. In the global regime, $X_j = g_j(P_j) + 0.65\,\eta_j$. In the target-only regime,

$$X_j = g_j(P_j) + 0.65\big(\sqrt{0.5}\,\eta_j + \sqrt{0.5}\,a_j(P_j)\epsilon_j\big),$$

where a_j is positive and nonconstant with $\mathbb{E}[a_j(P_j)^2] = 1$, so that only Y follows an additive-noise model. All noise variables are independent with unit-variance uniform distributions. The two regimes share the graphs, f, the g_j, and the noise of Y. For each d and regime, we use 5 graphs with 2 samples each, and the candidate columns are randomly permuted. The shared screen is computed on half of the sample (the fitting fold of the $50/25/25$ split), and every method receives the screened candidates and Y.

Sachs data. We use the 853 observations and the 17-edge consensus network as distributed with `bnlearn` [32]. For `IndRM`, all variables are standardized, and the screen ranks the 10 candidates of each target by the feature importance of a gradient-boosting regression of the target on them.

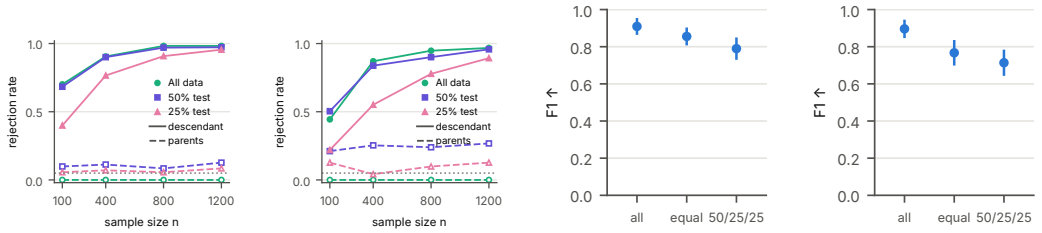
Configurations. Both variants of `IndRM` use kernel ridge regression with an RBF kernel and median-heuristic bandwidth. In the small-graph benchmark, the ridge is chosen by generalized cross-validation over 31 log-spaced values in $[10^{-6}, 10^{2}]$; in the large-graph benchmark and on Sachs, it is fixed at 0.1, and $K = 3$. The permutation tests of `IndRM` use level $\beta = 0.05$ and $B = 500$ resamples with a Matérn-$3/2$ `HSIC` kernel with median-heuristic length scale (small graphs), or $B = 199$ resamples with an RBF `HSIC` kernel (large graphs and Sachs). The threshold screen uses the Gaussian kernels of Appendix A. `DirectLiNGAM` uses the `lingam` package and `CAM` the `R` package `CAM`. `RESIT` uses random-forest regression in the small-graph benchmark and the regression and `HSIC` test of `IndRM` otherwise. `CMB` and `PCD-by-PCD` use G^2 tests on three-bin quantile discretizations. Their level and maximal conditioning-set size are 0.05 and 2 in the large-graph benchmark, 0.01 and 3 on Sachs, and 0.05 without limit in the small-graph benchmark. Global methods are fitted to all variables, and the parents of the target are read off the estimated graph. We found no official implementation of `CMB` or `PCD-by-PCD`, so we implemented both from the original papers with the help of an LLM (Claude, Anthropic); the code is part of our released implementation. We do not include `HLCD` [18], for which we found no author-released implementation.

Threshold screen. With the constants of Section 4.1, the screen of `IndRM-TS` retains on average 97% of the nonempty candidate sets in the small-graph benchmark (all of them in 83 of 100 datasets), 99.8% in the large-graph benchmark, and all of them on Sachs.

Additive-noise check on Sachs. For each non-root target, we regress the target on its reference parents with five-fold cross-fitting, repeated three times, using kernel ridge regression with fixed and with cross-validated ridge, gradient boosting, and extra trees. We test independence of the residuals and the reference parents with an `HSIC` permutation test (999 permutations), with p-values Holm-adjusted over targets. All four regressions reject independence for 7 of the 9 non-root targets.

D Additional Experimental Results

D.1 Kernel Ablation Study



(a) RBF kernel (b) Polynomial kernel (c) RBF kernel (d) Polynomial kernel

Figure 3: Kernel ablation on the small-graph benchmark. (a, b) `HSIC` rejection rates as in Figure 2a. (c, d) F_1 scores of `IndRM` with all data and with equal and $50/25/25$ splits; SHD and FDR are given in Table 3.

Figure 3 repeats Figures 2a and 2b with RBF and inhomogeneous quadratic polynomial `HSIC` kernels. With all data, `IndRM` attains $F_1 = .91$ (RBF) and .90 (polynomial), compared with .92 for the Matérn kernel.

D.2 False discovery rates

Table 2: False discovery rates: means over 100 datasets (small graphs) and 40 datasets per regime (large graphs), and pooled over the 11 targets (Sachs). The FDR of an empty estimate is 0.

		Large graphs		
Method	Small graphs	Target-only ANM	Global ANM	Sachs
`IndRM`	.04	.27	.29	.84
`IndRM-TS`	.51	.27	.27	.62
`CAM`	.14	.02	.00	.40
`RESIT`	.40	.43	.32	.61
`DirectLiNGAM`	.51	.70	.67	.55
`CMB`	.00	.23	.27	.00
`PCD-by-PCD`	.00	.44	.45	.00

D.3 Small-graph results

Table 3: Small-graph benchmark ($n = 3000$, 100 datasets): the values shown in Figure 1a and Figure 2b, with means and 95% confidence intervals. The last six rows use other `HSIC` kernels (Appendix D).

Method	SHD ↓	FDR ↓	F_1 ↑
`IndRM` (all data)	0.21 [0.12, 0.30]	0.04 [0.00, 0.08]	0.92 [0.88, 0.96]
`IndRM` (equal split)	0.39 [0.27, 0.51]	0.05 [0.01, 0.08]	0.87 [0.82, 0.91]
`IndRM` (50/25/25)	0.55 [0.40, 0.70]	0.08 [0.04, 0.13]	0.80 [0.75, 0.86]
`IndRM-TS` (all data)	1.50 [1.36, 1.64]	0.51 [0.43, 0.59]	0.42 [0.36, 0.48]
`IndRM-TS` (equal split)	1.46 [1.33, 1.60]	0.50 [0.42, 0.58]	0.42 [0.36, 0.48]
`IndRM-TS` (50/25/25)	1.43 [1.30, 1.57]	0.49 [0.41, 0.57]	0.42 [0.36, 0.48]
`CAM`	0.61 [0.46, 0.76]	0.14 [0.09, 0.20]	0.78 [0.73, 0.84]
`RESIT`	1.39 [1.19, 1.59]	0.40 [0.32, 0.48]	0.52 [0.44, 0.60]
`DirectLiNGAM`	1.67 [1.48, 1.86]	0.51 [0.43, 0.60]	0.37 [0.29, 0.45]
`CMB`	1.14 [0.94, 1.34]	0.00 [0.00, 0.00]	0.34 [0.25, 0.43]
`PCD-by-PCD`	1.09 [0.90, 1.28]	0.00 [0.00, 0.00]	0.36 [0.26, 0.45]
RBF `HSIC` kernel			
`IndRM` (all data)	0.24 [0.14, 0.34]	0.06 [0.01, 0.10]	0.91 [0.86, 0.95]
`IndRM` (equal split)	0.43 [0.30, 0.56]	0.06 [0.02, 0.09]	0.86 [0.81, 0.90]
`IndRM` (50/25/25)	0.59 [0.44, 0.75]	0.11 [0.05, 0.16]	0.79 [0.73, 0.85]
polynomial `HSIC` kernel			
`IndRM` (all data)	0.27 [0.16, 0.38]	0.07 [0.02, 0.12]	0.90 [0.85, 0.95]
`IndRM` (equal split)	0.57 [0.42, 0.72]	0.09 [0.04, 0.14]	0.77 [0.70, 0.84]
`IndRM` (50/25/25)	0.72 [0.56, 0.89]	0.15 [0.09, 0.21]	0.71 [0.64, 0.78]

D.4 Large-graph results by dimension

Table 4: Large-graph benchmark by regime and number of candidates d: SHD/FDR/F_1, means over 10 datasets.

Method	$d = 10$	$d = 20$	$d = 50$	$d = 100$
target-only ANM				
`IndRM` (all data)	.80/.27/.84	.80/.27/.84	.80/.27/.84	.87/.28/.83
`IndRM-TS` (all data)	.80/.27/.84	.80/.27/.84	.80/.27/.84	.80/.27/.84
`IndRM-TS` (50/25/25)	.98/.30/.80	.70/.22/.86	.87/.26/.83	.85/.28/.83
`CAM`	1.40/.00/.33	1.20/.00/.47	.80/.00/.67	1.00/.08/.65
`RESIT`	1.60/.37/.64	1.60/.36/.71	2.00/.52/.56	2.70/.49/.32
`DirectLiNGAM`	4.00/.62/.45	4.40/.73/.37	4.60/.75/.35	4.40/.72/.40
`CMB`	1.00/.17/.69	1.10/.30/.78	1.00/.26/.83	1.20/.18/.65
`PCD-by-PCD`	2.00/.44/.68	2.30/.52/.63	1.90/.44/.70	1.40/.34/.74
global ANM				
`IndRM` (all data)	.85/.28/.83	.87/.28/.83	.97/.29/.81	.97/.29/.81
`IndRM-TS` (all data)	.80/.27/.84	.80/.27/.84	.80/.27/.84	.80/.27/.84
`IndRM-TS` (50/25/25)	.75/.21/.85	1.07/.31/.79	.93/.29/.81	.83/.26/.83
`CAM`	1.40/.00/.33	1.20/.00/.47	.90/.00/.60	1.00/.00/.57
`RESIT`	1.10/.30/.79	1.60/.39/.72	1.40/.33/.75	1.50/.27/.75
`DirectLiNGAM`	4.00/.68/.43	3.90/.68/.44	4.10/.68/.45	3.70/.64/.50
`CMB`	.70/.13/.81	1.70/.39/.72	1.10/.26/.80	1.30/.28/.70
`PCD-by-PCD`	2.10/.49/.67	2.10/.47/.68	2.00/.43/.70	1.70/.39/.70

D.5 Computation Cost

Table 5: Average elapsed time on the Sachs data, in seconds per target (one CPU core). Methods fitted once to the whole graph are charged their fit time divided by 11.

Method	Time (s)
`IndRM`	42.4
`IndRM-TS`	15.0
`CAM`	0.7
`RESIT`	9.9
`DirectLiNGAM`	0.2
`CMB`	< 0.1
`PCD-by-PCD`	< 0.1

Table 5 reports the average wall-clock time per target on the Sachs data, measured on one core of an AMD EPYC 9554 that was shared with other jobs, so the times are approximate. `IndRM` is more expensive than the baselines because it fits a regression and runs resampling tests for each of the 42 candidate sets; the threshold screen avoids the resampling. All experiments ran as single-core jobs on two shared nodes with AMD EPYC 7763 and 9554 processors, and each run of `IndRM` needed less than 1 GB of memory. With all data, `IndRM` took on average 4.3 minutes per small-graph dataset and 5.3 minutes per large-graph dataset, and `IndRM-TS` 2.3 and 2.1 minutes. All reported experiments, including the sample-split variants and the kernel ablation, took about 50 CPU hours.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: See Section 3 and Section 4.2.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 3 discusses when the no-backward condition holds and gives a counterexample (Example 1); Appendix B.5 shows that all parents of Y must be observed.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: The assumptions are stated in the main text and full proofs are provided in the appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The method, data generation, evaluation metric, baselines, command, and default settings are described in the main text and experimental appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

Answer: [Yes]

Justification: Code that reproduces all experiments, figures, and tables, including the data-generating processes and the baselines, is available at `https://github.com/fschur/who-caused-y`, with instructions in its README. The Sachs data are publicly available (we use the copy distributed with `bnlearn`); we could not identify a license, but the authors gave written permission for their use in this work.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: The paper specifies the splits, kernels, hyperparameters, baselines, and evaluation protocol needed to understand the reported results.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: Figure 1, Figure 2b, the kernel ablation, and the small-graph table report means with 95% normal-approximation confidence intervals over datasets, which capture the variability of the randomly drawn graphs, mechanisms, and samples (Appendix C). The rejection rates in Figure 2a and the per-dimension and false-discovery-rate tables report means only. The Sachs data form a single dataset, so we report no intervals there.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix D.5 reports the hardware (single cores of shared AMD EPYC 7763 and 9554 nodes), the memory of `IndRM` (below 1 GB per job), its average run time per dataset, the total compute of all reported experiments (about 50 CPU hours), and the run time per target of every method on the Sachs data.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: To the best of our knowledge, the work uses standard simulations and benchmark data and does not deviate from the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [N/A]

Justification: We do not think local causal discovery has any direct societal impact.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: The paper does not release a high-risk model, scraped dataset, or other asset that would require special misuse safeguards.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: We cite the Sachs data [30] and use the copy and consensus network distributed with `bnlearn` [32]; we could not identify a license, but the authors gave written permission for its use in this work.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

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

13. **New assets**

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

    Answer: [Yes]

    Justification: We release our code at `https://github.com/fschur/who-caused-y` under the MIT license; its README documents the installation and the commands that reproduce every figure and table.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs are not part of the proposed method or its theory. We used an LLM as an assistant for coding, including our implementations of `CMB` and `PCD-by-PCD` (Appendix C), for drafting and checking proofs, and for editing. The authors verified all code, proofs and text and take full responsibility for them.

Guidelines:

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