Information-Theoretic Causal Discovery in Topological Order

Sascha Xu° **Sarah Mameche**° **Jilles Vreeken**
CISPA Helmholtz Center for Information Security

Abstract

Identifying causal relationships is a cornerstone task in science, but most data-driven methods offer ambiguous results or require restrictive assumptions. Recent work on the basis of information theory shows promising results across many domains, but leaves open how to provably identify causal graphs. Here, we develop a general information-theoretic framework called Topic for causal discovery in topological order. Topic is based on the universal measure of Kolmogorov complexity and is fully identifiable. We show that Topic's guarantees extend to both the i.i.d. and non-i.i.d. continuous settings. Our evaluations on continuous, time series, and interventional data show that Topic, using domain-specific approximations of Kolmogorov complexity, learns faithful topological orderings and frequently outperforms specialized methods.

1 INTRODUCTION

Answering causal questions is a central part of scientific inquiry. Understanding causal pathways in biological systems, for example, is important to reason about interventions on such systems, such as modern gene editing technologies [Dominguez et al., 2016]. However, effective discovery of causal networks from observational data only [Pearl, 2009] remains an ongoing challenge.

Methods for causal discovery use conditional independence tests [Spirtes et al., 2001] or greedy score-based search [Chickering, 2002], but these can only infer the underlying causal graph up to the Markov equivalence class. Additional assumptions on the causal model are necessary to obtain a fully oriented Directed Acyclic Graph (DAG). Prior work [Bühlmann et al., 2014, Zheng et al., 2018] usually considers single-domain settings under restrictive assumptions, such as continuous additive noise models [Peters et al., 2014].

In addition, real-world data is often not identically distributed (i.i.d.) but instead heterogeneous or non-stationary. For example, in gene knockdown or overexpression experiments in biology, targeted interventions change the expression of certain genes, resulting in non-i.i.d. datasets [Meinshausen et al., 2016]. These settings require yet again specialized methods and assumptions to discover causal networks [Huang et al., 2020, Squires et al., 2020, Perry et al., 2022].

In this work, we explore an *information-theoretic* approach to causal discovery. It takes inspiration from the principle of algorithmically independent mechanisms [Janzing and Schölkopf, 2010], which postulates that the true causal model compresses the data most effectively. Using this principle, methods for various i.i.d. [Compton et al., 2020, Xu et al., 2022, Jalaldoust et al., 2022] and non-i.i.d. settings [Mameche et al., 2023] have proved effective. Currently, however, there is limited research [Mian et al., 2021] on integrating information-theoretic scores into a general search framework for Directed Acyclic Graphs (DAGs).

To this end, we propose Topic, a unified framework for information-theoretic causal discovery. Topic is inspired by the universal concept of Kolmogorov complexity and, using domain-specific scores, applicable to both the i.i.d. and non-i.i.d. settings. We derive an approach that proceeds in a topological order of the true graph and subsequently discovers fully oriented causal graphs. We show under which conditions Topic is generally identifiable and then examine identifiability for two specific instantiations, one assuming i.i.d. causal additive models, the other based non-i.i.d. settings with mechanism shifts.

°Equal contribution.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

We evaluate TOPIC in continuous and time-series, i.i.d. and non-i.i.d. simulations and on real-world lung cancer data [Statnikov et al., 2015]. We find that TOPIC performs favorably to domain-specific methods, showcasing its flexibility and efficacy as an information-theoretic framework for causal discovery.

2 PRELIMINARIES

We first describe our setting and the information-theoretic causal discovery ideas we will use.

2.1 Problem Setup

Given a set of variables $X \in \mathcal{X}^p$ with domain $\mathcal{X} = \mathbb{R}$, we are interested in the underlying causal structure in the form of a *directed acyclic graph* (DAG) G that dictates how the data is generated in terms of causes and effects. That is, we assume that the joint distribution P_X is Markov w.r.t. a DAG G with nodes $1, \dots p$ and edges (X_i, X_j) in G whenever X_i is a direct cause of X_j. We denote the direct predecessors of a node X_j in G as pa_j^G. Throughout, we assume causal sufficiency, i.e., no unobserved common causes exist for any pair X_i and X_j. We refer to Lauritzen [1996] and Pearl [2009] for a formal introduction to Markov properties as well as d-separation in graphs.

We distinguish between two settings. In the *homogeneous* setting, we obtain an i.i.d. sample $X \in \mathbb{R}^{p \times n}$ with n samples, respectively a time series sampled at n discrete time points. In the *heterogeneous* setting, we obtain multiple i.i.d. datasets $\mathcal{C} = \{X^1, \dots, X^m\}$ which we refer to as contexts c. Each context c can differ in distribution P^c due to mechanism shifts of a set of nodes I_c^*, that is,

$$P^c_{X_1,\dots,X_p} = \prod_{i \notin I_c^*} P_{X_i|pa_i^G} \Big(\prod_{i \in I_c^*} P^c_{X_i|pa_i^G} \Big) ,$$

As is common, we hereby assume that there is a fixed underlying causal graph G^* in all contexts and interventions modify the conditionals. Our goal in both settings is to infer from data a graph G equal to G^*.

2.2 Information-Theoretic Causal Discovery

Inferring causal structures using purely statistical notions of independence is limited to Markov Equivalence classes of G [Hauser and Bühlmann, 2013]. To offer stronger identifiability guarantees, we turn to information-theoretic measures of independence, in particular, the *algorithmic mutual information* (AMI) [Chaitin, 1975]. It is defined through the Kolmogorov complexity K, which defines a complexity of an object with string description $x \in \{0,1\}^*$ as the length $K(x)$, in bits, of the shortest program that generates x on a universal Turing Machine which then halts [Li and Vitányi, 2009]. For two binary strings x and y, the algorithmic mutual information is given by

$$I_A(x; y) = K(x) + K(y) - K(x, y) .$$

When strings x, y are independent and provide no advantage in compressing them together, the AMI tends to zero, i.e., $I_A \stackrel{+}{=} 0$ holds up to a program of constant length. One can extend the definitions of K and I_A to distributions, as well as define a conditional version as $I_A\big(P_X; P_Y \mid P_Z\big) = K\big(P_X \mid P_Z\big) - K\big(P_X \mid P_Y, P_Z\big)$.

3 FRAMEWORK

We now introduce our TOPIC framework for **Top**ological Ordering Based **I**nformation-Theoretic **C**ausal Discovery. A topological order is defined as a unique node mapping function $T = \mathcal{V} \to \{1, \dots, p\}$, where for all edges (X, Y) in the true graph G^* : $T(X) < T(Y)$. Furthermore, we consider partial topological orders T^k, where only the first k nodes have been assigned a position. We denote that a node X is not yet assigned to a topological order by $T(X) = -1$.

To infer a causal graph, we use an oracle Ω that iteratively provides the next node in a valid topological order. Given a partial topological order T^{k-1} and a partially inferred graph G up to the k-th node, i.e. $\forall (X, Y) \in G^*, T^k(X) < k : (X, Y) \in G$, the oracle $\Omega(G, T^{k-1})$ returns a node Z such that $T(Z) = k$ is in accordance with the true graph G^*. We now show how to infer the true graph G^* using the oracle Ω, and then how to create the oracle itself based on information-theoretic principles.

3.1 Algorithm

TOPIC initializes the inferred graph G with no edges and an empty topological order T^0 with $k = 1$. We iterate once per node for a total of p times, where for each iteration k we perform in order

1. Call the oracle $\Omega\left(G, T^{k-1}\right)$ to obtain the next node X with $T(X) = k$.
2. Add all *outgoing* edges (X, Y) to G that compress, i.e. add edges (X, Y) for which $T(Y) > k$ and
$$K\left(P_Y \mid \mathrm{pa}_Y^G \cup X\right) \stackrel{+}{<} K\left(P_Y \mid \mathrm{pa}_Y^G\right) .$$
3. Remove all *incoming* edges (Z, X) that are redundant, i.e. remove (Z, X) from G for which
$$K\left(P_X \mid \mathrm{pa}_X^G \cup Z\right) \stackrel{+}{=} K\left(P_X \mid \mathrm{pa}_X^G\right) .$$

We will now show under which conditions TOPIC infers the true graph G^*.

Assumption 3.1. *[Faithfulness] Given p random variables with true causal graph G^*, a pair of variables X and Y is algorithmically independent conditioned on a set of variables Z, i.e. $I_A(P_X; P_Y \mid Z) \stackrel{+}{=} 0$, if and only if they are d-separated in G^*.*

Faithfulness is a standard assumption in causal discovery. Here, it implies that a true edge $(X, Y) \in G^*$ always compresses, i.e.

$$\forall G, (X, Y) \notin G : K\left(Y \mid \mathrm{pa}_Y^G \cup X\right) \stackrel{+}{<} K\left(Y \mid \mathrm{pa}_Y^G\right) \ ,$$

as there is no parent set without X that could d-separate X and Y. Therefore, by testing all *outgoing* edges, we have perfect recall of all true edges, i.e. after iteration k, $\forall (X, Y) \in G^*, T(X) \leq k : (X, Y) \in G$.

Next, we deal with the fact that whilst the recall is perfect, the precision may not be, in other words false positive edges may be added to G. To this end, we test all *incoming* edges for redundancy in Step 3. Under Algorithmic Independence of Conditionals, it is guaranteed that all false positive edges are pruned.

Assumption 3.2. *[Algorithmic Independence of Conditionals] The distribution of two random variables X and Y conditioned on their true causal parents in G^* is algorithmically independent , i.e.*

$$I_A\left(P_{X|pa_X^*}; P_{Y|pa_Y^*}\right) \stackrel{+}{=} 0 \ .$$

Assumption 3.2 states that any pair X and Y conditioned on their true parents is algorithmically independent. An equivalent reformulation thereof is that the total Kolmogorov complexity of the joint distribution $P_{X_1,\ldots,X_p}$ is equal up to an additive constant to the Kolmogorov complexity of each variable conditioned on its true parents.

Hence, we can reduce a set of nodes pa_X^G to the true parents pa_X^* by removing all redundant parents, i.e. parents that do not compress, provided that $\mathrm{pa}_X^G \supseteq \mathrm{pa}_X^*$. This is ensured through faithfulness, as we have perfect recall of all true edges. By pruning all incoming edges in Step 3 that do not compress, we are left with the true parents pa_X^*.

Theorem 3.3. *[DAG Identifiability] Under the Algorithmic Independence of Conditionals and Faithfulness, given an oracle Ω, TOPIC recovers the true causal graph G^*.*

We provide a proof of Theorem 3.3 in Appendix A.2. Initially, the conditions of the oracle Ω are met by a source of G^*, which is guaranteed to exist for a DAG. Faithfulness ensures that TOPIC identifies all true edges, which in turn allows us to use the oracle again and advance one step further in the topological order. The Algorithmic Independence of Conditionals ensures that pa_X^G is pruned to pa_X^* for node X with $T(X) = k$ so that we are left with the true graph G^*.

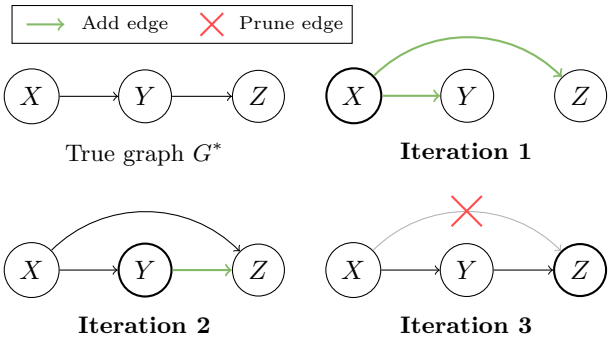


Figure 1: *Discovery of a chain G^*.* TOPIC proceeds in topological order (X, Y, Z) and adds edges that compress (green) and prunes edges that are redundant and hence no longer compress (red).

We illustrate the TOPIC algorithm in Figure 1 for a chain $X \to Y \to Z$. Using Ω, we identify the source of the graph X. From X, we add the edge (X, Y) as it compresses, as well as the false positive edge (X, Z) due to the information flow from X to Z. Now, G contains all true parents of Y, so that Ω identifies Y as the next node. For Y, we add the edge (Y, Z) as it compresses. Finally, we arrive at Z, where we prune the false positive edge (X, Z), since (Y, Z) is sufficient to explain Z per Assumption 3.2. Thus, TOPIC has correctly identified the true causal chain $X \to Y \to Z$.

3.2 Oracle

The oracle Ω is the key to TOPIC's identifiability guarantees. We now show that using Kolmogorov Complexity we can not only add and remove causal edges, but also identify a node for which all true parents have been accounted for, and which is thus eligible to be next in the topological order.

We base our oracle on the compression gain of an edge (X, Y) in graph G, defined as

$$g(X, Y; G) = K\left(P_Y \mid pa_Y^G\right) - K\left(P_Y \mid pa_Y^G \cup \{X\}\right) \ .$$

The gain quantifies the amount of information that is saved by adding the edge (X, Y) to the graph G. Intuitively, the compression gain in the causal direction outweighs the one in the anti-causal direction. Hence, we construct the oracle $\Omega(G, T^{k-1})$ as follows

$$\underset{T(X)=-1}{\arg\max} \left(\min_{T(X)=-1} g(X, Y; G) - g(Y, X; G) \right) \ . \quad (1)$$

Above we compare, for all not yet covered nodes X and Y ($T(X), T(Y) = -1$), the compression gain of the edge (X, Y) with the gain of the edge (Y, X). The delta $\Delta_{X,Y}(G) = g(X, Y; G) - g(Y, X; G) > 0$ indicates that X is a parent/ancestor of Y, i.e., X compresses Y better than vice versa, provided no unaccounted confounders are present. By taking the node X with the *maximal worst-case* Δ we isolate that node for which no parents are unaccounted for in G, so that setting $T(X) = k$ is valid with regard to G^*.

Assumption 3.4. *[Information-Theoretic Identifiability] Let X be a resolved node of G, i.e., $pa_X^G = pa_X^{G^*}$, and Y be a descendant of X in G^*. If all confounders of X and Y are accounted for in G, i.e., $pa_Y^G \supseteq (pa_X^* \cap pa_Y^*)$, then it holds that*

$$\begin{aligned} &I_A\left(P_X \mid pa_X^G;\, P_Y \mid pa_Y^G \cup \{X\}\right) \\ &\overset{+}{<} I_A\left(P_Y \mid pa_Y^G;\, P_X \mid pa_X^G \cup \{Y\}\right) \ . \end{aligned}$$

Assumption 3.4 states that information in the causal direction (X, Y) is better compressed than in the anti-causal direction (Y, X), provided that all parents of X and joint confounders are accounted for. It ensures that the oracle Ω selects a node X, for which $\mathrm{pa}_X^G = \mathrm{pa}_X^*$. Then the delta to any node Y, which has to be a descendant or independent of X, is positive, i.e.,

$$g(X, Y; G) \overset{+}{\geq} g(Y, X; G) \Leftrightarrow \Delta_{X,Y}(G) \overset{+}{\geq} 0 \ .$$

On the other hand, if there is a parent $Z \in \mathrm{pa}_X^*$ of X that is not accounted for in G with $T(Z) \geq k$, then the delta $\Delta_{X,Z}(G)$ is negative, i.e., $\Delta_{X,Z}(G) < 0$. The oracle Ω thus selects a node X for which it is guaranteed that all parents are accounted for in G, and which is thus eligible to be next in the topological order.

Theorem 3.5. *Under Assumption 3.4, the oracle Ω returns a node X for which $pa_X^G = pa_X^*$ and $T(X) = k$ is valid with regard to G^* provided that G contains all edges up to the k-th node.*

We provide a proof of Theorem 3.5 in the Appendix A.3. It outlines general conditions under which the oracle is provably correct. In Sections 4, we show under which conditions the information-theoretic identifiability as per Assumption 3.4 holds for specific causal models in the continuous i.i.d. and non-i.i.d. settings.

3.3 Score

Lastly, we discuss how to overcome the uncomputability of the Kolmogorov Complexity using lossless compression algorithms. Minimum Description Length (MDL) [Grünwald, 2007] is a statistically sound approximation from above to Kolmogorov Complexity. Two part MDL separates it into model complexity $L(M)$, which measures the cost of parameters in bits, and the cost of encoding the data with said model $L(D \mid M)$. In causality, MDL is used to great success for the continuous [Marx and Vreeken, 2017], discrete [Budhathoki and Vreeken, 2018] and time series domains [Jalaldoust et al., 2022] amongst others. The goal to minimize the description length as an upper bound to the Kolmogorov Complexity as

$$K(P) \leq L(M) + L(D \mid M) \ .$$

The MDL principle automatically balances the trade-off between model complexity and data fit. In practice, we use MDL to substitute the Kolmogorov Complexity in the score $g(X, Y; G)$. Marx and Vreeken [2022] show that in the limit of $n \to \infty$, an appropriate two-part MDL score on expectation gives the same inference results as one that has access to the Kolmogorov Complexity itself.

Significance. In addition, MDL allows for seamless integration of significance testing for causal edges. The no hyper-compression inequality [Grünwald, 2007] shows that the probability of obtaining a compression gain of t bits due to chance is less than 2^{-t}. This allows us to determine whether a gain is significant at level α by converting it into a practical threshold, $t = -\log_2(\alpha)$. Assume we have a model M and we add an edge to it to form M'. If the cost of the new model $L(M') + L(D \mid M') < L(M) + L(D \mid M) - t$, then the gain is significant and the edge will be added to the graph G. In practice, this helps us to avoid adding false positive edges due to limited samples.

4 INSTANTIATION

In the following, we justify how we instantiate the oracle Ω and the corresponding edge score g in our domains of interest. In particular, we examine specific causal models for which we can guarantee information-theoretic identifiability as in Assumption 3.4.

4.1 Homogeneous Domain

Assumption 4.1. *[Additive SCM] We consider a structural causal model where each variable Y is generated as a sum of its parents and noise as per*

$$Y = \sum_{X \in pa_Y^*} f_{X,Y}(X) + N_Y \ . \tag{2}$$

Additive SCMs are a common assumption in the literature [Peters et al., 2017] and their identifiability is well studied [Shimizu et al., 2006, Hoyer et al., 2008]. In particular, only in the causal direction one

can separate the effect Y into a function of the cause $f(X)$ and an independent noise term N_Y, whereas in the anti-causal direction no such decomposition exists (apart from the linear Gaussian case). This independence of cause and effect given the cause implies that $I_A(P_X; P_Y \mid X) \stackrel{+}{=} 0$, whilst in the anti-causal direction no such independence holds, i.e. $I_A(P_X \mid Y; P_Y) \stackrel{+}{>} 0$.

To ensure the guaranteed correctness of the oracle Ω, the independence in causal direction and dependence in anti-causal direction resp. has to hold for a resolved node X of G, i.e. $pa_X^G = pa_X^{G^*}$. For an additive SCM as per Eq. (2), we show in the Appendix A.4 that the model in the causal direction corresponds to a post-nonlinear model, which is known to be identifiable apart from certain pathological cases [Zhang and Hyvärinen, 2009, Peters et al., 2014].

Theorem 4.2. *Under the additive SCM as per Eq.* (2)*, any descendant Y of a resolved node X with $pa_X^G = pa_X^{G^*}$ can be expressed as a post-nonlinear model of X, where Assumption 3.4 holds if the post-nonlinear model is identifiable.*

Implementation. In practice, we model the local functions $f_{X,Y}$ using cubic splines, where we denote the parameters as $\theta_{X,Y}$. Cubic splines are well versed in smoothly approximating non-linear functions and used in the state-of-the-art causal discovery methods [Bühlmann et al., 2014, Mian et al., 2021]. We now describe the corresponding MDL score for use in Topic.

We start by encoding the model $L(M)$. It consists of the causal graph G and the local functions $\theta_{X,Y}$. We assign each parameter of a spline $\theta_{X,Y}$ a constant cost of r bits, i.e. $L(\theta_{X,Y}) = r \cdot ||\theta_{X,Y}||_0$, and each edge of the graph G also a cost of r bits, i.e. $L(G) = r \cdot ||G||_0$, where we choose $r = 2$ akin to the Akaike Information Criterion [Akaike, 1974]. To encode the data under the model, i.e. $L(D \mid M)$, we model the residual noise using a Gaussian $p(r) \sim \mathcal{N}(0, \sigma^2)$, as

$$r_i = y_i - \sum_{X \in \mathrm{pa}_Y^G} \hat{f}_{X,Y}(x_i)\ , L(D \mid M) = \sum_{i=1}^{n} \log(p(r_i))\ .$$

We thus approximate the Kolmogorov complexity of a variable $K(Y \mid \mathrm{pa}_Y^G)$ as

$$\approx \sum_{X \in \mathrm{pa}_Y^G} r \cdot (||\theta_{X,Y}||_0 + ||\mathrm{pa}_Y^G||_0) + \sum_{i=1}^{n} \log(p(r_i))\ .$$

We instantiate Topic in the continuous domain using this MDL score to compute $\hat{g}$, and compare it to state-of-the-art methods in Section 6. As a framework, Topic seamlessly accommodates different model classes resp. scores, and can even be employed for non-i.i.d. data, as we will show next.

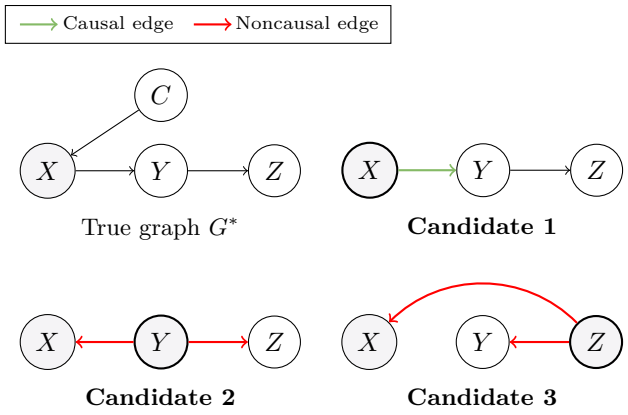


Figure 2: *Discovery of G^* in heterogeneous settings.* Given data with mechanism shifts (gray), Topic selects sources that need few, independent shifts (green) avoiding those that lead to dependent shifts (red).

4.2 Heterogeneous Domain

We now consider the setting with data from different environments. For that, let us revisit the causal chain shown in Fig. 2. We now observe X, Y and Z in multiple contexts, in one of which a causal mechanism change for X occurs (gray). For example, consider an (idealized) scenario where variables represent gene expression levels measured in different conditions, and a knock-down intervention affects gene X in one condition, represented by a categorical variable C. As no interventions affect the remaining variables, their generating process remains the same in all contexts (white) when we include causal edges (top row). In contrast, when the causal edge is not included (bottom row), Y and C are no longer d-separated in G^*. Considering either Y or Z as a source will, therefore, introduce an additional mechanism change (gray).

We first show that this observation extends to larger graphs. To this end, we require the following faithfulness assumption that true mechanism changes show in the observed distributions.

Assumption 4.3. *[Shift Faithfulness] For each i and any two environments c, c', when $i \in I^*_{c'}$, then*

$$P^c_{X_i|pa_i^G} \neq P^{c'}_{X_i|pa_i^G}\ .$$

Under this assumption, we can show that the observation in Fig. 2 generally holds for any resolved node X and variables Z downstream from X.

Lemma 4.4. *[Path Shifts] Given a resolved node X and node Z with a directed path from X to Z in G^*, when Assumption 4.3 holds, $P_{X|Z}$ and P_Z both reflect the true mechanism changes of X,*

$$P^c_{X|pa_X^G} \neq P^{c'}_{X|pa_X^G} \Rightarrow P^c_Z \neq P^{c'}_Z$$

and similarly for $P_{X|Z}$ for all pairs c, c'.

The above suggests that the misdirected edge $X \leftarrow Z$ results in additional mechanism changes and hence worse compression, therefore we can extend our score $\hat{g}^c$ and oracle Ω to accommodate data from multiple contexts. Given a partially inferred graph G and estimated intervened nodes I, we extend our score as

$$\hat{g}(X, Y; G, I) = \hat{g}(X^o, Y^o; G) + \Big(\sum_{c: Y \in I_c} \hat{g}(X^c, Y^c; G) \Big) .$$

Hereby we apply the score separately to contexts with changed conditionals and jointly to all remaining contexts, here denoted as X^o, Y^o.

To guarantee that this score can be used to identify sources, our central assumption is the following.

Assumption 4.5. *[Independent Mechanism Shift] We assume that causal mechanism shifts occur independently,*

$$\mathbb{I}\big(P^c_{X|pa^G_X} \neq P^{c'}_{X|pa^G_X}\big) \perp\!\!\!\perp \mathbb{I}\big(P^c_{Y|pa^G_Y} \neq P^{c'}_{Y|pa^G_Y}\big)$$

across any two contexts $c \neq c'$ for all $i \neq j$.

We also need to ensure that *not all* variables undergo shifts in all environments, respectively, that *some* shifts do exist. For a pair of contexts c, c', denoting the probability that mechanism change for node i occurs as

$$p_i^{c,c'} := \mathcal{P}[\mathbb{I}\big(P^c_{X_i|pa^G_i} \neq P^{c'}_{X_i|pa^G_i}\big) = 1] ,$$

we state this assumption as follows.

Assumption 4.6. *[Sparse Mechanism Shift] We assume that the probability p_i of a mechanism change between any two contexts occurring is bounded away from 0 and 1 for all i.*

This assumption was proposed to replace the i.i.d. assumption implicit in standard causal modeling for non-i.i.d. causal models [Perry et al., 2022].

Theorem 4.7. *Under Assumptions 4.3, 4.5 and 4.6, Assumption 3.4 holds with high probability as $|\mathcal{C}| \to \infty$.*

Implementation. Instantiating the interventional variant $\hat{g}$ needs two components: an observational score $\hat{g}$, as well as a way to estimate the intervention targets I. For the former, we use the MDL-based score proposed in Section 4.1 for consistency. To infer intervention targets, we need a means to test conditionals for (in)equality. For this purpose, we leverage the state-of-the-art Kernel Conditional Independence Test (KCI) [Zhang et al., 2011a].

5 RELATED WORK

Most approaches in causal DAG discovery can be categorized along two axes: whether they classify as constraint- or score-based, with a recent trend towards continuous optimization; and whether they address a single i.i.d. dataset or multi-context data, with a recent shift towards non-i.i.d. settings. As the field is gaining substantial research attention, we point to the most prominent approaches here and refer to, e.g., Zanga and Stella [2023] for an in-depth overview.

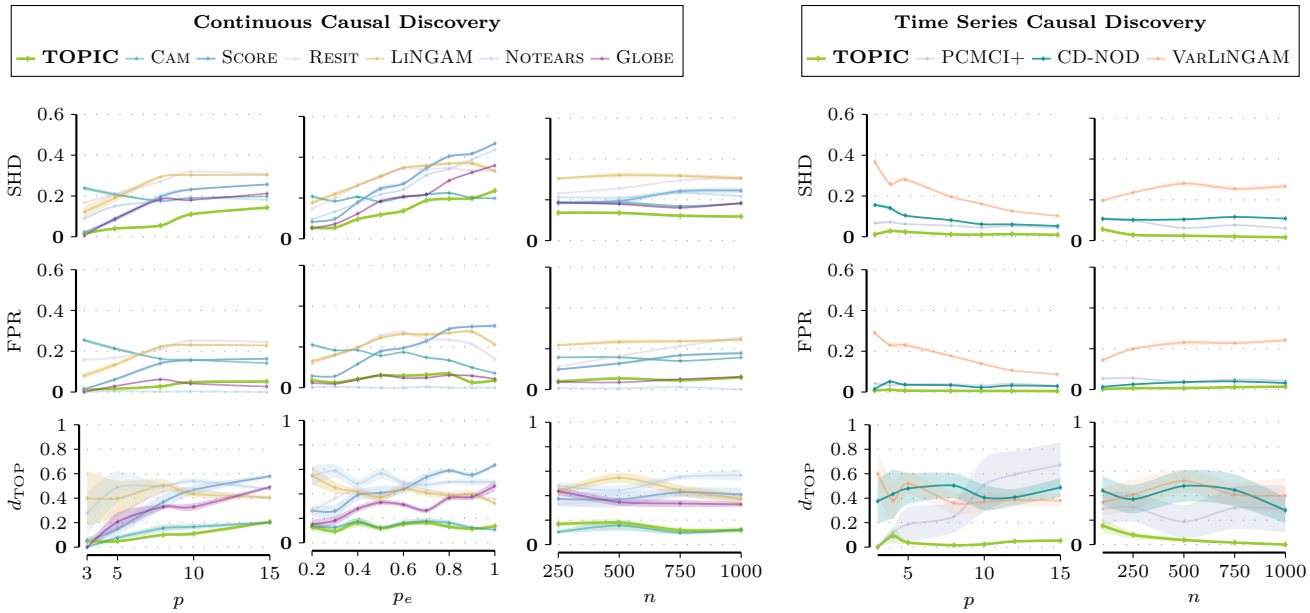
Constraint-based. Classical constraint-based methods include the well-known PC algorithm [Spirtes et al., 2001], which applies with a respective independence test in the continuous [Zhang et al., 2011b], discrete [Marx and Vreeken, 2019] and time series settings [Runge, 2020a]. For instance, PCMCI+ [Runge, 2020b] instantiates PC with the momentary conditional independence test (MCI) to capture time-lagged causal relationships.

Score-based. Score-based methods such as GES [Chickering, 2002] employ greedy search over the space of Markov Equivalence classes of DAGs. To recover all edge directions, methods such as LiNGAM [Shimizu et al., 2006] and Resit [Hoyer et al., 2008] assume an additive noise model and test for non-gaussianity resp. independence of residuals. The extension VarLiNGAM [Hyvärinen et al., 2010] enables causal discovery in the time series domain with a linear model.

Ordering-based. A recent trend is first learning a topological order and then estimating the fully oriented DAG. Cam [Bühlmann et al., 2014], Score [Rolland et al., 2022] and NOGAM [Montagna et al., 2023] topological order estimators for an additive SCM as in Eq. (2), and then add resp. prune edges using Lasso regression. For all methods, careful data-dependent tuning of the Lasso regularizer is required. Instead, Reisach et al. [2023] propose to use variance/R^2 respectively to determine causal order. Generally, all frameworks separately deal with edge selection and topological ordering and are exclusively applicable to the continuous domain.

Other Methods. Recently, methods based on continuous, neural network-based optimization have gained popularity. Notears [Zheng et al., 2018] formulates causal discovery as a continuous optimization task to learn DAGs, while Yu et al. [2019] propose to use graph neural networks to learn the causal structure. Like LiNGAM, Notears has a time series equivalent called DyNOTEARS [Pamfil et al., 2020], but which is restricted to linear models only.

Lastly, information-theoretic methods have gained traction in the field. Gao and Aragam [2021] derive



(a) DAG size (b) Edge density (c) Sample size (d) DAG size (e) Time points

Figure 3: *Causal Discovery with TOPIC.* On i.i.d. continuous (left) and time series (right) datasets, TOPIC's instantiation learns accurate topological orders (bottom), avoids both spurious and mis-oriented edges (middle), and thus outperforms domain-specific methods (top).

an entropy-based condition which allows identifiable causal discovery, but relies on the assumption of a single causal parent that is responsible for the majority of entropy generation. Most related to our work is the GLOBE algorithm [Mian et al., 2021] that greedily optimizes an MDL-score. While performing well in practice, it lacks identifiability guarantees for non-tree structured graphs.

Multi-Context. For non-i.i.d. data, frameworks such as CD-NOD offer extensions of constraint-based methods for multi-context data [Huang et al., 2020, Mooij et al., 2016] using conditional independence testing in an augmented graphical model. JPCMCI [Günther et al., 2023] treats multi-context time series data in a similar vein while also allowing for temporal confounders. Meanwhile, UT-IGSP [Squires et al., 2020] offers a hybrid score- and constraint-based approach. Although able to give tighter guarantees in the presence of interventions, the methods only identify equivalence classes. Recent work therefore leverages the sparse shift principle to discover additional causal directions [Perry et al., 2022], but does not address scalable DAG search.

6 EXPERIMENTS

In the following, we evaluate TOPIC on three domains: i.i.d. continuous, time series, and interventional data.

6.1 Experimental Setup

In our synthetic experiments, we generate Erdös-Renyi DAGs G with p nodes where edges are drawn with probability p_e. For the continuous case, we draw n samples using the structural model $X_i = f(\text{pa}_i^G) + N_i$ where f is an additive polynomial function with additive Gaussian noise $N_i \sim \mathcal{N}(0, 1)$.

For multivariate time series, we generate a Window Causal Graph (WCG) G_τ with maximum time lag τ and use the TIGRAMITE package to sample n timepoints as $X_i^t = f(\text{pa}_i^t) + N_i^t$ with non-linear f, additive Gaussian N_i^t, and where pa_i^t denotes the causes of X_i^t in G_τ, which can either be lagged parents X_j^t with $t \leq \tau$ or contemporaneous X_j^t with $j \neq i$ for $t = 0$.

In the non-i.i.d. scenario, we retain the same data generators and in addition sample i intervention targets in each of the m contexts. An intervention on node X_i replaces its structural equation f in the particular context by a constant c, and we then draw n samples in m under the updated equation model. Overall, this results in a parameter configuration (p, n, p_e, m, i).

We evaluate the methods in terms of the following metrics. The Structural Hamming Distance (SHD) is a standard measure quantifying the similarity of the true (G^*) and discovered (G) graph. We also report False Positive Rates (FPR) over directed edges between node pairs to assess whether the methods find

anti-causal or spurious edges. We finally evaluate the correct topological ordering using the topological order divergence d_{TOP} proposed in Rolland et al. [2022], which counts true edges whose direction strictly disagrees with the discovered topological order T,

$$d_{\text{TOP}}(T, G^*) = \sum_{i=1,\dots,p} \sum_{j:T(i)\geq T(j)} \mathbb{I}[(i,j) \in G^*].$$

We normalize the SHD and d_{TOP} measures to attain values between 0 and 1 (lower is better).

6.2 Homogeneous Data

We compare against the state-of-the-art methods in causal DAG discovery shown in Fig. 3, with configuration $(p = 8, n = 1000, p_e = 0.5)$ where we change one data parameter at a time. The methods PC, FCI, and GES generally performed worse in our experiments such that we omit them from presentation. NOTEARS and LINGAM appear with overall worst performance which we attribute to their strict linearity assumptions. We observe that in terms of the SHD (top), RESIT and LINGAM perform worse on sparse graphs, as opposed to GLOBE and SCORE who degrade as the graph becomes dense. CAM achieves close performance to TOPIC for fully connected graphs, while in the remaining settings TOPIC performs best.

In particular, TOPIC distinguishes itself through a low false positive rate (middle). Here, the advantage of automatic model/edge selection through MDL is apparent, compared to the Lasso-based methods CAM and SCORE. Only RESIT performs well with a similar false positive rate, but generally learns worse topological orders, emphasizing the advantage of topological-ordering-based approaches. Amongst these, TOPIC and CAM generally learn more accurate orders in terms of d_{TOP} than score-matching based SCORE (bottom). Particularly, compared to GLOBE which is also MDL based, TOPIC's inclusion of a topological order leads to significant improvements in accuracy.

Time Series. In the time series domain, we compare to PCMCI+, CD-NOD and VARLINGAM, where we omit DYNOTEARS from the presentation due to inferior results. We report the results in Fig. 3d and 3e. Regarding SHD, only PCMCI+ is competitive to our method, whilst CD-NOD and especially the linear VARLINGAM perform worse. For an increasing number of nodes, TOPIC remains stable, whilst PCMCI+ degrades, especially with regard to the topological ordering (bottom). On the other hand, TOPIC's topological ordering becomes increasingly accurate with more sampled time points (right), as there the MDL approximation of Kolmogorov complexity improves. Overall, TOPIC is also well versed for time series data and outperforms specialized methods such as PCMCI+.

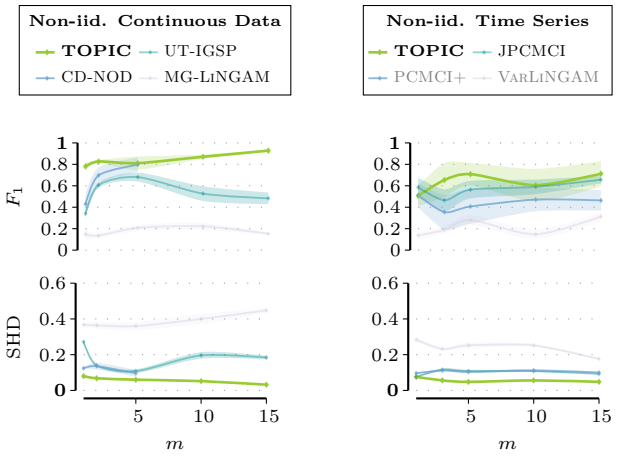


Figure 4: *Non-i.i.d. Causal Discovery with TOPIC.* On heterogeneous data, TOPIC performs well both in the continuous (left) and time series case (right), with edge recall increasing in the number of contexts (top).

6.3 Heterogeneous Data

Lastly, we test TOPIC in the presence of interventions for both continuous and time series data. We report the results comparing against JPCMCI [Günther et al., 2023] in the time series domain, as well as UT-IGSP Squires et al. [2020] and MG-LINGAM [Shimizu, 2012] for continuous non-i.i.d. data in Fig. 4. We find that TOPIC works very well for continuous data with mechanism shifts (left), outperforming its competitors by a wide margin, especially under many independent mechanism shifts.

In the time series domain (right), JPCMCI is competitive with TOPIC for few shifts, with the latter performing better in terms of the F_1, but worse in terms of the topological ordering. Again, when the number of shifts increases, TOPIC is able to improve its performance, while JPCMCI runs into scalability issues due to its costly conditional independence tests [Runge, 2020b]; this is similarly the case for CD-NOD, which did not scale beyond $m = 5$ contexts (left). Overall, the empirical results back up Theorem 4.7, which shows that in the limit of contexts $m \to \infty$, the identifiability condition of TOPIC is satisfied.

6.4 Real-World Data

To evaluate the practical applicability of TOPIC, we now assess its performance on realistic biological data. The REGED dataset encompasses re-simulated gene expression levels of lung cancer patients, and includes three networks with resp. 5, 15 and 500 nodes and a labeled ground truth DAG [Statnikov et al., 2015]. We report the results of all methods that could process the

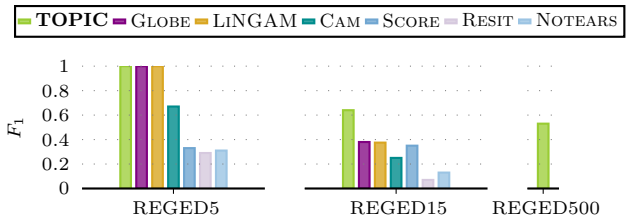


Figure 5: *Experiments on gene regulatory networks.* We report F_1 scores of all methods that finish within 2 days. TOPIC obtains the highest F_1 and is the only method to scale up to 500 variables.

data within two days in Fig. 5.

On the five-node network, TOPIC, GLOBE, and LINGAM all recover the true graph G^*, whilst CAM, SCORE, RESIT and NOTEARS achieve F_1 scores between 0.3-0.6 resp. For the 15-node network, TOPIC has the highest F_1 score of **0.64**, with the next best methods GLOBE, LINGAM and SCORE all scoring around 0.4. Finally, on the 500-node network, only TOPIC was able to finish within the timeout and achieves an F_1 score of 0.5, showcasing the scalability and accuracy of TOPIC to large real-world networks.

7 CONCLUSION

We introduced the TOPIC framework for causal discovery in topological order. Inspired by the universal measure of Kolmogorov complexity, TOPIC offers complementary approaches and identifiability guarantees for both the i.i.d. continuous and non-i.i.d. multi-context setting. On synthetic and real-world benchmarks, TOPIC outperforms specialized methods on continuous, time series, and interventional data.

Limitations. TOPIC relies on Kolmogorov complexity, which is uncomputable. To approximate it with MDL, we have to define a model class and respective encoding costs. If the real-world data does not adhere to the assumptions of the model class, the method may fail to adequately estimate the complexity. Hence, the resp. instantiation of TOPIC is sensitive to the choice of the model class and parameter encoding. Another limiting factor is the assumption of causal sufficiency and faithfulness. In practice, both assumptions may not hold. Consequently, due to the presence of hidden confounders, the method may infer spurious edges due to a hidden common cause unobserved in the data. Furthermore, failure to recall true edges due to a violation of faithfulness leads to an incomplete graph and can in the worst case affect the learning of the topological order.

Future Work. For future work, we plan to further employ TOPIC in the discrete domain and examine the identifiability guarantees for the time series setting. Another promising direction is extending TOPIC with MDL-based latent confounding detection, such as by Kaltenpoth and Vreeken [2019]. In addition, we aim to relax our assumptions further, e.g., by using triplet faithfulness [Marx et al., 2021] in place of classic faithfulness.

References

Antonia A. Dominguez, Wendell A. Lim, and Lei S. Qi. Beyond editing: repurposing crispr–cas9 for precision genome regulation and interrogation. *Nature Reviews Molecular Cell Biology*, 17 (1):5–15, 2016. doi: 10.1038/nrm.2015.2. URL `https://doi.org/10.1038/nrm.2015.2`.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Peter Bühlmann, Jonas Peters, and Jan Ernest. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014.

Nicolai Meinshausen, Alain Hauser, Joris M. Mooij, Jonas Peters, Philip Versteeg, and Peter Bühlmann. Methods for causal inference from gene perturbation experiments and validation. 113(27):7361–7368, 2016.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. 21(1), 2020. ISSN 1532-4435.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 1039–1048. PMLR, 03–06 Aug 2020.

Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. 2022.

Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10): 5168–5194, 2010.

Spencer Compton, Murat Kocaoglu, Kristjan Greenewald, and Dmitriy Katz. Entropic causal inference: Identifiability and finite sample results. *Advances in Neural Information Processing Systems*, 33:14772–14782, 2020.

Sascha Xu, Osman Mian, Alexander Marx, and Jilles Vreeken. Inferring cause and effect in the presence of heteroscedastic noise. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2022.

Amirkasra Jalaldoust, Kateřina Hlaváčková-Schindler, and Claudia Plant. Causal discovery in hawkes processes by minimum description length. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6978–6987, 2022.

Sarah Mameche, David Kaltenpoth, and Jilles Vreeken. Learning causal models under independent changes. In *Proceedings of Neural Information Processing Systems (NeurIPS)*. PMLR, 2023.

Osman A Mian, Alexander Marx, and Jilles Vreeken. Discovering fully oriented causal networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8975–8982, 2021.

Alexander Statnikov, Sisi Ma, Mikael Henaff, Nikita Lytkin, Efstratios Efstathiadis, Eric R. Peskin, and Constantin F. Aliferis. Ultra-scalable and efficient methods for hybrid observational and experimental local causal pathway discovery. *Journal of Machine Learning Research*, 16(100):3219–3267, 2015.

S.L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Clarendon Press, 1996. ISBN 9780191591228.

Alain Hauser and Peter Bühlmann. Jointly interventional and observational data: Estimation of interventional markov equivalence classes of directed acyclic graphs. 77, 03 2013. doi: 10.1111/rssb.12071.

Gregory J Chaitin. A theory of program size formally identical to information theory. *Journal of the ACM (JACM)*, 22(3):329–340, 1975.

Ming Li and Paul Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer Science & Business Media, 2009.

Peter Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

Alexander Marx and Jilles Vreeken. Telling cause from effect using mdl-based local and global regression. In *2017 IEEE international conference on data mining (ICDM)*, pages 307–316. IEEE, 2017.

Kailash Budhathoki and Jilles Vreeken. Origo: causal inference by compression. *Knowledge and Information Systems*, 56(2):285–307, 2018.

Alexander Marx and Jilles Vreeken. Formally justifying mdl-based inference of cause and effect. In *AAAI Workshop on Information-Theoretic Causal Inference and Discovery (ITCI'22)*, 2022.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

K Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proc. 25th Conference on Uncertainty in Artificial Intelligence (UAI2009)*, 2009.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 804–813, Arlington, Virginia, USA, 2011a. AUAI Press. ISBN 9780974903972.

Alessio Zanga and Fabio Stella. A Survey on Causal Discovery: Theory and Practice. *arXiv e-prints*, art. arXiv:2305.10032, May 2023. doi: 10.48550/arXiv.2305.10032.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813, 2011b.

Alexander Marx and Jilles Vreeken. Testing conditional independence on discrete data using stochastic complexity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 496–505. PMLR, 2019.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1388–1397. Pmlr, 2020a.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 1388–1397. PMLR, 03–06 Aug 2020b.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(56):1709–1731, 2010. URL `http://jmlr.org/papers/v11/hyvarinen10a.html`.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pages 18741–18753. PMLR, 2022.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *Conference on Causal Learning and Reasoning*, pages 726–751. PMLR, 2023.

Alexander Reisach, Myriam Tami, Christof Seiler, Antoine Chambaz, and Sebastian Weichwald. A scale-invariant sorting criterion to find a causal order in additive noise models. *Advances in Neural Information Processing Systems*, 36:785–807, 2023.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International conference on machine learning*, pages 7154–7163. PMLR, 2019.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Paul Beaumont, Konstantinos Georgatzis, and Bryon Aragam. DYNOTEARS: Structure Learning from Time-Series Data. *arXiv e-prints*, art. arXiv:2002.00498, February 2020. doi: 10.48550/arXiv.2002.00498.

Ming Gao and Bryon Aragam. Efficient bayesian network structure learning via local markov boundary search. *Advances in Neural Information Processing Systems*, 34:4301–4313, 2021.

Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint Causal Inference from Multiple Contexts. *arXiv e-prints*, art. arXiv:1611.10351, November 2016. doi: 10.48550/arXiv.1611.10351.

Wiebke Günther, Urmi Ninad, and Jakob Runge. Causal discovery for time series from multiple datasets with latent contexts. In Robin J. Evans and Ilya Shpitser, editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine*

Learning Research, pages 766–776. PMLR, 31 Jul–04 Aug 2023.

Shohei Shimizu. Joint estimation of linear non-gaussian acyclic models. *Neurocomputing*, 81:104–107, 2012. ISSN 0925-2312.

David Kaltenpoth and Jilles Vreeken. We are not your real parents: Telling causal from confounded using mdl. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 199–207. SIAM, 2019.

Alexander Marx, Arthur Gretton, and Joris M. Mooij. A weaker faithfulness assumption based on triple interactions. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 451–460. PMLR, 27–30 Jul 2021.

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016. URL `http://jmlr.org/papers/v17/14-518.html`.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, we provide a problem description, state all assumptions, provide instantiation details and pseudo code.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes, we provide a analysis of the complexity of our algorithm.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes, we provide all source code and data generators in the supplementary material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes we state all assumptions in the main text.
   (b) Complete proofs of all theoretical results. Yes we provide both proof overviews in the main paper and full proofs in the appendix.
   (c) Clear explanations of any assumptions. Yes, we provide a detailed explanation of all assumptions when introducing them.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes, we provide all source code and data generators in the supplementary material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes, we describe how all metrics were computed and hyperparameters for all methods were chosen.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes, we provide confidence intervals of 1 standard deviation for all presented experiments.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes, we run all experiments on a CPU cluster.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes.
   (b) The license information of the assets, if applicable. Yes.
   (c) New assets either in the supplemental material or as a URL, if applicable. Yes.
   (d) Information about consent from data providers/curators. Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

A Theory for the Observational Domain

We provide the proofs for our theorems here, where we assume familiarity with d-separation and refer to Lauritzen [1996] and Pearl [2009] for background on graphical causal models.

A.1 Faithfulness Assumption

Assumption 3.1. *[Faithfulness] Given p random variables with true causal graph G^*, a pair of variables X and Y is algorithmically independent conditioned on a set of variables Z, i.e. $I_A(P_X; P_Y \mid Z) \stackrel{+}{=} 0$, if and only if they are d-separated in G^*.*

Lemma 1. *Under the faithfulness assumption, the compression gain of a true edge (X, Y) is positive, i.e. $g(X_X, X_Y; G) > 0$ for any graph G that does not contain it.*

Proof. If X and Y are connected in the true graph G^*, then no set Z exists that d-separates X and Y. Hence, $I_A(P_X; P_Y \mid Z) \stackrel{+}{>} 0$ for any set Z. Then, it holds that

$$0 \stackrel{+}{<} I_A(P_X; P_Y \mid \mathrm{pa}_Y^G) = K\left(P_Y|\mathrm{pa}_Y^G\right) - K\left(P_Y|\mathrm{pa}_Y^G \cup \{X\}\right) = g(X_X, X_Y; G) \; .$$

□

A.2 Proof of the Graph Identifiability Theorem 3.3

Assumption 3.2. *[Algorithmic Independence of Conditionals] The distribution of two random variables X and Y conditioned on their true causal parents in G^* is algorithmically independent , i.e.*

$$I_A\big(P_{X|pa_X^*}; P_{Y|pa_Y^*}\big) \stackrel{+}{=} 0 \; .$$

Theorem 3.3. *[DAG Identifiability] Under the Algorithmic Independence of Conditionals and Faithfulness, given an oracle Ω, TOPIC recovers the true causal graph G^*.*

Proof. We begin by showing that TOPIC proceeds in a topological order of the true graph G^*. To this end, we inductively show that the oracle Ω's conditions are met in each iteration, hence resulting in a valid node with $T(X_k) = k$ for $k = 1, \ldots, p$.

Base case: We initialize the graph G as an empty graph. A true source X of the graph G^* fulfills the conditions of the oracle Ω. That is, all parents of X are accounted for in G. Since G^* is a directed acyclic graph, there exists at least one source node for Ω to select.

Inductive step: Assume that at iteration k we have selected a node Z such that $T(Z) = k$, and G contains all edges $(X, Y), T(X) < k$ that are in the true graph G^*. By the faithfulness assumption and Lemma 1, for any outgoing edge from Z, $(Z, Y) \in G^*, T(Y) > k$, the gain $g(X_Z, X_Y; G) > 0$ is positive and (Z, Y) hence added to the graph G. Thus, for the step $k + 1$, it holds that $\forall (X, Y) \in G^*, T(X) < k + 1 : (X, Y) \in G$. The oracle Ω is again admissible and will return a node X_{k+1} such that $T(X_{k+1}) = k + 1$.

We have established the TOPIC proceeds in a topological order T of the true graph G^*. We now show that the true graph G^* is recovered.

Recall of all true edges: TOPIC adds all edges (X, Y) to the graph G that compress, i.e. $g(X_X, X_Y; G) > 0$, and that are compatible with the topological order, i.e. $T(X) < T(Y)$. By definition, as T is a valid order for the true graph G^*, for all edges $(X, Y) \in G^*$ it holds that $T(X) < T(Y)$. Furthermore, by Lemma 1, the compression gain of a true edge (X, Y) is positive, i.e. $g(X_X, X_Y; G) > 0$ for any graph G that does not contain it. Hence, TOPIC adds a superset of the true edges, i.e. $G^* \subseteq G$.

Removal of all redundant edges: We now show that TOPIC removes all redundant edges. To this end, we use the Algorithmic Independence of Conditionals, which states that the Kolmogorov complexity of a superset of parents $\mathrm{pa}_X^G \supseteq \mathrm{pa}_X^*$ is equivalent to the complexity under the true parents, i.e.

$$K(P_X \mid \mathrm{pa}_X^G) \stackrel{+}{=} K(P_X \mid \mathrm{pa}_X^*) \; .$$

As per Lemma 1, we find a superset of all true edges by adding all edges that compress. Furthermore, it holds at iteration k that $\forall (X,Y) \in G, T(X) < k : (X,Y) \in G^*$. Hence, $\text{pa}_X^G \supseteq \text{pa}_X^*$. By removing all elements of $Z \in \text{pa}_X^G$, for which

$$K(P_X \mid \text{pa}_X^G \setminus \{Z\}) \stackrel{+}{=} K(P_X \mid \text{pa}_X^G) \ ,$$

we remove only non-parent nodes of X, as by the faithfulness for a true parent Z it holds that

$$K(P_X \mid \text{pa}_X^G \setminus \{Z\}) \stackrel{+}{>} K(P_X \mid \text{pa}_X^G) \ .$$

Therefore, we prune in iteration k the parent set pa_X^G to the true parent set pa_X^*, so that in the end for all nodes X it holds that $\text{pa}_X^G = \text{pa}_X^*$ and as such $G = G^*$. □

A.3 Oracle

Assumption 3.4. *[Information-Theoretic Identifiability] Let X be a resolved node of G, i.e., $pa_X^G = pa_X^{G^*}$, and Y be a descendant of X in G^*. If all confounders of X and Y are accounted for in G, i.e., $pa_Y^G \supseteq (pa_X^* \cap pa_Y^*)$, then it holds that*

$$\begin{aligned} &I_A\left(P_X \mid pa_X^G; P_Y \mid pa_Y^G \cup \{X\}\right) \\ \stackrel{+}{<} &I_A\left(P_Y \mid pa_Y^G; P_X \mid pa_X^G \cup \{Y\}\right) \ . \end{aligned}$$

We base our oracle on the compression gain of an edge (X,Y) in graph G, defined as

$$g(X,Y;G) = K\left(P_Y \mid pa_Y^G\right) - K\left(P_Y \mid pa_Y^G \cup \{X\}\right) \ .$$

With that in mind, we construct the oracle $\Omega(G, T^{k-1})$ as follows

$$\underset{T(X)=-1}{\arg\max} \left(\min_{T(X)=-1} g(X,Y;G) - g(Y,X;G) \right) \ . \tag{3}$$

Theorem 3.5. *Under Assumption 3.4, the oracle Ω returns a node X for which $pa_X^G = pa_X^*$ and $T(X) = k$ is valid with regard to G^* provided that G contains all edges up to the k-th node.*

Proof. We show that the oracle returns a node X such that $T(X) = k$, if given a partial topological order T^{k-1} and a graph G that contains all edges $(X,Y), T(X) < k$ that are in the true graph G^*. To this end, we distinguish between two cases: $\text{pa}_X^G = \text{pa}_X^*$, i.e. all parents are accounted for or $\exists Z : Z \notin \text{pa}_X^G \wedge Z \in \text{pa}_X^*$, i.e. a parent is missing.

All parents accounted for: : Let X be a node for which $\text{pa}_X^G = \text{pa}_X^*$ and consider any remaining node Y, $T(Y) > k, Y \neq X$. If Y is d-separated from X given pa_X^G, it holds that both $I_A(P_X \mid \text{pa}_X^G; P_Y \mid \text{pa}_Y^G \cup \{X\}) = 0$ and $I_A(P_X \mid \text{pa}_X^G \cup \{Y\}; P_Y \mid \text{pa}_Y^G) = 0$, as they are independent and their parents can not be collider nodes as it holds that $T(Z) < k$ for all $Z \in \text{pa}_Y^G$ and for all $Z \in \text{pa}_X^G$. Hence, if Y is d-separated from X given pa_X^G, the delta in this case is $\Delta_{X,Y} = 0$.

Let Y be not d-separated from X given pa_X^G. Then Y is a descendant of X in the true graph G^*, as all ancestor nodes of X are already accounted for in G. We first show that Assumption 3.4 applies to X. We note that for any confounder Z of X and Y, i.e. $Z \in \text{pa}_X^* \cap \text{pa}_Y^*$, must come before X in the order so that $T(Z) < k$. By the completeness of G up to k, all edges (Z,Y) are in the graph G, hence fulfilling the requirement that no confounding path is opened by adding X.

Hence, the edge (X, Y) is identifiable through Kolmogorov complexity, i.e.

$$\begin{aligned}
I_A\left(P_X \mid pa_X^G;\, P_Y \mid pa_Y^G \cup \{X\}\right) &\overset{+}{<} I_A\left(P_Y \mid pa_Y^G;\, P_X \mid pa_X^G \cup \{Y\}\right) \\
K(P_X \mid pa_X^G) - K(P_X \mid pa_X^G \cup \{Y\}) &\overset{+}{<} K(P_Y \mid pa_Y^G) - K(P_Y \mid pa_Y^G \cup \{X\}) \\
K(P_X \mid pa_X^G) + K(P_Y \mid pa_Y^G \cup \{X\}) &\overset{+}{<} K(P_Y \mid pa_Y^G) + K(P_X \mid pa_X^G \cup \{Y\}) \\
0 \overset{+}{<} \left(K(P_Y \mid pa_Y^G) - K(P_Y \mid pa_Y^G \cup \{X\})\right) &- \left(K(P_X \mid pa_X^G) - K(P_X \mid pa_X^G \cup \{Y\})\right) \\
0 &\overset{+}{<} g(X, Y; G) - g(Y, X; G) \\
0 &\overset{+}{<} \Delta_{X,Y} \,.
\end{aligned}$$

That is, the compression gain of the causal edge (X, Y) outweighs the compression gain of the anti-causal edge (Y, X) if $\mathrm{pa}_X^G = \mathrm{pa}_X^*$, i.e. $\Delta_{X,Y} > 0$. Hence, if X is a node for which $\mathrm{pa}_X^G = \mathrm{pa}_X^*$, for any admissible graph G

$$\min_{Y, T(Y)=-1} \Delta_{X,Y}(G) \overset{+}{\geq} 0$$

.

A parent is missing: On the other hand, let $\mathrm{pa}_X^G \neq \mathrm{pa}_X^*$. Consider a missing node $V \in \mathrm{pa}_X^* \setminus \mathrm{pa}_X^G$. Either V is resolved in the graph G, i.e. $pa_V^G = pa_V^*$, or we recursively find an ancestor node $Z \in pa_V^* \setminus pa_V^G$ that is. For that node Z there exists an unaccounted path from Z to X in the true graph G^* so that they are not algorithmically independent, i.e.

$$I_A\left(P_X \mid \mathrm{pa}_X^G; P_Z \mid pa_G^Z\right) \overset{+}{\neq} 0 \,.$$

As X is anti-causal for Z, and all parents of Z are accounted for in G, by Assumption 3.4 it holds that

$$g(X, Z; G) \overset{+}{<} g(Z, X; G) \;\Leftrightarrow\; \Delta_{X,Z}(G) \overset{+}{<} 0 \,.$$

Hence, if X is a node for which $\mathrm{pa}_X^G \neq \mathrm{pa}_X^*$, for any admissible graph G

$$\min_{Y, T(Y)=-1} \Delta_{X,Y}(G) \overset{+}{<} 0 \,.$$

Using the oracles definition from Eq. (3), which defines the returned node X as

$$\arg\max_{T(X)=-1} \left(\min_{T(X)=-1} g(X, Y; G) - g(Y, X; G) \right) \,,$$

this shows that for any node X it must hold that $\mathrm{pa}_X^G = \mathrm{pa}_X^*$, provided such a node X exists. As G^* is a directed acyclic graph and G is complete up to node $T(X) = k$, there exists a valid topological order and hence a compatible node X with $T(X) = k$ to progress the partial order T^{k-1}. □

A.4 Proof of Theorem 4.2

Assumption 4.1. *[Additive SCM] We consider a structural causal model where each variable Y is generated as a sum of its parents and noise as per*

$$Y = \sum_{X \in pa_Y^*} f_{X,Y}(X) + N_Y \,. \tag{2}$$

Theorem 4.2. *Under the additive SCM as per Eq. (2), any descendant Y of a resolved node X with $pa_X^G = pa_X^{G^*}$ can be expressed as a post-nonlinear model of X, where Assumption 3.4 holds if the post-nonlinear model is identifiable.*

Proof. Given a set of variables that follow a causal structure G^* adhering to an additive SCM as per Eq. (2). Let X be a resolved node of G with $pa_X^G = pa_X^{G^*}$, and Y be a descendant of X in G^* and all confounders of X and Y be accounted for in G, i.e. $\text{pa}_Y^G \supseteq (\text{pa}_X^* \cap \text{pa}_Y^*)$.

Consider the SCM of Y given the parents of Y and X as

$$Y = \sum_{Z \in \text{pa}_Y^*} f_{Z,Y}(Z \mid \text{pa}_Y^G \cup \{X\}) + N_Y \ .$$

We now show that the causal direction corresponds to a post non-linear noise model, which is known to be separable into independent functions and noise terms in the causal direction only [Zhang and Hyvärinen, 2009].

We first separate the parent set of Y into two sets: $pa_1 = \{Z \in \text{pa}_Y^G \vee Z \notin an_X^* \mid Z \in \text{pa}_Y^*\}$, i.e. all parents that are either accounted for or non-ancestors of X, and the remaining parent set $pa_2 = \text{pa}_Y^* \setminus pa_1$ that directly depends on X. We first show that each parent in pa_1 is independent of X conditioned on pa_Y^G. We begin by dropping all parents Z which are already accounted for in G, i.e. $T(Z) < k \wedge Z \in \text{pa}_Y^G$, as the term $f_{Z,Y}(Z \mid Z)$ is deterministic. For all $Z \in pa_1$ that are not yet accounted for in G, i.e. $Z \notin \text{pa}_Y^G$, we distinguish between three cases:

1. Z **is independent of** X: If Z is independent of X in the true graph G^*, $f_{Z,Y}(Z)$ also is independent of X. Furthermore, there can not be a collider for Z and X modeled in G as no outgoing edges from X have been fitted. Hence, $f_{Z,Y}(Z)$ is independent of X.

2. Z **is an ancestor of** X **in** G^*: This is a contradiction, as G must be complete up to X. Hence, the edge (Z, Y) must be in the graph, which contradicts the assumption that (Z, Y) is not accounted for in G.

3. Z **and** X **are confounded in** G^*: Let Z and X be confounded in the true graph G^*. Then there exists a set of variables C, $\forall V \in C : T(V) < k$ that d-separates Z and X, so that $I_A(P_Z; P_X \mid C) = 0$.

   For all $V \in C$ there must be an unblocked path to Y, so that they are algorithmically dependent, i.e. $I_A(P_V; P_Y \mid \{V\}) \overset{+}{>} 0$. Therefore C must be contained in the parents of Y in G, i.e. $C \subseteq \text{pa}_Y^G$, as all non-d-separated pairs of variables that have $T(V) < k$ are included in G. Consequently, X and Z are d-separated through pa_Y^G so that $f_{Z,Y}(Z \mid \text{pa}_Y^G)$ is independent of X.

Hence, all terms in pa_1 are independent of X, allowing us to compose an additive independent noise term together as $N_Y^G = N_Y + \sum_{Z \in pa1} f_{Z,Y}(Z \mid \text{pa}_Y^G)$.

What remains are the terms in pa_2, i.e. the parents of Y that directly depend on X. If $pa_2 = \{X\}$, then Y is an additive noise model $Y = f_{X,Y}(X) + N_Y^G$, which is identifiable as outlined by Hoyer et al. [2008]. If pa_2 contains variables Z which are ancestors of X in the true graph G^*, then those variable Z can in turn be represented as an additive noise model $Z = f_{X,Z}(X) + N_Z^G$ using the same distinction as above. Therefore, we can express Y as

$$Y = \sum_{Z \in pa_2} f_{Z,Y}(f_{X,Z}(X) + N_Z^G) + N_Y^G$$

which is a post-nonlinear model, and hence identifiable apart from certain pathological cases [Zhang and Hyvärinen, 2009, Peters et al., 2014]. In the case that it is identifiable, it holds that

$$I_A\left(P_X \mid \text{pa}_X^G;\, P_Y \mid \text{pa}_Y^G \cup \{X\}\right) \overset{+}{=} 0 \ ,$$

whereas in the anti-causal direction, no models exist that can separate the effect Y into independent noise terms so that

$$I_A\left(P_X \mid \text{pa}_X^G \cup \{Y\};\, P_Y \mid \text{pa}_Y^G\right) \overset{+}{>} 0 \ .$$

Hence, Assumption 3.4 holds for the additive SCM. □

B Theory for the Multi-Context Domain

We first provide background on a common representation of multi-context data through an augmented graphical model, allowing to translate changes of causal conditionals to d-separations in the augmented graph.

B.1 Augmented Causal Graphical Model

We recall our generating distribution in multiple contexts as follows.

Definition B.1 (Multi-Context Generating process)**.** *The joint distribution P^c over $X_1, \dots, X_p$ in a context c can be written as*

$$P^c_{X_1,\dots,X_p} = \prod_{i \notin I^*_c} P_{X_i|pa_i^G} \Big(\prod_{i \in I^*_c} P^c_{X_i|pa_i^G} \Big),$$

That is, each context c results from causal mechanism shifts of a subset of variable indices $I^*_c \subseteq \{1, \dots, p\}$.

Equivalently, to represent mechanism changes I^*_c jointly with the causal graph G, we use the following augmented graphical causal model [Huang et al., 2020].

Definition B.2 (Augmented Causal Graph)**.** *Given a collection of causal models $\{(G, P^c)\}_{c \in \mathcal{C}}$ in a set of contexts $\mathcal{C}$ with shared causal graph G and distribution P^c as in Def. B.1, the augmented causal model $(G', P^c_{X \cup \{C\}})$ consists of*

(i) a categorical index variable C with support $\mathcal{C}$

(ii) a causal DAG with vertices $\{1, \dots, p\} \cup \{C\}$ and edges

$$\{(i,j) \in G\} \cup \{(C,i) \mid \exists c, c' : P^c_{X_i|pa_i^G} \neq P^{c'}_{X_i|pa_i^G}\}.$$

.

We restate our shift faithfulness assumption for ease of access here.

Assumption 4.3. *[Shift Faithfulness] For each i and any two environments c, c', when $i \in I^*_{c'}$, then*

$$P^c_{X_i|pa_i^G} \neq P^{c'}_{X_i|pa_i^G}.$$

The shift faithfulness is stated for the causal model in Definition B.1 and corresponds to faithfulness with C in the augmented graph. Together with the causal Markov property, it allows connecting distribution changes of conditionals across contexts to d-separation statements in the above augmented causal graph as follows.

Lemma B.3. *For any node i and node set $P \subseteq \{1, \dots, p\}$, the conditional $P_{X_i|X_P}$ changes if and only if X_i is d-connected to C given the nodes X_P in G', $j \perp\!\!\!\perp_{G'} C \mid P$.*

This holds as a direct consequence of the Markov property and faithfulness in the augmented graph [Perry et al., 2022]. We move on to our theoretical results.

B.2 Proof of Lemma 4.4 on Mechanism Shifts along Paths

Lemma 4.4. *[Path Shifts] Given a resolved node X and node Z with a directed path from X to Z in G^*, when Assumption 4.3 holds, $P_{X|Z}$ and P_Z both reflect the true mechanism changes of X,*

$$P^c_{X|pa_X^G} \neq P^{c'}_{X|pa_X^G} \Rightarrow P^c_Z \neq P^{c'}_Z$$

and similarly for $P_{X|Z}$ for all pairs c, c'.

Proof. Assume there is a pair of contexts c, c' such that $P^c_{X_i|pa_i^G} \neq P^{c'}_{X_i|pa_i^G}$ holds.

Then the augmented G' with an auxiliary node C contains an edge towards node i by construction. Therefore, no matter the conditioning set X_P (where $P \subseteq \{1, \dots, p\}$ is a subset of node indices), i and C will be d-connected. The result then follows under (shift) faithfulness and Lemma B.3.

□

B.3 Proof of Theorem 4.7 on Multi-Context Identifiability

Assumption 4.5. *[Independent Mechanism Shift] We assume that causal mechanism shifts occur independently,*

$$\mathbb{I}\big(P^c_{X|pa^G_X} \neq P^{c'}_{X|pa^G_X}\big) \perp\!\!\!\perp \mathbb{I}\big(P^c_{Y|pa^G_Y} \neq P^{c'}_{Y|pa^G_Y}\big)$$

across any two contexts $c \neq c'$ for all $i \neq j$.

Assumption 4.5 implies that existence of the edge $C \to i$ provides no information on the existence of an edge $C \to j$ for $i \neq j$. To show that this allows to decide the orientation of edges in our topological ordering algorithm, we assume that there are sufficiently many such edges and that not all variables undergo mechanism shifts.

Assumption 4.6. *[Sparse Mechanism Shift] We assume that the probability p_i of a mechanism change between any two contexts occurring is bounded away from 0 and 1 for all i.*

We move to the result that information-theoretic identifiability as in Assumption 3.4 holds for our score given sufficiently many contexts.

Theorem 4.7. *Under Assumptions 4.3, 4.5 and 4.6, Assumption 3.4 holds with high probability as $|\mathcal{C}| \to \infty$.*

Proof. Let X be a resolved node of G with $pa^G_X = pa^{G^*}_X$, and Y be a descendant of X in G^* and all confounders of X and Y be accounted for in G, i.e. $\mathrm{pa}^G_Y \supseteq (\mathrm{pa}^*_X \cap \mathrm{pa}^*_Y)$. We recall our score as

$$\hat{g}(X, Y; G, I) = \hat{g}(X^o, Y^o; G) + \Big(\sum_{c: Y \in I_c} \hat{g}(X^c, Y^c; G) \Big),$$

and similarly $\hat{g}(Y, X; G, I)$ in the anti-causal direction.

By Lemma 4.4, the inferred targets I_c in the anti-causal direction are the same as in the causal direction but in addition include $Y \in I_c$ whenever $X \in I_c$. Therefore, if there exists a c such that $Y \notin I^*_c$ but $X \in I^*_c$, the score $\hat{g}(Y, X; G, I)$ will contain one interventional more term $\hat{g}(X^c, Y^c; G)$ and thus be strictly larger. For this to work, we also need to ensure there is no confounder Z, $Z \in I^*_c$ that overlaps with the change of X and Y.

Therefore, we bound the probability of the event that X_i changes, but neither X_j nor any of the confounders P change similarly as in Perry et al. [2022].

For a given pair c, c' of contexts, this event occurs with probability

$$\begin{aligned}
&\mathcal{P}\big[\mathbb{I}(P^c_i \neq P^{c'}_i) = 1; \mathbb{I}(P^c_j = P^{c'}_j) = 1; \mathbb{I}(P^c_k = P^{c'}_k) = 1, \forall k \in P\big] \\
&= p_i^{c,c'}(1 - p_j^{c,c'}) \prod_{k \in P} (1 - p_k^{c,c'}) \\
&\leq p_i^{c,c'}(1 - p_j^{c,c'})(1 - \max_{k \in P} p_k^{c,c'})^{|P|}
\end{aligned}$$

where $P \subseteq \{1, \ldots, p\}$ are the indices of the path-confounders accounted for in G. Above, we independence of mechanism shifts (Assumption 4.5) and a worst-case bound for the variable k with the highest shift probability $p_k^{c,c'}$.

We lower bound its inverse probability to ensure that the above event occurs in at least one pair of contexts. Using a union bound over all contexts and path-confounders leads to the following overall expression,

$$\begin{aligned}
&\mathcal{P}\big[\ \nexists c, c' : \big[\mathbb{I}(P^c_i \neq P^{c'}_i) = 1; \mathbb{I}(P^c_j = P^{c'}_j) = 1; \mathbb{I}(P^c_k = P^{c'}_k) = 1, \forall k \in P\big]\big] \\
&\geq 1 - \big((1 - \max_{c,c'} p_j^{c,c'}) \min_{c,c'} p_i^{c,c'} (1 - \max_{c,c',k \in P} p_k^{c,c'})\big)^{\lfloor |\mathcal{C}|/2 \rfloor |P|} \\
&\geq 1 - \big((1 - p_j^{\mathrm{UB}}) p_i^{\mathrm{LB}} (1 - \max_{k \in P} p_k^{\mathrm{UB}})\big)^{\lfloor |\mathcal{C}|/2 \rfloor |P|},
\end{aligned}$$

where we substitute in the worst-case probabilities with lower (upper) bounds $p^{\mathrm{LB}}, p^{\mathrm{UB}}$. Finally, identifiability is achieved in the limit using the assumption that bounds p_i away from zero and 1 for all i (Assumption 4.6).

□

Algorithm 1: TOPIC

Input: Data $X \in \mathbb{R}^{n \times p \times m}$, edge score function $\hat{g}$, edge significance test $\hat{t}$
Output: Causal DAG G
$G \leftarrow \varnothing$
$T(i) \leftarrow \infty$ for each $i \in V$
for $k = 1, \ldots, p$ **do**
  worst $\delta = \infty$
  for $i \in V, T(i) \geq k$ **do**
    $\delta_i =$ BESTEDGEGAINTO(i, G);
  node at $k \leftarrow \arg\min_i \delta_i$ ▷ *Step 1*
  T(node at k) $\leftarrow k$
  ADDEDGESFROM(node at k, G) ▷ *Step 2*
  PRUNEEDGESTO(node at k, G) ▷ *Step 3*
return G

C Implementation Details

We outline the iterative algorithm TOPIC that identifies the causal graph in a topological order. It uses an information-theoretic scoring criterion $\hat{g}((X_i, X_j); G)$ scoring a directed edge $X_i \to X_j$ under the current graph G, as well as a significance test $\hat{t}((X_i, X_j); G)$.

TOPIC has the following steps. We initialize the inferred graph G as a causal graph without any edges. In addition, we maintain a set of nodes $A = \{1, \ldots, p\}$, from which we remove nodes in topological order. Now, we perform p iterations until all nodes are resolved.

1. Select that node i for which the best attainable score gain is smallest, in the sense that
$$\arg\max_{i \in A} \left(\min_{j \in A, j \neq i} \hat{g}((X_i, X_j); G) - \hat{g}((X_j, X_i); G) \right) .$$
2. For all *outgoing* edges $(X_i, X_j) \notin G$ with $j \in A$, add the edge to G if the gain is significant, using the test $\hat{t}((X_i, X_j); G)$.
3. For all *incoming* edges $(X_h, X_i) \in G$, prune the edge if removal results in significant score improvement. Remove i from A.

Complexity We now examine the complexity of TOPIC with regard to the number of variables p and the number of samples n. We perform a total of p iterations. In the k-th iteration, we evaluate the gain of $\frac{(d-k)^2+d}{2}$ possible edges. To this end, we need to fit a model with at most d variables regressing onto Y. The complexity of model fitting depends on the model class, which we shall denote as $T(n)$. Solving cubic splines for each variable X_i in the additive model has $O(n)$ complexity, where we have at most d variables. Hence, the total complexity of TOPIC with cubic splines is $O(d^4 n)$. In practice, for sparse graphs with $\text{pa}_X^* << p$, the complexity is $O(d^3 n)$.

D Time Series Implementation

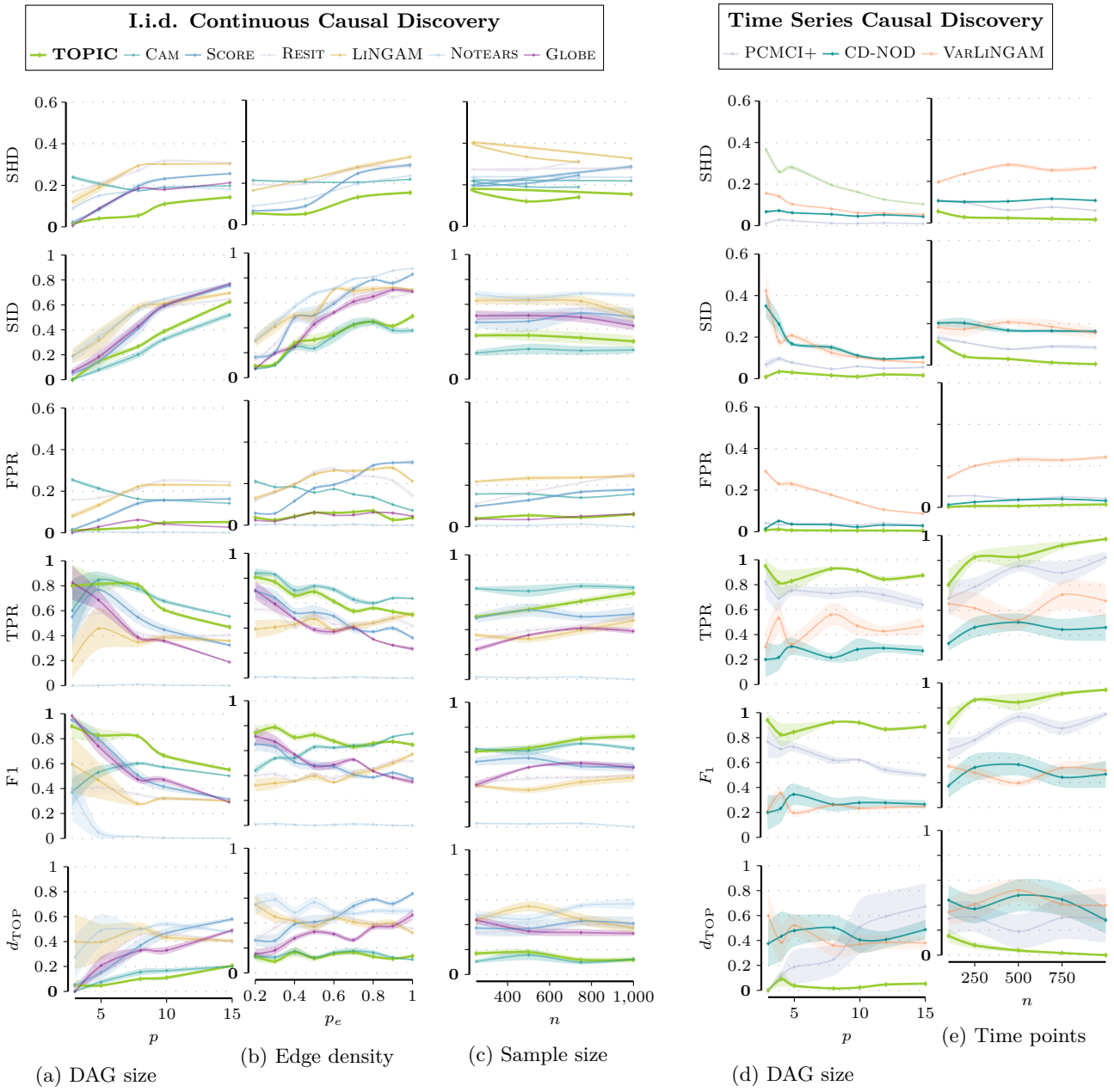
The above score can be straightforwardly applied to observational time series, where X is a real-valued process of length n observed over time, and X_t denotes an observation at time t. The necessary modifications include replacing the causal parents in the SCM by its timed variant

$$Y_t = \sum_{X_t \in \text{pa}_Y^*} f_{X_t, Y_t}(X_t) + N_{Y_t} ,$$

where pa_Y^* can include both instantaneous and lagged relationships. That is, G includes edges $((t, i), (t', j))$ with either $i \neq j$ or $t \neq t'$, where we assume a given maximum time lag $t' - t \leq \tau$.

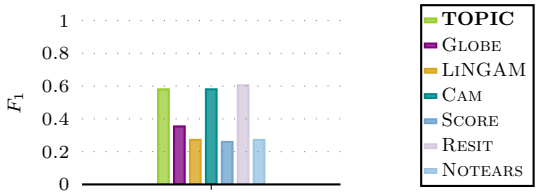
E Supporting Experiments

Extended Results Below, we extend Fig. 3 to include the full evaluation metrics for each method.



(a) DAG size (b) Edge density (c) Sample size (d) DAG size (e) Time points

Bivariate Causal Discovery To see whether TOPIC can effectively discover causal directions in the bivariate case, we also run the methods on the Tübingen Cause-Effect pairs [Mooij et al., 2016], a real-world benchmark comprising 108 cause-effect pairs with known ground truth. As shown below, TOPIC performs on par with the best competitors SCORE and RESIT.


1
0.8
0.6
0.4
0.2
0
F_1
TOPIC
Globe
LiNGAM
Cam
Score
Resit
Notears