
When Additive Noise Meets Unobserved Mediators: Bivariate Denoising Diffusion for Causal Discovery

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Distinguishing cause and effect from bivariate observational data is a foundational
2 problem in many disciplines, but challenging without additional assumptions. *Ad-*
3 *ditive noise models* (ANMs) are widely used to enable sample-efficient bivariate
4 causal discovery. However, conventional ANM-based methods fail when unob-
5 served mediators corrupt the causal relationship between variables. This paper
6 makes three key contributions: first, we rigorously characterize why standard ANM
7 approaches break down in the presence of unmeasured mediators. Second, we
8 demonstrate that prior solutions for hidden mediation are brittle in finite sample
9 settings, limiting their practical utility. To address these gaps, we propose *Bi-*
10 *variate Denoising Diffusion* (BiDD) for causal discovery, a method designed to
11 handle latent noise introduced by unmeasured mediators. Unlike prior methods
12 that infer directionality through mean squared error loss comparisons, our approach
13 introduces a novel independence test statistic: during the noising and denoising
14 processes for each variable, we condition on the other variable as input and evaluate
15 the independence of the predicted noise relative to this input. We prove asymptotic
16 consistency of BiDD under the ANM, and conjecture that it performs well under
17 hidden mediation. Experiments on synthetic and real-world data demonstrate
18 consistent performance, outperforming existing methods in mediator-corrupted
19 settings while maintaining strong performance in mediator-free settings.

20 1 Introduction

21 Determining the causal direction between two variables ($X \rightarrow Y$) is fundamental to scientific domains
22 ranging from genomics to economics. However, traditional discovery methods, such as constraint-
23 based [64, 63] and scoring-based methods [9, 29, 19] can only identify causal graphs up to an
24 equivalence class, leaving them unable to distinguish the causal direction between a variable pair.
25 Additional assumptions are necessary to enable bivariate discovery [49], and they mostly fall under
26 three categories: (1) *the location scale noise model* (LSNMs), (2) *the principle of independent*
27 *mechanisms* (ICM), and (3) the additive noise model.

28 LSNMs express the outcome Y with heteroskedastic, multiplicative noise relative to the treatment X ,
29 i.e. $Y = f(X) + g(X)\varepsilon$, where $\varepsilon \perp\!\!\!\perp X$. While LSNMs allow for increased flexibility, existing ap-
30 proaches require additional parametric assumptions for identifiability [65, 67, 20, 7]. ICM approaches
31 assume that the marginal distribution of the cause and the conditional mechanism generating the effect
32 are independent components of the *data-generating process* (DGP) [60, 24]. While they impose no
33 explicit functional form, these methods rely on unverifiable structural asymmetries [42, 21], often fail
34 under non-invertible mechanisms [22], and often lack theoretical guarantees [65]. In contrast, *additive*
35 *noise models* offer unique advantages for bivariate discovery, allowing for consistent recovery of
36 causal directions without strong parametric assumptions [70], permitting sample complexity charac-

37 terization under Gaussian noise [73], and enabling polynomial-time guarantees for global discovery
 38 on large graphs [50]. These properties have spurred both methodological advances [38, 15, 68, 16]
 39 and real-world applications [57, 30].

40 However, these strengths vanish when hidden variables corrupt the observed causal relationships—a
 41 near-ubiquitous scenario in real-world systems like biomedicine [30] and economics [2]. Indeed, as
 42 [51] point out, although the joint distribution of all variables may admit an ANM, the joint distribution
 43 over a subset which excludes some mediators may not allow for an ANM (see Appendix D.3). To
 44 the best of our knowledge, despite the rapid advances in statistical tests that handle unobserved
 45 confounding of causal pairs, [23, 36, 37, 69, 33], only one bivariate discovery method [6] addresses
 46 the problem of unobserved mediators. However, [6] provides no correctness guarantees, requires
 47 nonlinearity, and has poor empirical performance (Section 5). This leaves a glaring gap in practical
 48 bivariate causal discovery.

49 **Contributions.** In this paper, we propose *bivariate denoising diffusion* (BiDD), a causal direction
 50 identification method that works for general ANM, even in the presence of unobserved mediators.
 51 Our contributions are fourfold:

- 52 • **Analysis of Unmeasured Mediators:** We first introduce the ANM-UM, a novel approach for
 53 modeling unobserved mediators (Section 2). We then characterize how unobserved mediators
 54 break the ANM assumption over observed variables, finding that this occurs if and only if there are
 55 nonlinear mechanisms induced after the initial transformation of the cause (Lemma 2.3).
- 56 • **Failure-Mode of Existing Methods:** We first categorize conventional ANM-based methods into
 57 three types: Residual-Independence, MSE-Minimization, and Score-Matching based (Section 3).
 58 For each category, we show that existing methods will fail to correctly recover the directionality
 59 when unobserved mediators break the ANM assumption (Lemmas 3.1-3.4). We then analyze the
 60 only method developed to handle hidden mediation, discussing potential issues.
- 61 • **Diffusion Methodology and Gaurantees** We develop BiDD, a practical alternative to existing
 62 ANM based methods, hypothesizing that the noise predictions from a conditional diffusion model
 63 will be less dependent on the condition when the condition is the cause, rather than the effect
 64 (Section 4). We show a consistency result under the assumption of an ANM (Theorem 4.2), and
 65 conjecture that a similar result may hold in the ANM-UM setting.
- 66 • **Comprehensive Evaluation:** We extensively evaluate BiDD on synthetic data, demonstrating
 67 that only our approach is able to achieve uniformly strong performance across DGPs with linear,
 68 nonlinear noninvertible, and nonlinear invertible mechanisms (Section 5.2). We then validate BiDD
 69 on a large real world dataset, the Tubingen Cause-Effect pairs [44], where it achieves comparable
 70 results to the best baselines, highlighting BiDD’s robustness across diverse domains (Section 5.2).

71 2 Problem Setup

72 In this work, we focus on the discovery of the causal direction between a causal pair (X, Y) ,
 73 which is generated by an *ANM with Unobserved Mediators* (ANM-UM). In this section, we first
 74 formally introduce the structural causal model describing ANM-UM. We then establish identifiability
 75 conditions and characterize when ANM-UM cannot be simplified to standard ANMs. A complete
 76 notation table is included in Appendix A.

77 Under ANM-UM, the outcome Y is generated from cause X through unobserved mediators $\{Z_i\}$
 78 (Figure 1), with each Z_i introducing independent noise while remaining unmeasured. Formally, given
 79 T unmeasured mediators, the DGP between X and Y can be described as follows:

$$\begin{cases} Z_1 = f_1(X) + \varepsilon_1, \\ Z_2 = f_2(Z_1) + \varepsilon_2, \\ \vdots \\ Z_T = f_T(\text{Pa}(Z_T)) + \varepsilon_T, \\ Y = f_{T+1}(\text{Pa}(Y)) + \varepsilon_{T+1}, \end{cases} \quad (2.1)$$

80 where $X, \{\varepsilon_i\}_{i=1}^{T+1}$ are mutually independent. The functions $\{f_1, \dots, f_{T+1}\}$ can be linear or nonlin-
 81 ear, and the ε can be arbitrary (Gaussian or non-gaussian).

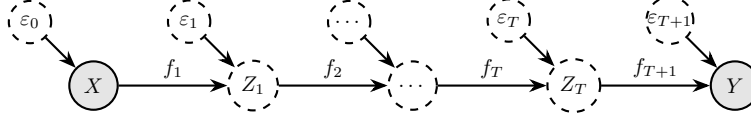


Figure 1: ANM-UM (Eq 2.1), where mediators Z_1, \dots, Z_T and noises $\varepsilon_0, \dots, \varepsilon_{T+1}$ are all unobserved.

Assumption 2.1 (ANM-UM Setting). Suppose X, Y follow ANM-UM described by Eq. (2.1). Then, we assume: 1) no observed confounders among $X, \{Z_i\}_i$, and Y ; 2) acyclicity; 3) no selection bias (noise independence is preserved in the data collection process); 4) $X \not\perp\!\!\!\perp Y$, i.e., f is nonzero almost everywhere (otherwise, $X \perp\!\!\!\perp Y$, detectable via simple independence testing).

The conventional ANM and the related Post-Nonlinear (PNL) Model [70] are special cases of ANM-UM. ANM corresponds to zero mediators (i.e., $T = 0$), while PNL corresponds to one mediator (i.e., $T = 1$) and no additive noise on Y (i.e., $\varepsilon_{T+1} = 0$). Our ANM-UM also generalizes the Cascade Additive Noise Model (CANM) [6], which assumes all functions $\{f_1, \dots, f_{T+1}\}$ are nonlinear. See Appendix D.1 for details.

Prior work (Theorem 1, Cai et al. [6]) shows that certain (X, Y) distributions admit both $X \rightarrow Y$ and $Y \rightarrow X$ ANM-UM representations, rendering the causal direction unidentifiable without further constraints. This occurs only in pathological cases, such as when the functions f are linear and the noises are Gaussian. We thus impose:

Assumption 2.2 (Identifiability Constraint). Under Eq. (2.1) and Assump. 2.1, no backward ANM-UM exists where $X = g(Y, \hat{\varepsilon}_1, \dots, \hat{\varepsilon}_T) + \hat{\varepsilon}_{T+1}$ with $Y, \hat{\varepsilon}_1, \dots, \hat{\varepsilon}_{T+1}$ mutually independent.

Appendix D.2 provides explicit constraints on the backward mechanism g and noise terms $\{\hat{\varepsilon}_i\}_{i \in [T]}$ that preclude non-identifiability under Assumption 2.2.

While CANM [6] requires all mediators to be nonlinear, ANM-UM permits identifiability with unobserved mediators even under linear transformations, reducing to standard ANM when the causal effect admits an additive decomposition:

$$Y = A_1(X) + A_2(\varepsilon_1, \dots, \varepsilon_T) + \varepsilon_{T+1} \quad (2.2)$$

where functions A_1 and A_2 (determined by f_1, \dots, f_T) are separable without X - ε interaction terms. For example, in a 3 variable ANM-UM $X \rightarrow Z_1 \rightarrow Y$, if f_1 is nonlinear and f_2 is linear, it reduces to the ANM setting, whereas if f_1, f_2 are both nonlinear, it does not (see Appendix D.3). Lemma 2.3 (proof in Appendix D.4) formalizes this: ANM-UM is irreducible to ANM if and only if there exists a mediator Z_i such that Y depends nonlinearly on Z_i :

Lemma 2.3 (Irreducible ANM-UM and Nonlinear Mediator). Under ANM-UM (Eq. (2.1)) and Assump.s 2.1 and 2.2, Y does not admit a decomposition in Eq. (2.2) if and only if there exists a mediator Z_i such that $Y = h(Z_i) + \tilde{\varepsilon}$ for some nonlinear h , with $\tilde{\varepsilon} \perp\!\!\!\perp Z_i$. Additionally, we call such a mediator Z_i a nonlinear mediator.

3 Failure-Modes of Prior Work

In this section, we illustrate how both standard ANM methods and one existing hidden mediator approach fail under ANM-UM settings (Eq.(2.1) and Assumption 2.1), assuming identifiability (Assumption 2.2) and irreducibility of ANM-UM (Lemma 2.3).

3.1 Traditional ANM-based Bivariate Methods

Existing methods mostly fall into three categories: 1) *Residual-Independence* (RI): identify the cause via an independent residual, 2) *Score-Matching*: identify the effect via conditions on the score function, 3) *MSE-Minimization*: identify the cause via the smallest residual. For each class, we present its core decision rule and construct ANM-UM counterexamples where it fails.

Residual-Independence Key methods include DirectLiNGAM [62], its nonparametric generalization RESIT [50], and PNL [70]. The former two leverage ANM-induced residual independence

asymmetries via a common decision rule (Decision Rule E.2): if the residual from regressing Y on X is independent of X but the residual from regressing X on Y depends on Y , we conclude X causes Y , and vice versa. If both residuals are independent or dependent, no conclusion can be drawn.

As a counterexample, consider the following ANM-UM: $X \sim \mathcal{N}(0, 1)$, $Z = X^2 + \varepsilon_1$, $Y = Z^2 + \varepsilon_2$, with $\varepsilon_1, \varepsilon_2 \sim \mathcal{N}(0, 1)$. The residual $e_1 := Y - E[Y|X]$ can be simplified as

$$e_1 = Y - E[Y|X] = Y - E[(X^4 + \varepsilon_1^2 + \varepsilon_2) + (2\varepsilon_1 X^2)] = \varepsilon_1^2 + \varepsilon_2 + 2\varepsilon_1 X^2 - 1. \quad (3.1)$$

We observe $e_1 \not\perp\!\!\!\perp X$. Consequently, Decision Rule E.2 does not return the correct causal directionality and fails to identify $X \dashrightarrow Y$. We formalize this intuition in Lemma 3.1 (proof in Appendix E.3):

Lemma 3.1 (Regression Residual-Independence Fails). *Assuming a consistent estimator for regression residuals and access to infinite data, Decision Rule E.2 fails to identify the correct causal direction when at least one mediator is nonlinear.*

PNL assumes a more complicated structure between X, Y : $Y = f(g(X) + \varepsilon_1)$, where g, ε_1, f are the nonlinear effect, independent noise, and invertible post-nonlinear distortion, respectively. As ε_1 can be represented as the difference $f^{-1}(Y) - g(X)$, [70] proposes to identify the causal direction by recovering independent noise. If they can find functions l_1, l_2 such that $e_1 \perp\!\!\!\perp X$ for $e_1 = l_2(Y) - l_1(X)$, then they say that the causal hypothesis $X \dashrightarrow Y$ ‘holds’ (Decision Rule E.3).

While valid for restricted ANM-UM cases (e.g., single nonlinear mediator), this approach may fail with multiple mediators due to f ’s invertibility requirement (Lemma 3.2, proof in Appendix E.5).

Lemma 3.2 (PNL Residual-Independence Fails). *Assuming a consistent ICA residual estimator and access to infinite data, Decision Rule E.3 fails to recover the correct causal direction when there exists at least one non-invertible nonlinear mediator.*

Prior work [62, 50, 70] propose alternative rules to compare measures of dependence, rather than independence, to improve finite sample performance (see Appendix E.4 for more details). However, empirically, we find this heuristic often fails (see Section 5).

Score-Matching The original score-matching method SCORE [55] (with several followup works [39, 59] leveraging the same fact) relies on the assumption of Gaussian noise and nonlinear mechanisms to identify the effect via a condition on the Jacobian of the score function ($\nabla \log p(X, Y)$). Montagna et al. [38] prove that SCORE can fail to correctly decide causal direction when the noise is non-Gaussian, proposing NoGAM as a noise agnostic solution for nonlinear ANM. They further extend NoGAM to Adascore [40], which they prove correctly recovers the causal direction for all identifiable ANM.

Adascore identifies the causal direction by proving that only the residual from nonparametrically regressing the effect onto the cause is a consistent estimator of a particular expression involving the score (Rule E.4). However, their theory relies on the estimated residual being independent from the cause, which, as demonstrated in Eq. (3.1) may be false in some ANM-UM. Thus, Decision Rule E.4 fails to identify $X \dashrightarrow Y$. We formalize this intuition in Lemma 3.3 (proof in Appendix E.6)

Lemma 3.3 (Score-Matching Fails). *Assuming a consistent estimator of the conditional expectation and access to infinite data, Decision Rule E.4 fails to recover the correct causal direction when there exists at least one nonlinear mediator.*

MSE Minimization Key methods include CAM [5], NoTEARS [71] and GOLEM [46], and NoTEARS-MLP [72]. The causal direction is determined by comparing prediction error: whichever variable better predicts the other (lower MSE) is designated the cause (Rule E.5). While effective in some synthetic settings, this rule suffers from two key flaws: 1) standardizing degrades performance [53], and 2) the L_2 loss is only lower in the causal direction under restrictive variance conditions [47], which may not hold under ANM-UM (Lemma 3.4, proof in Appendix E.7). Intuitively, the causal direction becomes unidentifiable when R^2 -sortability vanishes (i.e., equal prediction errors in both directions), a problematic limitation since DGPs may exhibit arbitrary R^2 values [54].

Lemma 3.4 (MSE-Minimization Fails). *Assuming a consistent estimator of the conditional expectation and access to infinite data, Rule E.5 fails to recover the correct causal direction when $E[\text{Var}[X|Y]] < E[\text{Var}[Y|X]]$.*

3.2 Hidden Mediator Method—CANM

Assuming nonlinear mechanisms, CANM [6] uses a variational autoencoder (VAE) framework to: 1) learn latent noise via VAE ($X, Y \rightarrow \mathcal{N}(\mu, \sigma^2)$), 2) compare the *evidence lower bound* (ELBO, Eq. (3.2)) scores for both causal directions, and 3) infer causation via higher ELBO (Rule E.6). While CANM [6] succeeds on synthetic non-invertible Gaussian DGPs, it lacks theoretical guarantees, even for vanilla ANM. Our experiments (Section 5) show failure cases with: 1) linear/invertible mechanisms, 2) non-Gaussian noise (often exhibiting posterior collapse, see Appendix E.9).

As VAE training often encounters posterior collapse in practice, next we examine the behavior of CANM under this phenomenon. Posterior collapse causes CANM’s learned $\mathcal{N}(\mu, \sigma^2)$ to degenerate to $\mathcal{N}(0, 1)$. This eliminates the KL term in ELBO and reduces the objective to the sum of negative entropy of X and the conditional log-likelihood of $Y|X$ (Eq. (3.3)):

$$\text{ELBO}_{X \rightarrow Y} = \mathbb{E}[\log p(x)] + -\beta \text{KL}(q_\phi(n | x, y) || p(n)) + \mathbb{E}_{n \sim q_\phi(n|x,y)} \log p(\varepsilon = y - f(x, n; \theta)) \quad (3.2)$$

$$= -H(X) + \mathbb{E} \log p_{Y|X}(\varepsilon = y - f(x; \theta)). \quad (3.3)$$

When posterior collapse occurs, CANM is provably inconsistent for ANM-UM where this sum is not higher for the causal direction (Lemma 3.5, proof in Appendix E.10):

Lemma 3.5 (CANM Fails). *Assuming infinite data and a consistent estimator of the conditional expectation, Rule E.6 fails to recover the causal direction if posterior collapse occurs and the expected conditional log-likelihood minus the entropy is higher in the causal direction.*

4 Bivariate Causal Discovery Using Diffusion

In this section, we develop our conditional diffusion-based method for distinguishing between cause and effect generated by the ANM-UM. We first warm up by developing intuition in the linear setting about when denoising leads to predicted noise that is independent of one of its input variables. We then spell out a decision rule for deciding the causal direction that leverages the developed intuition, providing theoretical guarantees of correctness under certain restrictions of the ANM-UM. We end by introducing a practical method for denoising-diffusion for bivariate discovery, BiDD, and providing its computational complexity.

4.1 Denoising and Independence

To better understand what asymmetries may arise from denoising in the causal vs. anticausal direction, we start with a simplified setup, restricting ANM-UM to only linear mechanisms without unobserved mediators. We let the DGP of X, Y follow

$$Y = X + \varepsilon_1, \quad X \perp\!\!\!\perp \varepsilon_1,$$

where at least one of X, ε_1 is non-Gaussian (to ensure identifiability [62]). In the denoising process, we inject independent Gaussian noise into both X and Y , obtaining the noised terms

$$\tilde{X} = X + \varepsilon_X \quad \text{and} \quad \tilde{Y} = Y + \varepsilon_Y, \quad \varepsilon_X, \varepsilon_Y \sim \mathcal{N}(0, 1). \quad (4.1)$$

Now, in the denoising process, we aim to find the best estimators $f_{\varepsilon_X}, f_{\varepsilon_Y}$ such that MSE losses

$$(\varepsilon_Y - f_{\varepsilon_Y}(\tilde{Y}, X))^2 \quad \text{and} \quad (\varepsilon_X - f_{\varepsilon_X}(\tilde{X}, Y))^2 \quad (4.2)$$

are minimized. Intuitively, the unnoised variable contains information about the noised one, so including it can enhance noise prediction and reduce the loss. However, this inclusion may also introduce dependence between the predicted noise and the unnoised variable. Crucially, we expect this dependence to differ between the causal and anticausal directions, providing a signal for identifying the correct causal direction. Specifically, we expect that the independence test outcomes for the pairs $(X, f_{\varepsilon_Y}(\tilde{Y}, X))$ and $(Y, f_{\varepsilon_X}(\tilde{X}, Y))$ to differ. We now formalize this intuition.

Causal Direction: Denoising \tilde{Y} and Testing Independence between X and $f_{\varepsilon_Y}(\tilde{Y}, X)$ Given infinite data, the best estimators $f_{\varepsilon_Y}^*, f_{\varepsilon_X}^*$ of the MSE loss converges to the conditional expectation [38]. This implies that the prediction of injected ε_Y equals

$$\hat{\varepsilon}_Y = \mathbb{E}[\varepsilon_Y | \tilde{Y}, X]. \quad (4.3)$$

208 Substituting $Y = X + \varepsilon_1$ into $\tilde{Y} = Y + \varepsilon_Y$, we have:

$$\tilde{Y} - X = \varepsilon_1 + \varepsilon_Y. \quad (4.4)$$

209 Next, we will show that $\tilde{Y} - X$ is a sufficient statistics for $\mathbb{E}[\varepsilon_Y | \tilde{Y}, X]$, i.e., $\mathbb{E}[\varepsilon_Y | \tilde{Y}, X] = \mathbb{E}[\varepsilon_Y | \tilde{Y} - X]$. To see this, we observe that since $X \perp\!\!\!\perp \varepsilon_1$ and $X \perp\!\!\!\perp \varepsilon_Y$, we have $\varepsilon_Y \perp\!\!\!\perp X | \varepsilon_1 + \varepsilon_Y \implies \varepsilon_Y \perp\!\!\!\perp X | \tilde{Y} - X$. This implies that

$$\hat{\varepsilon}_Y = \mathbb{E}[\varepsilon_Y | \tilde{Y}, X] = \mathbb{E}[\varepsilon_Y | X, \tilde{Y} - X] = \mathbb{E}[\varepsilon_Y | \tilde{Y} - X] = \mathbb{E}[\varepsilon_Y | \varepsilon_1 + \varepsilon_Y], \quad (4.5)$$

212 where the second equality is due to the parametrization of \tilde{Y} and X ; the third equality is due to
213 $\varepsilon_Y \perp\!\!\!\perp X | \varepsilon_1 + \varepsilon_Y \implies \varepsilon_Y \perp\!\!\!\perp X | \tilde{Y} - X$, and the last equality is due to Eq. (4.4).

214 Now, as our conditional expectation in Eq. (4.5) is shown to consist of terms entirely independent of
215 X , we have that our predicted noise is independent of the un-noised conditioning variable:

$$\hat{\varepsilon}_Y \perp\!\!\!\perp X. \quad (4.6)$$

216 **Anticausal Direction: Denoising \tilde{X} and Testing Independence between Y and $f_{\varepsilon_X}(\tilde{X}, Y)$** In
217 the anticausal direction, we repeat the same calculation and observe that the noise prediction is no
218 longer independent of the input unnoised variable. First, substituting $Y = X + \varepsilon_1$ into $\tilde{X} = X + \varepsilon_X$,
219 we obtain

$$\tilde{X} - Y = -\varepsilon_1 + \varepsilon_X. \quad (4.7)$$

220 We note that the same argument in the causal direction no longer works here as $\tilde{X} - Y$ is not a
221 sufficient statistic for $\mathbb{E}[\varepsilon_X | \tilde{X}, Y]$. In fact, we can show that

222 **Lemma 4.1.** $\hat{\varepsilon}_X = \mathbb{E}[\varepsilon_X | \tilde{X}, Y] \not\perp\!\!\!\perp Y$.

223 The proof of Lemma 4.1 (Appendix E.11) proceeds by contradiction. While prior diffusion-based
224 approaches have focused on leveraging diffusion to estimate the Jacobian of the score function [59],
225 to our knowledge we are the first to point out an asymmetry arising from the independence of the
226 predicted noise. Although the intuition is developed on a simple linear DGP, we hypothesize that
227 the same argument generalizes to nonlinear DGPs, leading to more dependent predicted noise in the
228 anticausal direction.

229 4.2 Theoretical Guarantees

230 Building on the intuition that we developed in Section 4.1, we build a decision rule that identifies
231 the correct causal direction according to which denoising process (denoising \tilde{Y} or \tilde{X}) leads to a
232 prediction that is less dependent on the unnoised variable.

233 **Decision Rule 1** (Bivariate Denoising Diffusion (BiDD)). *Let $\hat{\varepsilon}_Y = \epsilon_{Y,\theta}(\tilde{Y}, X)$, $\hat{\varepsilon}_X = \epsilon_{X,\theta}(\tilde{X}, Y)$
234 be the predictions of the noise added to Y, X , respectively. Given a mutual information estimator
235 $MI(\cdot)$, if $MI(\hat{\varepsilon}_Y, X) \leq MI(\hat{\varepsilon}_X, Y)$, conclude that X causes Y , else conclude that Y causes X .*

236 When the ANM-UM reduces to ANM (i.e., when Lemma 2.3 does hold), we can guarantee the
237 correctness of Decision Rule 1 (Theorem 4.2, proof in Appendix E.12).

238 **Theorem 4.2** (Consistency of Decision Rule 1). *Suppose X, Y follow Eq. (2.1), Assumptions 2.1 and
239 2.2 hold, and no nonlinear mediator exists. Then, given a consistent mutual information estimator
240 and infinite data, Decision Rule 1 correctly recovers the causal direction between X, Y .*

241 We conjecture that Decision Rule 1 remains consistent for cases when the ANM-UM does not reduce
242 to ANM, such as when there is a nonlinear mediator. We validate this conjecture empirically (Section
243 5), finding that our approach performs well across a wide variety of DGPs.

244 4.3 BiDD: A Practical Bivariate Denoising Diffusion Approach

245 Guided by the intuition developed in the linear case, we now present *BiDD*, a practical method for
246 inferring causal direction based on asymmetries in the independence of denoising estimates.

247 BiDD fits two conditional diffusion models, one for each direction. For $B \rightarrow A$, we corrupt A with
248 noise and train to recover it given B , and vice versa. We then compare dependence between predicted
249 noise and the condition, choosing the direction with lower dependence.

Noise Prediction We train a neural network to reconstruct the Gaussian noise injected into a noised sample \tilde{A}_t , conditioned on B . Our training follows the standard denoising diffusion framework of Ho et al. [17] and its conditional extensions [56].

Let $\{\alpha_t\}_{t=1}^T$ denote a fixed noise schedule and let $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ be its cumulative product. For a variable A and a diffusion timestep t , we define the noised version:

$$\tilde{A}_t = \sqrt{\bar{\alpha}_t} A + \sqrt{1 - \bar{\alpha}_t} \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 1). \quad (4.8)$$

Given (\tilde{A}_t, B, t) , the model $\varepsilon_{A,\theta}$ is trained to minimize the noise prediction loss:

$$L_{\text{CDM}} = \mathbb{E}_{A,B,\varepsilon,t} \left[\|\varepsilon - \varepsilon_\theta(\tilde{A}_t, B, t)\|^2 \right], \quad t \sim \text{Unif}(\{1, \dots, T\}).$$

At each iteration, we sample a timestep t , generate \tilde{A}_t , and update ε_θ by minimizing L_{CDM} with stochastic gradient descent over E epochs. This yields a trained model $\varepsilon_{A,\theta}$, which predicts $\hat{\varepsilon}_A = \varepsilon_{A,\theta}(\tilde{A}_t, B, t)$.

Dependence Testing After training, we evaluate the model on the test set. For each diffusion timestep $t = 1, \dots, T$, we generate noised inputs to estimate the dependence between the predicted noise and the conditioning variable. Specifically, for each test sample (A_i, B_i) , we construct k noised versions $\{\tilde{A}_{t,1}^{(i)}, \dots, \tilde{A}_{t,k}^{(i)}\}$ by sampling independent noise $\varepsilon \sim \mathcal{N}(0, 1)$ k times.

We then apply the trained model to obtain noise predictions $\hat{\varepsilon}_{A,i,j} = \varepsilon_{A,\theta}(\tilde{A}_{t,j}^{(i)}, B_i, t)$ for each noised sample. The mutual information $MI_{A,t}$ is computed between the predicted noises and the conditioning variable B as $MI_{A,t} = MI(\hat{\varepsilon}_A, B)$.

We repeat the procedure in the reverse direction by training a second model that predicts $\hat{\varepsilon}_B$ from noised B and conditioning on A , and compute $MI_{B,t}$ analogously.

Inferring Causal Direction To determine the causal direction, we compare $MI_{A,t}$ and $MI_{B,t}$ for each timestep t . We count how often one direction yields a lower mutual information value and select the direction that does so more frequently across timesteps. A formal description of this procedure is provided in Algorithm 1 in Appendix G.2.

While our theoretical framework assumes sample splitting between training and testing, we find in practice that using the full dataset for both training and dependence estimation often improves performance, consistent with observations from prior work [20]. Therefore, we empirically evaluate two variants: BiDD_{Test}, which uses a held-out test set for dependence estimation, and BiDD_{Total}, which uses the full dataset. Additional implementation details, including learning rate, optimizer, noise schedule, and estimator configuration, are provided in Appendix G.2.

Computational Complexity The computational complexity of BiDD involves two main stages. Firstly, the training of two conditional denoising diffusion models, each for E epochs over m_{train} training samples. If C_{step} denotes the cost of a single neural network training step (forward pass, loss computation, backward pass, and parameter update) per sample, this stage has a complexity of $O(E \cdot m_{\text{train}} \cdot C_{\text{step}})$. Secondly, the inference stage as per Decision Rule 1 requires generating noise predictions for $l = k \cdot m_{\text{eval}}$ evaluation samples per timestep T in the model (costing $O(m_{\text{eval}} \cdot k \cdot T \cdot C_{\text{fwd}})$, where C_{fwd} is the neural network forward pass cost) and computing two mutual information (MI) estimates. If $C_{\text{MI},l}$ is the cost for one MI estimation on l samples, this adds $C_{\text{MI},l}$ to the inference cost. The overall computational complexity of BiDD is thus $O(E \cdot m_{\text{train}} \cdot C_{\text{step}} + T(m_{\text{eval}} \cdot k \cdot C_{\text{fwd}} + C_{\text{MI},l}))$, which is typically dominated by the $O(E \cdot m_{\text{train}} \cdot C_{\text{step}})$ training component.

5 Experimental Results

We evaluate BiDD on synthetic data with linear, nonlinear invertible, and nonlinear non-invertible mechanisms, as well as a real-world dataset [58]. BiDD achieves state-of-the-art and consistent performance across settings, while most baselines performs poorly in at least one setting.

5.1 Setup

Synthetic Data Details. We produce synthetic bivariate causal pairs under the following ANM-UM (Eq 2.1), with varying causal mechanisms, exogenous noise distributions, sample size, and number

Method Noise	Linear	Neural Net		Quadratic		Tanh	
	Unif.	Gauss.	Unif.	Gauss.	Unif.	Gauss.	Unif.
BiDD _{Total}	0.77	.93	.90	1.00	1.00	0.80	.93
BiDD _{Test}	0.73	.97	.97	1.00	1.00	0.60	0.83

Table 1: Accuracy of BiDD_{Total} and BiDD_{Test} across different transformation-noise combinations, with *no mediators* and $n = 1000$.

Method Noise	Linear	Neural Net		Quadratic		Tanh	
	Unif.	Gauss.	Unif.	Gauss.	Unif.	Gauss.	Unif.
BiDD _{Total}	0.83	<u>0.87</u>	<u>0.97</u>	1.00	1.00	0.80	<u>0.83</u>
BiDD _{Test}	0.80	<u>0.87</u>	1.00	1.00	1.00	0.63	0.77
CANM	0.10	0.93	0.87	1.00	1.00	0.50	0.10
Adascore	<u>0.93</u>	0.73	0.77	0.43	0.13	0.67	1.00
NoGAM	1.00	0.43	0.43	0.00	1.00	0.63	1.00
SCORE	1.00	0.73	0.57	0.43	1.00	0.50	1.00
DagmaL	0.13	0.17	0.10	0.00	0.00	0.00	0.07
CAM	0.03	0.77	0.80	1.00	1.00	0.93	0.13
PNL	0.73	0.83	0.70	0.83	0.83	0.67	0.70
RESIT	<u>0.93</u>	0.70	0.67	1.00	1.00	<u>0.87</u>	1.00
DLiNGAM	1.00	0.13	0.13	0.10	0.40	<u>0.17</u>	1.00
Var-Sort	0.43	0.57	0.63	0.47	0.57	0.33	0.60

Table 2: Accuracy of methods across different transformation-noise combinations, with *one mediator* and $n = 1000$. **Bold** indicates best, underline indicates second-best.

of mediators. We use linear mechanisms with randomly drawn coefficients; we use both invertible (tanh) and non-invertible (quadratic, neural networks with randomly initialized weights [32, 25, 16]) nonlinear mechanisms. We use both Uniform and Gaussian noise (excluding the linear Gaussian case to ensure identifiability). We standardize the data to mean 0 and variance 1 to ensure that the simulated data is sufficiently challenging; methods are evaluated on 20 randomly generated seeds in each experimental setting. See Appendix G.1.1 for details on specific parameters used for each DGP.

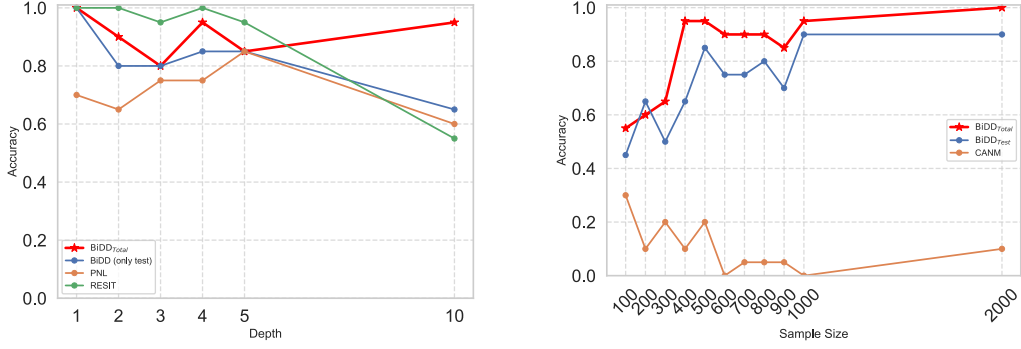
Real-World Data Details. To confirm the real-world applicability of our approach, we test BiDD on the Tübingen Cause-Effect dataset [44], a widely used bivariate discovery benchmark that consists of 99 causal pairs that may have unobserved mediators. Due to runtime issues with baselines, we subsample the dataset of each causal pair by randomly selecting up to $n = 3000$ data points.

Baselines and Evaluation. We benchmark BiDD against a mix of classical and SOTA methods: we compare against three Residual-Independence methods (DirectLiNGAM, RESIT, PNL), three Score-Matching methods (SCORE, NoGAM, Adascore), two MSE-Minimization methods (DagmaLinear [3], CAM), and the only hidden mediator method in the literature (CANM). We include the heuristic algorithm Var-Sort, which exploits artifacts common to simulated ANMs [53], to show that BiDD performance is not driven by such shortcuts. Similar to [44], we use the *accuracy* for *forced decisions*, which corresponds to forcing the compared methods to decide the causal direction.

5.2 Results

Synthetic Data.

We first examine the performance of BiDD_{Total} in the traditional ANM setting (no unobserved mediator): Table 1 shows results for BiDD_{Total} on data generated by different mechanism-noise combinations and sample size $n = 1000$. We observe the robust performance of BiDD_{Total}, achieving $\geq 77\%$ accuracy across all mechanisms. This empirically confirms the theoretical correctness guarantee given in Section 4 (Theorem 4.2).



(a) $n = 1000$ with increasing number of mediators. (b) One unobserved mediator, with increasing n .

Figure 2: Tanh mechanism, Uniform noise setting.

Method	BiDD _{total}	BiDD _{test}	CANM	CAM	Adascore	Entropy
Accuracy	0.64	0.60	0.47	0.56	0.06	0.36
Method	DagmaLinear	DirectLiNGAM	NoGAM	RESIT	PNL	SCORE
Accuracy	0.30	0.51	0.69	0.62	0.61	<u>0.65</u>

Table 3: Accuracy for Tübingen Causal Pairs dataset, $n = 3000$

We now examine how BiDD_{Total} performs under unmeasured mediators: in Table 2 we display results for different mechanism-noise combinations, each generated with one unobserved mediator (i.e., $Y = f_2(f_1(X) + \varepsilon_1) + \varepsilon_2$) and sample size $n = 1000$. We observe the robustness of BiDD_{Total}, as it achieves $\geq 80\%$ accuracy across all experimental setups, getting the first or second best accuracy 5/7 times. In contrast, all baselines except PNL and RESIT perform extremely poorly ($\leq 50\%$) in at least two settings. PNL’s performance is significantly lower ($\sim 10\% - 20\%$) than BiDD_{Total} in almost every setting, while RESIT struggles in the neural network setting ($\leq 70\%$). The other hidden mediator method, CANM, performs poorly ($\leq 50\%$) for invertible mechanisms (linear, tanh), even for Gaussian noise, which is consistent with our analysis of CANM’s behavior under posterior collapse (see Appendix E.9). The degraded baseline performance when the ANM assumption is violated highlights the limited applicability of current bivariate ANM methods.

In Figure 2, we investigate how BiDD_{Total} performs under fixed sample size ($n = 1000$) and varying depth (Figure 2a), or fixed depth (one mediator) and varying sample size (Figure 2b), in the tanh mechanism, uniform noise setting. In Figure 2b we observe that as the number of mediators increases, the performance of RESIT and PNL both degrade (to $\sim 60\%$), while BiDD_{Total} remains performant ($\sim 95\%$). This shows that the performance of Residual-Independence based methods (RESIT and PNL) is sensitive to the number of mediators, while our denoising diffusion approach remains robust.

In Figure 2b we observe the consistency of BiDD_{Total}, as its accuracy approaches 100% while CANM does not improve, and in fact seems to decrease in performance. This points to CANM being inconsistent in settings with unmeasured mediators, rather than merely having finite sample issues.

Real-world Data The results of Tübingen dataset are presented in Table 3: BiDD_{total} (64%) performs comparable to the best baselines, NoGAM (69%) and SCORE (65%), outperforming the rest of the methods. This confirms the robustness of BiDD across a diverse range of real-world setups.

Discussion. Future work includes extending BiDD to be robust to latent confounding, and analyzing its potential consistency in settings where ANM-UM cannot be reformulated as an ANM.

References

- [1] Aczél, J. *On applications and theory of functional equations*. Academic Press, 2014.
- [2] Addo, P. M., Manibialoa, C., and McIsaac, F. Exploring nonlinearity on the co2 emissions, economic production and energy use nexus: a causal discovery approach. *Energy Reports*, 7: 6196–6204, 2021.
- [3] Bello, K., Aragam, B., and Ravikumar, P. DAGMA: Learning DAGs via M-matrices and a Log-Determinant Acyclicity Characterization. In *Advances in Neural Information Processing Systems*, 2022.
- [4] Bloebaum, P., Janzing, D., Washio, T., Shimizu, S., and Schoelkopf, B. Cause-effect inference by comparing regression errors. In Storkey, A. and Perez-Cruz, F. (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 900–909. PMLR, 09–11 Apr 2018. URL <https://proceedings.mlr.press/v84/bloebaum18a.html>.
- [5] Bühlmann, P., Peters, J., and Ernest, J. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6), December 2014. ISSN 0090-5364. doi: 10.1214/14-AOS1260. URL <http://arxiv.org/abs/1310.1533>. arXiv:1310.1533 [cs, stat].
- [6] Cai, R., Qiao, J., Zhang, K., Zhang, Z., and Hao, Z. Causal discovery with cascade nonlinear additive noise models. *arXiv preprint*, arXiv:1905.09442, 2019.
- [7] Cai, R., Ye, J., Qiao, J., Fu, H., and Hao, Z. Fom: Fourth-order moment based causal direction identification on the heteroscedastic data. *Neural Networks*, 124:193–201, 2020. ISSN 0893-6080. doi: <https://doi.org/10.1016/j.neunet.2020.01.006>. URL <https://www.sciencedirect.com/science/article/pii/S0893608020300083>.
- [8] Casella, G. and Berger, R. L. *Statistical Inference*. Duxbury, 2nd edition, 2002.
- [9] Chickering, D. M. Learning Equivalence Classes of Bayesian Network Structures. *Journal of Machine Learning Research*, 2013.
- [10] Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., and Carin, L. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. (arXiv:1903.10145), June 2019. doi: 10.48550/arXiv.1903.10145. URL <http://arxiv.org/abs/1903.10145>. arXiv:1903.10145 [cs].
- [11] Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.
- [12] Halmos, P. R. *Measure Theory*. D. Van Nostrand Company, 1950.
- [13] He, J., Spokoyny, D., Neubig, G., and Berg-Kirkpatrick, T. Lagging inference networks and posterior collapse in variational autoencoders. (arXiv:1901.05534), January 2019. doi: 10.48550/arXiv.1901.05534. URL <http://arxiv.org/abs/1901.05534>. arXiv:1901.05534 [cs].
- [14] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. β -vae: Learning basic visual concepts with a constrained variational framework. 2017.
- [15] Hiremath, S., Maasch, J., Gao, M., Ghosal, P., and Gan, K. Hybrid top-down global causal discovery with local search for linear and nonlinear additive noise models. *NeurIPS 2024*, 2024. URL <https://arxiv.org/abs/2405.14496>. <https://arxiv.org/abs/2405.14496>.
- [16] Hiremath, S., Ghosal, P., and Gan, K. Losam: Local search in additive noise models with mixed mechanisms and general noise for global causal discovery. *UAI 2025*, 2025. URL <https://arxiv.org/abs/2410.11759>. <https://arxiv.org/abs/2410.11759>.
- [17] Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

- [18] Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, 2008.
- [19] Huang, B., Zhang, K., Lin, Y., Schölkopf, B., and Glymour, C. Generalized Score Functions for Causal Discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1551–1560, London United Kingdom, July 2018. ACM. ISBN 978-1-4503-5552-0. doi: 10.1145/3219819.3220104. URL <https://dl.acm.org/doi/10.1145/3219819.3220104>.
- [20] Immer, A., Schultheiss, C., Vogt, J. E., Schölkopf, B., Bühlmann, P., and Marx, A. On the Identifiability and Estimation of Causal Location-Scale Noise Models, June 2023. URL <http://arxiv.org/abs/2210.09054>. arXiv:2210.09054 [cs, stat].
- [21] Janzing, D. and Scholkopf, B. Causal Inference Using the Algorithmic Markov Condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, October 2010. ISSN 0018-9448, 1557-9654. doi: 10.1109/TIT.2010.2060095. URL <http://ieeexplore.ieee.org/document/5571886/>.
- [22] Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniušis, P., Steudel, B., and Schölkopf, B. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183:1–31, May 2012. ISSN 00043702. doi: 10.1016/j.artint.2012.01.002. URL <https://linkinghub.elsevier.com/retrieve/pii/S0004370212000045>.
- [23] Janzing, D., Peters, J., Mooij, J., and Schölkopf, B. Identifying confounders using additive noise models. *arXiv preprint arXiv:1205.2640*, 2012.
- [24] Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, Cambridge, Massachusetts, 2017. ISBN 978-0-262-03731-0. URL <https://mitpress.mit.edu/9780262037310/elements-of-causal-inference/>.
- [25] Ke, N., Chiappa, S., Wang, J., Bornschein, J., Goyal, A., Rey, M., Weber, T., Botvinick, M., Mozer, M., and Rezende, D. Learning to induce causal structure. *International Conference on Learning Representations*, 2023.
- [26] Kelly, M., Longjohn, R., and Nottingham, K. The UCI Machine Learning Repository. <https://archive.ics.uci.edu>, 2025. Accessed 22 May 2025.
- [27] Kolmogorov, A. N. Three approaches to the quantitative definition of information'. *Problems of information transmission*, 1(1):1–7, 1965.
- [28] Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Physical Review E*, 69(6):066138, June 2004. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.69.066138. URL <https://link.aps.org/doi/10.1103/PhysRevE.69.066138>.
- [29] Lam, W.-Y., Andrews, B., and Ramsey, J. Greedy Relaxations of the Sparsest Permutation Algorithm, 2022. URL <https://proceedings.mlr.press/v180/lam22a/lam22a.pdf>.
- [30] Lee, J. J., Srinivasan, R., Ong, C. S., Alejo, D., Schena, S., Shpitser, I., Sussman, M., Whitman, G. J., and Malinsky, D. Causal determinants of postoperative length of stay in cardiac surgery using causal graphical learning. *The Journal of Thoracic and Cardiovascular Surgery*, pp. S002252232200900X, August 2022. ISSN 00225223. doi: 10.1016/j.jtcvs.2022.08.012. URL <https://linkinghub.elsevier.com/retrieve/pii/S002252232200900X>.
- [31] Lehmann, E. L. and Casella, G. *Theory of Point Estimation*. Springer Texts in Statistics. Springer, 2nd edition, 2003. ISBN 978-0-387-98502-3.
- [32] Lippe, P., Cohen, T., and Gavves, E. Efficient Neural Causal Discovery without Acyclicity Constraints, February 2022. URL <http://arxiv.org/abs/2107.10483>. arXiv:2107.10483 [cs, stat].
- [33] Liu, M., Sun, X., Qiao, Y., and Wang, Y. Causal discovery via conditional independence testing with proxy variables. *ICML 2024*, 2024. URL <https://arxiv.org/pdf/2305.05281>.

- 440 [34] Maeda, T. N. and Shimizu, S. Causal Additive Models with Unobserved Variables. In *Pro-*
 441 *ceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 10,
 442 2021.
- 443 [35] Marx, A. and Vreeken, J. Identifiability of cause and effect using regularized regression. In
 444 *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery &*
 445 *Data Mining*, pp. 852–861, 2019.
- 446 [36] Miao, W., Geng, Z., and Tchetgen Tchetgen, E. Identifying causal effects with proxy vari-
 447 ables of an unmeasured confounder. 2018. URL [https://academic.oup.com/biomet/](https://academic.oup.com/biomet/article-abstract/105/4/987/5073056?redirectedFrom=fulltext)
 448 [article-abstract/105/4/987/5073056?redirectedFrom=fulltext](https://academic.oup.com/biomet/article-abstract/105/4/987/5073056?redirectedFrom=fulltext).
- 449 [37] Miao, W., Hu, W., Ogburn, E., and Zhou, X.-H. Identifying effects of multiple treatments in
 450 the presence of unmeasured confounding. *Journal of American Statistical Association*, 2022.
 451 URL [https://www.tandfonline.com/doi/full/10.1080/01621459.2021.2023551#](https://www.tandfonline.com/doi/full/10.1080/01621459.2021.2023551#abstract)
 452 [abstract](https://www.tandfonline.com/doi/full/10.1080/01621459.2021.2023551#abstract).
- 453 [38] Montagna, F., Noceti, N., Rosasco, L., Zhang, K., and Locatello, F. Causal Discovery with Score
 454 Matching on Additive Models with Arbitrary Noise. In *Proceedings of the 2nd Conference*
 455 *on Causal Learning and Reasoning*. arXiv, April 2023. URL [http://arxiv.org/abs/2304.](http://arxiv.org/abs/2304.03265)
 456 [03265](http://arxiv.org/abs/2304.03265). arXiv:2304.03265 [cs, stat].
- 457 [39] Montagna, F., Noceti, N., Rosasco, L., Zhang, K., and Locatello, F. Scalable Causal Discovery
 458 with Score Matching. In *Proceedings of the 2nd Conference on Causal Learning and Reasoning*.
 459 arXiv, April 2023. URL <http://arxiv.org/abs/2304.03382>. arXiv:2304.03382 [cs, stat].
- 460 [40] Montagna, F., Faller, P. M., Bloebaum, P., Kirschbaum, E., and Locatello, F. Score matching
 461 through the roof: linear, nonlinear, and latent variables causal discovery. *arXiv preprint*
 462 *arXiv:2407.18755*, 2024.
- 463 [41] Mooij, J. and Janzing, D. Distinguishing between cause and effect. *JMLR Workshop and*
 464 *Conference Proceedings*, 6:147–156, 2010.
- 465 [42] Mooij, J., Janzing, D., Peters, J., and Schölkopf, B. Regression by dependence minimization
 466 and its application to causal inference in additive noise models. In *Proceedings of the 26th*
 467 *annual international conference on machine learning*, pp. 745–752, 2009.
- 468 [43] Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. Distinguishing cause
 469 from effect using observational data: methods and benchmarks. *Journal of Machine Learning*
 470 *Research*, 17(32):1–102, 2016.
- 471 [44] Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. Distinguishing Cause
 472 from Effect Using Observational Data: Methods and Benchmarks. *Journal of Machine Learning*
 473 *Research*, 17, 2016.
- 474 [45] Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. Distinguishing cause
 475 from effect using observational data: methods and benchmarks. *Journal of Machine Learning*
 476 *Research*, 17(32):1–102, 2016.
- 477 [46] Ng, I., Ghassami, A., and Zhang, K. On the role of sparsity and dag constraints for learning
 478 linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.
- 479 [47] Park, G. Identifiability of additive noise models using conditional variances. *Journal of Machine*
 480 *Learning Research*, 21(75):1–34, 2020.
- 481 [48] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z.,
 482 Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani,
 483 A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. *PyTorch: an imperative*
 484 *style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA,
 485 2019.
- 486 [49] Pearl, J. Myth, Confusion, and Science in Causal Analysis. 2000.

- [50] Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. Causal Discovery with Continuous Additive Noise Models, April 2014. URL <http://arxiv.org/abs/1309.6779>. arXiv:1309.6779 [stat].
- [51] Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.
- [52] Pham, T., Maeda, T. N., and Shimizu, S. Causal additive models with unobserved causal paths and backdoor paths. *arXiv preprint arXiv:2502.07646*, 2025.
- [53] Reischach, A., Seiler, C., and Weichwald, S. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34: 27772–27784, 2021.
- [54] Reischach, A. G., Tami, M., Seiler, C., Chambaz, A., and Weichwald, S. A Scale-Invariant Sorting Criterion to Find a Causal Order in Additive Noise Models. In *37th Conference on Neural Information Processing Systems*. arXiv, October 2023. URL <http://arxiv.org/abs/2303.18211>. arXiv:2303.18211 [cs, stat].
- [55] Rolland, P., Cevher, V., Kleindessner, M., Russel, C., Scholkopf, B., Janzing, D., and Locatello, F. Score Matching Enables Causal Discovery of Nonlinear Additive Noise Models. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.
- [56] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models . pp. 10674–10685, June 2022. doi: 10.1109/CVPR52688.2022.01042. URL <https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042>.
- [57] Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10(1):2553, December 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-10105-3. URL <http://www.nature.com/articles/s41467-019-10105-3>.
- [58] Sachs, K., Perez, O., Pe’er, D., Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.
- [59] Sanchez, P., Liu, X., O’Neil, A. Q., and Tsaftaris, S. A. Diffusion models for causal discovery via topological ordering. *arXiv preprint arXiv:2210.06201*, 2022.
- [60] Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On Causal and Anticausal Learning.
- [61] Shimizu, S., Hoyer, P. O., Hyvarinen, A., and Kerminen, A. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.
- [62] Shimizu, S., Inazumi, T., Sogawa, Y., Hyvarinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., Bollen, K., and Hoyer, P. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr):1225–1248, 2011.
- [63] Spirtes, P. An Anytime Algorithm for Causal Inference. In *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*, volume R3, pp. 278–285. PMLR, 2001.
- [64] Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*, volume 81 of *Lecture Notes in Statistics*. Springer New York, New York, NY, 2000. ISBN 978-1-4612-7650-0 978-1-4612-2748-9. doi: 10.1007/978-1-4612-2748-9. URL <http://link.springer.com/10.1007/978-1-4612-2748-9>.
- [65] Tagasovska, N., Chavez-Demoulin, V., and Vatter, T. Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery. In *International Conference on Machine Learning*, pp. 9311–9323. PMLR, 2020.

- 536 [66] Turing, A. M. et al. On computable numbers, with an application to the entscheidungsproblem.
537 *J. of Math*, 58(345-363):5, 1936.
- 538 [67] Xu, S., Mian, O. A., Marx, A., and Vreeken, J. Inferring cause and effect in the presence of
539 heteroscedastic noise. In *International Conference on Machine Learning*, pp. 24615–24630.
540 PMLR, 2022.
- 541 [68] Xu, Z., Li, Y., Liu, C., and Gui, N. Ordering-based causal discovery for linear and nonlinear
542 relations. In *Proceedings of the 37th Conference on Neural Information Processing Systems*
543 (*NeurIPS*), 2024.
- 544 [69] Yuan, Y. and Qu, A. De-confounding causal inference using latent multiple-mediator pathways.
545 *Journal of the American Statistical Association*, 119(547):2051–2065, 2023. doi: 10.1080/
546 01621459.2023.2240461. URL <https://doi.org/10.1080/01621459.2023.2240461>.
- 547 [70] Zhang, K. and Hyvarinen, A. On the Identifiability of the Post-Nonlinear Causal Model.
548 *Uncertainty in Artificial Intelligence*, 2009.
- 549 [71] Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags with no tears: Continuous
550 optimization for structure learning. *Advances in neural information processing systems*, 31,
551 2018.
- 552 [72] Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. Learning sparse nonparametric
553 dags. In *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425. Pmlr,
554 2020.
- 555 [73] Zhu, Z., Locatello, F., and Cevher, V. Sample Complexity Bounds for
556 Score-Matching: Causal Discovery and Generative Modeling, 2023. URL
557 [https://proceedings.neurips.cc/paper_files/paper/2023/file/
558 0a3dc35a2391cabcb59a6b123544e3db-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/0a3dc35a2391cabcb59a6b123544e3db-Paper-Conference.pdf).

559 Appendix

560 A Notation

561	$\text{Pa}(x_i)$	The set of parent vertices of x_i .
562	Z_i	The i 'th Mediator
563	f_i	Arbitrary function, generating Z_i
564	ε_i	An independent noise term sampled from an arbitrary distribution.
565	$X \perp\!\!\!\perp Y$	X is independent of Y
566	$X \not\perp\!\!\!\perp Y$	X is not independent of Y
567	$X \rightarrow Y$	X is a parent of Y in the ANM-UM
568	$X \dashrightarrow Y$	X is a parent of Y in the ANM-UM
569	$[T]$	Set of integers $\{1, \dots, T\}$
570	$\{Z_i\}_i$	Set of Mediators $\{Z_1, \dots, Z_T\}$
571	A_i	i 'th data point
572	$\{A_i, B_i\}_{i=1}^n$	Collection of n data points
573	$\mathbb{E}[Y X]$	Conditional expectation of Y given X
574	f^{-1}	The inverse function of f
575	∇	Gradient operator
576	$\text{Var}[Y X]$	Variance of Y given X
577	\tilde{X}	Noised version of X
578	f_{ε_X}	Predictor of the noise added to X
579	$f_{\varepsilon_X}^*$	The best predictor of the noise added to X . Best means lowest MSE.
580	$MI(\cdot)$	Empirical mutual-information estimator
581	ε_A	Noise added to $\tilde{A} = A + \varepsilon_A$
582	$\varepsilon_{A,\theta}$	A learned model for predicting ε_A with parameters θ
583	$O(\cdot)$	Big-O (Landau) notation for asymptotic upper bound
584	\exists	There exists
585	∂	Partial derivative
586	$\mathbb{E}_X[\cdot]$	Expectation given X
587	$T[\cdot]$	Minimal-sufficient statistic
588	$\frac{\partial f}{\partial x}$	Partial derivative with respect to x
589	$\Delta f(u)$	Difference of function shifted by a constant c at u , $\Delta f(u) = f(u + x) - f(u)$
590	f_X	Probability density function of X
591	$g_{X,Y}$	Joint density of X and Y
592	σ	Sigma algebra
593	$\text{Cov}(X, Y)$	Covariance of X and Y
594	$\ X\ ^2$	Euclidian norm of X
595	$\mathcal{N}(0, 1)$	Normal distribution with mean 0 and variance 1
596	$\mathcal{U}(-1, 1)$	Continuous uniform distribution between -1 and 1
597	$\text{Unif}(\{1, \dots, N\})$	Discrete uniform distribution between N elements
598	$\mathbf{1}$	Indicator function
599	\mathcal{D}	Bivariate Dataset, consisting of n observations of A and B

B Further Discussion of Related Bivariate Methods

In this section, we further clarify the difference in modeling assumptions employed by ANM-based methods, LSM-based methods, and ICM-based methods.

Location Scale Noise Models. LSMs follow $Y = f(X) + g(X)\varepsilon$, where $X \perp\!\!\!\perp \varepsilon$, allowing the noise term ε to be scaled and shifted for each X value, according to $g(X)$. This increased flexibility models the possibility of heteroskedastic noise (noise dependent on the input X).

LSMs are similar to ANM-UM, in that both include the vanilla ANM ($Y = f(X) + g(X)\varepsilon$) as a subcase. However, we note that while LSMs and ANM-UM overlap, ANM-UM cover many cases which LSMs do not. For example, $y = e^{x+\varepsilon_1} + \varepsilon_2$ can be modeled as a ANM-UM, while LSMs do not admit such a representation. In general, ANM-UM can admit much more complicated joint distributions because they attempt to account for unobserved mediators: this introduces multiple (rather than one) independent noise distributions, with multiple transformations (rather than one) of the original input.

Additionally, methods developed to exploit LSMs have several drawbacks. They either lack theoretical guarantees, or they require parametric assumptions than the general LSM case. For example, [67] require linear mechanisms, while [20, 7] require Gaussian noise for correctness results.

Principle of Independent Mechanism Approaches The independence of cause and mechanism postulate [60] (ICM) states that the cause X should be independent of the mechanism that maps X to the effect Y . More concretely, this means that $X \dashrightarrow Y$ only if the shortest description of $P_{X,Y}$ is given by separate descriptions of $P_{Y|X}$ and P_X , in the sense that knowing P_X does not enable a shorter description of $P_{Y|X}$ (and vice versa) [22]. Here description length is understood in the sense of algorithmic information ("Kolmogorov complexity") [27].

The overall ICM postulate is a true generalization of the ANM-UM (as well as the LSM), as the functional mechanisms f_1, \dots, f_{T+1} and noise distributions $\varepsilon_1, \dots, \varepsilon_{T+1}$ (which make up the data generating process from X to Y) do not change for different input distributions of X . However, concrete methods developed to exploit ICM have many drawbacks. These generally follow from the fact that Kolmogorov complexity is known generally to not be computable [66]. Therefore, methods use approximations or proxies of the Kolmogorov complexity to develop heuristic approaches.

For example, [41] substitute the minimum message length principle, [65] use quantile scoring as a proxy for Kolmogorov complexity, and [35] leverage a condition on the parameter size of the true causal model implied by the Kolmogorov construction function. In general, methods based on the ICM do not come with strong identifiability results [35].

C Further Discussion of Related Global Methods

Recent work in global causal discovery has grappled with the issue of unobserved mediators, to various degrees of success. The multivariate version of Adascore [40] is the first score-matching method to handle unobserved confounders, but the authors clearly state that it fails to correctly recover the graph when unobserved mediators are present (See Examples 4, 5, and 7 in [40]). [34] showed that, under a further restriction of the ANM, where both the causal effects and error terms are additive (causal additive models, CAM, a subcase of ANM-UM), it is possible to recover the correct causal edge when all parents of a variable are measured, and otherwise leave the causal edge undecided if an unobserved mediator is a parent of an observed variable. In a recent extension of this work, [52] has shown that, under the CAM restriction, the correct causal edge can be recovered if the unmeasured mediator that is a parent of an observed variable is embedded in certain types of global graphical structures. However, neither of the latter two works comment on the general bivariate case involving unobserved mediators (AMM-UM), where additional global information may not be present.

Our bivariate method for handling unobserved mediators (BiDD) is motivated by the drawbacks of current ANM-based global discovery methods, as either they cannot handle unobserved mediators at all, fail to recover edges under hidden mediation, or can only do so under very narrow global graphical structures. Future work can incorporate BiDD as a subroutine in a global ANM-based discovery method, providing utility in real-world systems where hidden mediation abounds.

D Problem Setup

D.1 Relation of ANM-UM to ANM, PNL, and CANM

Note that ANM-UM can be represented as $Y = f_{T+1}(f_T(\dots) + \varepsilon_T) + \varepsilon_{T+1}$, where the last term inside the \dots is $f_1(X) + \varepsilon_1$.

The traditional ANM [50] models $Y = f(X) + \varepsilon_1$, where the key constraint is that $X \perp\!\!\!\perp \varepsilon_1$. If $T = 0$ for the ANM-UM, i.e., there are no unobserved mediators, then f_1 and f_{T+1} coincide, and Eq 2.1 reduces to $Y = f_1(X) + \varepsilon_1$, which is exactly the ANM.

The PNL [70] models $Y = g(f(X) + \varepsilon_1)$ where $X \perp\!\!\!\perp \varepsilon_1$ and g is an invertible nonlinear transformation. If $T = 1$ for the ANM-UM, f_2 is nonlinear and invertible, and $\varepsilon_{T+1} = 0$, then Eq 2.1 reduces to $Y = f_2(f_1(X) + \varepsilon_1)$, which is exactly the PNL.

The CANM [6] models $Y = f_{T+1}(f_T(\dots) + \varepsilon_T) + \varepsilon_{T+1}$ where all f_1, \dots, f_{T+1} are nonlinear. If all f_1, \dots, f_{T+1} in Eq 2.1 are nonlinear, then ANM-UM reduces to the CANM.

D.2 Identifiability Discussion

We first note that any ANM-UM (Eq 2.1) can be represented equivalently as

$$Y = F(X, \varepsilon_1, \dots, \varepsilon_T) + \varepsilon_{T+1} \quad (\text{D.1})$$

where $X, \varepsilon_1, \dots, \varepsilon_{T+1}$ are all mutually independent, and $F = f_{T+1}(f_T(\dots) + \varepsilon_T)$. For the ANM-UM to be identifiable, we require that there is no backwards ANM-UM that fits the anticausal direction $X \dashrightarrow Y$, i.e. there does not exist $G, \hat{\varepsilon}_1, \dots, \hat{\varepsilon}_{T+1}$ such that

$$X = G(Y, \hat{\varepsilon}_1, \dots, \hat{\varepsilon}_T) + \hat{\varepsilon}_{T+1} \quad (\text{D.2})$$

where $G = g_{T+1}(g_T(\dots) + \hat{\varepsilon}_T)$ and, additionally, $Y, \hat{\varepsilon}_1, \dots, \hat{\varepsilon}_{T+1}$ are mutually independent. Theorem 1 from [6] shows that for any $G, \hat{\varepsilon}_1, \dots, \hat{\varepsilon}_{T+1}$ which satisfy Eq D.2, we have that $Y, \hat{\varepsilon}_1, \dots, \hat{\varepsilon}_{T+1}$ are mutually independent (and thus ANM-UM unidentifiable) if and only if, $\hat{\varepsilon}_{T+1}$ takes a very particular form:

$$p_{\hat{\varepsilon}_{T+1}}(\hat{\varepsilon}_{T+1}) = \int e^{2\pi i \hat{\varepsilon}_{T+1} \cdot \nu} \frac{\int \int p(x) p(n) p_{\varepsilon_{T+1}}(y - f(x, n)) e^{-2\pi i x \cdot \nu} dn dx}{p(y) \int p(\hat{n}) e^{-2\pi i g(y, \hat{n}) \cdot \nu} d\hat{n}} d\nu \quad (\text{D.3})$$

where $n = \{\varepsilon_1, \dots, \varepsilon_T\}$ and $\hat{n} = \{\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_T\}$

Therefore, to ensure identifiability, Assumption 2.2 requires that for any such $G, \hat{\varepsilon}_1, \dots, \hat{\varepsilon}_{T+1}$ which satisfy Eq D.2, $\hat{\varepsilon}_{T+1}$ does not satisfy Eq D.3.

Cai et al. [6] show that, when ANM-UM can be represented as a traditional ANM (i.e., ANM-UM satisfies conditions in Lemma 2.3), the Assumption of 2.2 reduces to known identifiability constraints. For example, in Corollary 1 of [6], they show that if all mechanisms f_1, \dots, f_{T+1} in the ANM-UM are linear, then Assumption 2.2 reduces to requiring that at least one of $X, \{\varepsilon_i\}$ are non-Gaussian. This is exactly the constraint described by [61]. In Corollary 2 of [6], they show that if $T = 0$ in ANM-UM (no unobserved mediators), then Assumption 2.2 reduces to requiring that X, f_1, ε_1 satisfy the differential equation described in the identifiability assumptions of the general ANM in [18].

D.3 Nonlinear ANM Not Always Closed Under Marginalization

Let $X \rightarrow Z_1 \rightarrow Y$, which corresponds to the ANM-UM $X, Z_1 = f_1(X) + \varepsilon_1, Y = f_2(Z_1) + \varepsilon_2$. As $Y = f_2(f_1(X) + \varepsilon_1) + \varepsilon_2$, it is straightforward that Y follows an ANM if and only if $f_2(f_1(X) + \varepsilon_1)$ can be decomposed into the addition of a function of X and a function of ε_1 , i.e. $f_2(f_1(X) + \varepsilon_1) = A_1(X) + A_2(\varepsilon_1)$. Note that this follows the form of Pexider's equation $C(x + y) = D(x) + E(y)$ - it is known [1] that if C, D, E satisfy this equation, C, D, E must all be linear functions. Therefore, if f_2 is nonlinear, the ANM-UM does not reduce, and if f_2 is linear, then the ANM-UM does reduce.

689 D.4 Proof of Lemma 2.3

690 **Lemma 2.3** (Irreducible ANM-UM and Nonlinear Mediator). *Under ANM-UM (Eq. (2.1)) and*
 691 *Assump.s 2.1 and 2.2, Y does not admit a decomposition in Eq. (2.2) if and only if there exists a*
 692 *mediator Z_i such that $Y = h(Z_i) + \tilde{\varepsilon}$ for some nonlinear h , with $\tilde{\varepsilon} \perp\!\!\!\perp Z_i$. Additionally, we call such*
 693 *a mediator Z_i a nonlinear mediator.*

694 *Proof.* Suppose that there exists a mediator Z_i such that $Y = h(Z_i) + \tilde{\varepsilon}$ where h is nonlinear, and
 695 $Z_i \perp\!\!\!\perp \tilde{\varepsilon}$. Then, \exists some function F such that $Y = h(F(X, \varepsilon_1, \dots, \varepsilon_{i-1}) + \varepsilon_i) + \tilde{\varepsilon}$. Suppose for
 696 contradiction that Y admits an additive decomposition, i.e. $Y = h(F(X, \varepsilon_1, \dots, \varepsilon_{i-1}) + \varepsilon_i) +$
 697 $\tilde{\varepsilon} = A_1(X) + A_2(\varepsilon_1, \dots, \varepsilon_i) + A_3(\tilde{\varepsilon})$. Note that this implies $h(F(X, \varepsilon_1, \dots, \varepsilon_{i-1}) + \varepsilon_i) + \tilde{\varepsilon} =$
 698 $A_1(X) + A_2(\varepsilon_1, \dots, \varepsilon_i)$, which follows the form of Pexider’s equation, $C(x + y) = D(x) + E(y)$.
 699 It is known that if C, D, E satisfy this equation, C, D, E must all be linear functions [1]. However,
 700 this contradicts that h is nonlinear. Therefore, Y does not admit a decomposition in Eq 2.2.

701 Suppose that Y does not admit a decomposition in Eq 2.2. Suppose for contradiction that there does
 702 not exist a nonlinear mediator Z_i . This implies that f_2, \dots, f_{T+1} are all linear functions. Then, Y can
 703 be written as a linear function of Z_1 and noise terms $\varepsilon_2, \dots, \varepsilon_{T+1}$, i.e. $Y = \alpha Z_1 + \sum_{i=2}^{T+1} \beta_i \varepsilon_i$. Then,
 704 $Y = \alpha f_1(X) + \alpha \varepsilon_1 + \sum_{i=2}^{T+1} \beta_i \varepsilon_i$ which follows the additive decomposition in Eq 2.2. Therefore,
 705 there must exist a nonlinear mediator Z_i .

706 □

707 E Failure-mode of Prior Work + Proof of Lemma 4.1, Theorem 4.2, 708 Experimental Analysis of CANM

709 E.1 Decision Rules

710 **Decision Rule E.2** (Regression Residual-Independence). *Let e_1, e_2 be the residuals obtained from*
 711 *regressing Y onto X , and X onto Y (respectively). If $e_1 \perp\!\!\!\perp X, e_2 \not\perp\!\!\!\perp Y$, then conclude X causes Y .*
 712 *If $e_1 \not\perp\!\!\!\perp X, e_2 \perp\!\!\!\perp Y$, then conclude Y causes X . If $e_1 \perp\!\!\!\perp X, e_2 \perp\!\!\!\perp Y$, conclude neither causes each*
 713 *other. Otherwise, do not decide.*

714 **Decision Rule E.3** (Nonlinear ICA Residual-Independence). *Check to see if the hypothesis $X \dashrightarrow Y$*
 715 *holds and the hypothesis $Y \dashrightarrow X$ holds. If only one hypothesis holds, we conclude that one is the*
 716 *causal direction. If they both hold, conclude there is no causal relationship. Otherwise, do not decide.*

717 **Decision Rule E.4** (Adascore Score-matching). *Let r_1, r_2 be the residuals obtained*
 718 *from regressing Y onto X , and X onto Y (respectively). Let s_1, s_2 be the val-*
 719 *ues obtained by plugging r_1, r_2 into $\mathbb{E}[(\mathbb{E}[\partial_Y \log p(X, Y) | r_1] - \partial_Y \log p(X, Y))^2]$ and*
 720 $\mathbb{E}[(\mathbb{E}[\partial_X \log p(X, Y) | r_2] - \partial_X \log p(X, Y))^2]$ *respectively. If $s_1 = 0, s_2 \neq 0$, conclude that*
 721 *$X \dashrightarrow Y$. If $s_1 \neq 0, s_2 = 0$, conclude that $Y \dashrightarrow X$. Else, do not decide.*

722 **Decision Rule E.5** (MSE-Minimization). *Let $Loss_{X \dashrightarrow Y}, Loss_{Y \dashrightarrow X}$ be the MSE obtained from*
 723 *predicting Y from X , and X from Y respectively. Then if $Loss_{X \dashrightarrow Y} < Loss_{Y \dashrightarrow X}$, conclude that*
 724 *$X \dashrightarrow Y$, else conclude that $Y \dashrightarrow X$.*

725 **Decision Rule E.6** (ELBO-Maximization). *Let $ELBO_{X \dashrightarrow Y}, ELBO_{Y \dashrightarrow X}$ be the ELBOs obtained*
 726 *from training a VAE to predict Y from X , and X from Y respectively. Then if $ELBO_{X \dashrightarrow Y} >$*
 727 *$ELBO_{Y \dashrightarrow X}$, conclude that $X \dashrightarrow Y$, else conclude that $Y \dashrightarrow X$.*

728 E.2 Decision Rule Discussion

729 We note that while the Decision Rules E.2-E.6 are generally representative of each type of ANM-
 730 based bivariate methods (Regression Residual-Independence, Score-Matching, MSE-Minimization,
 731 etc.), there exist subclasses of methods in each category. For example, while Decision Rule E.4
 732 reflects the method Adascore [40] (and NoGAM [38] to some extent), the score-matching method
 733 SCORE [55] leverages a slightly different condition on the score function to recover the causal
 734 direction. In our analysis, we choose each Decision Rule to reflect the methodology of the most
 735 general (and typically most recent) method developed in each category. For example, we choose to

736 focus on Adascore over SCORE, as Adascore handles linear, nonlinear, and non-Gaussian ANM,
737 while SCORE requires nonlinear Gaussian ANM.

738 E.3 Proof that Regression Residual-Independence Fails (Lemma 3.1)

739 **Lemma 3.1** (Regression Residual-Independence Fails). *Assuming a consistent estimator for re-*
740 *gression residuals and access to infinite data, Decision Rule E.2 fails to identify the correct causal*
741 *direction when at least one mediator is nonlinear.*

742 *Proof.* We note that Decision Rule E.2 identifies the causal direction if and only if the residual
743 obtained from regressing Y onto X is independent of X , i.e. $e_1 \perp\!\!\!\perp X$. Suppose for contradiction that
744 $e_1 \perp\!\!\!\perp X$. Then, we have that for $e_1 = Y - g(X) = h(\varepsilon_1, \dots, \varepsilon_{T+1})$, where $g(X) = E[Y|X]$, $e_1 \perp\!\!\!\perp$
745 X . Therefore, we can rewrite $Y = g(X) + h(\varepsilon_1, \dots, \varepsilon_{T+1})$. However, this leads to a contradiction:
746 if there is at least one nonlinear mediator, then by Lemma 2.3, the ANM-UM underlying X, Y does
747 not admit an additive decomposition. Instead, $Y = A_1(X) + A_2(\varepsilon_{T+1}) + A_3(X, \varepsilon_1, \dots, \varepsilon_T)$, where
748 $A_3(X, \varepsilon_1, \dots, \varepsilon_T)$ contains nonlinear interaction between X and noise terms ε . Therefore, $e_1 \not\perp\!\!\!\perp X$,
749 and therefore Decision Rule E.2 fails to identify the causal direction.

750 □

751 E.4 Discussion of Residual-Dependence Comparisons

752 We note that despite Decision Rules E.2 being framed in terms of the outcomes of independence tests,
753 most implementations of Residual-Dependence tests leverage the comparison of test statistic values
754 that measure dependence, rather than strictly comparing the outcome of independence tests. For
755 example, DirectLiNGAM [61] estimates and compares the mutual information, while RESIT [50]
756 estimates and compares the p -value of the HSIC independence test ([11]). We analyze the empirical
757 performance of such an approach in Section 5, as the baselines we use (DirectLiNGAM, RESIT)
758 leverage these dependence comparisons to boost performance.

759 E.5 Proof that Post-Nonlinear Residual Independence Fails (Lemma 3.2)

760 **Lemma 3.2** (PNL Residual-Independence Fails). *Assuming a consistent ICA residual estimator and*
761 *access to infinite data, Decision Rule E.3 fails to recover the correct causal direction when there*
762 *exists at least one non-invertible nonlinear mediator.*

763 *Proof.* As there exists at least one non-invertible nonlinear mediator (i.e., \exists a function f_t where
764 $t \geq 2$ and f_t non-invertible and nonlinear in the ANM-UM generating Y from X), we note that
765 by Lemma 2.3 we can rewrite Y as $Y = A_1(X) + A_2(\varepsilon_1, \dots, \varepsilon_{T+1}) + A(f_1(X) + \varepsilon_1, \dots, \varepsilon_{T+1})$,
766 where $A_3(\cdot)$ produces nonlinear interaction between X and ε , and $A_3(\cdot)$ is non-trivial (non-zero),
767 and non-invertible in $f_1(X) + \varepsilon_1$.

768 We note that Decision Rule E.3 identifies the causal direction if and only if the causal hypothesis
769 $X \dashrightarrow Y$ holds, and the hypothesis $Y \dashrightarrow X$ does not hold. We note that the causal hypothesis
770 $X \dashrightarrow Y$ holds if and only if \exists functions l_1, l_2 such that for $e_1 = l_2(Y) - l_1(X)$, $e_1 \perp\!\!\!\perp x$. Suppose for
771 contradiction that \exists functions l_1, l_2 such that for $e_1 = l_2(Y) - l_1(X)$, $e_1 \perp\!\!\!\perp x$. Note that e_1 must be
772 some function of noise terms ε , i.e., $e_1 = h(\varepsilon_1, \dots, \varepsilon_{T+1}) = h(\varepsilon)$. Note that l_2 must be invertible,
773 as otherwise it would contradict that Y is a proper function of X and noise terms ε , as there exists an
774 original DGP $Y = F(X, \varepsilon_1, \dots, \varepsilon_{T+1})$.

775 Suppose l_2^{-1} is linear. Then, we can write $Y = \alpha(l_1(X)) + \alpha(e_1)$. However, this contradicts the
776 non-triviality of $A_3(\cdot)$. Then l_2^{-1} must be nonlinear and invertible. However, that contradicts the fact
777 that $A_3(\cdot)$ is non-invertible in $f(X) + \varepsilon_1$. Therefore, there cannot exist functions l_1, l_2 such that for
778 $e_1 = l_2(Y) - l_1(X)$, $e_1 \perp\!\!\!\perp x$. Therefore Decision Rule E.2 fails to identify the causal direction.

779 □

780 E.6 Proof Score-Matching Fails (Lemma 3.3)

781 **Lemma 3.3** (Score-Matching Fails). *Assuming a consistent estimator of the conditional expectation*
 782 *and access to infinite data, Decision Rule E.4 fails to recover the correct causal direction when there*
 783 *exists at least one nonlinear mediator.*

784 *Proof.* We note that for Decision Rule E.4 to correctly identify the causal direction, it requires that
 785 $\mathbb{E} \left[(\mathbb{E} [\partial_Y \log p(X, Y) | r_1] - \partial_Y \log p(X, Y))^2 \right] = 0$. Notably, [40] shows (Proposition 4) that this
 786 holds if and only if for $Y = A_1(X) + A_2(U)$, we have $A_2(U) \perp\!\!\!\perp X$. However, as there is at least
 787 one nonlinear mediator, then by Lemma 2.3, the ANM-UM underlying X, Y admits the following
 788 decomposition $Y = A_1(X) + A_2(\varepsilon_{T+1}) + A_3(X, \varepsilon_1, \dots, \varepsilon_T)$, where $A_3(X, \varepsilon_1, \dots, \varepsilon_T)$ contains
 789 nonlinear interaction between X and noise terms ε , and is non-trivial. Therefore, by it follows
 790 from Proposition 4 of [40] that $\mathbb{E} \left[(\mathbb{E} [\partial_Y \log p(X, Y) | r_1] - \partial_Y \log p(X, Y))^2 \right] \neq 0$, and therefore
 791 Decision Rule E.4 fails to recover the right causal direction. In fact, [40] explicitly states that their
 792 method fails to recover causal relationships when unobserved mediators occur (see Examples 4, 5, 7
 793 in [40]).

794 □

795 E.7 Proof that MSE-Minimization Fails (Lemma 3.4)

796 **Lemma 3.4** (MSE-Minimization Fails). *Assuming a consistent estimator of the conditional ex-*
 797 *pectation and access to infinite data, Rule E.5 fails to recover the correct causal direction when*
 798 $E[\text{Var}[X|Y]] < E[\text{Var}[Y|X]]$.

799 *Proof.* We note that under infinite data the optimal estimator of the MSE

$$\text{MSE}(f) = \mathbb{E}[(Y - f(X))^2],$$

800 converges to the conditional expectation

$$f^*(X) = \mathbb{E}[Y | X].$$

801 This implies that as the sample size n goes to infinity, the MSE converges to the expected conditional
 802 variance of $Y|X$:

$$\begin{aligned} \mathbb{E}_{\mathbb{X}}[\text{MSE}(f)] &= \mathbb{E}_{\mathbb{X}}[(Y - f^*(X))^2 | X] \\ &= \mathbb{E}_{\mathbb{X}}[(Y - \mathbb{E}[Y | X])^2 | X] \\ &= \mathbb{E}_{\mathbb{X}}[\text{Var}(Y | X)] \end{aligned}$$

803 Therefore, if $E[\text{Var}[X|Y]] < E[\text{Var}[Y|X]]$, this implies that $\text{Loss}_{X \rightarrow Y} > \text{Loss}_{Y \rightarrow X}$, which
 804 implies that Decision Rule E.5 fails to recover the causal direction.

805 We note that $E[\text{Var}[X|Y]] < E[\text{Var}[Y|X]]$ can occur if the R^2 -sortability favors the anti-causal
 806 direction. We note that the coefficient of determination R^2 is a simple function of the expected
 807 conditional variance, when the variables are standardized:

$$\begin{aligned} R_{X \rightarrow Y}^2(f^*) &= 1 - \frac{\mathbb{E}[(Y - f^*)^2]}{\text{Var}(Y)} = 1 - \mathbb{E}_{\mathbb{X}}[\text{Var}(Y | X)] \\ R_{Y \rightarrow X}^2(f^*) &= 1 - \frac{\mathbb{E}[(X - f^*)^2]}{\text{Var}(X)} = 1 - \mathbb{E}_{\mathbb{Y}}[\text{Var}(X | Y)] \end{aligned}$$

808 [54] show that linear ANM may or may not always be R^2 -sortability; as linear ANM are a subset of
 809 ANM-UM, this justifies our claim that MSE-Minimization methods will fail on some ANM-UM.

810 □

811 E.8 Necessary Constraints for MSE-Minimization methods

812 We note that the conditions under which MSE-Minimization actually does actually correspond to
 813 causal direction identification, i.e., the conditions that ensure $E[\text{Var}[X|Y]] > E[\text{Var}[Y|X]]$, have

been discussed in other work. For example, [4] show that under the assumption of independence between the function relating cause and effect, the conditional noise distribution, and the distribution of the cause, as well as a close to deterministic causal relation, the errors are smaller in the causal direction. [35] build up on [4], and show that, under the assumption that the best anti-causal model requires at least as many parameters as the causal model (leveraging Kolmogorov’s structure function), the regression errors should be smaller in the causal direction. However, these assumptions are quite distinct from our ANM-UM setting, and we leave further investigation to future work.

E.9 Failure Mode of VAE

CANM uses a variational autoencoder (VAE) framework to decide causal direction by picking the direction with the lower ELBO. The training objective used consists of three parts: the log likelihood of x (which does not depend on the model parameters θ , the KL divergence of the latent code, and the reconstruction error [6, Equation 4].

Which of these terms dominates the loss is highly dependent on the training procedure for the VAE, since training VAEs is known to suffer from posterior collapse [13], which we observed during running the method. Figure 3 shows a decomposition of the training loss of the VAE for CANM, consisting of the KL-divergence term in the latent space and the reconstruction error. The KL divergence is close to zero. The high reconstruction error indicates that the model is ignoring the latent code, just using X to reconstruct Y .

Different mitigation strategies for mitigating posterior collapse have been proposed in the literature. Among them, the line following β -VAE, which introduces a factor in front of the KL term, and uses a scheduling of the β part during training [10, 14].

In our experiments, we found good results for training the VAE with a cyclical beta annealing schedule, following [10].

Figure 3 shows that depending on the training, the reconstruction error (constant β) or the KL divergence (cyclical β) can dominate the loss function.

As shown in Table 4, cyclical scheduling failed to improve performance for the invertible cases (*tanh* and *linear*) under uniform noise. The algorithm predicts the *opposite* causal direction with high probability ($\geq 90\%$). While achieving a better reconstruction after improved training, the accuracy in the *tanh* + *Gaussian* setting declined using the new training schedule. Both phenomena imply that it is exploiting a heuristic signal rather than truly recovering the correct causal orientation.

Method Noise	Linear Unif.	Neural Net Gauss.	Unif.	Quadratic Gauss.	Unif.	Tanh Gauss.	Unif.
CANM	0.10	0.93	0.87	1.00	1.00	0.50	0.10
CANM (constant β)	0.00	0.97	0.83	1.00	0.90	0.83	0.10

Table 4: Accuracy of CANM variants across different transformation–noise combinations, with *one mediator* and $n = 1000$. **Bold** indicates best.

The original CANM paper proposes selecting the number of latent variables by comparing model likelihoods across different dimensionalities. In our implementation, we instead provide CANM with the ground-truth number of mediators as a hyperparameter, which defines the size of its latent space. In contrast, BiDD does *not* require knowledge of the number of mediators.

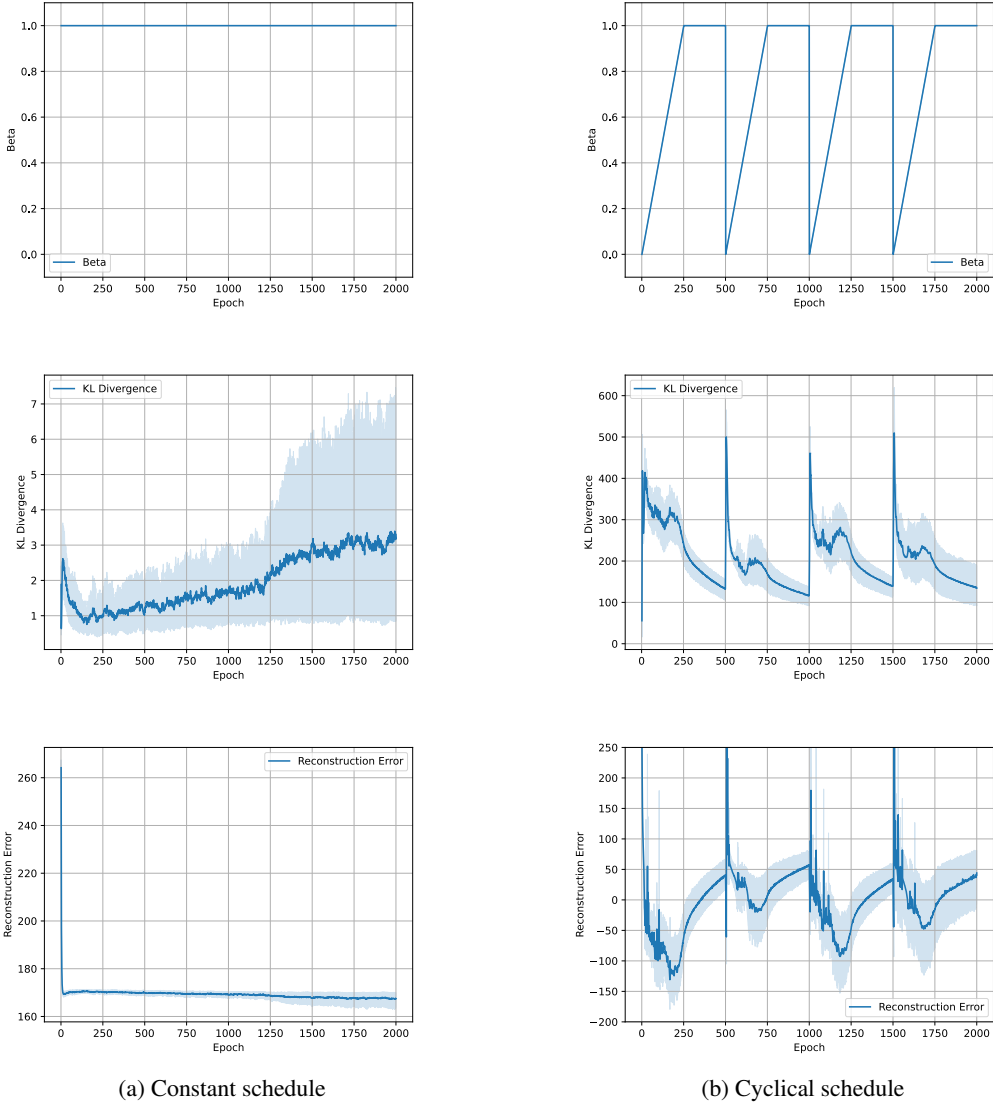


Figure 3: Comparison between different training schedules for β in adapted β VAE for the tanh uniform noise case. From top to bottom: $\beta(t)$, KL divergence, example reconstructions. Mean and 95% bootstrap confidence interval (20 runs) from estimating ε_Y from \tilde{Y} and X .

848 E.10 Proof that CANM Fails (Lemma 3.5)

849 **Lemma 3.5** (CANM Fails). *Assuming infinite data and a consistent estimator of the conditional*
 850 *expectation, Rule E.6 fails to recover the causal direction if posterior collapse occurs and the expected*
 851 *conditional log-likelihood minus the entropy is higher in the causal direction.*

852 *Proof.* We note that Decision Rule E.6 identifies the causal direction when posterior collapse occurs if
 853 and only if that expected conditional log-likelihood minus the entropy is lower in the causal direction.
 854 As we assume it is higher in the causal direction, this implies that Decision Rule E.6 must fail. \square

855 E.11 Proof of Lemma 4.1 - Anticausal Direction in Linear ANM

856 **Lemma 4.1.** $\hat{\varepsilon}_X = \mathbb{E}[\varepsilon_X | \tilde{X}, Y] \not\perp Y$.

857 *Proof.* This proof proceeds through the following steps:

- 858 1. First, we will restate the DGP of $X, Y, \varepsilon_X, \tilde{X}$, and all assumptions.
- 859 2. Then, we will show that, given (\tilde{X}, Y) , the minimal sufficient statistic (MSS) T for
860 $\mathbb{E}[\varepsilon_X | \tilde{X}, Y]$ is the identity function or data itself, i.e. $T(\tilde{X}, Y) = (\tilde{X}, Y)$.
- 861 3. We will then show that the minimal sufficient statistic being the full data vector (\tilde{X}, Y)
862 implies that $\mathbb{E}[\varepsilon_X | \tilde{X}, Y] \not\perp Y$.

863 Step 1: Restate DGP and Assumptions

864 Note, X, Y follow

$$Y = X + \varepsilon_1, \quad X \perp\!\!\!\perp \varepsilon_1. \quad (\text{E.1})$$

865 We require Assumption 2.1 and 2.2); note that 2.2) stipulates that ε_1 is non-Gaussian, for identifiability.
866 We inject independent Gaussian noise into both X , obtaining the noised term \tilde{X} :

$$\tilde{X} = X + \varepsilon_X, \quad \varepsilon_X \sim \mathcal{N}(0, 1). \quad (\text{E.2})$$

867 Additionally, we require that $X, \varepsilon_1, \varepsilon_X$ are random variables with everywhere-positive, absolutely
868 continuous and differentiable densities f_X, f_{ε_1} , and f_{ε_X} .

869 Step 2: Show that (\tilde{X}, Y) is MSS for $\mathbb{E}[\varepsilon_X | \tilde{X}, Y]$

870 This step in the proof proceeds by contradiction. We assume for contradiction that \exists MSS
871 $T[(\tilde{X}, Y)] \neq (\tilde{X}, Y)$.

872 Then, by the Lehmann-Scheffe Minimal Sufficiency Criterion [31], there exists infinite pairs of points
873 $(\tilde{x}_1, y_1) \neq (\tilde{x}_2, y_2)$ such that the following holds:

$$T(\tilde{x}_1, y_1) = T(\tilde{x}_2, y_2), \text{ and } \frac{f_{\tilde{X}, Y, \varepsilon_X}(\tilde{x}_1, y_1, e)}{f_{\tilde{X}, Y, \varepsilon_X}(\tilde{x}_2, y_2, e)} = g(\tilde{x}_1, y_1, \tilde{x}_2, y_2), \quad (\text{E.3})$$

874 where $f_{\tilde{X}, Y, \varepsilon_X}$ is the joint density of $\tilde{X}, Y, \varepsilon_X$, and (\tilde{x}_1, y_1, e) are particular values of $\tilde{X}, Y, \varepsilon_X$.

875 We use the change of variables

$$X = \tilde{X} - \varepsilon_X, \quad (\text{E.4})$$

$$\varepsilon_1 = Y - X = Y - \tilde{X} + \varepsilon_X, \quad (\text{E.5})$$

876 to re-parameterize the joint density $f_{\tilde{X}, Y, \varepsilon_X}$:

$$f_{\tilde{X}, Y, \varepsilon_X}(\tilde{x}, y, e) = f_X(\tilde{x} - e) f_{\varepsilon_1}(y - \tilde{x} + e) f_{\varepsilon_X}(e), \quad (\text{E.6})$$

877 where the equality follows from the fact that $X, \varepsilon_1, \varepsilon_X$ are mutually independent. Plugging Eq E.6
878 into the likelihood ratio in Eq E.3 we get:

$$\frac{f_X(\tilde{x}_1 - e) f_{\varepsilon_1}(y_1 - \tilde{x}_1 + e)}{f_X(\tilde{x}_2 - e) f_{\varepsilon_1}(y_2 - \tilde{x}_2 + e)} = g(\tilde{x}_1, y_1, \tilde{x}_2, y_2). \quad (\text{E.7})$$

879 Let us fix $\tilde{x}_1, y_1, \tilde{x}_2, y_2$ such that the $T(\tilde{x}_1, y_1) = T(\tilde{x}_2, y_2)$. Now, by taking the logarithm of the
880 ratio in Eq E.7, and then taking the derivative with respect to e , we get:

$$\frac{\partial f}{\partial e} \log(g(\tilde{x}_1, y_1, \tilde{x}_2, y_2)) = 0 = \frac{\partial f}{\partial e} \left[\log \left(\frac{f_X(\tilde{x}_1 - e) f_{\varepsilon_1}(y_1 - \tilde{x}_1 + e)}{f_X(\tilde{x}_2 - e) f_{\varepsilon_1}(y_2 - \tilde{x}_2 + e)} \right) \right] \quad (\text{E.8})$$

$$\begin{aligned} &= \frac{\partial f}{\partial e} \phi(\tilde{x}_1 - e) + \frac{\partial f}{\partial e} \psi(y_1 - \tilde{x}_1 + e) \\ &\quad - \frac{\partial f}{\partial e} \phi(\tilde{x}_2 - e) - \frac{\partial f}{\partial e} \psi(y_2 - \tilde{x}_2 + e) \end{aligned} \quad (\text{E.9})$$

881

$$\implies [\phi'(\tilde{x}_1 - e) + \psi'(y_1 - \tilde{x}_1 + e)] - [\phi'(\tilde{x}_2 - e) + \psi'(y_2 - \tilde{x}_2 + e)] = 0 \quad (\text{E.10})$$

882 where $\phi'(\cdot) = \frac{\partial f}{\partial e} \log(f_X(\cdot))$ and $\psi'(\cdot) = \frac{\partial f}{\partial e} \log(f_{\varepsilon_1}(\cdot))$. Now, we do a series of change of variables:
883 first, let

$$c = \tilde{x}_1 - \tilde{x}_2 \quad (\text{E.11})$$

$$d = (y_1 - \tilde{x}_1) - (y_2 - \tilde{x}_2) \quad (\text{E.12})$$

$$u = \tilde{x}_2 - e \quad (\text{E.13})$$

$$t = y_2 - u, \quad (\text{E.14})$$

$$\implies 0 = [\phi'(u + c) - \phi'(u)] + [\psi'(y_2 - u + d) - \psi'(y_2 - u)] \quad (\text{E.15})$$

884

885 Now, let $t = y_2 - u$. Then,

$$0 = [\phi'(u + c) - \phi'(u)] + [\psi'(t + d) - \psi'(t)] \quad (\text{E.16})$$

$$0 = \Delta\phi'(u) + \Delta\psi'(t), \quad (\text{E.17})$$

886 where the Δ operator denotes the difference a function at its original input and shifted input ($\phi'(u)$
887 shifted by constant c , $\psi'(t)$ shifted by constant d). Note Eq E.17 satisfies the form of Pexider's
888 equation:

$$f(x + y) = h(x) + g(y) \quad (\text{E.18})$$

889 where $f(u + t) = 0$, $h(u) = \Delta\phi'(u)$, $g(t) = \Delta\psi'(t)$. It is known [1] that continuous solutions to
890 Pexider's equation satisfy

$$f(o) = c * o + a + b \quad (\text{E.19})$$

$$h(o) = c * o + b \quad (\text{E.20})$$

$$g(o) = c * o + a. \quad (\text{E.21})$$

891 As $f(u + t) = 0$ this implies that

$$\Delta\phi'(u) = K \quad (\text{E.22})$$

$$\Delta\psi'(t) = -K. \quad (\text{E.23})$$

892 where K is a function of the fixed inputs $(\tilde{x}_1, y_1, \tilde{x}_2, y_2)$.893 Now, we focus on $\psi'(\cdot)$. Note that

$$\Delta\psi'(t) = \psi'(t + d) - \psi'(t) = -K. \quad (\text{E.24})$$

$$\implies \psi'(t + d) = \psi'(t) - K \quad (\text{E.25})$$

$$= \psi'(t) + g(d) \quad (\text{E.26})$$

894 where g is chosen such that $g(d) = K$. Eq E.26 again follows Pexider's equation. This implies that
895 $\psi'(t)$ is a linear function of t , i.e.,

$$\psi'(t) = At + B \quad (\text{E.27})$$

$$\implies \psi'(e) = Ae + \tilde{B} \quad (\text{E.28})$$

896 as $t = y_2 - \tilde{x}_2 + e$, y_2, \tilde{x}_2 are constants, and A, B, \tilde{B} are functions of constants. Now, we undo the
897 derivative and log transform taken in Eq E.8:

$$\psi'(e) = \frac{\partial f}{\partial e} \log(f_{\varepsilon_1}(e)) = Ae + \tilde{B} \quad (\text{E.29})$$

$$\implies f_{\varepsilon_1}(e) = D * \exp\left(\frac{A}{2}e^2 + \tilde{B}e + C\right), \quad (\text{E.30})$$

898 where A, \tilde{B}, C, D are all constants. We note that under the regularity conditions assumed in Step
899 1, it is known that any density proportional to the exponential of a quadratic polynomial must be
900 Gaussian [8]. Therefore, it follows that f_{ε_1} is a Gaussian distribution. However, this contradicts our
901 assumption in Step 1 that ε_1 is not a Gaussian. Therefore, the MSS T for $\mathbb{E}[\varepsilon_Y | \tilde{X}, Y]$ must equal the
902 full data vector, i.e., $T(\tilde{X}, Y) = (\tilde{X}, Y)$.

903 **Step 3: Show that if MSS $T = (\tilde{X}, Y)$, then $\mathbb{E}[\varepsilon_X | \tilde{X}, Y] \not\perp Y$**

904 Let $U = \mathbb{E}[\varepsilon_Y | \tilde{X}, Y]$, and define the sigma-algebras $\mathcal{G} = \sigma(\tilde{X}, Y)$, $\mathcal{H} = \sigma(\tilde{X})$, $\mathcal{K} = \sigma(Y)$. Note
 905 that as (\tilde{X}, Y) are minimal sufficient statistics for U , \mathcal{G} is the smallest possible sigma field under
 906 which U is measurable [12]. Suppose for contradiction that $U \perp Y$. Then, U is measurable under a
 907 sigma-field \mathcal{C} that is a strict subset of \mathcal{G} , i.e., $\mathcal{C} = \sigma(T(\tilde{X}, Y)) \subsetneq \mathcal{G}$. However, this contradicts the
 908 minimality of \mathcal{G} . Therefore, it follows that $\mathbb{E}[\varepsilon_X | \tilde{X}, Y] \not\perp Y$.

909 □

910 E.12 Proof of Diffusion Correctness for ANM

911 **Theorem 4.2** (Consistency of Decision Rule 1). *Suppose X, Y follow Eq. (2.1), Assumptions 2.1 and*
 912 *2.2 hold, and no nonlinear mediator exists. Then, given a consistent mutual information estimator*
 913 *and infinite data, Decision Rule 1 correctly recovers the causal direction between X, Y .*

914 *Proof.* We note that if no nonlinear mediator exists, then by Lemma 2.3 X, Y can be represented by a
 915 standard ANM, where $Y = f(X) + \varepsilon_1$, $X \perp \varepsilon_1$ and Assumptions 2.1, 2.2 still hold. We additionally
 916 require that X, ε_1 are random variables with everywhere-positive, absolutely continuous densities
 917 f_X, f_{ε_1} , and f_{ε_X} . We further assume that f is continuous and three-times differentiable. Note again,
 918 that we inject independent Gaussian noise into both X and Y , obtaining the noised terms

$$\tilde{X} = X + \varepsilon_X \quad \text{and} \quad \tilde{Y} = Y + \varepsilon_Y, \quad \varepsilon_X, \varepsilon_Y \sim \mathcal{N}(0, 1). \quad (\text{E.31})$$

919

920 We note that, under infinite data, Decision Rule 1 correctly recovers the causal direction if and only
 921 if both of the following statements hold (written equivalently in terms of mutual information and
 922 independence):

$$MI(\hat{\varepsilon}_Y, X) = 0 \iff \mathbb{E}[\varepsilon_Y | \tilde{Y}, X] \perp X \quad (\text{E.32})$$

$$MI(\hat{\varepsilon}_X, Y) > 0 \iff \mathbb{E}[\varepsilon_X | \tilde{X}, Y] \not\perp Y \quad (\text{E.33})$$

923 We will first show the causal direction (Eq E.32 holds), then the anticausal direction (Eq E.33 holds).

924 E.12.1 Causal Direction

925 We will find a minimal sufficient statistic (MSS) for $\mathbb{E}[\varepsilon_Y | \tilde{Y}, X]$. It will have the property
 926 $T[(\tilde{Y}, X)] \perp X$. Then, as $\varepsilon_Y \perp X$, we will conclude that Eq E.32 holds.

927 Note, the joint density $g_{\tilde{Y}, X, \varepsilon_Y}$ can be written as

$$g_{\tilde{Y}, X, \varepsilon_Y}(\tilde{y}, x, e) = f_X(x) f_{\varepsilon_1}(\tilde{y} - f(x) - e) f_{\varepsilon_Y}(e) \quad (\text{E.34})$$

928 using the change of variables

$$\tilde{Y} = f(X) + \varepsilon_1 + \varepsilon_Y, \quad (\text{E.35})$$

929 as $X, \varepsilon_1, \varepsilon_Y$ are mutually independent. Now, for any pairs of distinct points $(x_1, \tilde{y}_1) \neq (x_2, \tilde{y}_2)$
 930 consider the likelihood ratio:

$$\frac{g_{\tilde{Y}, X, \varepsilon_Y}(\tilde{y}_1, x_1, e)}{g_{\tilde{Y}, X, \varepsilon_Y}(\tilde{y}_2, x_2, e)} = \frac{f_X(x_1) f_{\varepsilon_1}(\tilde{y}_1 - f(x_1) - e) f_{\varepsilon_Y}(e)}{f_X(x_2) f_{\varepsilon_1}(\tilde{y}_2 - f(x_2) - e) f_{\varepsilon_Y}(e)} \quad (\text{E.36})$$

$$= \frac{f_X(x_1) f_{\varepsilon_1}(\tilde{y}_1 - f(x_1) - e)}{f_X(x_2) f_{\varepsilon_1}(\tilde{y}_2 - f(x_2) - e)} \quad (\text{E.37})$$

931 Let $T[(\tilde{Y}, X)] = \tilde{Y} - f(X)$. Then, we restrict to distinct pairs of points $(x_1, \tilde{y}_1) \neq (x_2, \tilde{y}_2)$ such
 932 that $T(x_1, \tilde{y}_1) = T(x_2, \tilde{y}_2) = T$. Then, Eq E.37 reduces to

$$= \frac{f_X(x_1) f_{\varepsilon_1}(T(x_1, \tilde{y}_1) - e)}{f_X(x_2) f_{\varepsilon_1}(T(x_2, \tilde{y}_2) - e)} \quad (\text{E.38})$$

$$= \frac{f_X(x_1)}{f_X(x_2)} \cdot \frac{f_{\varepsilon_1}(-e)}{f_{\varepsilon_1}(T - e)} \quad (\text{E.39})$$

933 Given this decomposition, where e depends on x, y only through the statistic T , it follows from the
 934 the Lehmann-Scheffe Minimal Sufficiency criterion [31] that T is MSS for $\mathbb{E}[\varepsilon_Y|\tilde{Y}, X]$. Then, we
 935 have

$$\mathbb{E}[\varepsilon_Y|\tilde{Y}, X] = \mathbb{E}[\varepsilon_Y|T(\tilde{Y}, X)] \quad (\text{E.40})$$

$$= \mathbb{E}[\varepsilon_Y|\varepsilon_1 + \varepsilon_Y]. \quad (\text{E.41})$$

936 As $X, \varepsilon_1, \varepsilon_Y$ are all jointly independent, it follows that $X \perp\!\!\!\perp \mathbb{E}[\varepsilon_Y|\tilde{Y}, X]$.

937 E.12.2 Anticausal Direction

938 If f is linear, then by Lemma 4.1 we have $\mathbb{E}[\varepsilon_X|\tilde{X}, Y] \not\perp\!\!\!\perp Y$.

939 Suppose f is nonlinear.

940 Let $h(\tilde{X}, Y) := \mathbb{E}[\varepsilon_X|\tilde{X}, Y]$. This proof will proceed in the following steps.

- 941 1. We will characterize $h(\tilde{X}, Y)$, showing that it must be a non-trivial function of Y and \tilde{X} .
- 942 2. We will outline all possible cases (and subcases), in which $h(\tilde{X}, Y)$ is a non-trivial function
 943 of Y .
- 944 3. For each case (and associated subcases) we will show that $h(\tilde{X}, Y)$ is a function of a noise
 945 term dependent on Y .
- 946 4. We conclude that, as $h(\tilde{X}, Y)$ is always a function of noise dependent on Y , $h(\tilde{X}, Y) \not\perp\!\!\!\perp Y$.

947 Note that $h(\tilde{X}, Y)$ can be decomposed as $h(\tilde{X}, Y) = A_1(\tilde{X}) + A_2(Y) + A_3(\tilde{X}, Y)$, where A_3
 948 contains only (linear or nonlinear) interaction between \tilde{X} and Y , while A_1, A_2 are univariate.

949 In the graphical model (see Figure 4), there is an active path $Y \rightarrow \varepsilon_X$ when conditioning on the
 950 collider \tilde{X} . Due to d-separation rules [64], it follows that $\varepsilon_X \not\perp\!\!\!\perp Y|\tilde{X}$. Note that this implies that,
 951 under regularity conditions assumed above, the conditional distribution $P(\varepsilon_X|B, C) \neq P(\varepsilon_X|C)$
 952 on a set of positive measure. This implies that $\mathbb{E}[\varepsilon_X|\tilde{X}, Y] \neq \mathbb{E}[\varepsilon_X|\tilde{X}]$, and therefore $h(\tilde{X}, Y)$ is a
 953 non-trivial function of Y . Therefore, at least one of A_2, A_3 is non-trivial. Similarly, as $\varepsilon_X \not\perp\!\!\!\perp \tilde{X}|Y$,
 954 at least one of A_1, A_3 must be non-trivial.

955 We will now walk through the following 3 cases: 1) that A_3 is non-trivial, 2) that A_3 is trivial and A_2
 956 is non-trivial and f is invertible, and 3) that A_3 is trivial and A_2 is non-trivial and f is non-invertible.
 957 In each case we will show that $h(\tilde{X}, Y)$ is a function of a noise term dependent on Y . Case 2 will
 958 have 4 subcases.

959 Suppose A_3 is non-trivial. This implies that there exist interaction terms between \tilde{X}, Y . Suppose
 960 for contradiction that $h(\tilde{X}, Y)$ is not dependent on a noise term dependent on Y . Then, it implies
 961 that A_1, A_2 somehow cancel out all Y terms in A_3 . This leads to a contradiction, since the space
 962 of additive functions—i.e., those expressible as a linear combination of univariate functions of
 963 each variable—cannot represent interaction terms, which require non-additive combinations such
 964 as products of variables. Therefore, $h(\tilde{X}, Y)$ is a non-trivial function of Y , making it a non-trivial
 965 function of noise term $\varepsilon_1, \varepsilon_Y \not\perp\!\!\!\perp Y$.

966 Suppose A_3 is trivial. Then, A_1 and A_2 must be non-trivial.

967 Suppose f is invertible. There are 4 possible subcases. In each case, X is modelled by a different
 968 function of Y , and different noise. We list each case explicitly:

- 969 1. $X = f_1(Y) + e_1, Y \perp\!\!\!\perp e_1$
- 970 2. $X = f_2(Y, e_2), Y \perp\!\!\!\perp e_2$
- 971 3. $X = f_3(Y, e_3), Y \not\perp\!\!\!\perp e_3$
- 972 4. $X = f_4(Y) + e_4, Y \not\perp\!\!\!\perp e_4$

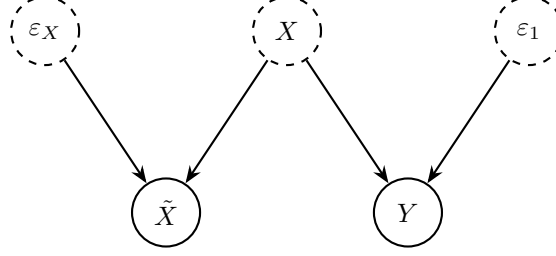


Figure 4: Dependence in the Anticausal Direction: $\tilde{X} = X + \varepsilon_X$ and $Y = f(X) + \varepsilon_1$. All root nodes are drawn from independent noise.

We note that, in Case 2 and 3 functions f_2 and f_3 induce nonlinear interactions between their inputs Y, e_2 and e_3 .

Note that Case 1 cannot occur, as it violates Assumption 2.2 by allowing for the existence of a backwards model $X \rightarrow Y$ with additive noise.

Let's assume Case 2. Then,

$$h(\tilde{X}, Y) = A_1(f_2(Y, e_2) + \varepsilon_X) + A_2(Y).$$

As f_2 induces nonlinear interaction between Y and e_2 , where $e_2 \perp\!\!\!\perp Y, e_2 \perp\!\!\!\perp \varepsilon_X$, the collection of terms in A_1 containing Y, e_2 cannot be equal to a univariate function of Y . Therefore, the residual $r = A_1(f_2(Y, e_2) + \varepsilon_X) - A_2(Y)$ must be dependent on both Y and e_2 .

Let's assume Case 3. Then

$$h(\tilde{X}, Y) = A_1(f_3(Y, e_3) + \varepsilon_X) + A_2(Y).$$

Note that as f_3 induces nonlinear interaction between Y and e_3 , e_3 needs to be a function of ε_1 - if it would be only a function of Y , that would imply a deterministic relationship between X and Y , which contradicts our setup. Due to the nonlinear dependence between Y and e_3 in f_3 , the e_3 term can not be canceled out by the univariate function $A_2(Y)$. Therefore, $h(\tilde{X}, Y)$ must contain the noise term $e_3, e_3 \not\perp\!\!\!\perp Y$.

Let's assume Case 4. Then

$$h(\tilde{X}, Y) = A_1(f_4(Y) + e_4 + \varepsilon_X) + A_2(Y) \tag{E.42}$$

Similar to the argument in Case 3, although $e_4 \not\perp\!\!\!\perp Y$, e_4 cannot solely be a function of Y - this would again imply a deterministic relationship between X, Y . Therefore, e_4 must also be a function of ε_1 . Then, the residual noise $r = A_1(f_4(Y) + e_4 + \varepsilon_X) + A_2(Y) \neq 0$, and $r \not\perp\!\!\!\perp Y$.

Therefore, $h(\tilde{X}, Y)$ is always a function of a noise term dependent on Y .

Suppose f is non-invertible. Then there exists no backwards model where X can be written as a function of Y and noise. Therefore, A_1 cannot be written as a function of Y , noise and ε_X . Then, the residual $r = A_1 + A_2$ must contain Y , which means that $h(\tilde{X}, Y)$ contains Y (which contains ε_1).

We have shown, either that h always contains ε_1 , or that h contains a different term dependent on Y . Taken together our results imply that $Y \not\perp\!\!\!\perp h(\tilde{X}, Y) \implies \mathbb{E}[\varepsilon_X | \tilde{X}, Y] \not\perp\!\!\!\perp Y$.

□

Method	Linear	Neural Net		Quadratic		Tanh		Average
Noise	Unif.	Gauss.	Unif.	Gauss.	Unif.	Gauss.	Unif.	
Full data, voting								
HSIC(1)	0.80	0.95	0.90	1.00	1.00	<u>0.85</u>	0.80	<u>0.85</u>
HSIC(.5)	<u>0.85</u>	0.95	0.85	1.00	0.80	0.90	0.80	0.83
HSIC(2)	0.80	<u>0.90</u>	0.90	1.00	1.00	<u>0.85</u>	0.80	0.86
NPEET(3)	<u>0.85</u>	0.95	0.85	1.00	0.70	0.80	0.95	0.81
NPEET(5)	0.90	0.95	0.85	1.00	0.70	0.75	0.95	0.81
NPEET(10)	0.90	<u>0.90</u>	0.85	1.00	0.60	0.80	0.95	0.82
Full data, mean								
HSIC(1)	0.80	0.95	0.95	1.00	1.00	0.80	0.70	0.84
HSIC(.5)	<u>0.85</u>	<u>0.90</u>	0.90	0.80	0.75	0.90	0.70	0.79
HSIC(2)	0.70	0.95	1.00	1.00	1.00	0.70	0.65	0.83
NPEET(3)	0.90	0.95	0.85	1.00	0.65	0.70	0.95	0.81
NPEET(5)	0.90	0.95	0.85	1.00	0.55	0.75	0.95	0.81
NPEET(10)	<u>0.85</u>	0.95	0.80	1.00	0.45	0.75	0.95	0.79
Test data, voting								
HSIC(1)	0.80	0.85	0.80	1.00	<u>0.95</u>	0.75	0.70	0.79
HSIC(.5)	0.75	0.90	0.85	1.00	0.75	0.75	0.75	0.76
HSIC(2)	0.65	0.90	0.90	1.00	1.00	0.80	0.65	0.80
NPEET(3)	0.75	0.85	0.80	1.00	0.70	0.70	<u>0.90</u>	0.79
NPEET(5)	0.70	0.85	0.85	1.00	0.80	0.60	<u>0.90</u>	0.78
NPEET(10)	0.75	<u>0.90</u>	0.85	1.00	0.85	0.70	<u>0.90</u>	0.81
Test data, mean								
HSIC(1)	0.75	0.85	0.90	1.00	<u>0.95</u>	0.65	0.65	0.79
HSIC(.5)	0.75	0.95	0.90	<u>0.95</u>	0.60	<u>0.85</u>	0.70	0.77
HSIC(2)	0.75	0.85	1.00	1.00	1.00	0.60	0.65	0.81
NPEET(3)	0.75	0.80	0.85	1.00	0.45	0.75	0.75	0.74
NPEET(5)	0.80	0.80	0.80	1.00	0.40	0.70	0.75	0.73
NPEET(10)	0.80	0.95	0.90	1.00	0.55	0.70	0.95	0.80

Table 5: Accuracy of various MI estimators across transform–noise combinations. Best scores per column are in **bold**, second–best are underlined. Accuracy is averaged over 20 runs per mechanism. HSIC(s): Factor s applied to kernel width. NPEET(k): Number of neighbors k

F Ablations

F.1 Mutual information estimate

F.1.1 Setup

To test the sensitivity of BiDD to the choice of mutual information estimator, we test its performance under two different mutual information estimators. We test three different sets of hyperparameters for each of them.

For HSIC, in our main experiments, we heuristically pick the width of the Gaussian kernel as the median in the distance metric. As an ablation, we scale this width by .5 and 2.

As an alternative estimator, we use the NPEET package, which implements the non-parametric estimator by Kraskov et al. [28], using k-nearest neighbors approach. We vary the number of neighbors $k = 3, 5, 10$.

We use the same experimental setup as in our previous synthetic experiments, setting $n = 1000$, and using a test split of 20%.

For every run we report two accuracy criteria: (i) *voting*, which counts a direction as correct if the majority of timesteps agree; (ii) a *mean*, which averages the MI over all timesteps and picks the lower-dependence direction.

Method	Linear	Neural Net		Quadratic		Tanh	
Noise	Unif.	Gauss.	Unif.	Gauss.	Unif.	Gauss.	Unif.
With conditioning:							
BiDD _{Total}	0.83	<u>0.87</u>	<u>0.97</u>	1.00	1.00	0.80	0.83
BiDD _{Test}	<u>0.80</u>	<u>0.87</u>	1.00	1.00	1.00	<u>0.63</u>	<u>0.77</u>
No conditioning:							
BiDD _{Total}	0.15	0.90	0.75	<u>0.30</u>	0.35	0.60	0.05
BiDD _{Test}	0.15	0.85	0.80	0.05	<u>0.55</u>	0.40	0.10

Table 6: Conditioning is necessary for identification across different mechanisms: Results for same setup as Table 2, but with modified training objective without conditioning. We report the mean accuracy over 20 runs.

F.1.2 Results

Complete results appear in Table 5.

Overall robustness. Every configuration achieves at least 76% mean accuracy, rising to 79% or better when the full data set is used for scoring.

Estimator–mechanism interactions. HSIC dominates on the *quadratic + uniform-noise* mechanism, whereas NPEET is best on *tanh + uniform*. These preferences are consistent across both voting and mean rules.

Voting vs. mean. Discrepancies between the two decision rules widen on the test split. For example, NPEET with $k=3$ misses the *quadratic + uniform* case under the mean rule but is working good under voting. Conversely, HSIC occasionally loses 5% – 10% on *tanh + normal* when switching from voting to mean.

Hyperparameters for estimators. Within each estimator family, hyper-parameter choices have second-order impact: HSIC’s wider bandwidth ($\times 2$) and NPEET’s larger neighbourhoods ($k=5$ or 10) bring modest, but consistent, improvements.

F.2 Conditioning vs Non-Conditioning

F.2.1 Setup

To test how conditioning impacts the performance of our methods, we provide an ablation study where we train on an unconditional loss. That means that the diffusion model does not have access to the conditioning variable B anymore, and needs to predict the noise ε only from \tilde{A} and t . We replace the original loss function:

$$L_{\text{CDM}} = \mathbb{E}_{A,B,\varepsilon,t} \left[\|\varepsilon - \varepsilon_{\theta}(\tilde{A}_t, B, t)\|^2 \right], \quad t \sim \text{Unif}(\{1, \dots, T\})$$

with the unconditional one:

$$L_{\text{DM}} = \mathbb{E}_{A,\varepsilon,t} \left[\|\varepsilon - \varepsilon_{\theta}(\tilde{A}_t, t)\|^2 \right], \quad t \sim \text{Unif}(\{1, \dots, T\}),$$

and keep the setup and training identical.

F.2.2 Results

Conditioning is important Table 6 shows the importance of conditioning. While accuracy in the *neural-network* setting remains similar, it deteriorates across all other mechanisms, and the model often selects wrong directions for both *linear + uniform* and *tanh + uniform*. These results indicate that conditioning the diffusion model is an important part for reliable causal discovery in our framework.

1042 G Experimental details

1043 G.1 Evaluation Data

1044 G.1.1 Synthetic Data

1045 To generate the synthetic data, we used different link functions in combination with different noise
1046 types.

Details for the link functions

$$\text{Quadratic: } f(x) = (x)^2 + \varepsilon,$$

$$\text{Tanh: } f(x) = \tanh(x + o) + \varepsilon, \quad o \sim \mathcal{U}(-1, 1),$$

$$\text{Linear: } f(x) = ax + o + \varepsilon, \quad a \sim \mathcal{U}(-5, 5) \\ o \sim \mathcal{U}(-3, 3),$$

$$\text{Neural network: } f(x) = \tanh(\mathbf{x} \mathbf{w}_{\text{in}}^\top + \mathbf{1} \mathbf{b}_{\text{h}}^\top) \mathbf{w}_{\text{out}} + \varepsilon,$$

1047 where the weight vectors lie in \mathbb{R}^h and are sampled i.i.d. from

$$\mathbf{w}_{\text{in}}, \mathbf{b}_{\text{h}}, \mathbf{w}_{\text{out}} \sim \mathcal{U}(-5, 5).$$

1048 The parameters $o, a, \mathbf{w}_{\text{in}}, \mathbf{b}_{\text{h}}, \mathbf{w}_{\text{out}}$ are randomly drawn for each mediator and run.

1049 **Noise types** We evaluated BiDD on two different noise types: uniform and Gaussian. For noise
1050 inside the mediators (ε_1 till ε_T in Figure 1), we used noise with mean 0 and variance .5. For noise
1051 generating X and Y (ε_0 and ε_{T+1} in Figure 1), we used noise with mean 0 and variance 1.

1052 **Data generating process** We used the same noise type for generating the cause and noise in
1053 the process for our experiments to generate X and Y . For each mediator, we redrew the random
1054 parameters for the link functions. More specifically, to synthesise a single cause–effect observation
1055 (X, Y) with T latent mediators, we begin by drawing the cause $X \sim \mathcal{D}(0, 1)$, where $\mathcal{D}(\mu, \sigma^2)$
1056 denotes the chosen base noise family (e.g. \mathcal{N} or \mathcal{U}). Setting $Z_0 = X$, we traverse a chain of T
1057 unobserved mediators. For each index $j = 1, \dots, T$ we independently (i) sample a noise term
1058 $\varepsilon_j \sim \mathcal{D}(0, 0.5)$ and (ii) select a link function g_j at random. The mediator is then evaluated as

$$Z_j = g_j(Z_{j-1}) + \varepsilon_j.$$

1059 After the final mediator we draw $\varepsilon_{T+1} \sim \mathcal{D}(0, 1)$ and an additional link function f_{T+1} , generating
1060 the effect variable by

$$Y = f_{T+1}(Z_T) + \varepsilon_{T+1}.$$

1061 Finally, both X and Y are centered and rescaled to unit variance. Repeating this procedure indepen-
1062 dently yields an i.i.d. dataset that follows an additive-noise model with T unobserved mediators.

1063 G.1.2 Real-world Data

1064 We evaluated the performance on the first 99 pairs of the Tübingen dataset [45], loaded from the
1065 causal-discovery-toolbox package. For each pair, we randomly subsampled 3000 data points
1066 in each run for faster execution. We calculated the accuracy as the simple average over all 99 pairs.

1067 G.2 Implementation Details

1068 We coded our experiments using python 3.11.11 with PyTorch 2.5.1 [48], and ran the experi-
1069 ments on AWS g4dn.xlarge ec2 instances.

1070 G.3 Details for BiDD

1071 G.3.1 Diffusion Background

1072 BiDD works by training a diffusion model in both direction and selecting the direction with the
1073 lower noise prediction. We will now describe the details of training the diffusion model, estimating
1074 the mutual information, and deciding on causal direction. We formalized all three components as
1075 subroutines in Algorithm 1.

1076 **Training Conditional Diffusion Model** TrainConditionalDiffusion in Algorithm 1 adapts
1077 the denoising-diffusion training loop of Ho et al. [17, Alg. 1] to our conditional setting. In our
1078 training, we perform a single, full-dataset update per epoch.

1079 **Mutual Information Estimation** After training the diffusion model, we use
1080 EstimateMutualInformation from Algorithm 1 to estimate the mutual information between the
1081 condition and the predicted noise.

1082 Note that in order to improve the finite sample performance of the mutual information estimators, the
1083 test set input to Algorithm 1 can be oversampled. This will lead to datapoints with identical conditions
1084 B , but differently noised version \tilde{A} . We use an oversampling factor $k = 10$ in our experiments.

1085 **Deciding causal direction** After obtaining $\{MI_{A,t}\}_{i=1}^T$ (denoising \tilde{A}) and $\{MI_{B,t}\}_{i=1}^T$ (denoising
1086 \tilde{B}) for both directions, we decide the direction of the causal dependence.

1087 **Voting rule** For the voting rule, we conclude $B \rightarrow A$ iff $MI_{A,t} < MI_{B,t}$ for the ma-
1088 jority of timesteps. We use this rule in the main body of the paper. We formalize it in
1089 CompareMutualInformation in Algorithm 1.

1090 **Mean rule** For the mean rule, we conclude $B \rightarrow A$ iff $\frac{1}{T} \sum_i MI_{A,i} < \frac{1}{T} \sum_i MI_{B,i}$. We present
1091 results on this decision rule in Appendix F.

1092 G.3.2 Implementation details

1093 All models were implemented in PyTorch. Training employed the AdamW optimizer. We used a
1094 Hilbert–Schmidt Independence Criterion (HSIC) implementation in PyTorch. For the alternative
1095 mutual-information estimator evaluated in Appendix F, we used the NPEET package.

1096 **Mutual-information estimation** Dependence between the predicted noise and the conditioning
1097 variable is quantified with HSIC [11], which we treat as a surrogate for mutual information. HSIC
1098 is computed with a Gaussian kernel whose bandwidth is selected via the median pairwise-distance
1099 heuristic. The robustness of our results to the choice of dependence measure is examined in Ap-
1100 pendix F.

1101 **Hyperparameters of training** In our diffusion setup, we use $T = 256$ timesteps, scheduling β
1102 linearly from $\beta_{\min} = 0.0001$ to $\beta_{\max} = 0.02$. For stochastic gradient descent, we use the AdamW
1103 optimizer with cosine annealing. We set an initial learning rate of 0.0001 and decay to 0.00001. We
1104 train our model for a total of 4000 epochs.

1105 **Model Architecture** The diffusion model for BiDD uses an MLP architecture to predict $\varepsilon(\tilde{A}_t, B, t)$.
1106 Each of the three inputs is first fed through a dedicated input projection: a $1 \rightarrow 512$ linear layer for
1107 \tilde{A}_t , a small conditioning network for B , and a sinusoidal time embedding followed by two SiLU-
1108 activated linear layers for t . The resulting $(512 + 4 + 512) = 1028$ -d feature vector is concatenated
1109 and processed by two residual MLP blocks, each expanding to two times width and returning to
1110 the original size. A final projection ($1028 \rightarrow 512 \rightarrow 1$) produces the noise estimate $\varepsilon_\theta(\tilde{A}_t, B, t)$.
1111 We document the exact layer setup in Table 7. Overall, the model contains **9,260,893** learnable
1112 parameters.

Algorithm 1 BiDD: Causal-direction discovery via conditional diffusion

1: **function** TRAINCONDITIONALDIFFUSION
Require: Training set $\{(A^{(i)}, B^{(i)})\}_{i=1}^n$, epochs E , total timesteps T , noise schedule $\{\bar{\alpha}_t\}_{t=1}^T$
Ensure: Trained model ε_θ
2: **for** $e = 1, \dots, E$ **do**
3: Independently sample timesteps $t^{(i)} \sim \mathcal{U}\{1, \dots, T\}$ for all $i = 1, \dots, n$
4: Draw noises $\varepsilon^{(i)} \sim \mathcal{N}(0, 1)$ for all i
5: Construct noised inputs
$$\tilde{A}_t^{(i)} = \sqrt{\bar{\alpha}_{t^{(i)}}} A^{(i)} + \sqrt{1 - \bar{\alpha}_{t^{(i)}}} \varepsilon^{(i)} \quad \text{for } i = 1, \dots, n$$

6: Compute loss
$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^n \|\varepsilon^{(i)} - \varepsilon_\theta(\tilde{A}_t^{(i)}, B^{(i)}, t^{(i)})\|^2$$

7: Take gradient step on θ
8: **end for**
9: **return** ε_θ
10: **end function**

11: **function** ESTIMATEMUTUALINFORMATION
Require: Test set $\{A^{(i)}, B^{(i)}\}_{i=1}^n$, epochs E , timesteps T , mutual information estimator MI , noise schedule $\{\bar{\alpha}_t\}_{t=1}^T$, trained diffusion model ε_θ
Ensure: Series of MI estimations $\{MI_t\}_{t=1}^T$
12: **for** $t = 1, \dots, T$ **do**
13: Draw random noise $\varepsilon^{(i)} \sim \mathcal{N}(0, 1)$ for all i .
14: Generate $\tilde{A}_t^{(i)}$ from test set as $\tilde{A}_t^{(i)} = \sqrt{\bar{\alpha}_t} A^{(i)} + \sqrt{1 - \bar{\alpha}_t} \varepsilon^{(i)}$
15: Predict $\hat{\varepsilon}_A^{(i)} = \varepsilon_\theta(\tilde{A}_t^{(i)}, B^{(i)}, t)$
16: Calculate $MI_t = MI(\{\hat{\varepsilon}_A^{(i)}\}_{i=1}^n, \{B^{(i)}\}_{i=1}^n)$
17: **end for**
18: **return** $\{MI_t\}_{t=1}^T$
19: **end function**

20: **function** COMPAREMUTUALINFORMATION
Require: Two MI sequences $\{MI_{A,t}\}_{t=1}^T$ and $\{MI_{B,t}\}_{t=1}^T$
Ensure: Chosen causal direction ($A \rightarrow B$ or $B \rightarrow A$)
21: $v \leftarrow \sum_{i=1}^T \mathbf{1}\{MI_{A,i} < MI_{B,i}\}$
22: **if** $v > T/2$ **then**
23: **return** $B \rightarrow A$
24: **else**
25: **return** $A \rightarrow B$
26: **end if**
27: **end function**

28: **function** DECIDEDIRECTIONUSINGBiDD($\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}, E, T, MI, \{\bar{\alpha}_t\}$)
29: $\varepsilon_{A|B} \leftarrow \text{TRAINCONDITIONALDIFFUSION}(\mathcal{D}_{\text{train}}, E, T, \{\bar{\alpha}_t\})$
30: $\{MI_{A,t}\} \leftarrow \text{ESTIMATEMUTUALINFORMATION}(\mathcal{D}_{\text{test}}, E, T, MI, \{\bar{\alpha}_t\}, \varepsilon_{A|B})$
31: $\varepsilon_{B|A} \leftarrow \text{TRAINCONDITIONALDIFFUSION}(\text{swap}(\mathcal{D}_{\text{train}}), E, T, \{\bar{\alpha}_t\})$
32: $\{MI_{B,t}\} \leftarrow \text{ESTIMATEMUTUALINFORMATION}(\text{swap}(\mathcal{D}_{\text{test}}), E, T, MI, \{\bar{\alpha}_t\}, \varepsilon_{B|A})$
33: **return** $\text{COMPAREMUTUALINFORMATION}(\{MI_{A,t}\}, \{MI_{B,t}\})$
34: **end function**

1113 G.4 Baseline Implementation

1114 We imported CANM from the code provided in Cai et al. [6]. We did not use the described method
1115 to find the correct number of mediators, but instead called the method with the correct number of
1116 mediators. We trained the VAE for 2000 epochs.

Stage	Operation / Activation	Output shape
<i>Input projections</i>		
\tilde{A}_t proj.	Linear ($1 \rightarrow 512$)	$(B, 512)$
B proj.	Linear ($1 \rightarrow 16$) + ReLU	$(B, 16)$
	Linear ($16 \rightarrow 32$) + ReLU	$(B, 32)$
	Linear ($32 \rightarrow 4$)	$(B, 4)$
t embedding	Sinusoidal ($1 \rightarrow 16$)	$(B, 16)$
	Linear ($16 \rightarrow 512$) + SiLU	$(B, 512)$
	Linear ($512 \rightarrow 512$) + SiLU	$(B, 512)$
CONCAT	—	$(B, 1028)$
<i>Residual MLP blocks (repeat twice)</i>		
Hidden	Linear ($1028 \rightarrow 2056$) + SiLU	$(B, 2056)$
	Linear ($2056 \rightarrow 1028$)	$(B, 1028)$
Residual add	$x \leftarrow x + \text{block}(x)$	$(B, 1028)$
<i>Output projection</i>		
Output proj.	Linear ($1028 \rightarrow 512$) + SiLU	$(B, 512)$
	Linear ($512 \rightarrow 1$)	$(B, 1)$

Table 7: Layer specification for the BiDD denoising model predicting $\varepsilon(\tilde{A}_t, B, t)$. B is batch size. SiLU is the Sigmoid-weighted Linear Unit, ReLU is the Rectified Linear Unit.

1117 DirectLiNGAM and RESIT were imported from the `lingam` package. CAM, SCORE and NoGAM
1118 were imported from the `dodiscover` package. CAM-UV was imported from the `lingam` package.
1119 Adascore was imported using the `causal-score-matching` package. Var-Sort was implemented
1120 using NumPy. DagmaL was imported from the `dagma` package. PNL was implemented following the
1121 logic in the `causal-learn` library, but with slight modifications in order to execute model training
1122 on GPU for faster execution.

We list the hyperparameters we used in Table 8.

Method	Hyperparameters
DirectLiNGAM	None
CAM	prune=False
RESIT	RandomForestRegressor(max_depth=4)
SCORE	prune=False
NoGAM	n_crossval=2, prune=False
var_sort	None
entropy_knn	k=100, base=2
PNL	None
AdaScore	alpha_orientation=0.05 alpha_confounded_leaf=0.05 alpha_separations=0.05
DagmaL	lambda1=0.05, T=4, mu_init=1 s=[1, 0.9, 0.8, 0.7], mu_factor=0.1

Table 8: Hyperparameters used for each method.

1123

1124 G.5 Runtime Results

1125 We provide the runtime for mean runtime for 80 independent runs across all mechanisms (linear,
1126 tanh, neural network, quadratic) and noises (Gaussian, Uniform). The runtime experiments were
1127 conducted on AWS g4dn.xlarge ec2 instances with GPU support, with no parallelization.

1128 All methods that do not rely on stochastic-gradient training finish in under 2 seconds. Among the
1129 baselines, PNL is an order of magnitude slower because it trains a small neural network. The two
1130 mediator-aware approaches, BiDD and CANM, show comparable run times.

1131 For PNL, BiDD, and CANM the wall-clock time is governed mainly by (i) the size of the training set,
 1132 (ii) the complexity of the model, and (iii) the number of training epochs. We didn't exhaustively tune
 1133 model size, epoch count, or code efficiency, so runtimes could be reduced with further optimisation.

Method	Runtime (s)
BiDD	172.0
CANM	145.5
Adascore	1.25
NoGAM	0.23
SCORE	0.28
DagmaL	0.85
CAM	0.15
PNL	35.80
RESIT	0.43
DLiNGAM	0.01
Var-Sort	0.00

Table 9: Average runtimes of methods for $n = 1000$ in seconds.

1134 G.6 Asset information

1135 All external code we import is open-source under permissive licences: `lingam`, `dodiscover`,
 1136 `causal-learn`, and `NPEET` are MIT; `causal-score-matching` is MIT-0; `dagma` is Apache-2.0.

1137 We use the Tübingen cause-effect dataset curated by Mooij et al. [43]. Several of its variable pairs
 1138 originate from datasets released by Kelly et al. [26] in the UCI Machine Learning Repository.