Self-Compatibility: Evaluating Causal Discovery without Ground Truth

Philipp M. Faller[1,2] **Leena Chennuru Vankadara**[2] **Atalanti A. Mastakouri**[2]

Francesco Locatello[3] **Dominik Janzing**[2]

[1]Karlsruhe Institute of Technology, Germany
[2]Amazon Research Tübingen, Germany
[3]Institute of Science and Technology Austria

Abstract

As causal ground truth is incredibly rare, causal discovery algorithms are commonly only evaluated on simulated data. This is concerning, given that simulations reflect preconceptions about generating processes regarding noise distributions, model classes, and more. In this work, we propose a novel method for *falsifying* the output of a causal discovery algorithm in the *absence of ground truth*. Our key insight is that while statistical learning seeks stability across *subsets of data points*, causal learning should seek stability across *subsets of variables*. Motivated by this insight, our method relies on a notion of compatibility between causal graphs learned on different *subsets of variables*. We prove that detecting incompatibilities can falsify wrongly inferred causal relations due to violation of assumptions or errors from finite sample effects. Although passing such compatibility tests is only a necessary criterion for good performance, we argue that it provides strong evidence for the causal models whenever compatibility entails strong implications for the joint distribution. We also demonstrate experimentally that detection of incompatibilities can aid in causal model selection.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 INTRODUCTION

Causal relationships are often formalized as directed acyclic graphs (DAGs) (Pearl, 2009), or more general graphical models which also account for hidden confounders (Spirtes et al., 2000) and cycles (Bongers et al., 2021). Discovering causal relations is an important problem in science, which has led to the development of a diverse range of methods to infer causal graphs from passive observations. These methods are based on various approaches, such as Bayesian priors (Heckerman, 1995), independence testing (Spirtes et al., 2000; Lam et al., 2022), additive noise assumptions (Shimizu et al., 2006; Hoyer et al., 2008a), generalizations thereof (Zhang and Hyvärinen, 2009; Strobl et al., 2016), or various implementations of the so-called Independence of Mechanism assumption (Daniusis et al., 2010; Marx and Vreeken, 2017).

Causal discovery algorithms are typically evaluated primarily on simulated data. This is because causal ground truth is incredibly rare as it often necessitates real-world experiments. These experiments can be not only expensive and potentially unethical, but also frequently infeasible or even ill-defined from the outset (Spirtes and Scheines, 2004). Despite promising performance on simulated data, and in spite of numerous results on the identifiability of causal DAGs or parts thereof from passive observations, skepticism about the applicability of these algorithms on real data is warranted. This is because these algorithms are built on assumptions such as faithfulness, additive noise, post-nonlinear models, or independence principles, all of which can be violated in practice. Indeed, recent studies on causal discovery methods reveal a disconcerting reality about their applicability to real-world datasets (Huang et al., 2021; Reisach et al., 2021). Accordingly, the value of existing algorithms for down-

stream causal inference tasks is unclear and debated (Imbens, 2020).

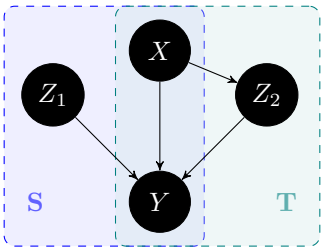
In this paper we propose a novel methodology for the falsification of the outputs of causal discovery algorithms on real data *without access to causal ground truth*. The key idea involves testing the *compatibility* of the causal models inferred by the algorithm when applied to different subsets of variables. In essence, while statistical learning aims for stability across different subsets of data points, we argue that causal discovery should aim to achieve stability across different subsets of variables. We prove that checking for incompatibilities provides a means of falsifying the outputs of causal discovery, as these incompatibilities indicate either violated causal discovery assumptions or finite sample effects that lead to non-negligible changes in the algorithm's output.

It is natural to ask if one can trust an algorithm if it satisfies such compatibility constraints. To address this question, we align with the theory of science by Popper (1959), according to which a hypothesis gathers evidence when numerous attempts to falsify it fail. We argue that, under a sufficiently strong notion of compatibility, the algorithm's outputs on various subsets of variables can entail strong implications for the joint distribution, thereby offering numerous opportunities for falsification.

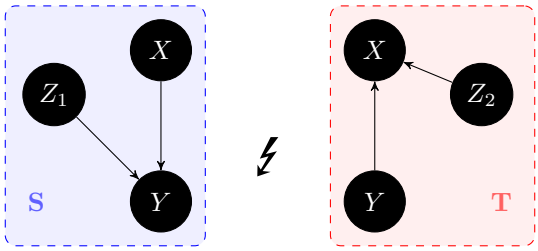
Our contributions. In this work we present a novel framework to evaluate causal graphs in the absence of ground truth. Specifically,

- we introduce two different notions of compatibility: interventional and graphical (Section 2.1). Under these definitions, we prove that if the assumptions of a causal discovery algorithm are met, its outputs are compatible in the population limit. Furthermore, we show for existing algorithms that they admit falsification using this approach (Section 2.2).
- We connect compatibility to the causal marginal problem and argue that compatibility can entail strong implications for potential joint distributions of given marginals, which can then be falsified statistically (Section 2.3).
- We introduce the incompatibility score for causal discovery which quantifies the *level of incompatibility* of the outputs of causal discovery. We argue based on stability arguments from learning-theory that the incompatibility score could serve as a proxy for measures like structural Hamming distance (SHD) which require access to ground truth (Section 3).
- We demonstrate that our score can potentially be used for model selection in simulation studies and on real-world data where ground truth knowledge is available. Our results show a significant correlation between the score and SHD (Section 4).

1.1 Motivational Example



(a) Compatible DAGs over S and T that admit a joint DAG over $S \cup T$.



(b) Incompatible DAGs over S and T

Figure 1: Each marginal causal models over S and T graphical implies a constraint for the edge $X - Y$ as it can only be directed in one way.

To illustrate our key ideas, we describe a simple example where a causal discovery algorithm results in causal graphs on different subsets of variables which are incompatible in a sense that we will explain now. Consider the directed acyclic graph (DAG) in figure 1a. Assume we are given the two subsets $S = \{X, Y, Z_1\}$ and $T = \{X, Y, Z_2\}$. We will call a causal model that represents only a subset of all relevant variables a *marginal causal model* in analogy to the marginal distribution. Assume the PC algorithm[1] (Spirtes et al., 2000) is applied to samples from the marginal distributions over S and T. Suppose the observed distribution violates faithfulness by satisfying $Y \perp\!\!\!\perp Z_2$ but otherwise only (conditional) independences hold if they are required by the Markov condition. PC then outputs the two DAGs in figure 1b, which entail the interventional statements: $p(y|do(x)) = p(y|x)$ (for S) versus $p(y|do(x)) = p(y)$ (for T). This is a contradiction unless $Y \perp\!\!\!\perp X$, which would be an additional violation of faithfulness in contrast to our assumption. Hence,

[1]For simplicity, we use the popular PC algorithm in this example and therefore implicitly assume that S and T are causally sufficient (Pearl, 2009) sets of variables. A similar construction can be made for the FCI algorithm (Spirtes et al., 2000), which does not assume sufficiency (see section A8 in the appendix).

the outputs of PC cannot both be correct given the observed joint distribution. Perhaps more strikingly, the marginal models in figure 1b have different orientations for the edge between X and Y. Accordingly, section 2.1 defines an *interventional* and a *graphical* notion of compatibility. Note, that an ordinary statistical cross-validation would not have discovered this inconsistency since we assumed $Y \perp\!\!\!\perp Z_2$ to hold in the population case, rather than assuming that a statistical test accepted independence due to a type II error.

2 COMPATIBILITY OF CAUSAL GRAPHS

Notation. Throughout this paper we will use $[n] := \{1, \dots, n\}$ for $n \in \mathbb{N}$, and for $i \in [n]$ we denote random variables with X_i and values of a variable with the respective lower case letter x_i. By slight abuse of notation we denote a vector of random variables in a set S by X_S and a vector of values of these variables with x_S. We denote the matrix containing $m \in \mathbb{N}$ values of all variables in a set S with $\mathbf{X}_S^m$. We sometimes omit m. $\mathbf{X}$ denotes the matrix containing vectors of values for all variables X_i for $i \in [n]$. For a probability distribution P over a set of variables V and $S \subset V$ we denote with P_S the marginal distribution over S and p be the probability density function of P where we assume for simplicity that there always exists a density with respect to a product measure.

Causal models. Although our proposed compatibility-based evaluation is in principle not restricted to any kind of causal model, we will focus our exposition on graphical models. Precisely, we will use acyclic directed mixed graphs (ADMGs) (Richardson, 2003; Evans and Richardson, 2014), partial ancestral graphs (PAGs) (Zhang, 2008) and completed partially directed acyclic graphs (CPDAGs) (Spirtes et al., 2000). In the main paper we focus on ADMGs, which include DAGs as special cases. Formal definitions of the remaining graphical models and of the causal semantics of graphical models can be found in section A7. Sections with the prefix A are in the appendix.

Definition 1 (ADMG) A *mixed graph* G consists of a finite set of nodes V and a set of directed edges $E \subseteq V \times V$ as well as a set of bidirected edges $B \subseteq \{\{X_i, X_j\} : X_i, X_j \in V\}$. If $(X_i, X_j) \in E$ we say there is a *directed edge* between X_i and X_j and we write $X_i \to X_j$. If $\{X_i, X_j\} \in B$ we say there is a *bidirected edge* and write $X_i \leftrightarrow X_j$. Directed and bidirected edges can occur together. A sequence of $l \in \mathbb{N}$ nodes $X_{i_1}, \dots, X_{i_l} \in V^l$ with any edge between X_{i_j} and $X_{i_{j+1}}$ for $j \in [l]$ is called an *undirected path* and it is called a *directed path* if all edges are in E and point towards the same direction. A mixed graph is called *acyclic directed mixed graph* (ADMG) if there is no directed path from X_i to itself for all $X_i \in V$. An ADMG with no bidirected edges is called *directed acyclic graph* (DAG).

Definition 2 (causal discovery algorithm) For the purpose of this paper, a *causal discovery algorithm* $\mathcal{A}$ takes i.i.d. data as input, i.e. a matrix $\mathbf{X}_S$ with $S \subseteq V$ containing samples from P_S, and outputs a ADMG, CPDAG or PAG over nodes in S, denoted by G_S or a special symbol $\perp$ to indicate that the algorithm itself detected a violation of assumptions.[2]

Throughout our work, we assume that all data has been generated by a single causal model.

Assumption 1 (existence of joint model) *Whenever we consider $k \in \mathbb{N}$ sets of variables S_i with $i \in [k]$ and distributions P_{S_i} we assume there is set V and a DAG $G = (V, E)$ such that there is a distribution P_V where G is the causal graph[3] of P_V and for all $i \in [k]$ we have $S_i \subseteq V$ and P_V has P_{S_i} as marginal distributions over S_i.*

2.1 Notions of Compatibility

We now introduce the concept of *compatibility* between causal graphs. We will discuss later how we can use compatibility of the outputs of a causal discovery algorithm (i.e. the *self-compatibility*) to falsify these outputs.

Definition 3 (compatibility notion) Let $(\mathcal{G}_V \cup \{\perp\})^*$ be the space of tuples[4] of the special token $\perp$ or graphs of some type (DAGs, CPDAGs, ADMGs, MAGs, PAGs) over subsets of a set V. Let $\mathcal{P}_V$ denote the space of probability distributions over V. A *compatibility notion* is a function

$$c : (\mathcal{G}_V \cup \{\perp\})^* \times \mathcal{P}_V \to \{0, 1\}.$$

For $k \in \mathbb{N}$ and $S_1, \dots, S_k \subseteq V$, the graphs $G_{S_1}, \dots, G_{S_k}$ are called compatible with respect to c and P_V if $c(G_{S_1}, \dots, G_{S_k}, P_V) = 1$.

In this work we will discuss two compatibility notions. We define an *interventional* compatibility notion, as

[2] In fact there are few algorithms that can output such a token. E.g. Ramsey et al. (2012) proposed an algorithm that can detect some violated assumptions itself. We will use the token in the proof of theorem 8.

[3] Note that the requirement that this data generation is formalized as a DAG is not a hard restriction, as CPDAGs, ADMGs, MAGs and PAGs naturally correspond to DAGs (when some variables of the DAG are unobserved).

[4] I.e. for any set A define $A^1 = A, A^{i+1} = A^i \times A$ for $i \in \mathbb{N}$ and $A^* = \cup_{i \in \mathbb{N}} A^i$.

we consider it a natural requirement of a causal discovery algorithm to make contradiction-free interventional statements. We also define a *graphical* notion of compatibility that has the advantage that it does not involve statistical decisions (as it does not directly depend on the distribution), but it raises conceptual problems due to implicit genericity assumptions that we discuss in section A9.2. We want to emphasize that other notions of compatibility (both, interventional and graphical) are conceivable and might be more appropriate in some situations.

Definition 4 (interventional compatibility) Let $S_1, \dots S_k$ be sets of variables for some $k \in \mathbb{N}$ and denote $S := \bigcup_{i\in[k]} S_i$. Let $G_{S_1}, \dots, G_{S_k}$ be causal models (DAGs, CPDAGs, ADMGs, MAGs, PAGs) and P_S be a probability distribution over S. Define the *interventional compatibility* via $c(G_{S_1}, \dots, G_{S_k}, P_S) = 1$ iff there exists a superset of nodes $V \supseteq S$, a DAG G_V over the variables in V and a distribution P_V such that

1. P_V has P_S as marginal over S,
2. P_V is Markovian w.r.t. to G_V and
3. for any $X_i, X_j \in S_l$ with $l \in [k]$ and any identification formula[5] in G_{S_l} holds: if the intervention $p(x_i|do(x_j))$ is identifiable in G_{S_l}, then the interventional probabilities coincide with the interventional probabilities $p^{G_V}(x_i|do(x_j))$ derived from G_V.

Further set $c(\dots, \perp, \dots) = 0$ regardless of the other arguments.

Interventional compatibility requires that the different causal graphs entail interventional probabilities that could come from a single joint causal model. Especially, this compatibility is with respect to a specific distribution. The second notion of compatibility we discuss in this work is the notion of graphical compatibility. For simplicity, we will only define graphical compatibility for ADMGs in the main paper, but in section A9 we will propose detailed definitions for the other graphical models mentioned in this work.

To define graphical compatibility, we first discuss what Pearl and Verma (1995) called the *latent projection* of a graphical model. Precisely, we will define the latent projection of an ADMG like Richardson et al. (2023).

[5]As we will describe in more detail in section A7, an interventional probability can often be expressed by different observational terms. These symbolic terms are a priori only related to the graph, but map a distribution to a probability. E.g. in figure 1a $p(y|do(x))$ can be identified by the backdoor formula either using $\emptyset$ and therefore $p \mapsto p(y|x)$ or with adjustment set $\{Z_1\}$ we get $p \mapsto \int p(y|x, z_1)p(z_1)\, dz_1$.

Definition 5 (latent ADMG) Let G be an ADMG with variables V and $S \subset V$. The *latent ADMG* $L(G, S)$ is the ADMG that contains all nodes in S, all edges between nodes in S and additionally

1. a directed edge between $X, Y \in S$ if there is a directed path from X to Y where all intermediate nodes are in $V \setminus S$
2. a bidirected edge between X, Y if there is a (undirected) path such that every non-endpoint is a non-collider in $V \setminus S$ and there are arrowheads towards X and Y on the incident edges on the path.

Definition 6 (graphical compatibility) Let $G_{S_1}, \dots, G_{S_k}$ be ADMGs over the $k \in \mathbb{N}$ sets of nodes $S_1, \dots, S_k$ respectively. Then we define *graphical compatibility* via $c(G_{S_1}, \dots, G_{S_k}) = 1$ iff there exists a set $V \supseteq \cup_{i=1}^k S_i$, and a DAG G_V such that $L(G_V, S_i) = G_{S_i}$ for all $i \in [k]$. Again, $c(\dots, \perp, \dots) = 0$ regardless of the other arguments.

Note, that neither interventional compatibility implies graphical compatibility nor vice versa. E.g. the empty graph is graphically compatible with its subgraphs, as graphical compatibility only depends on the distribution via the algorithm. But they are only interventionally compatible if the nodes have no causal effect onto each other in one Markovian joint graph. On the contrary, in section A10 we present an example where a non-generic distribution leads to interventionally compatible results that are not graphically compatible.

2.2 Falsifiability via Compatibility

In this section we demonstrate how compatibility between the outputs of an algorithm on different variables can be used to detect that either the assumptions of the algorithm are violated or there are finite sample effects in a way that actually change the output of an algorithm. We will start by defining the terms *observational falsifiability* and *self-compatibility*.

Definition 7 (observational falsifiability) The output of an algorithm $\mathcal{A}$ is *observationally falsifiable* with respect to a compatibility notion c if there exists a set of variables V, a distribution P_V and $k \geq 2$ subsets $S_1, \dots, S_k \subseteq V$ such that for all $\epsilon > 0$ there is an $m \in \mathbb{N}$ such that for all $m' \geq m$ we have

$$c(\mathcal{A}(\mathbf{X}_{S_1}^{m'}), \dots, \mathcal{A}(\mathbf{X}_{S_k}^{m'}), P_V) = 0 \tag{1}$$

with probability at least $1 - \epsilon$, where $\mathbf{X}_V^{m'}$ (and all submatrices) is drawn from[6] P_V. We call the left hand

[6]Note, that we now used finite sample versions of the graphs $\mathcal{A}(\mathbf{X}_{S_i}^{m'})$ but still the population version of the distribution P_V as last argument in $c(\dots, P_V)$. With the former, we want to emphasize that the graphs can be subject

side of equation (1) the *self-compatibility*[7] of $\mathcal{A}$ w.r.t. c, $S_1, \dots, S_k$, P_V and $\mathbf{X}_V^{m'}$.

In words, falsifiability means that there exists a joint distribution for which $\mathcal{A}$'s outputs on the subsets is incompatible according to c. Note, that c does not necessarily have to depend on P_V as the last parameter, like e.g. in definition 6. To illustrate this definition, recall that in section 1.1 we have constructed a distribution such that the PC algorithm produces interventionally incompatible results on $\{Z_1, X, Y\}$ and $\{X, Y, Z_2\}$ in the limit of infinite data. Therefore the output of PC is observationally falsifiable.

Remark 1 One might wonder whether the existence of a *single* incompatible distribution in definition 7 might be a too weak condition. In example 1 we will see, that we can nonetheless find algorithms that are not falsifiable in this sense at all. We expect algorithms that are *in principle* falsifiable to also be practically falsifiable. Quantifying how many distributions admit falsification is a difficult problem and would drastically increase the scope of this paper.

The following result, while not surprising, ensures that there are no incompatibilities in the limit of infinite data if all assumptions are met. Accordingly, if G is the causal DAG for a distribution P_V over variables in V, we require that for any $S \subset V$ for which P_S *meets the assumptions* of $\mathcal{A}$, and all $\epsilon > 0$, there exists an $m \in \mathbb{N}$ such that $\mathcal{A}(X_S^{m'}) = L(G, S)$ with probability at least $1 - \epsilon$ for all $m' \geq m$.

Lemma 1 *Let $S_1 \dots, S_k$ be $k \in \mathbb{N}$ sets of variables and P_V be a probability distribution over $V \supseteq \bigcup_{i \in [k]} S_i$ such that all P_{S_i} with $i \in [k]$ fulfil the assumptions of $\mathcal{A}$. Then for every $\epsilon > 0$ there is an $m \in \mathbb{N}$ such that $\mathcal{A}(\mathbf{X}_{S_1}^m), \dots, \mathcal{A}(\mathbf{X}_{S_k}^m)$ are interventionally and graphically compatible (w.r.t. to P) w.p. at least $1 - \epsilon$.*

All proofs are in section A8. Most causal discovery algorithms come with theoretical guarantees that their output is correct under some assumptions. But still, the following theorem shows for two exemplary algorithms that they are falsifiable (see section A8 for other algorithms).

Theorem 1 *The FCI algorithm and the Repetitive Causal Discovery[8] (RCD) algorithm are falsifiable w.r.t. interventional compatibility.*

The proof basically consists of constructing two distributions that have the same marginal distribution over two variables but would lead the algorithms to different causal models, similarly as in figure 1.

Now we have established that the algorithms in theorem 1 indeed have falsifiable outputs. But when all of their assumptions are met we can only "accidentally" falsify their output because of finite-sample effects, as lemma 1 shows. To illustrate that falsifiability is a non-trivial property, consider the following example.

Example 1 (Entropy Ordering) Define an algorithm $\mathcal{A}$ that orders nodes according to a simple criterion (e.g. starting from variables with lowest entropy[9]) and outputs an ADMG containing the complete DAG with respect to that order for directed edges and additionally bidirected edges between all nodes. The outputs of this algorithm on any subset will be graphically compatible, as the marginal models are the latent projection of the models on supersets and also interventionally compatible, as no interventional distribution is identifiable in any of the graphs.

This raises the question if there are properties that already imply the falsifiability of an algorithm. Indeed, in section A11 we will show that all causal discovery algorithms that are not "too simple" or "indecisive" can produce incompatible outputs on different sets of variables and hence, are falsifiable.

Remark 2 Note that our approach does not falsify *particular* causal graphs. Especially, an algorithm $\mathcal{A}$ may well output the ground truth graph on $\mathbf{X}$ but we can still find incompatible models on some subset. Due to lemma 1 we know that incompatibility of the outputs indicates that some assumptions are violated or there are errors due to finite sample effects. In this sense we would argue, that we cannot *trust* the causal discovery algorithm in this case even if $\mathcal{A}(\mathbf{X})$ happens to be the ground truth graph.[10]

2.3 Is Self-Compatibility a Strong Condition?

We will now show that even though typically more than one joint distribution can have the same marginal distributions over some sets S and T, there are cases where the assumptions of an algorithm $\mathcal{A}$ render the joint distribution unique, as otherwise the outputs of $\mathcal{A}$ on S, T and $S \cup T$ would be incompatible. If the outputs of an algorithm on the marginals predict a unique

to finite sample effects. With the latter we want express that the statistical difficulties with testing the equivalence of the interventional distributions are beyond the scope of this work.

[7] Note that compatibility refers to graphical models while self-compatibility is a property of algorithms.

[8] RCD (Maeda and Shimizu, 2020) relies on linear models with non-Gaussian noise like LiNGAM (Shimizu et al., 2006) but does not assume causal sufficiency.

[9] Indeed, suggestions like this have been made, motivated by misconceptions on thermodynamics, as criticised by Janzing (2019).

[10] This situation bears similarity to the Gettier problem in epistemology. Gettier (1963) argues that a person does not *know* a fact A even if A is true but the person based her belief in A on false assumptions.

joint distribution, any other potential joint distribution falsifies the outputs. On the other hand, if the joint data points do not contradict this unique joint, we count this, in the spirit of Popper (1959), as strong evidence in favor.

Definition 8 (merging-enabling algorithms) An algorithm $\mathcal{A}$ is said to *enable merging* distributions with respect to a notion of compatibility c if for some set of variables V and $k \geq 2$ there exist sets $S_1, \ldots, S_k \subset V$ and distributions $P_{S_1}, \ldots, P_{S_k}$ such that

1. there is exactly one joint distribution P_V such that for any $\epsilon > 0$ there is an $m \in \mathbb{N}$ such that for all $m' \geq m$ we get $c(\mathcal{A}(\mathbf{X}_V^{m'}), \mathcal{A}(\mathbf{X}_{S_1}^{m'}), \ldots, \mathcal{A}(\mathbf{X}_{S_k}^{m'}), P_V) = 1$, with probability at least $1 - \epsilon$ and
2. there exists a second distribution $\tilde{P}_V$ whose marginals coincide with all P_{S_i} for $i \in [k]$,

where $\mathbf{X}_{S_j}^{m'}$ is drawn from P_{S_j} for $j \in [k]$.

Condition 1 states that there is only one joint distribution P_V for which $\mathcal{A}$'s output on V does not contradict $\mathcal{A}$'s outputs on the subsets. Condition 2 states that there would be more than one possible extension of the marginal probabilities without the graphical information provided by $\mathcal{A}$. In other words, the condition of compatibility of causal models implies constraints for the joint distributions that result in a unique solution of the *causal* marginal problem, although the solution to the *probabilistic* marginal problem (Vorob'ev, 1962) is not unique in this case.

Theorem 2 (FCI enables merging) *The FCI algorithm enables merging w.r.t. graphical compatibility on PAGs.*

In section A13 we provide a similar statement for an idealized version of RCD[11]. In the proof we construct an example where the algorithms output the existence of an *unconfounded* edge. This edge rules out the existence of a confounding path and therefore implies a conditional independence that is not observable in the marginals. In the example of the proof, this independence together with the given marginals suffices to identify the joint distribution.

Although these statements only ensure the *existence* of cases where merging is possible, we think they illustrate the strength of the self-compatibility condition. In addition, we demonstrate empirically that the condition is often violated in section 4.

[11]The idealized version can output $\perp$ if it detects that the given distribution has not been generated by a linear additive noise model. As we defined this token to be incompatible with all ADMGs, condition 1 of definition 8 allows us to rule out such distributions.

Relationship to interventions. Note, that there is a close relationship between enabling merging and being able to predict the impact of interventions: let the node X_i describes a coin flip, triggering the intervention $do(X_k = x_k)$ on another node X_k in a set S when $X_i = 1$. When $\mathcal{A}$ allows the unique reconstruction of $P_{\{X_i, X_j\}}$ after being applied to $S \cup \{X_i\}$ and $S \cup \{X_j\}$, it implicitly provides an interventional probability via $p(x_j|do(x_k)) = p(x_j|x_i = 1)$ *for an observer who knows that* X_i *controls the intervention on* X_k.

3 INCOMPATIBILITY SCORE

In this section we propose a practical score, that quantifies "how incompatible" the outputs of an algorithm applied to different subsets of variables are. This score can be seen as a continuous relaxation of the binary notion of compatibility in definition 3. We propose to use this relaxation, as we showed in theorem 2 that self-compatibility can be a strong criterion and in practice it is often violated. This score is defined such that a perfect score indicates self-compatibility and in this sense the score can be used to *falsify* the outputs of a causal discover algorithm as we described before. But moreover, our experimental results in section 4 suggest that the continuous score could be used to *evaluate* causal discovery algorithms in the sense that it could be used as a "proxy" for the structural Hamming distance, which cannot be evaluated without ground truth knowledge. The first score in this section is based on the interventional compatibility notion (definition 4). We also present an incompatibility score based on graphical compatibility (definition 6) and in section A14 we discuss further details of the scores.

Su and Henckel (2022) proposed a parametric test for whether the interventional distributions of different adjustment sets agree. We use this to test whether the parent-adjustment sets derived from the marginal models $\mathcal{A}(\mathbf{X}_{S_1}), \ldots, \mathcal{A}(\mathbf{X}_{S_k})$ yield the same causal effect as the joint one in $\mathcal{A}(\mathbf{X})$.

Definition 9 (interventional score) For $k \in \mathbb{N}$, let $S_1, \ldots, S_k \subseteq V$. We define the *interventional incompatibility score* κ^I of $\mathcal{A}$ via

$$\kappa^I(\mathcal{A}, \mathbf{X}) := C^{-1} \sum_{\substack{X,Y \in V \\ X \neq Y}} T\left(X, Y, \mathcal{G}(\mathcal{A}, \mathbf{X})\right) \qquad (2)$$

where $\mathcal{G}(\mathcal{A}, \mathbf{X}) := \{\mathcal{A}(\mathbf{X}), \mathcal{A}(\mathbf{X}_{S_1}), \ldots, \mathcal{A}(\mathbf{X}_{S_k})\}$ and we define $T(X, Y, \mathcal{G}(\mathcal{A}, \mathbf{X})) = 1$ if

1. there are (at least two) different valid parent-adjustment sets for X and Y in $\mathcal{G}(\mathcal{A}, \mathbf{X})$ and the test from Su and Henckel (2022) rejects the hypothesis that they all entail the same causal effect, or

2. there is an $i \in [k]$ such that the parent-adjustment for X and Y is valid in $L(\mathcal{A}(\mathbf{X}), S_i)$ but not in $\mathcal{A}(\mathbf{X}_{S_i})$ or vice versa,

else $T(X, Y, \mathcal{G}(\mathcal{A}, \mathbf{X})) = 0$ and C is the number of pairs X, Y such that there is at least one graph in $\mathcal{G}(\mathcal{A})$ with a valid parent-adjustment set for X and Y (except for the cases where this graph is $\mathcal{A}(\mathbf{X})$ and the effect is not identifiable in any $L(\mathcal{A}(\mathbf{X}), S_i)$ for $i \in [k]$).

This definition counts the cases where the algorithm makes incompatible statements about any pairwise causal effect on different subsets, normalized by the number of cases where the algorithm does commit to any falsifiable statement on any subset at all.

The score defined in definition 9 is only applicable to linear models, as we build on the test of Su and Henckel (2022). As non-linear models are ubiquitous in practice, we also want to propose a score that can be applied in these settings. Graphical compatibility notions do not require any statistical test, thus we chose to build on them for this purpose. Definition 6 always refers to the existence of a joint model. Therefore we found it natural to check the compatibility of each marginal causal graph with the causal graph $\mathcal{A}(\mathbf{X})$ that an algorithm outputs on all available variables (even though, if we do not assume that the observed variables are causally sufficient, $\mathcal{A}$ can at most output a latent projection of the joint graph in assumption 1). Further, definition 6 requires that the latent projection and the marginal graph are identical. We relax this notion by taking the SHD between the joint model $\mathcal{A}(\mathbf{X})$ and a marginal model. The SHD is zero iff the graphs are identical.

Definition 10 (graphical incompatibility score) Let $k \in \mathbb{N}$ and $S_1, \dots, S_k \subseteq V$. We define the *graphical incompatibility score* κ^G of $\mathcal{A}$ via

$$\kappa^G(\mathcal{A}, \mathbf{X}) := \frac{1}{k} \sum_{i \in [k]} \mathrm{SHD}\left(L(\mathcal{A}(\mathbf{X}), S_i), \mathcal{A}(\mathbf{X}_{S_i})\right). \quad (3)$$

Relationship to stability. While, as previously discussed, it is not possible to guarantee that a causal discovery algorithm that achieves a low incompatibility score will accurately predict system behavior under interventions, we argue that the resulting models are at least *useful* due to their ability to predict statistical properties of *unobserved joint distributions*. This perspective is influenced by Janzing et al. (2023), who reconceptualizes causal discovery as a *statistical learning problem*. The key principle underlying this reconceptualization posits that causal models offer predictive value beyond predicting system behavior under interventions; they can also predict statistical properties of unobserved joint distributions. See section A12 for the formal setup and description of this idea and the associated learning problem.

Observe that, for a causal discovery method to achieve a low incompatibility score, it's output must remain largely unchanged under small modifications to the variable sets it is applied to. Following this idea, in section A12 we define a notion of stability of a causal discovery algorithm. Under this definition, we provide high-probability generalization bounds for causal models generated by stable causal discovery methods. The result demonstrates that stable algorithms, provably, generate useful causal models due to their ability to *generalize statistical predictions across variable sets*. Informally, it provides evidence that a low incompatibility constitutes a useful inductive bias for causal discovery. This is notably distinct from the standard setting in statistical learning, where algorithms that exhibit stability under small modifications to the data are known to *generalize across data points*.

4 EXPERIMENTS

We now explore the efficacy of the methods described in section 3 on real and simulated data. The details of the experiments can be found in section A15 and the source code under `https://github.com/amazon-science/causal-self-compatibility`. In the main paper we present results for the RCD algorithm, as it is one of the few available algorithms that does not assume causal sufficiency and its outputs are close[12] to the presented formalism based on ADMGs. Additional experiments with other algorithms are shown in section A16.

Model evaluation. As a first experiment we focus on a setting where we would expect a causal discovery algorithm to work reasonably well. We therefore generate 100 datasets with a linear model, uniform noise and potentially hidden variables. We use the incompatibility score κ^I from definition 9. The first insight from the plot in figure 2 is that interventional compatibility indeed is a strong condition, in the sense that even in this scenario we find many interventionally incompatible marginal graphs. Further, the plot in figure 2 shows a significant correlation between κ^I and the structural Hamming distance of the joint graph to the true graph. We suspected the density of the ground truth graph to influence both, the incompatibility score as well as the SHD. We also present the partial correlation between SHD and κ^I, adjusted for the average node degree of the ground truth graph,

[12] Note however, that the output of RCD does not strictly describe ADMGs, as the the algorithm does not differentiate between purely confounded relationships and confounding with an additional direct edge.

which stands at 0.52 with a p-value of 3×10^{-8}. This suggests that κ^I might be a useful proxy for SHD, which we cannot calculate in the absence of ground truth.

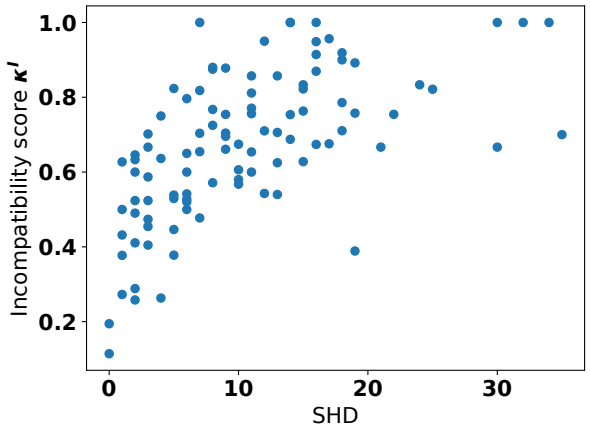


Figure 2: RCD on 100 datasets that fulfill its assumptions. The plot shows structural Hamming distance of estimated graphs $\hat{G}$ to the respective true graph G versus the interventional incompatibility score κ^I. As both are influenced by the degree of the true graph, we also calculated the partial correlation given the average node degree of the true graph, which is 0.52 with p-value $3 \cdot 10^{-8}$.

Model selection. As we have seen that the incompatibility score is correlated with the SHD of the joint model to the ground truth, we now want to investigate whether the score could potentially be used to guide model selection and parameter tuning. We used the interventional score to select hyperparameters for the RCD algorithm. RCD has three threshold values for different independence tests. For simplicity we only checked the two configurations, where all of them are set to either 0.1 or 0.001 respectively.[13]

In figure 3 we plot the difference in SHD between the estimated graphs and the ground truth graph, respectively, on the y-axis, where we always subtract the SHD of the algorithm with better κ^I from the SHD of the algorithm with worse κ^I. Analogously for κ^I on the x-axis. If the incompatibility score κ^I was a "perfect" selection criterion we would hope to see all points on or above the horizontal line. In fact, 68% are strictly above the line and 28% are below the line. Moreover, we can see, that in most cases where the incompatibility score picked the hyperparameters that produce a worse SHD, the scores of the hyperparameters were close[14] to each other.

Real data. Finally, we also used the score on the biological dataset presented by Sachs et al. (2005). We noted that all causal discovery algorithms that we tried performed quite poorly on this datasets compared to the algorithm's performance on simulated data. Our incompatibility score reflects this in the sense that in two out of four cases we get medium to bad incompatibility scores compared to the simulated experiments. The cases where the incompatibility score was good were the ones with the best results in terms of SHD and F_1 score. More details can be found in section A16.4.

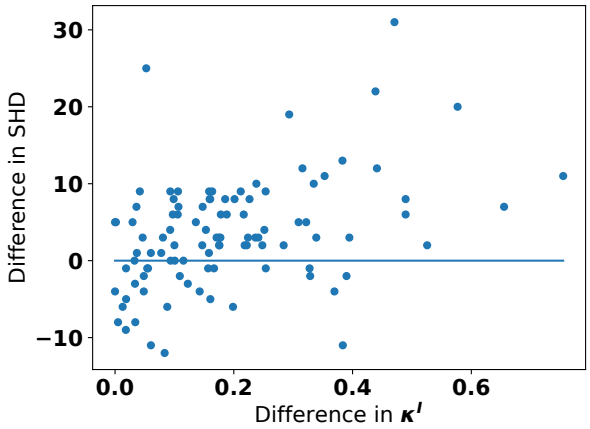


Figure 3: We chose between the hyperparameters $\alpha = 0.1$ and $\alpha = 0.001$ of RCD according to the incompatibility κ^I for 100 datasets. For 72% of datasets we picked the better model or an equally good model. In most cases where we picked the worse model in terms of SHD the difference in κ^I is small.

5 RELATED WORK

To the best of our knowledge we are the first to leverage compatibility constraints of marginal causal models to falsify the output of causal discovery algorithms.

Robustness of causal effects. While we proposed a method to falsify the underlying assumptions of causal discovery algorithms, there exists several methods for scenarios where causal directions are given and the goal is to estimate the strength of the treatment effect. E.g. Walter and Tiemeier (2009); Lu and White

[13]Indeed this is overly simplistic, as in actual applications one would pick the parameters from a grid.

[14]This, of course, raises the question which differences should be considered *significant*. The answer may depend on the particular downstream task—just as e.g. for SHD itself. Further, bootstrapping or permutation methods like the one proposed by Eulig et al. (2023) might be helpful to derive a meaningful baseline. We defer this to future work.

(2014); Oster (2019) propose to test whether a regression model is causal by dropping parts of the potential covariates and testing robustness. Similarly, Su and Henckel (2022) present the aforementioned parametric test for the case where a hypothetical graph is given.

Evaluating causal models. The gold standard of evaluating the quality of a causal model would be to conduct a randomized control trial. In contrast, our method allows for falsification in settings where experiments are infeasible. There are other methods to judge the quality of causal models that do not rely on ground truth data as well, but they are limited to special cases, such as falsification via Verma-constraints (Verma and Pearl, 2022) for instrument variables, tests that require parametric assumptions (Bollen and Ting, 1993; Daley et al., 2022) or the derivation of Bayesian uncertainty estimates (Claassen and Heskes, 2012). A common approach is to count the number of d-separation statements that are not reflected in the data (Textor et al., 2017; Reynolds et al., 2022), although this either requires ground truth again or is also subject to assumptions such as faithfulness. Eulig et al. (2023) propose to reject causal graphs that do not reflect the conditional independences in the data significantly better than a random baseline.

Causal marginal problem. Tsamardinos et al. (2012) use causal models on sets of variables S, T to predict conditional (in)dependences in $P_{S\cup T}$. For the toy scenario of a collider structure with three binaries, Gresele et al. (2022) studied compatibility of structural equations for the bivariate marginals, which amounts to falsifiability of causal statements of rung 3 in Pearl's ladder of causation (Pearl and Mackenzie, 2018). In a scenario where causal directions are known, Guo et al. (2023) define *out-of-variable generalization* as the capability of a machine learning algorithm to perform well across environments with different causal features, and use marginal observations to predict joint distributions. Janzing et al. (2023) infers a DAG G on $S := \bigcup_j S_j$ in order to predict properties of P_S from the set of all P_{S_j}, which admits falsifying the DAG without interventions via "test marginal distributions" $P_{S_{j+1}}$. Despite the connections, the key difference of these approaches to our method is, that we do not use marginal distributions to reconstruct an unobservable joint distribution. Instead, we propose to learn a causal model on the joint distribution but to falsify the output of the algorithm by learning marginal models.

6 CONCLUSION

We have proposed a method to falsify the output of causal discovery algorithms which does not rely on causal ground truth. It is based on compatibility of the output of an algorithm on different sets of variables. While this compatibility seems, at first glance, just as a weak sanity check, i.e. a weak necessary condition for providing good causal models, we have argued that there are cases where the compatibility requirement results in strong predictions for the joint distribution of variables which can be falsified from passive observations (and thus provide strong evidence in favor of causal models if all attempts of falsification fail).

This approach is limited as we can only *falsify* the outputs of causal discovery. Even though we argue in section 4 that our incompatibility score could be used as a proxy for SHD, we have no hard theoretical guarantees that ensure good performance for good scores (and we do not think that such guarantees can be proven without further assumptions). Further, our work provides no guidance as to which degree of self-compatibility is "good enough" or when the outputs of an algorithm should definitely be rejected.

Acknowledgements

Part of this work was done while Philipp M. Faller was an intern at Amazon Research Tübingen. Philipp M. Faller was supported by a doctoral scholarship of the Konrad Adenauer Foundation and the Studienstiftung des deutschen Volkes (German Academic Scholarship Foundation).

References

K. A. Bollen and K.-f. Ting. Confirmatory tetrad analysis. *Sociological methodology*, pages 147–175, 1993.

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

O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2: 499–526, 2002.

O. Bousquet, Y. Klochkov, and N. Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626. PMLR, 2020.

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

T. Claassen and T. Heskes. A bayesian approach to constraint based causal inference. *arXiv preprint arXiv:1210.4866*, 2012.

P. J. Daley, K. J. Resch, and R. W. Spekkens. Experimentally adjudicating between different causal accounts of bell-inequality violations via statistical model selection. *Physical Review A*, 105(4):042220, 2022.

P. Daniusis, D. Janzing, J. M. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *Proceedings of the 26th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 143–150. AUAI Press, 2010.

G. Darmois. Analyse générale des liaisons stochastiques. *Rev. Inst. Internationale Statist.*, 21:2–8, 1953.

P. Dawid. Decision-theoretic foundations for statistical causality. *Journal of Causal Inference*, 9(1):39–77, 2021. doi: doi:10.1515/jci-2020-0008. URL `https://doi.org/10.1515/jci-2020-0008`.

E. Eulig, A. A. Mastakouri, P. Blöbaum, M. Hardt, and D. Janzing. Toward falsifying causal graphs using a permutation-based test, 2023.

R. J. Evans and T. S. Richardson. Markovian acyclic directed mixed graphs for discrete data. 2014.

E. Gettier. Is knowledge justified true belief? *Analysis*, 23(6):121–123, 1963.

L. Gresele, J. Von Kügelgen, J. Kübler, E. Kirschbaum, B. Schölkopf, and D. Janzing. Causal inference through the structural causal marginal problem. In *International Conference on Machine Learning*, pages 7793–7824. PMLR, 2022.

S. Guo, J. Wildberger, and B. Schölkopf. Out-of-variable generalization. *arXiv preprint arXiv:2304.07896*, 2023.

D. Heckerman. A Bayesian approach to learning causal networks. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 285–295, San Francisco, CA, 1995. Morgan Kaufmann Publishers.

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

P. O. Hoyer, S. Shimizu, A. J. Kerminen, and M. Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2): 362–378, 2008b.

Y. Huang, M. Kleindessner, A. Munishkin, D. Varshney, P. Guo, and J. Wang. Benchmarking of data-driven causality discovery approaches in the interactions of arctic sea ice and atmosphere. *Frontiers in big Data*, 4:642182, 2021.

G. W. Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–79, December 2020.

D. Janzing. The cause-effect problem: Motivation, ideas, and popular misconceptions. In I. Guyon, R. Statnikov, and B. Bakir Batu, editors, *Cause Effect Pairs in Machine Learning*, pages 3–26. Springer, 2019.

D. Janzing, P. M. Faller, and L. C. Vankadara. Reinterpreting causal discovery as the task of predicting unobserved joint statistics. *arXiv preprint arXiv:2305.06894*, 2023.

W.-Y. Lam, B. Andrews, and J. Ramsey. Greedy relaxations of the sparsest permutation algorithm. In J. Cussens and K. Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1052–1062. PMLR, 01–05 Aug 2022.

X. Lu and H. White. Robustness checks and robustness tests in applied economics. *Journal of Econometrics*, 178:194–206, 2014. Annals Issue: Misspecification Test Methods in Econometrics.

T. N. Maeda and S. Shimizu. Rcd: Repetitive causal discovery of linear non-gaussian acyclic models with latent confounders. In *International Conference on Artificial Intelligence and Statistics*, pages 735–745. PMLR, 2020.

A. Marx and J. Vreeken. Telling cause from effect using mdl-based local and global regression. In *2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017*, pages 307–316, 2017.

S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25:161–193, 2006.

E. Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019.

J. Pearl. *Causality*. Cambridge university press, 2009.

J. Pearl and J. Mackenzie. *The book of why*. Basic Books, USA, 2018.

J. Pearl and T. S. Verma. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 789–811. Elsevier, 1995.

E. Perković, J. Textor, M. Kalisch, and M. H. Maathuis. A complete generalized adjustment criterion. In *Proceedings of the Thirty-First Conference*

on Uncertainty in Artificial Intelligence, pages 682–691, 2015.

J. Peters and P. Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3):771–799, 2015.

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

K. Popper. *The logic of scientific discovery*. Routledge, London, 1959.

J. Ramsey, J. Zhang, and P. L. Spirtes. Adjacency-faithfulness and conservative causal inference. *arXiv preprint arXiv:1206.6843*, 2012.

A. Reisach, C. Seiler, and S. Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27772–27784. Curran Associates, Inc., 2021.

R. J. Reynolds, R. T. Scott, R. T. Turner, U. T. Iwaniec, M. L. Bouxsein, L. M. Sanders, and E. L. Antonsen. Validating causal diagrams of human health risks for spaceflight: An example using bone data from rodents. *Biomedicines*, 10(9):2187, 2022.

T. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.

T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1):334–361, 2023.

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

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvarinen, Y. Kawahara, T. Washio, P. O. Hoyer, K. Bollen, and P. Hoyer. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr):1225–1248, 2011.

V. Skitovic. Linear combinations of independent random variables and the normal distribution law. *Select. Transl. Math. Stat. Probab.*, (2):211–228, 1962.

P. Spirtes and R. Scheines. Causal inference of ambiguous manipulations. *Philosophy of Science*, 71 (5):833–845, 2004. doi: 10.1086/425058.

P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, prediction, and search*. MIT press, 2000.

E. V. Strobl, P. L. Spirtes, and S. Visweswaran. Estimating and controlling the false discovery rate for the pc algorithm using edge-specific p-values. *arXiv preprint arXiv:1607.03975*, 2016.

Z. Su and L. Henckel. A robustness test for estimating total effects with covariate adjustment. In *Uncertainty in Artificial Intelligence*, pages 1886–1895. PMLR, 2022.

J. Textor, B. van der Zander, M. S. Gilthorpe, M. Liśkiewicz, and G. T. Ellison. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology*, 45(6):1887–1894, 01 2017. ISSN 0300-5771. doi: 10.1093/ije/dyw341. URL `https://doi.org/10.1093/ije/dyw341`.

J. Tian and J. Pearl. A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573, 2002.

S. Triantafillou and I. Tsamardinos. Score-based vs constraint-based causal learning in the presence of confounders. In *Cfa@ uai*, pages 59–67, 2016.

I. Tsamardinos, S. Triantafillou, and V. Lagani. Towards integrative causal analysis of heterogeneous data sets and studies. *The Journal of Machine Learning Research*, 13(1):1097–1157, 2012.

T. S. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 221–236. 2022.

N. Vorob'ev. Consistent families of measures and their extensions. *Theory Probab. Appl*, 7(2):147–163, 1962.

S. Walter and H. Tiemeier. Variable selection: current practice in epidemiological studies. *European journal of epidemiology*, 24:733–736, 2009.

J. Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9:1437–1474, 2008.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, Montreal, Canada, 2009.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 804–813. AUAI Press, 2011.

Y. Zheng, B. Huang, W. Chen, J. Ramsey, M. Gong, R. Cai, S. Shimizu, P. Spirtes, and K. Zhang. Causal-learn: Causal discovery in python. *arXiv preprint arXiv:2307.16405*, 2023.

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.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes.
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes.
   (b) Complete proofs of all theoretical results. Yes.
   (c) Clear explanations of any assumptions. Yes.
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.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes.
   (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.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes.
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. Not applicable.
   (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.

Self-Compatibility: Evaluating Causal Discovery without Ground Truth Appendix

A7 FORMAL DEFINITIONS

With our formalism we mainly follow Peters et al. (2017) and Pearl (2009).

Structural causal models Causality can be formalized mathematically by saying that all relationships between variables are governed by some deterministic functions except for some genuine, independent sources of randomness. We define structural models like Pearl (2009).

Definition 11 (structural causal model) Let $n \in \mathbb{N}$ and V be a set of random variables $X_1, \dots, X_n$, U be the a set of variables $N_1, \dots, N_n$ and P_U be a probability distribution over $N_1, \dots, N_n$ such that all $N_1, \dots, N_n$ are jointly independent. Let there be a set of (measurable) functions F such that for all $i \in [n]$ we have

$$X_i := f_i(PA_i, N_i),$$

where PA_i is some subset of $V \setminus \{X_i\}$ such that there are no cyclic dependencies between variables and f_i depends on all variables in PA_i. Then we call (V, U, F, P_U) a *structural causal model* (SCM).

Due to the acyclicity, the distribution P_U also entails a unique joint distribution over V, as each value x_i can be solved recursively until it only depends on noise terms. Accordingly, we define interventions as Pearl (2009).

Definition 12 (intervention) Let $\mathcal{S} = (V, U, F, P_U)$ be an SCM. The *intervention* $do(X_i = x_i^*)$ for some $i \in [n]$ is defined by inserting the fixed value x_i^* in all equations in F that depend on X_i, regardless of other variables in the model. Denote with $\mathcal{S}_{do(X_i=x_i)}$ the modified model with these equations. The *interventional distribution* is defined as the distribution that canonically arises from $\mathcal{S}_{do(X_i=x_i)}$ and we denote it with $P(X_S = x_S \mid do(X_i = x_i))$ for any $S \subseteq V$.

Graphical models Some of the causal aspects of a SCM can be represented graphically. We will first define the graphical structures and then discuss their connection to the causal semantics defined via SCMs. There are several popular graphical models for causality. We mostly follow Zhang (2008) and Perković et al. (2015) with our formalism. First recall the definitions of DAGs, mixed graphs and ADMGs from definition 1. We say there is an *arrowhead towards* X_j if there is an edge $X_j \to X_i \in E$ or $X_i \leftarrow X_j \in B$ and a *tail* towards X_j if there is an edge $X_i \leftarrow X_j \in E$. We further say X_i and X_j are *adjacent* if $X_i \to X_j \in E, X_j \to X_i \in E$ or $X_i \leftrightarrow X_j \in B$.

Definition 13 (maximal ancestral graphs) An *undirected path* is a sequence of nodes $X_0, \dots, X_k$ with $k \in \mathbb{N}$ such that either $X_i \to X_{i+1} \in E$, $X_{i+1} \to X_i \in E$ or $X_i \leftrightarrow X_{i+1} \in B$ for $i \in [k-1]$. A node X_i is called *collider* on an undirected path p if there are two edges with a head towards X_i on p. We say there is an *almost directed cycle* if there is a directed cycle from X_i to X_j and there is a bidirected edge $X_i \leftrightarrow X_j$. Let $L \subset V$ and call the first and the last nodes on a path the *endpoints* of a path. An *inducing path* relative to L is an undirected path p, such that every node on p that is in $V \setminus L$ except for the endpoints is a collider on p and an ancestor of one of the endpoints. A mixed graph G is called a *maximal ancestral graph* (MAG) if it contains no almost directed cycles and there is no inducing path between non-adjacent nodes.

Definition 14 (m-separation) Let $G = (V, E, B)$ be a mixed graph. An undirected path p between X_i and X_j in G is called *m-connecting* given a set $Z \subset V \setminus \{X_i, X_j\}$ if every non-collider on p is not in Z and every collider on p is an ancestor of a node (or is itself) in Z. If no undirected path between X_i and X_j is m-connecting given Z we say X_i and X_j are *m-separated* by Z and write $X_i \perp_G X_j \mid Z$. In a DAG m-separation reduces to well-known d-separation.

Graphical models and causality We will now connect the graphical models with causal semantics.

Definition 15 (causal DAG) Let P be a probability distribution over variables in V that has been generated by an SCM S as described above. We call a DAG G a *causal graph* of P if G contains a node for each variable in V and an edge from X_i to $X_j \in V$ iff $X_i \in PA_j$ in S.

Definition 16 (Global Markov condition) We say a probability distribution P over V fulfils the *global Markov condition* w.r.t. the DAG (MAG) G if for every two nodes $X_i, X_j \in V$ and set $Z \subseteq V \setminus \{X_i, X_j\}$ we have that $X_i \perp_G X_j \mid Z$ implies $X_i \perp\!\!\!\perp X_j \mid Z$. We also say P is *Markovian* w.r.t. G. If $X_i \perp\!\!\!\perp X_j \mid Z$ also implies $X_i \perp_G X_j \mid Z$ we say P is *faithful* to G.

A distribution that has been generated by an SCM is always Markovian w.r.t. its causal DAG.

Definition 17 (Markov equivalence) Two DAGs (MAGs) G_1, G_2 are *Markov equivalent* if for all nodes $X_i, X_j \in V$ and sets of nodes $Z \subseteq V \setminus \{X_i, X_j\}$ we have $X_i \perp_{G_1} X_j \mid Z$ iff $X_i \perp_{G_2} X_j \mid Z$. We call

$$[G] := \{G' : G \text{ and } G' \text{ are Markov equivalent}\}$$

the *Markov equivalence class* of G.

Definition 18 (CPDAG) Let G be a DAG. The *completed partially directed graph* C of $[G]$ is a mixed graph that contains a directed edge $X_i \to X_j$ iff this edge exists in all DAGs in $[G]$ and a bidirected edge $X_i \leftrightarrow X_j$ iff there is a DAG in $[G]$ with the edge $X_i \to X_j$ and a DAG in $[G]$ with the edge $X_i \leftarrow X_j$. We call C a *causal CPDAG* if G is a causal DAG.

The partial ancestral graph that we now introduce can also represent the presence of selection bias. In this paper we omit this part of the formalism (similarly to Zhang (2008)).

Definition 19 (PAG) Let G be a MAG over variables in a set V. The *partial ancestral graph* of $[G]$ is a graph $H = (V, E)$ with a symmetric set of edges $E \subseteq V \times V$ and a map end : $E \to \{>, -, \circ\}$ such that for $X_i, X_j \in V$

- $(X_i, X_j) \in E$ and $(X_j, X_i) \in E$ if X_i, X_j are adjacent in any graph in $[G]$
- end$(X_i, X_j) =$"$>$" iff there is an arrowhead from X_i to X_j in every graph in $[G]$
- end$(X_i, X_j) =$"$-$" iff there is *no* arrowhead from X_i to X_j in every graph in $[G]$
- end$(X_i, X_j) =$"$\circ$" else.

X_i and X_j are called adjacent if $(X_i, X_j) \in E$. We say there is a *directed edge* from X_i to X_j if end$(X_i, X_j) =$"$>$" and end$(X_j, X_i) =$"$-$" and then define directed paths if there are paths of directed edges. We say there is a *possibly directed path* from X_1 to X_k if there are nodes $X_1, \ldots, X_k$ such that X_i is adjacent to X_{i+1} and there is no arrowhead towards X_i for $i \in [k-1]$. A node X_j is a *possible descendant* of X_i if there is a possibly directed path from X_i to X_j. A node X_i is a collider on a path if there are two arrowheads towards X_i on that path. A collider path is a path, such that every non-endpoint is a collider. A direct edge is also a (trivial) collider path. A *definite non-collider* on a path is a node X_i that has at least one tail towards X_i on the path and a *definite status path* is path such that every node on the path is either a collider or a definite non-collider on this path.

Definition 20 (visible edge) Every edge in a DAG or CPDAG is *visible*. A directed edge $X_i \to X_j$ in a MAG (PAG) G is visible if there is a node X_k not adjacent to X_i and there is a collider path from X_k to X_i such that every non-endpoint of the path is a parent of X_j and the last edge towards X_i has an arrowhead at X_i.

Intuitively, a visible edge indicates an *unconfounded* edge, i.e. that there is no hidden confounder between the start and the endpoint of the edge.

Identification We now want to define what we mean with *identifiability*. The following definition is from Tian and Pearl (2002).

Definition 21 (identifiability) Let $\mathcal{S}$ be a set of SCMs and $S, S' \in \mathcal{S}$ be SCMs with the same causal graph. Any quantity of $Q(S)$ is *identifiable* in $\mathcal{S}$ if $Q(S) = Q(S')$ whenever P_S coincides with $P_{S'}$, where P_S denotes the probability distribution entailed by S.

Often one assumes that the causal graph G is known and then tries to derive the property $Q(S)$ from G and the distribution P_S. We want to make the distinction between the graphical properties of G and the distribution P_S a bit more explicit, as in our self-compatibility framework we deal with estimated graphs that might not

correctly represent the underlying data generation. In such a case, the identification formulae derived from such a graph may be incorrect and different identification formulae may even lead to contradicting results for the quantities of interest. Note, that the following formalism assumes for an interventional probability $p(x_j|do(x_i))$ that x_i, x_j are fixed.

Definition 22 (identification formula) Let S be an SCM with causal graph G and identifiable quantity $Q(S)$. A *identification formula* in G is a map $g : \mathcal{P} \to \mathbb{R}$, such that for a probability distribution P_S of S we have $g(P) = Q(S)$, where $\mathcal{P}$ denotes the space of probability distributions.

Note that if we have two different graphs G, G', we may have identification formulae g, g' for Q in G and in G' respectively but for a distribution P we get $g(P) \neq g'(P)$, as G and G' may be different graphs.

Latent projection In definition 5 we have already defined how to project ADMGs (and therefore also DAGs, as they are ADMGs without hidden confounders) to an ADMG that represent only a subset of variables. We will now define similar projections for MAGs, PAGs and CPDAGs, and start be defining a projection of a DAG to a MAG (see also (Zhang, 2008)). We mainly do this to connect MAGs and PAGs to the SCM formalism.

Definition 23 (latent MAG) Let G be a DAG with variables in V and $S \subseteq V$. The *latent MAG* $L^{\text{MAG}}(G, S)$ is the MAG that contains all nodes in S and and for $X_i, X_j \in S$

1. an edge between X_i and X_j iff there is an inducing path in G between them
2. an arrowhead at X_i (or X_j) iff the last edge of the inducing path has an arrowhead at X_i (X_j).

When no confusion with definition 5 can arise, we also just write $L(G, S)$. We call a MAG M a *causal MAG* if it is the latent projection of a causal DAG G.

Definition 24 (latent PAG) Let H be a PAG with variables in V, $S \subseteq V$ and M be some MAG in the equivalence class described by H. The *latent PAG* $L^{\text{PAG}}(H, S)$ is the PAG of $[L^{\text{MAG}}(M, S)]$. We call a PAG H a *causal PAG* if it is the latent projection of a causal MAG.

The latent PAG is well-defined, as all MAGs in the same equivalence class also have the same independence statements in S. Therefore, these independences are represented by the same PAG, i.e. defining the latent PAG via *some* MAG does not introduce arbitrariness.

It is not possible to define a projection operator for CPDAGs without assumptions about the subset S, as this model class cannot represent the presence of latent confounders. Nonetheless, we wanted to include them in our framework as popular causal discovery algorithms like PC and GES output CPDAGs. We chose to only define the projection operator for sets S that fulfil the causal sufficiency assumption, i.e. sets that do not contain two nodes $X_i, X_j \in S$ with common ancestor $L \in V \setminus S$ such that any intermediate nodes on paths from L to X_i or X_j are also in $V \setminus S$. In other words, the latent ADMG does not contain birected edges:

Definition 25 (latent CPDAG) Let G be a DAG with variables V and $S \subset V$ be a subset such that the latent ADMG $L(G, S)$ contains no bidirected edges. Then the *latent CPDAG* $L(G, S)$ is the CPDAG that represents the equivalence class $[L(G, S)]$.

Adjustment criteria The following theorems from the literature provide graphical criteria to identify interventional probabilities. The first one is Theorem 1 in (Tian and Pearl, 2002).

Theorem 3 (parent adjustment in ADMGs) *Let G be a causal ADMG over discrete variables in V and $X_i, X_j \in V$. If there is no bidirected edge connected to X_j, we have*

$$p(x_i|do(x_j)) = \sum_{pa_i} p(x_i|x_j, pa_j)p(pa_j).$$

The sum can easily be replaced by an integral for continuous variables with positive densities. Tian and Pearl (2002) also show the more general result in their Theorem 3:

Theorem 4 (generalised identifiability in ADMGs) *Let G be a causal ADMG over discrete variables in V and $X_i, X_j \in V$. The probability $p(x_i|do(x_j))$ is identifiable iff there is no bidirected path between X_j and any of X_j's children.*

For the other graphical models we considered, Perković et al. (2015) provided the following result in their Theorem 3.4, where we restrict ourselves to the case of adjustment between single variables. First we define the *forbidden set*.

Definition 26 (forbidden set) Let G be a causal DAG, CPDAG, MAG or PAG for the probability distribution P over V, $X_i, X_j \in V$. Denote with $\text{Forb}(X, Y, G)$ the set of possible descendant in G of any $W \in V \setminus \{X_j\}$ that lies on a possibly directed path from X_j to X_i and call this set the *forbidden set*.

Theorem 5 (generalised adjustment criterion) *Let G be a causal DAG, CPDAG, MAG or PAG for the probability distribution P over V, $X_i, X_j \in V$ and $Z \subseteq V \setminus \{X_i, X_j\}$. Then we have*

$$p(x_i|do(x_j)) = \begin{cases} p(x_i|x_j), & \text{if } Z = \emptyset \\ \int p(x_i|x_j, x_Z)p(x_Z)dx_Z, & \text{else} \end{cases}$$

if and only if

1. *every possibly directed path in G from X_j to X_i starts with a visible edge (see definition 20)*
2. *$Z \cap \text{Forb}(X, Y, G) = \emptyset$*
3. *all definite status paths that are not directed are m-separated by Z.*

Perković et al. (2015) also give a definition for a set that is a valid adjustment set, iff there is a valid adjustment set.

Definition 27 (canonical adjustment set) Let G be a causal DAG, CPDAG, MAG or PAG for the probability distribution P over V, $X_i, X_j \in V$. We call the set

$$\text{Adjust}(X, Y, G) = \text{PossAnc}(\{X, Y\}, G) \setminus (\text{Forb}(X, Y, G) \cup \{X, Y\})$$

the *canonical adjustment set*, where $\text{PossAnc}(\{X, Y\}, G)$ is the set of possible ancestors of X and Y.

A8 PROOFS FOR THE MAIN PAPER

A8.1 For Lemma 1

Proof Let $S_1 \dots, S_k$ be $k \in \mathbb{N}$ sets of variables and P_V be a probability distribution over $V \supseteq \bigcup_{i\in[k]} S_i$ such that all P_{S_i} with $i \in [k]$ fulfil the assumptions of $\mathcal{A}$. Further, let $\epsilon > 0$. Set

$$\delta := 1 - \sqrt[k]{1 - \epsilon}.$$

As every marginal distribution fulfils the assumptions of $\mathcal{A}$, we know that for every $i \in [k]$ there is a $m_i \in \mathbb{N}$ such that for all $m' > m_i$ we get $\mathcal{A}(X_{S_i}^{m'}) = L(G, S_i)$ with probability at least $1 - \delta$, where again G is the true causal DAG. Set $m^* = \max_{i\in[k]} m_i$. Then we get

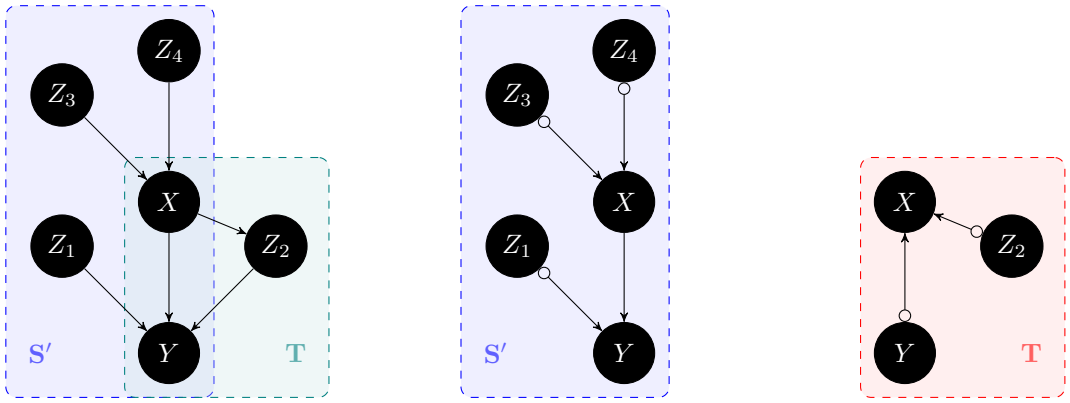
$$P(\exists i \in [k] : \mathcal{A}(X_{S_i}^{m*}) \neq L(G, S_i)) = 1 - \prod_{i\in[k]} P(\mathcal{A}(X_{S_i}^{m*}) = L(G, S_i)) \leq 1 - (1 - \delta)^k = \epsilon.$$

The graphical compatibility follows directly from definition 6. Similarly, the interventional compatibility follows from the fact that the algorithms find the latent projections of G. This renders them causal in the sense of theorems 4 and 5 and therefore the interventional probabilities coincide with the ones in G if they are identifiable.

A8.2 For Theorem 1

Proof We will prove the statement separately for the different algorithms. For FCI we explicitly construct a joint distribution such that the algorithms make contradicting interventional statements on different subsets, as our motivating example from section 1.1 almost suffices to show the statement. For RCD we show an example where two LiNGAM models with different linear coefficient between X_i and X_j generate the same marginal for X_i and X_j. In section section A11 we will show that this suffices to render RCD falsifiable. Precisely, in this proof we will show that RCD is non-bivariate (definition 32) and therefore theorem 6 implies that it is falsifiable.

For FCI we have to slightly modify the example from section 1.1 to render the edge $X \to Y$ visible (and therefore the interventional probability identifiable). To this end, assume we have the graph G shown in figure 4a. Assume that as before, all independences are given by d-separation in G except for the additional independence $Y \perp\!\!\!\perp Z_2$. On the set $S' = \{X, Y, Z_1, Z_3, Z_4\}$, FCI will find all edges between nodes in S' that are in G as shown in figure 4b (except for some circle marks). In this subgraph, the edge $X \to Y$ is visible as Z_3 (or Z_4) has an edge towards X but is not adjacent to X. Further, there are no non-causal paths (i.e. backdoor-paths) between X and Y. Then the empty set is a valid adjustment set according to theorem 5. On T, FCI will find the marginal model as in figure 4c for the same reason as in section 1.1. As there is an arrowhead towards X, the there is no effect from X to Y. Now we get the same interventional probabilities $p^{S'}(y|do(x)) = p(y|x) \neq p(y) = p^T(y|do(x))$ as in section 1.1, where $S' = S \cup \{Z_3, Z_4\}$. Therefore we have constructed a joint distribution such that we can find incompatible results on some subsets in the limit of infinite data.



(a) True Graph G (b) Graph $G_{S'}$ (c) Graph G_T

Figure 4: This modification of figure 1b renders the edge $X \to Y$ visible if FCI is applied to S' and thus shows that FCI is falsifiable.

For RCD we modify an example from Hoyer et al. (2008b). Precisely, their example V consists of a linear SCM with three nodes defined via

$$X_3 := N_3, \quad X_1 := \beta X_3 + N_1, \quad X_2 := \alpha X_1 + \gamma X_3 + N_2,$$

where N_1, N_2, N_3 are jointly independent non-Gaussian noise variables. Figure 5 shows two such models. Hoyer et al. (2008b) show that there are two models that cannot be distinguished on the marginal over $\{X_1, X_2\}$, one with structural coefficient α and one with $(\alpha\beta+\gamma)/\beta$ between X_1 and X_2. For example, we can set $\alpha = \beta = \gamma = 1$ and assume all noise variables to have the same distribution with zero mean and unit variance as visualized in figure 5a. This model generates the distribution P. Analogously, we define another distribution $\tilde{P}$ via a model with $\alpha = 2, \beta = 1, \gamma = -1$ (we require the noise terms to have the same distributions as in the previous model) shown in figure 5b. Then the joint behaviour between X_1 and X_2 can be described via the vector

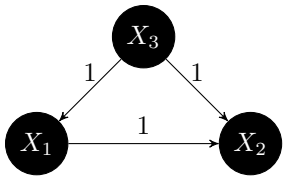
$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} N_1 \\ N_2 \\ N_3 \end{pmatrix} = \begin{pmatrix} N_1 + N_3 \\ N_1 + N_2 + 2N_3 \end{pmatrix}$$

for P, while for $\tilde{P}$ we get

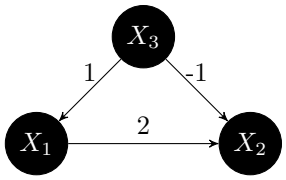
$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 2 & 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} N_1 \\ N_2 \\ N_3 \end{pmatrix} = \begin{pmatrix} N_1 + N_3 \\ 2N_1 + N_2 + N_3 \end{pmatrix}.$$

Since all noise terms have the same distribution, the vector $(X_1, X_2)^T$ has the same distribution in both cases.

Further, RCD will identify the respective model when all three nodes are observed. In other words, there are two distributions $P, \tilde{P}$ over X_1, X_2, X_3 (generated by the SCMs described above) that have identical marginals over X_1, X_2 but RCD results in different interventional probabilities $p(x_2|do(x_1)) \neq p'(x_2|do(x_1))$. Therefore RCD is non-bivariate (definition 32) and via theorem 6 it is falsifiable.



(a) Graph G_1 with structural coefficient 1 between X_1 and X_2.

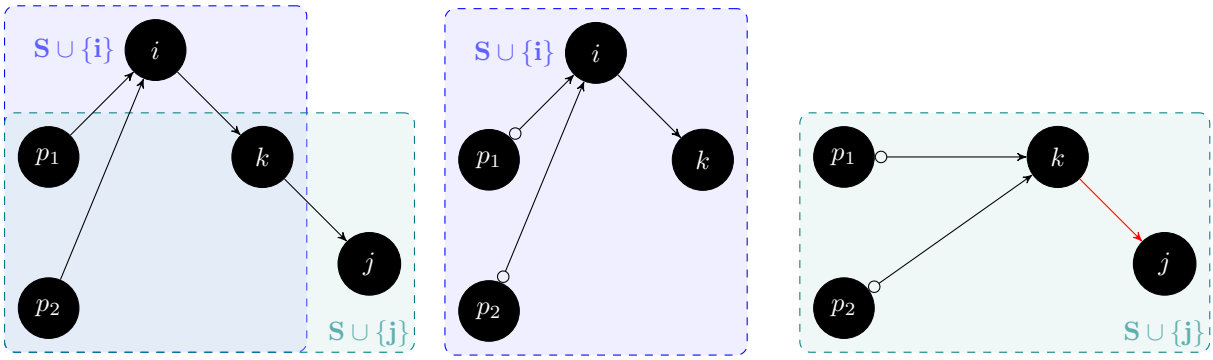


(b) Graph G_2 with structural coefficient 2 between X_1 and X_2.

Figure 5: Two causal models that fulfil the LiNGAM assumption, have the same marginal over X_1 and X_2 and different coefficient from X_1 to X_2.

A8.3 For Theorem 2

Proof The idea is that we apply FCI to data generated from the DAG in figure 6a on $S \cup \{i\}$ and $S \cup \{j\}$. In the first case, FCI identifies the direction $i \to k$. Likewise, FCI infers that $k \to j$ is visible, i.e. that k is an unconfounded cause of j in the second case. This unconfoundedness excludes a direct link between i and j, as otherwise i would be a confounder for k and j. Further, S also contains no common child of i and j. Consequently, S m-separates i and j. From these independences, we can reconstruct the entire joint distribution.



(a) True Graph G (b) Graph $G_{S\cup\{i\}}$ (c) Graph $G_{S\cup\{j\}}$

Figure 6: True DAG G and PAGs G_1 and G_2 over $S \cup \{i\}$ and $S \cup \{j\}$ respectively. The red edge indicate a visible edge. There cannot be any edge between i and j due to the visible edge between k and j.

More precisely, let $S := \{k, p_1, p_2\}$ be a set of nodes with $i, j \notin S$. Let G_1 and G_2 denote the PAGs given by the asymptotic outputs when FCI is applied to $S \cup \{i\}$ and $S \cup \{j\}$, respectively, where we assume that the distribution is Markovian to the joint PAG in figure 6a. As visualized in figure 6b, G_1 consists of the edges $p_1 \to i$, $p_2 \to i$ (with circle marks) and $i \to k$ (without circle, since this link is recognized as visible). Likewise, G_2 consists of the edges $p_1 \to j$, $p_2 \to j$ (with circle marks) and $k \to j$ (without circle), where the latter is a visible edge, as e.g. the edge between p_1 and k has an arrow head towards k and p_1 is not adjacent to j. This graph is shown in figure 6c. We now conclude that there cannot be a direct link between i and j as follows: $i \to j$ and $i \to k$ would create an inducing path w.r.t. i between k and j (which is ruled out by the visible edge), while $j \to i$ would be a directed cycle. We thus obtain $X_i \perp\!\!\!\perp X_j \,|X_S$, which enables constructing the joint distribution from the two marginals. Precisely, we get

$$p(i, k, j, p_1, p_2) = p(i|k, p_1, p_2)p(j|k, p_1, p_2)p(k, p_1, p_2),$$

where every factor on the right hand side is already given by the marginal distributions.

We now need to show that there is another distribution $\tilde{P}$ that has the same marginals, but is ruled out by the self-compatibility constraint. We start by restricting the example from figure 6a further to the linear Gaussian case. If we assume that all noise terms adhere to independent standard normal distributions and all structural

coefficients are one, we get a distribution that is Markovian and faithful to P and its covariance matrix is given via

$$\Sigma = ((I-A)(I-A)^T)^{-1} = \begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 3 & 3 \\ 1 & 1 & 3 & 4 & 4 \\ 1 & 1 & 3 & 4 & 5 \end{pmatrix},$$

where A is the adjacency matrix of the graph G. The marginal covariance matrices follow directly form this:

$$\Sigma_{\{p_1,p_2,i,k\}} = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 3 & 3 \\ 1 & 1 & 3 & 4 \end{pmatrix} \quad \Sigma_{\{p_1,p_2,k,j\}} = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 4 & 4 \\ 1 & 1 & 4 & 5 \end{pmatrix}$$

Finally, we define the matrix

$$\widetilde{\Sigma} = \begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 3 & 7/2 \\ 1 & 1 & 3 & 4 & 4 \\ 1 & 1 & 7/2 & 4 & 5 \end{pmatrix}.$$

$\widetilde{\Sigma}$ is also a symmetric, positive definite matrix and is therefore a valid covariance matrix. Further, it has the same marginal covariances over $\{p_1, p_2, i, k\}$ and $\{p_1, p_2, k, j\}$ as Σ. Therefore, the normal distribution with zero means and covariance matrix $\widetilde{\Sigma}$ is an example for a distribution $\tilde{P}$ that we were looking for.

A9 GRAPHICAL COMPATIBILITY

A9.1 Definitions for Further Graphical Models

Now we want to define graphical notions for the remaining types of graphical models.

Definition 28 (purely graphical compatibility) A compatibility notion c is called purely graphical if it does not depend on P_V, that is, it can also be written as a function

$$c : \mathcal{G}_V^* \to \{0, 1\}.$$

We will present several purely graphical compatibility notions for different classes of graphical models respectively. In definition 6 we have already defined graphical compatibility for ADMGs. We will now define graphical compatibility for the other graphical models that we considered.

Definition 29 (graphical compatibility of MAGs) Let $S_1, \ldots, S_k$ be $k \in \mathbb{N}$ sets of nodes and $G_{S_1}, \ldots, G_{S_k}$ be MAGs. Then we define graphical compatibility by the function c with $c(G_{S_1}, \ldots, G_{S_k}) = 1$ iff there exists a set $V \supseteq \cup_{j=1}^k S_j$, and an DAG G_V such that $L(G_V, S_j) = G_{S_j}$, where L is the latent projection from DAGs to MAGs as defined in definition 23.

Definition 30 (graphical compatibility of PAGs) Let $S_1, \ldots, S_k$ be $k \in \mathbb{N}$ sets of nodes and $G_{S_1}, \ldots, G_{S_k}$ be PAGs. Then we define graphical compatibility by the function c with $c(G_{S_1}, \ldots, G_{S_k}) = 1$ iff there exists a set $V \supseteq \cup_{j=1}^k S_j$, and a DAG G_V such that $L^{\text{PAG}}(L^{\text{MAG}}(G_V, V), S_j) = G_{S_j}$, i.e. if G_{S_j} represents the conditional independences of the DAG G_V over S_j.

The definition of latent PAGs is centered around conditional independences and the causal semantics is not as obvious as for ADMGs. Note, that this is due to the fact that PAGs represent the causal statements that are consistent for *all* models that entail the same independence structure (via theorem 5). Therefore we can also "safely" marginalise PAGs by only referencing the independence structure.

Definition 31 (graphical compatibility of CPDAGs) Let $S_1, \ldots, S_k$ be $k \in \mathbb{N}$ sets of nodes and $G_{S_1}, \ldots, G_{S_k}$ be CPDAGs. Then we define graphical compatibility by the function c with $c(G_{S_1}, \ldots, G_{S_k}) = 1$ iff there exists a set $V \supseteq \cup_{j=1}^k S_j$, and an DAG G_V such that $L(G_V, S_j) = G_{S_j}$, where L is the latent CPDAG as defined in definition 25.

A9.2 Critical Discussion of Graphical Compatibility

Here we show that a purely graphical criterion implicitly relies on a genericity condition which is close to faithfulness in spirit. Consider, the causal model $Y \to Z$ and recall that ADMG compatibility excludes that an additional variable X influences both Y and Z. This conclusion can be criticised, depending on the precise interpretation of $Y \to Z$. Assume our interpretation of $Y \to Z$ reads: "Y is an unconfounded cause of Z in the sense that $p(z|do(y)) = p(z|y)$". One can then construct a causal model with the complete DAG on X, Y, Z in which the relation between X and Z is confounded by a hidden variable W. In a linear Gaussian model, we can easily tune parameters such that the confounding bias that W induces on $P(Y, Z)$ cancels out with the confounding bias induced by Z. This way, we still have $p(z|do(y)) = p(z|y)$ despite the confounding paths. Excluding such a non-generic choice of parameters follows from faithfulness in a DAG with X, Y, Z, W, F_Y where F_Y controls randomized interventions on Y (the so-called 'regime-indicator variable' of Dawid (2021)). In our example, vanishing confounding bias corresponds to

$$F_Y \perp\!\!\!\perp Z \,|Y,$$

without d-separation. In other words, the conclusion that $X \to Y$ together with the unconfounded causal relation $Y \to Z$ is incompatible with the complete DAG on X, Y, Z relies on a genericity condition against which one may raise doubts. After all, it is problematic to benchmark causal discovery algorithms via methods that implicitly rely on principles close to faithfulness, if we on the other hand argue that assumptions like faithfulness are often violated. We cannot resolve this counter argument entirely. It may be reassuring, however, that also model classes that do not rely on faithfulness come to the same conclusion, that is, also exclude the direct link from X to Z. This will be shown in the proof of theorem 8.

A10 Relationship between Interventional and Graphical Compatibility

The next example shows a case, where interventional compatibility does not imply graphical compatibility.

Example 2 (Non-generic confounder) Let there be an ADMG G_1 with variables X, Y, Z. Let X consist of two components[15] X_1, X_2 with $X_1 \perp\!\!\!\perp X_2$ as in figure 7a. If Y only depends on X_1 and Z only on X_2 we get

$$\begin{aligned} p(z \mid do(y)) &= \sum_x p(x)p(z \mid y, x) \\ &= \sum_{x_1,x_2} p(x_1)p(x_2)p(z \mid y, x_2) \\ &= \sum_{x_2} p(x_2)p(z \mid y, x_2) \\ &= \sum_{x_2} p(x_2 \mid y)p(z \mid y, x_2) = p(z \mid y). \end{aligned}$$

Now assume we have an ADMG G_2 that only contains $X \to Y$, i.e. implicity rules out confounding. Especially, G_2 implies

$$p(z \mid do(y)) = p(z \mid y).$$

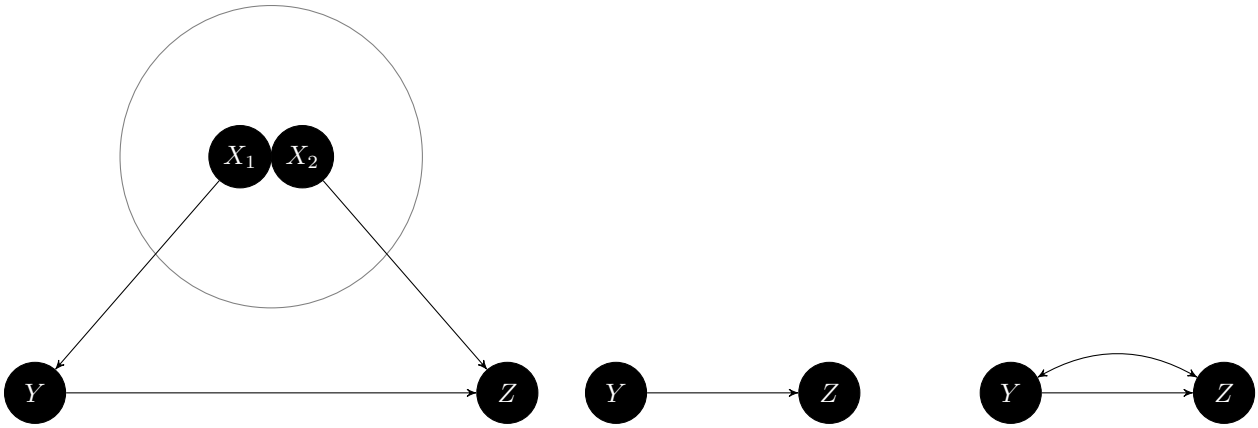
Therefore, the two models entail the same interventional statements and are interventionally compatible. Yet, G_1 and G_2 are not graphically compatible, as G_2 is different from the latent projection $L(G_1, \{X, Y\})$.

In the main paper we already mentioned that graphical compatibility also does not imply interventional compatibility. In the following example we want to illustrate this in more detail.

Example 3 (Empty graphs) Let G from figure 8a be the true underlying DAG for some distribution. Assume the distribution is faithful to the DAG. Further, let $\hat{G}$ and $\hat{G}_{X,Y}$ be the DAGs in figures 8b and 8c, respectively. Clearly, $\hat{G}$ and $\hat{G}_{X,Y}$ are graphically compatible. They imply

$$p(y|do(x)) = p(y).$$

[15] One might ask, whether it makes more sense to treat X_1 and X_2 as separate variables, instead of a single variables with two components. Indeed, summarizing X_1 and X_2 as a single variable may seem a bit artificial. Though we want to note, that this is merely an illustrative example. The same phenomenon would e.g. also occur for a scalar variable, where Y only depends on the first bit of the binary encoding and Z only depends on the second bit.



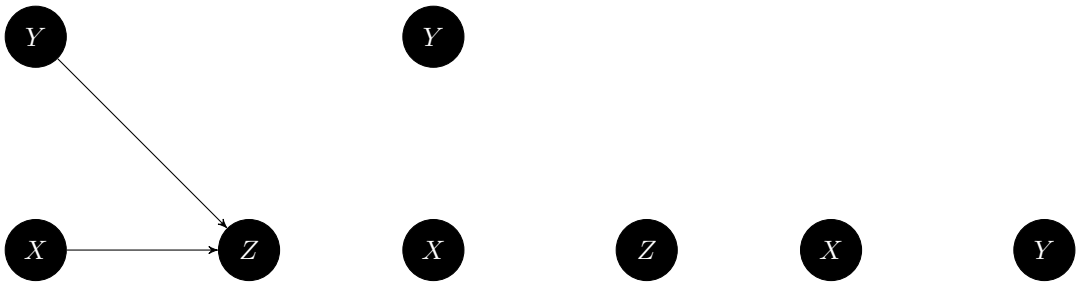
(a) Graph G_1 (b) Graph G_2 (c) Graph $L(G_1, \{X, Y\}$

Figure 7: Non-generic case where interventional compatibility does not imply graphical compatibility.

But further, $\hat{G}$ also entails (as Z is a valid conditioning set)

$$p(y|do(x)) = \sum_z p(y|x,z)p(z),$$

which is not equal to $p(y)$ for any distribution that is Markovian and faithful to G. Thus, $\hat{G}$ and $\hat{G}_{X,Y}$ are not interventionally compatible.



(a) True DAG G (b) Graph $\hat{G}$ (c) Marginal graph $\hat{G}_{X,Y}$

Figure 8: The graphical compatibility only depends indirectly on the data. For non-Markovian graphs, their graphical compatibility does not necessarily imply interventional compatibility.

Indeed, if we require that the graphs are Markovian w.r.t. to the distribution at hand, we get that graphical compatibility implies interventional compatibility.

Lemma 2 *Let $S_1, \dots S_k$ be sets of variables for some $k \in \mathbb{N}$ and denote $S := \bigcup_{i\in[k]} S_i$. Let $G_{S_1}, \dots, G_{S_k}$ be an ADMGs or PAGs and P_S be a probability distribution over S. Let further G_{S_i} be Markovian w.r.t. P_{S_i} for all $i \in [k]$. Then, if $G_{S_1}, \dots, G_{S_k}$ are graphically compatible, they are also interventionally compatible w.r.t. P_S.*

This follows directly from the construction of the latent projection in definitions 5 and 24.

In definition 4 we did not require the marginal graphs to be Markovian w.r.t. the distribution. With the following example we want to show why.

Example 4 (Non-Markovian model) Consider the linear Gaussian model with variables X, Y, Z and structural coefficients as in figure 9a. The graph $\hat{G}$ in figure 9b is clearly not the latent projection of G to $\{X, Z\}$. Yet, both, G (with the structural coefficients) and $\hat{G}$ imply

$$p(z|do(x)) = p(z).$$

Therefore they are interventionally compatible. With definition 4 we want to allow such non-generic cases, as long as the marginal model gets the interventional distributions "right" in the sense that they could have been created by a single joint model. As we have mentioned before, this might not be the only appropriate choice, depending on the downstream task of the causal model.

(a) True DAG G (b) Graph $\hat{G}$

Figure 9: A marginal model that is not Markovian w.r.t. to the distribution might still be interventionally compatible.

A11 WHICH ALGORITHMS ARE FALSIFIABLE?

In theorem 1 we have seen that FCI and RCD are observationally falsifiable. Other popular causal discovery algorithms that do not assume causal sufficiency are falsifiable as well. We deferred the following lemma to the appendix to keep the presentation in the main paper more concise.

Lemma 3 *The algorithms PC, GES (Chickering, 2002) and DirectLiNGAM (Shimizu et al., 2011) are falsifiable.*

Proof Recall the graph from figure 1a and assume that this is the ground truth graph G. Also recall that we constructed a distribution that is Markovian and faithful to G except for the additional independence $Y \perp\!\!\!\perp Z_2$. We have already discussed that the PC algorithm will find the graphs in figure 1b in the population limit, as these are the only graphs that capture the conditional independences on the the subsets X, Y, Z_1 and X, Y, Z_2 under the assumption of no hidden confounder.

Similarly, GES will find the same graphs in the limit of infinite data, as a graph that reflects exactly the independences in the distribution will have an optimal score[16] (Proposition 8 (Chickering, 2002)). Denote the graph that PC and GES find on S with G_S and the one over T with G_T. In these graphs we get $p^S(y|do(x)) = p(y|x) \neq p(y) = p^T(y|do(x))$ where $p^S(y|do(x))$ denotes the causal effect derived from G_S using the identification formula from theorem 5 and analogously for T. The inequality follows since we had $Y \perp\!\!\!\perp X$ otherwise, in contradiction to the assumption that $Y \perp\!\!\!\perp Z_2$ is the only independence that is not entailed by the true DAG.[17]

For DirectLiNGAM, the same argument as in the previous proof for RCD (visualized in figure 5) suffices.

As the unfalsifiable algorithm in example 1 showed, falsifiability is not a trivial property. This raises the question if the falsifiable algorithms can be characterised differently. With the following definition we want to exclude algorithms like the aforementioned ordering by entropy and indeed show that this is a sufficient criterion for falsifiability. The definition has two aspects: 1) a non-bivariate algorithm must be able to produce an output that allows identification of a causal effect 2) this output does not only depend on bivariate properties of X_i and X_j but also depends on the distribution of the other nodes.

Definition 32 (non-bivariate causal discovery) A causal discovery algorithm $\mathcal{A}$ is *non-bivariate* if there exists a set of variables V with $|V| > 2$, as well as probability distributions P_V and $\widetilde{P}_V$ over V whose marginalisations to a subset of two variables $\{X_i, X_j\} \subseteq V$ coincide (i.e. $P_{\{X_i,X_j\}} = \widetilde{P}_{\{X_i,X_j\}}$) such that for every $\epsilon > 0$ there is a $m \in \mathbb{N}$ where for all $m' \geq m$ the following conditions hold with probability at least $1 - \epsilon$:

1. $p(x_i|do(x_j))$ is identifiable in the estimated graphs $\mathcal{A}(\mathbf{X}^{m'})$ and $\mathcal{A}(\widetilde{\mathbf{X}}^{m'})$ and
2. there are identification formulae in $\mathcal{A}(\mathbf{X}^{m'})$ and $\mathcal{A}(\widetilde{\mathbf{X}}^{m'})$ respectively such that $p^{\mathcal{A}(\mathbf{X}^{m'})}(x_i|do(x_j)) \neq p^{\mathcal{A}(\widetilde{\mathbf{X}}^{m'})}(x_i|do(x_j))$,

[16] For simplicity we assume that the greedy search procedure always finds the optimal score.

[17] In this case we showed the falsifiability directly by constructing a joint distribution. But the example also shows that PC and GES are non-bivariate (as the marginal distribution over S and T could be interpreted as P and $\widetilde{P}$ from definition 32) and then theorem 6 could be applied.

where $\mathbf{X}^{m'}$ denotes a data matrix with m' samples drawn from P_V and $\widetilde{\mathbf{X}}^{m'}$ contains samples from $\widetilde{P}_V$ and $p^G(x_i \mid do(x_j))$ denotes the identification formula from G applied to probability distribution P.

Note that we did not require the distributions to fulfil the causal discovery assumptions of $\mathcal{A}$.

The following theorem asserts that for every non-bivariate causal discovery algorithm $\mathcal{A}$ there is at least one distribution P for which we can detect that P does not fulfil the causal discovery assumptions of $\mathcal{A}$ by applying $\mathcal{A}$ to a subset of variables. We observe:

Theorem 6 (non-bivariate implies falsifiable) *If $\mathcal{A}$ is non-bivariate, it is observationally falsifiable with respect to interventional compatibility.*

Proof Let the set of variables S and the distributions $P, \widetilde{P}$ be like in 32. Let $\tilde{S}$ with $S \cap \tilde{S} = \{i, j\}$ be such that $\tilde{S} \setminus \{X_i, X_j\}$ are variables of the same type as $S \setminus \{X_i, X_j\}$ with canonical one-to-one correspondence. We will assign different distributions to them now: define P' to be such that $\widetilde{P}$ is "copied" to $\tilde{S}$, i.e. X_i and X_j have the same distribution like in $\widetilde{P}$ and all nodes in $\tilde{S} \setminus \{X_i, X_j\}$ have the same distributions as their corresponding node in S has in $\widetilde{P}$. Then define the joint distribution via

$$P(\mathbf{X}_{S \cup \tilde{S}}) := P(\mathbf{X}_{S \setminus \{i,j\}} | X_i, X_j) P(X_i, X_j) P'(\mathbf{X}_{\tilde{S} \setminus \{i,j\}} | X_i, X_j). \tag{4}$$

One checks easily that its restrictions to S and $\tilde{S}$ coincide with $P(\mathbf{X}_S)$ and $\widetilde{P}(\mathbf{X}_S)$, respectively. By construction, the algorithm $\mathcal{A}$ will result on contradictory statements about the interventional distribution $p(x_i|do(x_j))$ when applied to $\mathbf{X}_S$, versus when applied to $\mathbf{X}_{\tilde{S}}$ in the limit of infinite data. Thus we have $c(\mathcal{A}(\mathbf{X}_S), \mathcal{A}(\mathbf{X}_{\tilde{S}}), P_{S \cup \tilde{S}}) = 0$.

We now want to consider an example that is similar to the one in example 1 but with very different conclusions.

Example 5 (DAG Entropy Ordering) Define an algorithm $\mathcal{A}$ that orders nodes according to their entropy and outputs the complete DAG with respect to that order.

This example is interesting for multiple reason. First, note that the algorithm in this example is non-bivariate in the sense of our definition. This illustrates that we refer to the non-bivariateness of interventional statements and not e.g. the resulting edges. Second, it highlights the importance of condition 1) of definition 32, as this is basically the only difference between the example above and example 1. Third, the following lemma shows that this naïve algorithm is indeed falsified for almost all distributions.

Lemma 4 (DAG via entropy order is almost always falsified) *Let $T := \{1, 2, 3\}$ and $P_{\{1,2,3\}}$ be generic in the sense that the entropies of all variables are different. Assume $H(X_1) < H(X_2) < H(X_3)$, without loss of generality, and assume the further genericity condition*

$$\sum_{x_1} p(x_3|x_2, x_1) p(x_1) \neq p(x_3|x_2). \tag{5}$$

Then the complete DAGs $G_{\{1,2\}}$ and $G_{\{1,2,3\}}$, obtained by entropy ordering of nodes, are interventionally incompatible.

Proof The left hand side of (5) is $p(x_3|do(x_2))$ in $\mathcal{A}(P_{\{1,2,3\}})$, while the right hand side is the same interventional probability in $\mathcal{A}(P_{\{2,3\}})$.

Note that P_T in lemma 4 is always Markovian to $\mathcal{A}(P_{\{1,2,3\}})$ because this is a complete DAG. It is thus notable that the algorithm falsifies itself although P_T is a distribution that is allowed by $\mathcal{A}(P_{\{1,2,3\}})$.

A12 FURTHER DISCUSSION OF THE RELATIONSHIP TO STABILITY

In this section, we motivate our incompatibility score using stability arguments from learning theory (Shalev-Shwartz et al., 2010). Observe that, for a causal discovery method to achieve a low incompatibility score, it's output must remain largely unchanged under small modifications to the variable sets it is applied to. Following this idea, under some notion of stability of a causal discovery algorithm, we now show that stable algorithms, provably, generate *useful causal models* in the sense described below.

While, it is not possible to guarantee that a causal discovery algorithm that achieves a low incompatibility score will accurately predict system behavior under interventions, we argue that the resulting models are at least

useful due to their ability to predict statistical properties of unobserved joint distributions. This perspective is influenced by Janzing et al. (2023), who reconceptualizes causal discovery as a *statistical learning problem*. The key principle underlying this reconceptualization posits that causal models offer predictive value beyond predicting system behavior under interventions; they can also predict statistical properties of unobserved joint distributions. To illustrate the main idea, consider the following example.

Consider a set of variables $\mathcal{X} = \{X_1, X_2, \cdots X_n\}$, and let $\mathcal{S}$ represent a collection of subsets of $\mathcal{X}$. A statistical property Q can be defined as a mapping from $\mathcal{S}$ to $\mathbb{R}$. For a given tuple $S_i = (X_{l_1}, X_{l_2}, \cdots X_{l_k})$, Q might indicate whether the conditional independence $X_{l_1} \perp\!\!\!\perp X_{l_2} | X_{l_3}, \cdots X_{l_k}$ holds, represented as $Q(S_i) = 1$ (if true) or $Q(S_i) = 0$ (if false).[18]

Causal models such as Directed Acyclic Graphs (DAGs) can thus be viewed as predictors of statistical properties, like conditional independences. The statistical property predicted by a causal model $\mathcal{M}$ is denoted as $\widehat{Q}_{\mathcal{M}} : \mathcal{S} \to \mathbb{R}$. [19]

The statistical prediction problem can now be outlined as follows: Given a set of m observations $\mathbb{S} = \{S_1, S_2, \cdots S_m)\} \in \mathcal{S}^m$, and a loss function $l : Q(\mathcal{S}) \times Q(\mathcal{S}) \to [0, 1]$, the goal of causal discovery is to learn a joint causal model over $\mathcal{X}$ that minimizes the loss $l(\widehat{Q}_{\mathcal{M}}(S), Q(S))$ on unobserved subsets $S \in \mathcal{S}$. As an illustrative example, the PC algorithm constructs a joint causal model across all variables—encoding conditional independences for any subset of $\mathcal{X}$—by evaluating conditional independences within a selected set of smaller subsets.

If we assume that the subsets are drawn according to a distribution over $\mathcal{S}$, we can invoke stability arguments from learning theory to guarantee generalization from observed to unobserved sets of variables for causal discovery algorithms that demonstrate stability under small modifications to variable sets. This is different from the standard setting in statistical learning, where algorithms that exhibit stability under small modifications to the data are known to *generalize across data points* (Shalev-Shwartz et al., 2010).

In order to formalize our discussion, let's define the concept of 'stability' in the context of causal discovery. While there are many related notions of stability in statistical learning theory, their relevance varies depending on the context. In our case, Leave-one-out (LOO) stability (Mukherjee et al., 2006), which requires stability of the loss when a single data point (or a subset of data points) is included or excluded from the training set, is particularly applicable. Nonetheless, to maintain a high-level perspective, we will introduce a strong, distribution-independent form of stability: uniform stability (Bousquet and Elisseeff, 2002), which encompasses other weaker forms of stability like LOO stability.

Definition 33 (Uniform Stability) A causal discovery algorithm A is said to be $\gamma-$uniformly stable with respect to the loss function l if, for any $S \in \mathcal{S}$, any subset $\mathbb{S} = \{S_1, S_2, \cdots S_m\} \subset \mathcal{S}^m$, and for any index $i \in [m]$, the following inequality holds:

$$|l(\widehat{Q}_{A_{\mathbb{S}}}(S), Q(S)) - l(\widehat{Q}_{A_{\mathbb{S}^{/i}}}(S), Q(S))| \leq \gamma,$$

where $S^{/i}$ denotes the set $\mathbb{S}$ replacing the element S_i by some $S_i' \in \mathcal{S}$.

With this definition, we can formally prove that stable causal discovery algorithms generate useful causal hypotheses in the sense of their ability to generalize statistical predictions across variable sets. To this end, we first introduce the notions of empirical and true risks and state the key result in Theorem 7.

Definition 34 (Empirical and Expected Risks) Under the reconceptualization of causal discovery as a learning problem, assuming that the subsets are drawn according to a distribution $\mathbb{L}$ over $\mathcal{S}$, the empirical ($\widehat{R}(\mathcal{M})$) and expected risks ($R(\mathcal{M})$) incurred by a causal model $\mathcal{M}$ are defined as:

$$\widehat{R}(\mathcal{M}) \coloneqq \frac{1}{n} \sum_{i=1}^{m} l(Q_{\mathcal{M}}(S_i), Q(S_i)), \quad R(\mathcal{M}) \coloneqq \mathbb{E}_{\mathbb{L}}[l(Q_{\mathcal{M}}(S_i), Q(S_i))].$$

Under this definition, one can leverage recent results from Bousquet et al. (2020) to derive the following high probability generalization bounds for uniformly stable causal discovery algorithms.

[18] Empirically, it can also indicate whether a given test T detects the conditional independence under consideration based on datasets of observations corresponding to the variable set.

[19] For a more detailed discussion, formalism, and justification of this problem we refer the reader to Janzing et al. (2023).

Theorem 7 (Generalization bounds for uniformly stable causal discovery) *Let A denote a causal discovery algorithm that is $\gamma-$uniformly stable with respect to the loss function l in the sense described above. Then there exists a constant c such that for any probability distribution $\mathbb{L}$ over $\mathcal{S}$ and for any $\delta \in (0,1)$, the following holds with probability at least $1-\delta$ over the draw of m subsets $S_1, S_2, \ldots, S_m$ according to $\mathbb{L}$:*

$$R(A(\mathbb{S})) - \widehat{R}(A(\mathbb{S})) \leq c\Big(\gamma \log(n) \log(1/\delta) + \frac{\sqrt{\log(1/\delta)}}{\sqrt{n}}\Big),$$

where $A(\mathcal{S})$ denotes the causal model output by the causal discovery algorithm A applied on the set $\mathbb{S} = \{S_1, S_2, \cdots, S_m\}$.

The result demonstrates that stable algorithms, provably, generate useful causal models due to their ability to *generalize statistical predictions across variable sets*. Informally, it provides evidence that a low incompatibility constitutes a useful inductive bias for causal discovery. This is notably distinct from the standard setting in statistical learning, where algorithms that exhibit stability under small modifications to the data are known to *generalize across data points*.

A13 ADDITIONAL RESULTS ABOUT MERGING

In theorem 2 we have seen that FCI enables what we have called *merging* in definition 8. We now want to show that an idealized version of RCD also has this property. Repetitive Causal Discovery (RCD) (Maeda and Shimizu, 2020) is a based on the LiNGAM-assumptions and is able to also infer the presence of latent confounders. It assumes a linear model with independent non-Gaussian noise, where confounding is modelled by some shared noise-variables. Explicitly, it reads:

$$X_i = \sum_j \beta_{ij} X_j + \sum_j \epsilon_{ij} W_j + N_i, \tag{6}$$

where all N_i, W_j are independent noise variables (with non-zero variance). The variables W_j, which are shared by at least two X_i, describe the confounding.

Definition 35 (idealized RCD) We define *idealized RCD* to be a causal discovery algorithm that has an additional output token $\perp$, that indicates that idealized RCD 'abstain from a decision', i.e. it indicates that (idealized RCD estimated[20] that) the distribution cannot be generated via equation (6). Otherwise, it draws an arrow $X_i \to X_j$ whenever $\beta_{ij} \neq 0$, and a bidirected link whenever they share a variable W_k, that is, there exists a k such that $\epsilon_{ik} \neq 0$ and $\epsilon_{jk} \neq 0$ according to the usual RCD algorithm.

Theorem 8 (idealized RCD enables merging) *Idealized RCD enables merging with respect to graphical consistency.*

Proof For the variables X_1, X_2, X_3, assume that idealized RCD outputs $X_1 \to X_2$ (without confounding) asymptotically when applied to data from $P_{\{1,2\}}$. Assume further, that it outputs $X_2 \to X_3$ (without confounding) when applied to data from $P_{\{2,3\}}$. We now show that applying RCD to all three variables can only yield compatible results if $X_1 \perp\!\!\!\perp X_3 \,|X_2$ and thus

$$P(X_1, X_2, X_3) = P(X_1, X_2) P(X_3 | X_2).$$

Consequently, the joint distribution follows uniquely from the two marginal distributions. The proof builds heavily on the theorem of Darmoir-Skitovic (Darmois, 1953; Skitovic, 1962), which entails that for any set of independent non-Gaussian variables $Y_1, \ldots, Y_d$ with non-zero variance, we have

$$\Big(\sum_j a_j Y_j\Big) \quad \perp\!\!\!\perp \quad \Big(\sum_j b_j Y_j\Big) \quad \Rightarrow a_j \cdot b_j = 0 \quad \forall j.$$

In other words, two linear combinations can only be independent if they share none of the variables.

[20] We assume that the idealized algorithm estimates this in an "oracle"-fashion, i.e. we do not discuss how this could be estimated.

If the output of RCD on the joint distribution $P_{\{1,2,3\}}$ is compatible with the two marginal models, it needs to be described by a model of the form (6) with causal ordering $1, 2, 3$. For any different order, causal directions would not be compatible with the marginal models. Further, RCD would output $\perp$ if the joint model was not of the form (6), which we also count as incompatibility. Hence, compatibility entails a joint model of the following form:

$$X_1 = \sum_j \epsilon_{1j} W_j + N_1 \tag{7}$$

$$X_2 = \beta_{21} X_1 + \sum_j \epsilon_{2j} W_j + N_2 \tag{8}$$

$$X_3 = \beta_{31} X_1 + \beta_{32} X_2 + \sum_j \epsilon_{3j} W_j + N_3. \tag{9}$$

For all j we have $\epsilon_{1j}\epsilon_{2j} = 0$ otherwise the causal relation $X_1 \to X_2$ would be confounded and RCD could not output an unconfounded link as bivariate model. This can be seen as follows. RCD only outputs an unconfounded link $X_1 \to X_2$ if $X_2 - \alpha X_1$ is independent of X_1 for some α. This can only be true for $\alpha = \beta_{21}$, otherwise both expressions contain N_1. Further, $X_2 - \beta_{21} X_1 = \sum_j \epsilon_{2j} W_j + N_2$ can only be independent of $\sum_j \epsilon_{1j} W_j + N_1$ if the linear combinations share no W_j. In a similar way we conclude that $\epsilon_{2j}\epsilon_{3j} = 0$: We first check that $X_3 - \alpha X_2$ can only be independent of X_2 for $\alpha = \beta_{32}$, otherwise N_2 appears in both expressions. Further, $X_3 - \beta_{32} X_2$ can only be independent from X_2 if they share none of the variables W_j.

The variables W_j thus fall into three classes: those that appear in only one of the variables X_1, X_2, X_3 and those that are shared by X_1 and X_3. The former ones can be absorbed into the noise variables N_j. We thus simplify (7) to (9) to

$$X_1 = \sum_j \epsilon_{1j} W_j + N_1 \tag{10}$$

$$X_2 = \beta_{21} X_1 + N_2 \tag{11}$$

$$X_3 = \beta_{31} X_1 + \beta_{32} X_2 + \sum_j \epsilon_{3j} W_j + N_3. \tag{12}$$

In a similar way as we have repeatedly argued, $X_3 - \alpha X_2$ can only be independent of X_2 if $\alpha = \beta_{31}$. We obtain

$$X_3 - \beta_{32} X_2 = \beta_{31} \sum_j \epsilon_{1j} W_j + \sum_j \epsilon_{31} W_j + \beta_{31} N_1 + N_3. \tag{13}$$

This expression can only be independent of X_1 if $\beta_{31} = 0$, otherwise N_1 appears in both expressions. The remaining term $\sum_j \epsilon_{31} W_j + N_3$ can only be independent of X_1 if there is no W_j shared by X_1 and X_2, i.e., $\epsilon_{1j}\epsilon_{3j} = 0$ for all j. However, then the respective linear combination of all W_j that appear only in X_1 can be merged with N_1, and the others with N_3. Thus, we end up with the structural equations

$$X_1 = N_1 \tag{14}$$

$$X_2 = \beta_{21} X_1 + N_2 \tag{15}$$

$$X_3 = \beta_{32} X_2 + N_3, \tag{16}$$

which implies $X_1 \perp\!\!\!\perp X_3 \,|X_2$.

We now want to construct a different distribution $\tilde{P}$ that has the same marginals as P, to show that without the self-compatibility constraint the solution to the statistical marginal problem is not unique. We will construct an SCM with the edges $X_1 \leftarrow X_2 \to X_3$ and $X_1 \to X_3$ and violate the assumption of a linear model with additive noise.

We first reconstruct the marginal $P_{1,2}$ using a construction from proposition 4.1 by Peters et al. (2017). Define the conditional cumulative distribution function

$$F_{Y|x}(y) := P(Y \leq |y \mid X = x)$$

and further define

$$F^{-1}_{Y|x}(n_Y) := \inf\{y \in \mathbb{R} : F_{Y|x}(y) \geq n_Y\}.$$

Now set X_1 via the structural equation

$$X_1 = f(X_2, \tilde{N}_1),$$

where $f(x_2, n_1) = F^{-1}_{Y|x}(n_1)$ and $\tilde{N}_1$ is uniformly distributed over $[0, 1]$ and independent from X_2 and $P_2 = \tilde{P}_2$. By construction we get $\tilde{P}_{1,2} = P_{1,2}$.

We now set X_3 via the structural equation

$$X_3 = \beta_{32} X_2 + F^{-1}_{N_3}(\tilde{N}_1),$$

where β_{32} is the underlying structural coefficient between X_2 and X_3 from P and $F^{-1}_{N_3}$ is the quantile function of the noise term N_3. Again, by construction we get the same marginal as in P. But clearly, in $\tilde{P}$ we do not have $X_1 \perp\!\!\!\perp X_3 | X_2$, as the noise term of X_3 is a deterministic function of the noise term of X_1.

A14 PRACTICAL EVALUATION OF COMPATIBILITY

A14.1 Details of Interventional Incompatibility Score

We have already introduced our interventional compatibility score in definition 9. In this section we want to shortly elaborate on this score to avoid confusion.

Su and Henckel (2022) propose to falsify interventional statements of a linear causal model by comparing the interventional distributions entailed by different adjustment sets. More precisely: in many cases a causal DAG G implies several sets $C_1, \dots, C_k$ for $k > 2$ satisfying the backdoor criterion Pearl (2009) such that

$$p(y \mid do(x)) = \sum_{c_i} p(y \mid x, c_i) p(c_i),$$

for all $i \in [k]$. Figure 1a is a (rather trivial) example, as we could use $\emptyset$ and $\{Z_1\}$. The set of parents of X is always such an adjustment set if the effect is identifiable (Tian and Pearl, 2002) in an ADMG. If G is the true causal model, we can decide whether a set C is a valid adjustment set by graphical criteria (see theorem 5). Su and Henckel (2022) now propose a statistical test to reject the null hypothesis H_0 that all sets under consideration yield the same interventional distribution. Precisely, for sets of variables $C_1, \dots, C_k$ the hypothesis reads

$$H_0 : \quad \beta_{X_i, X_j \cdot C_1} = \dots = \beta_{X_i, X_j \cdot C_k},$$

where $\beta_{X_i, X_j \cdot C_l}$ denotes the partial regression coefficient for X_i regressed on X_j and all variables in C_l for some $l \in [k]$. Their test assumes linear causal models and Gaussian noise.

We build on their work in the following sense: instead of using one causal model G to derive multiple adjustment sets, we use adjustment sets of *different* causal models, learned on different subsets of variables. Just as in their work, they should all yield the same interventional distributions. For simplicity, we use parent-adjustment for each marginal model in the case of RCD and the canonical adjustment set for FCI. Note, that in the case where all models are correct, each adjustment set in a marginal model will be a valid adjustment set in the joint model. But that the marginal parent sets are valid adjustment sets in the joint model is neither sufficient nor necessary for the test to accept.

A14.2 Further Practical Considerations

In the following we want to specify some aspects of definitions 9 and 10 that have not been discussed in detail and also slightly modify definition 10 further to tackle some practical issues.

Sampling of subsets. In definitions 9 and 10 we have assumed the subsets are given. In practice, we sample them randomly. But we did not sample from the set of all possible substes for the following reason: if a subset is very small, some algorithms like the FCI algorithm often give uninformative outputs in the sense that most edges are unoriented. On the other hand, if subsets are large and differ only by few nodes, the resulting marginal models usually do not differ much. Therefore, in all experiments, we uniformly drew subsets S_i from the set of subsets with $|S_i| = \lceil |V|/2 \rceil$ for $i = 1, \dots, l$ for some $l \in \mathbb{N}$.

Canonical adjustment and false positives In our experiments with FCI we noted that many tests indicated a difference in the interventional distributions, despite the fact that the graphical models seemed to be Markovian and graphically compatible. This seemed to be due to the fact that the canonical adjustment set is usually quite large, rendering the problem statistically hard. In cases where there was no possibly directed causal path (and therefore the model graphically implied no dependence for any adjustment set) we chose to replace the canonical adjustment set with a randomly chosen subset of size one.

Marginalisation of RCD outputs As we have noted before, the outputs of RCD are not ADMGs, as RCD cannot differentiate between the case, where a node X has a direct edge to Y *and* they have an unobserved hidden confounder or where they are confounded without any directed edge between them. During marginalisation we treated these bidirected edges accordingly, and drew additional edges if one node is a *potential* ancestor or an additional bidirected edge if two nodes are *potentially* confounded.

Causal sufficiency. In practice, many algorithms like the PC algorithm or GES rely on the assumption that all causally relevant variables are observed, i.e. that there is no common cause between two observed variables that is itself unobserved and all directed paths from this common cause to the observed nodes only contains unobserved nodes. Clearly, if we sample the subsets uniformly from all subsets with the same size, causal sufficiency will often be violated for these subsets, even if it holds for V. Consequently, if we detect an incompatibility between $\mathcal{A}(\mathbf{X})$ and $\mathcal{A}(\mathbf{X}_S)$ we cannot know whether this indicates an actual error or it is simply due to the newly introduced hidden confounder. Still, in section 4 we will show experiments with PC and GES that indicate that the graphical incompatibility score might help in model selection for these algorithms. The interventional criterion seems to be more sensitive to these violations of sufficiency.

CPDAGs and hidden confounders. In definition 25 we have only defined the latent projection of DAG to a CPDAG on causally sufficient subsets. But as we have just discussed, we will not only look at causally sufficient subsets. In the experiments, we simply calculated the latent ADMG and deleted its bidirected edges. The resulting graph is a DAG again and we proceeded with its respective CPDAG.

A14.3 Runtime of the Incompatibility Scores

Let $\mathcal{A}$ be an algorithm, k be the number of considered subsets, m be the number of samples, $f(n, m)$ be the worst-case run time of $\mathcal{A}$ on n nodes and m samples, $g(k, n, m)$ be the time of the test by Su and Henckel (2022) and $h(n)$ be the time to calculate the latent projection from a graph with n nodes. We then get a run time in

$$\mathcal{O}(f(n,m) + k \cdot f(\lceil n/2 \rceil, m) + n^2 \cdot (g(k,n,m) + k \cdot h(n))), \quad \mathcal{O}(f(n,m) + k \cdot (f(\lceil n/2 \rceil, m) + n^2 + h(n))),$$

for κ^I and κ^G respectively. With our simple implementation, $h(n) \in \mathcal{O}(n^3)$. As g is polynomial in k, n and m, the run time in our main experimental setting is dominated by f which is exponential in n for FCI and RCD.

A15 EXPERIMENTAL DETAILS

A15.1 Data Generation

For our first experiments, we generated synthetic data. We first sampled a random ground truth graph using the Erdos-Renyi model (with number of nodes $n + h$ and expected degree d, where h is the number of potential hidden confounders). For each node X we then define a functional model from the class of linear models. For the linear model we drew the parameters uniformly from $[-1, -0.1] \cup [0.1, 1]$. We then apply an additive noise term which is either drawn from a standard normal distribution or a uniform distribution with zero mean and unit variance. We then randomly pick d nodes as observed nodes and marginalise out the others.

In all experiments we used graphs with 10 nodes and expected degree 2, as well as linear Gaussian and linear uniform structural equations. For all incompatibility scores we drew 40 subsets uniformly from the set of all subsets of size 5. For experiments with RCD and FCI we set $h = 3$, otherwise $h = 0$. For all experiments we draw 1000 samples from the SCM.

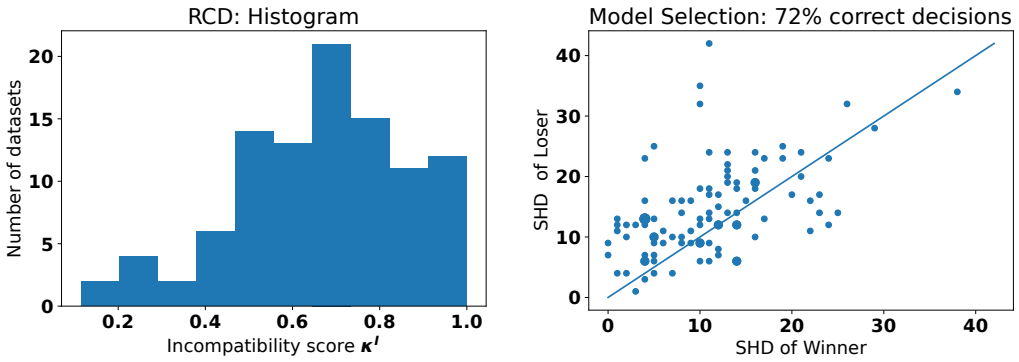


Figure 10: (Left) A histogram of κ^I values for RCD on 100 linear uniform datasets. (Right) The incompatibility score κ^I as metric for model selection with RCD and $\alpha = 0.1$ or $\alpha = 0.001$. We picked strictly better parameters for 68% of datasets and for 28% we picked strictly worse parameters. Overall, in 72% of datasets we picked better or equally good parameters.

A15.2 Other Details

We evaluated the SHD of CPDAGs and PAGs with respect to the respective CPDAG and PAG of the ground truth graph and for PAGs, according to the definition of Triantafillou and Tsamardinos (2016). For the test of Su and Henckel (2022) in the interventional compatibility score we always chose the confidence level 0.001. For RCD and FCI we also set all confidence thresholds to 0.001, unless stated otherwise. Similarly, for PC and GES we set the parameters of the algorithms to $\alpha = 0.01$ and $\lambda = 0.01$, respectively, unless stated otherwise. For all algorithms, we used the implementation from the `causal-learn` python package (Zheng et al., 2023).

The computations were done on an Intel Core i5-5200U CPU with 8 GB RAM or an Apple M1 Pro with 32 GB of RAM. All experiments can be run in less than a day.

A16 ADDITIONAL EXPERIMENTS

A16.1 Additional Plots

In the experiments in figures 2 and 3 we have studied the behaviour of the RCD algorithm and the interventional incompatibility score κ^I. The plots in figure 10 do not contain novel insights about the experiments. Yet, the visualization emphasises slightly different aspects. For comparability, we also report them for the following experiments. The "Winners" in the right plot are determined as follows:

$$\text{Winner} = \text{argmax}_{\mathcal{A} \in \{\text{RCD}^{\alpha=0.1}, \text{RCD}^{\alpha=0.001}\}} \kappa^I(\mathcal{A}, \mathbf{X}).$$

The loser are the respective other parameters. The defenitions of "Winners" and "Losers" in the following plots is analogous.

A16.2 Model Evaluation

In section 4 we have already seen that κ^I is correlated with the SHD to the ground truth graph for RCD on linear non-Gaussian data. As a next step, we repeated this experiment with the FCI algorithm, where we used linear Gaussian data and the Fisher Z test for conditional independence. In figure 12 we can see that this also yields a significant partial correlation (given the average node degree of the ground truth graph), albeit not as strong as for RCD. Recall, that we considered the partial correlation given the average node degree of the ground truth graph, as we suspect the density of the ground truth graph to affect both, the SHD and the incompatibility score.

We also repeated the experiments with the graphical score κ^G. Figures 14 and 15 show that we also get a significant correlation for the graphical criterion.

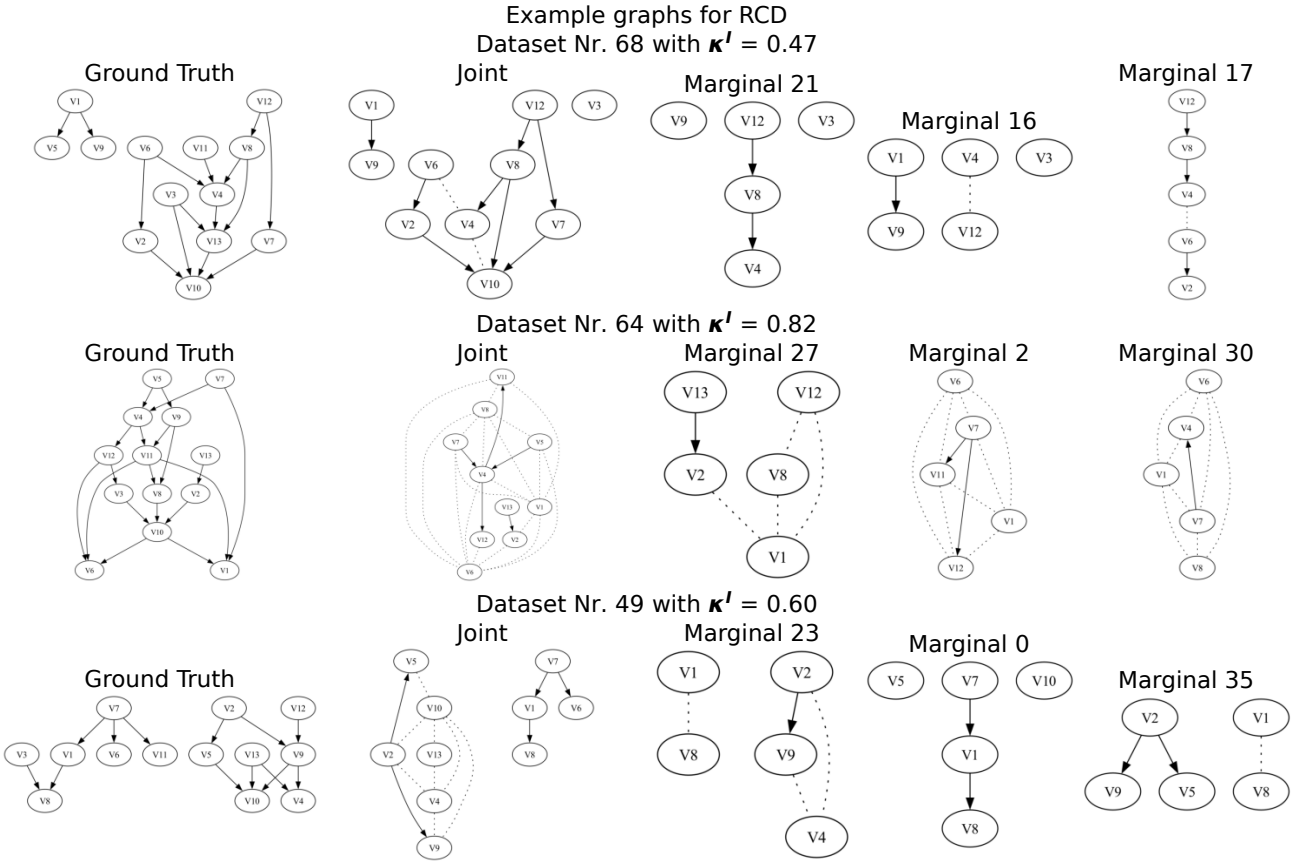


Figure 11: Randomly drawn example graphs from the experiment shown in figure 2. The figure shows for three randomly picked datasets the ground truth graph (including hidden variables), the joint graph found be the algorithm on all variables and some randomly drawn marginal graphs, i.e. graphs that the algorithm found on subsets of variables. Dotted edges indicate that RCD could not infer the direction of the edge.

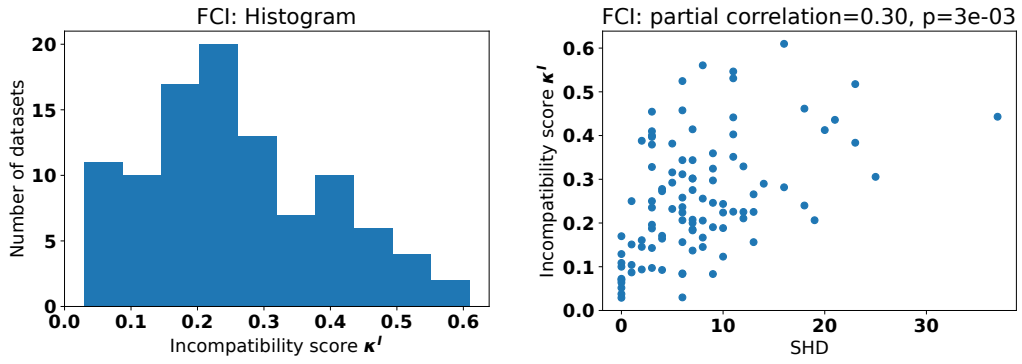


Figure 12: (Left) A histogram of κ^I values for FCI on 100 linear Gaussian datasets. (Right) The structural Hamming distance of estimated graphs $\hat{G}$ to the respective true graph G is on the x-axis and on the y-axis the incompatibility score κ^I. The figure shows a significant correlation between κ^I and the SHD.

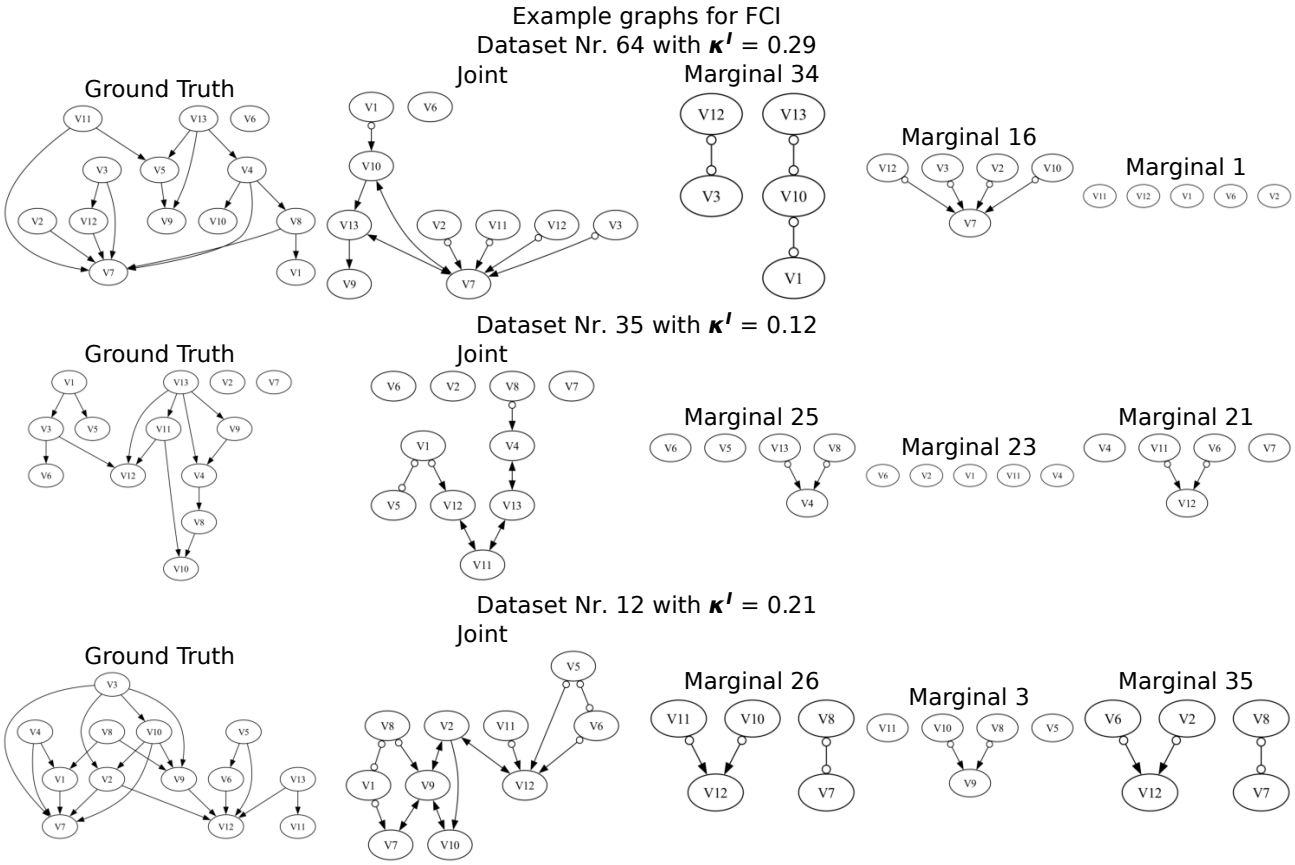


Figure 13: Randomly drawn example graphs from the experiment shown in figure 12. The figure shows for three randomly picked datasets the ground truth graph (including hidden variables), the joint graph found by the algorithm on all variables and some randomly drawn marginal graphs, i.e. graphs that the algorithm found on subsets of variables.

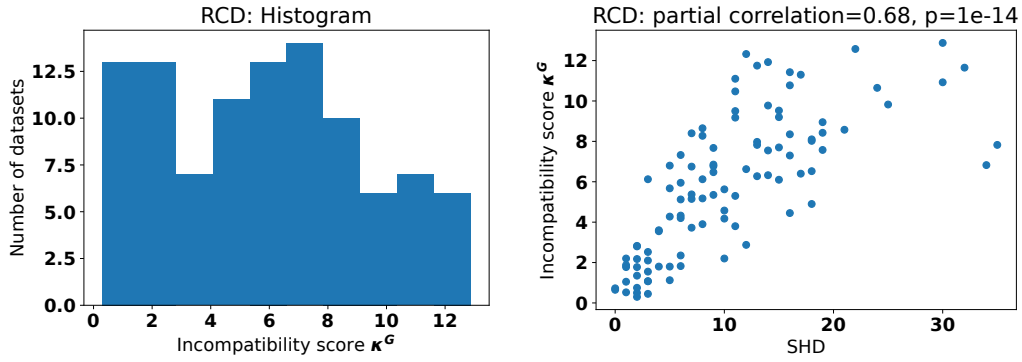


Figure 14: (Left) A histogram of κ^G values for RCD on 100 datasets with linear models and uniform noise. (Right) The structural Hamming distance of estimated graphs $\hat{G}$ to the respective true graph G is on the x-axis and on the y-axis the incompatibility score κ^G. The figure shows a significant correlation between κ^G and the SHD.

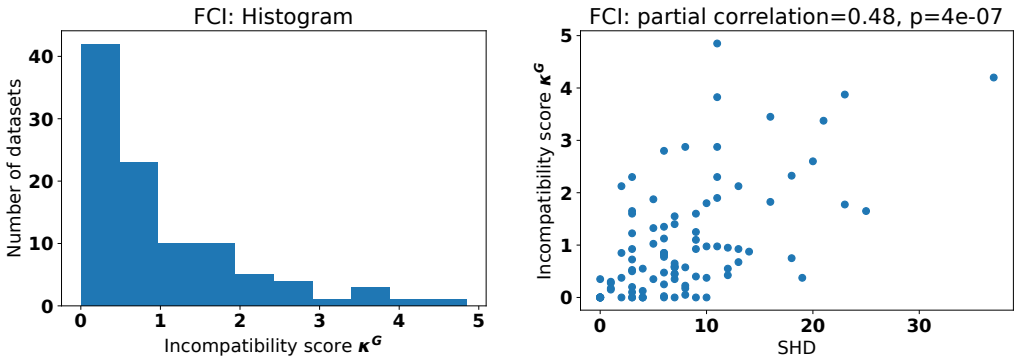


Figure 15: (Left) A histogram of κ^G values for FCI on 100 datasets with linear models and Gaussian noise. (Right) The structural Hamming distance of estimated graphs $\hat{G}$ to the respective true graph G is on the x-axis and on the y-axis the incompatibility score κ^G. The figure shows a significant correlation between κ^G and the SHD.

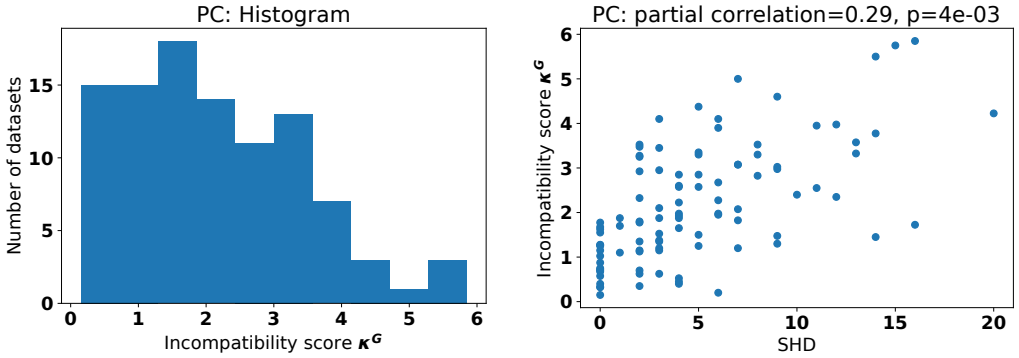


Figure 16: The structural Hamming distance of estimated graphs $\hat{G}$ to the respective true graph G is on the x-axis and on the y-axis the incompatibility score κ^G for PC.

In the next line of experiments we wanted to see whether this correlation also occurs in the setting where algorithms assume causal sufficiency and with the graphical criterion κ^G. We therefore generated 100 datasets (as described above) and estimated graphs $\hat{G}$ with two different causal discovery methods, namely PC and GES. In figures 16 and 17 we see again that this yields a significant partial correlation. In figures 18 and 19 we can see that also the correlation between κ^G and the bounds in the structural Interventional distance (SID) (Peters and Bühlmann, 2015) is significant. (Recall that SID is only defined for DAGs and Peters and Bühlmann (2015) proposed to calculate the bounds on the SID over all DAGs in the equivalence class described by a CPDAG). Yet, figure 20 shows no correlation between SHD and κ^I (in contrast to our experiments with RCD and FCI in figures 2 and 12). This seems to suggest that κ^I might not be suitable to be used with algorithms that assume causal sufficiency without further modifications.

A16.3 Model Selection

We repeated the experiment from figure 3 with the FCI algorithm. The FCI algorithm has exactly one hyper-parameter, namely the threshold α below which the p-values of the conditional independence tests lead to a rejection of the null hypothesis. Again, we chose between $\alpha = 0.1$ and $\alpha = 0.001$. Figure 23 shows that we picked the better (or equally good) parameter in 76 % of the datasets. Precisely, 62% are strictly better, while 24% are strictly worse.

Figure 24 shows the same experiment as in figure 3 but this time with the graphical score κ^G. In this setting we pick the strictly better parameters in 69% of the datasets and a strictly worse parameters in 27%.

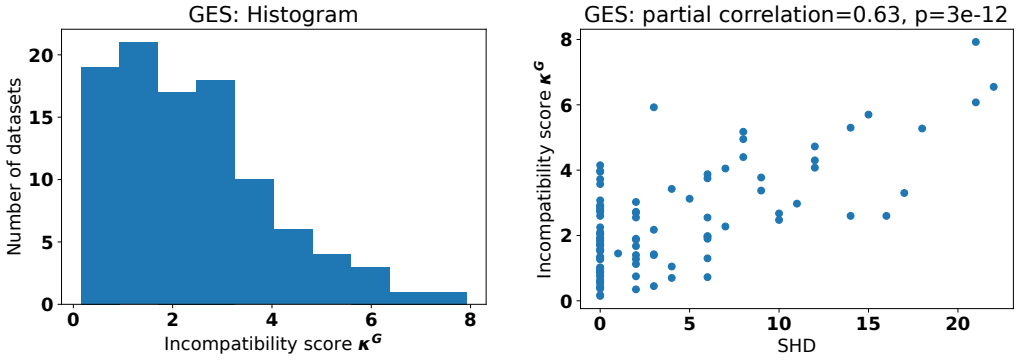


Figure 17: (Left) A histogram of κ^G values for GES on 100 datasets with linear models and Gaussian noise. (Right) The structural Hamming distance of estimated graphs $\hat{G}$ to the respective true graph G is on the x-axis and on the y-axis the incompatibility score κ^G.

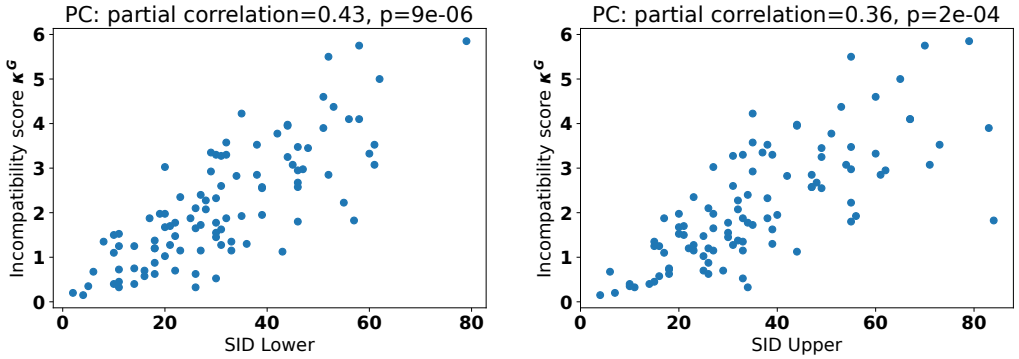


Figure 18: The bounds on the structural interventional distance of estimated CPDAGs $\hat{G}$ to the respective true graph G is on the x-axis and on the y-axis the incompatibility score κ^G for PC. The plot shows a significant correlation between SID and κ^G.

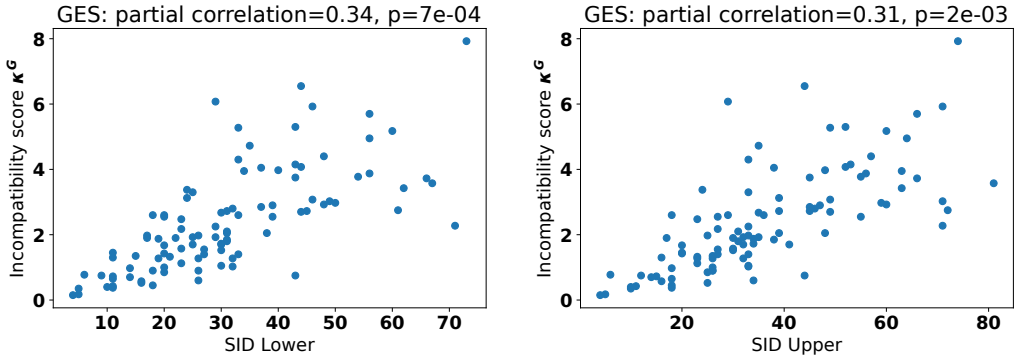


Figure 19: The bounds on the structural interventional distance of estimated CPDAGs $\hat{G}$ to the respective true graph G is on the x-axis and on the y-axis the incompatibility score κ^G for GES. The plot shows a significant correlation between SHD and κ^G.

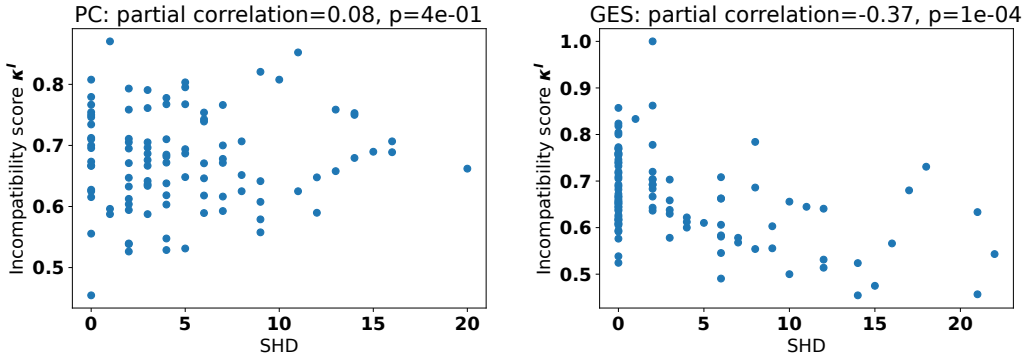


Figure 20: The structural Hamming distance of estimated graphs $\hat{G}$ to the respective true graph G is on the x-axis and on the y-axis the incompatibility score κ^I for PC and GES. In contrast to e.g. figures 2 and 12 we do not see a positive correlation.

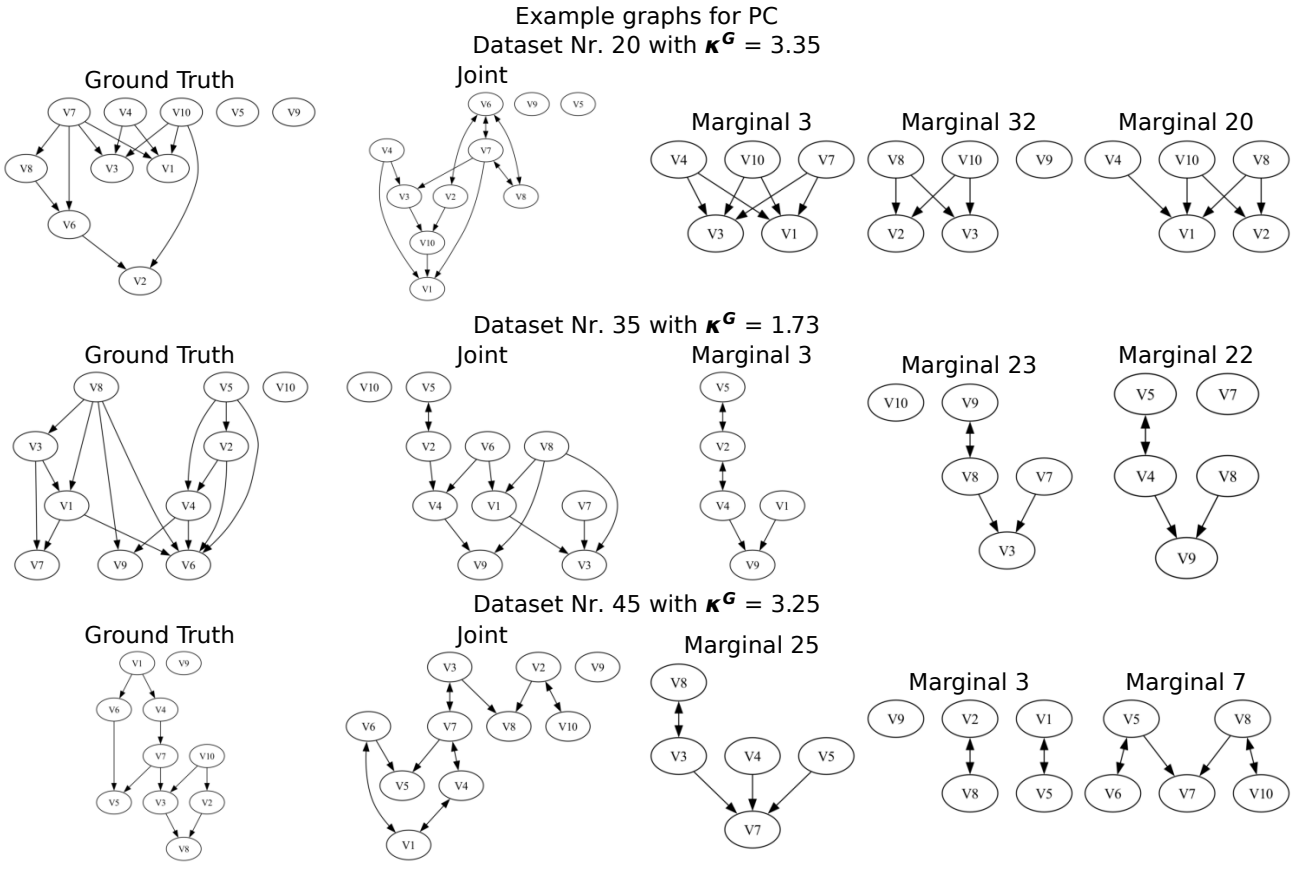


Figure 21: Randomly drawn example graphs from the experiment shown in figure 16. The figure shows for three randomly picked datasets the ground truth graph, the joint graph found by the algorithm on all variables and some randomly drawn marginal graphs, i.e. graphs that the algorithm found on subsets of variables.

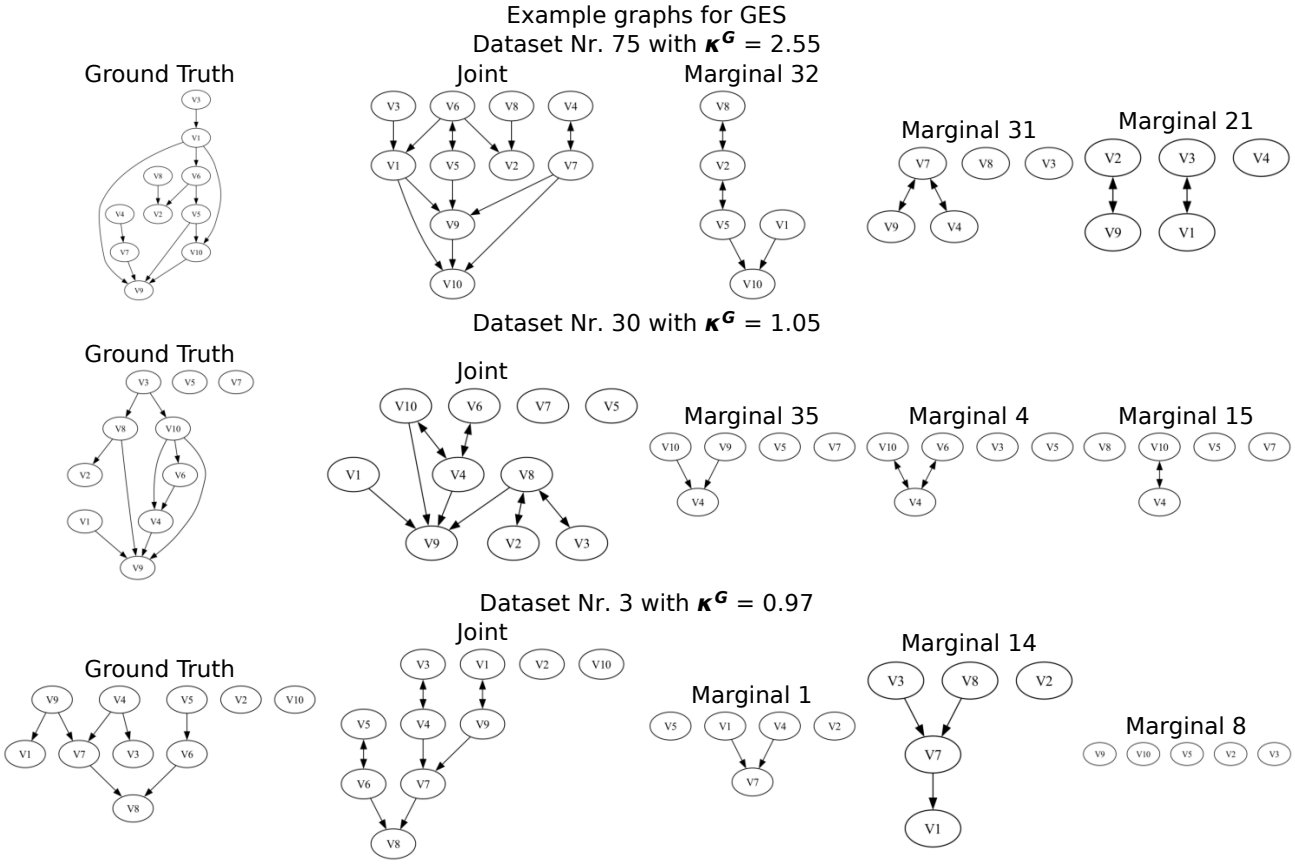


Figure 22: Randomly drawn example graphs from the experiment shown in figure 16. The figure shows for three randomly picked datasets the ground truth graph, the joint graph found by the algorithm on all variables and some randomly drawn marginal graphs, i.e. graphs that the algorithm found on subsets of variables.

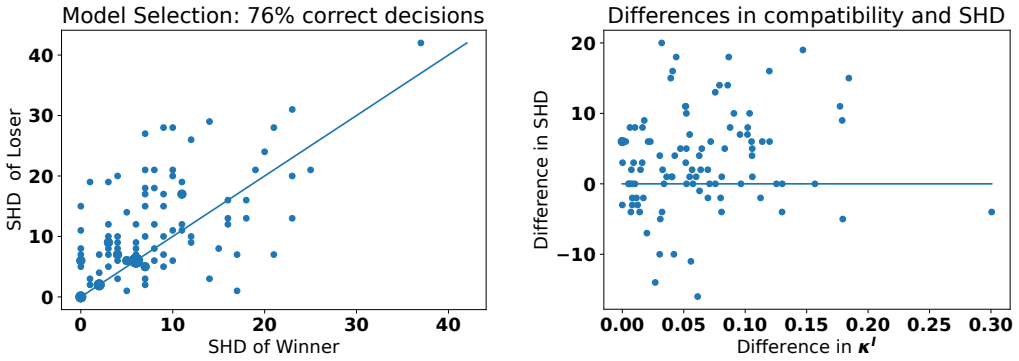


Figure 23: The incompatibility score κ^I as metric for model selection FCI with $\alpha = 0.1$ and $\alpha = 0.001$ analogously to the main paper. We picked the model with the better κ^I score as winner and report the SHD of the winner and of the loser on x-axis and y-axis respectively. In 62% of datasets we picked the strictly better model, while 24% are strictly worse.

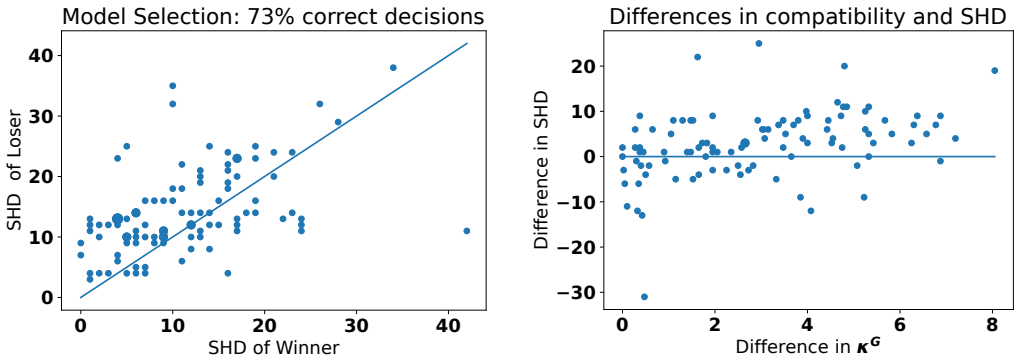


Figure 24: The incompatibility score κ^G as metric for model selection RCD with $\alpha = 0.1$ and $\alpha = 0.001$ analogously to the main paper. We picked the model with the better κ^G score as winner and report the SHD of the winner and of the loser on x-axis and y-axis respectively. In 69% of datasets we picked the strictly better model, while 27% are strictly worse.

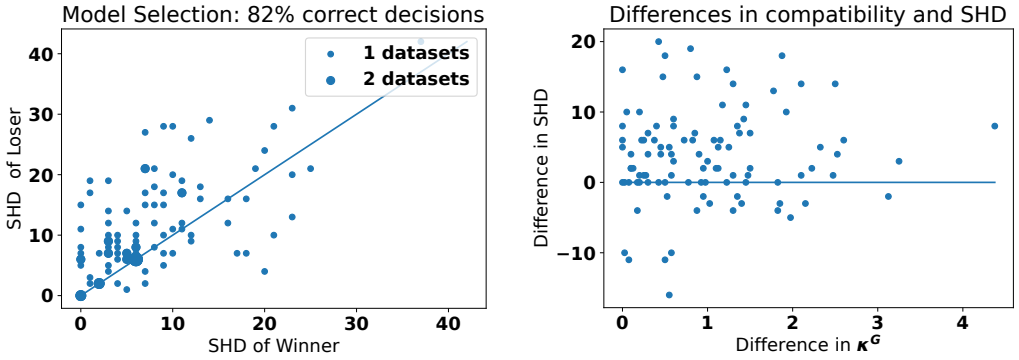


Figure 25: The incompatibility score κ^G as metric for model selection FCI with $\alpha = 0.1$ and $\alpha = 0.001$ analogously to the main paper. We picked the model with the better κ^G score as winner and report the SHD of the winner and of the loser on x-axis and y-axis respectively. In 68% of datasets we picked the strictly better model, while 18% are strictly worse. The dot size indicates that several points overlap.

In figure 25 we can see the same experiment with FCI and κ^G. Here we got the strictly better parameter in 68% of datasets and the strictly worse parameter in 18%.

Again, we also used PC and GES as algorithms that assume causal sufficiency. Analogously to FCI, PC has only the α-threshold as parameter. We repeated an analogous experiment as above with the PC algorithm with $\alpha = 0.1$ and $\alpha = 0.001$. The plots in figure 26 suggest that again, the incompatibility score is an effective selection criterion and here we even make correct decisions 73% of the cases.

Eventually, we used the graphical score κ^G to select between different *algorithms*, in contrast to the previous experiments where we picked hyperparameters. Note, that such a comparison would not be possible between FCI and RCD, as the scores of PAGs and ADMGs are not a priori comparable. But PC and GES both output CPDAGs, which is why we chose them for these experiments. Figure 29 does not show a similarly good performance as before, as we chose the strictly better or algorithm in 43% of datasets, while we picked the worse algorithm in 25%. Analogously, figure 30 shows that we picked the strictly better model w.r.t. the lower bound on the SID given by the CPDAG in 34% of the cases and the strictly worse model in 27% of the cases. Although w.r.t. to the upper bound on the SID we picked a worse model in 31% of the cases and the better model only in 28% of the datasets. Figure 32 seems to suggest that κ^I cannot be used for model selection with algorithms that assume causal sufficiency (without further modifications): 60% of the points are on or above the line, but only 28% are strictly above the line and 40% are strictly below the line. This is in line with what we expected after the experiment shown in figure 20.

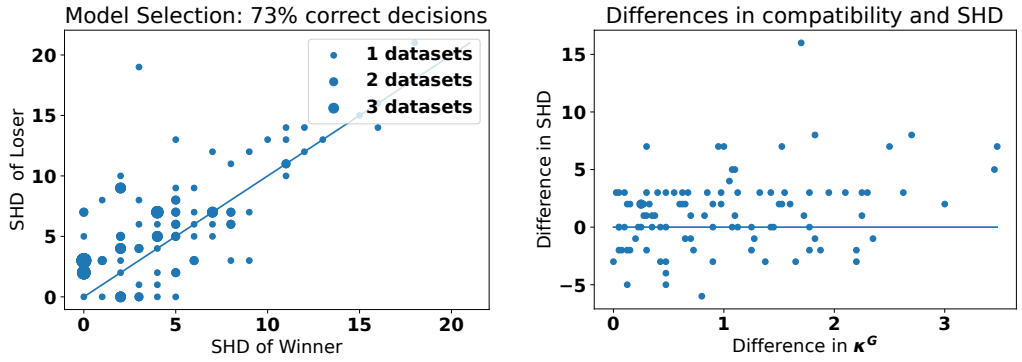


Figure 26: We used the incompatibility score κ^G for model selection of the PC algorithm with $\alpha = 0.1$ and $\alpha = 0.001$. We picked the strictly better parameter in 58% of datasets and the strictly worse in 27%. The dot size indicates that several points overlap.

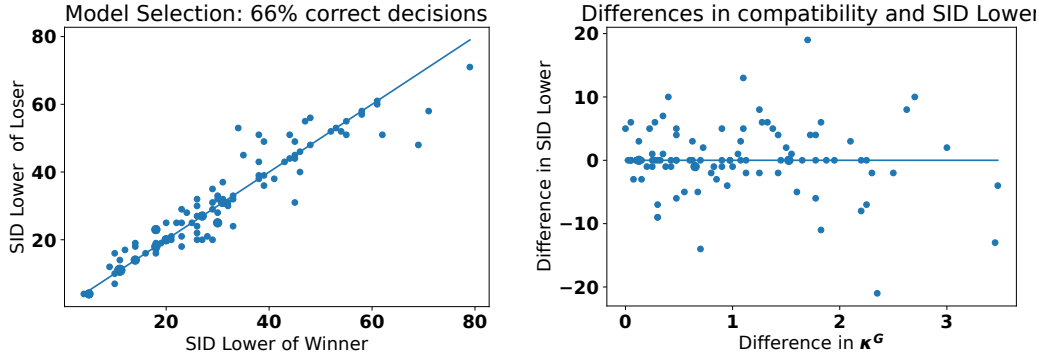


Figure 27: We used the incompatibility score κ^G for model selection of the PC algorithm with $\alpha = 0.1$ and $\alpha = 0.001$. We picked the strictly better parameter w.r.t. the lower bound on the SID in 33% of datasets and the strictly worse in 34%.

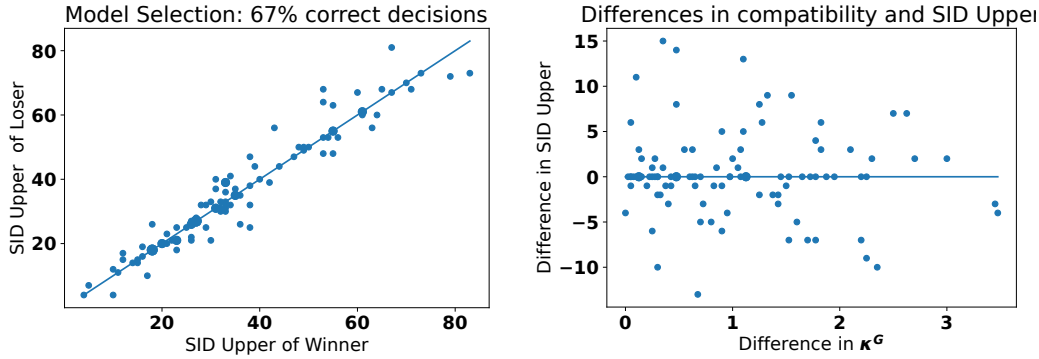


Figure 28: We used the incompatibility score κ^G for model selection of the PC algorithm with $\alpha = 0.1$ and $\alpha = 0.001$. We picked the strictly better parameter w.r.t. the upper bound on the SID in 32% of datasets and the strictly worse in 33%.

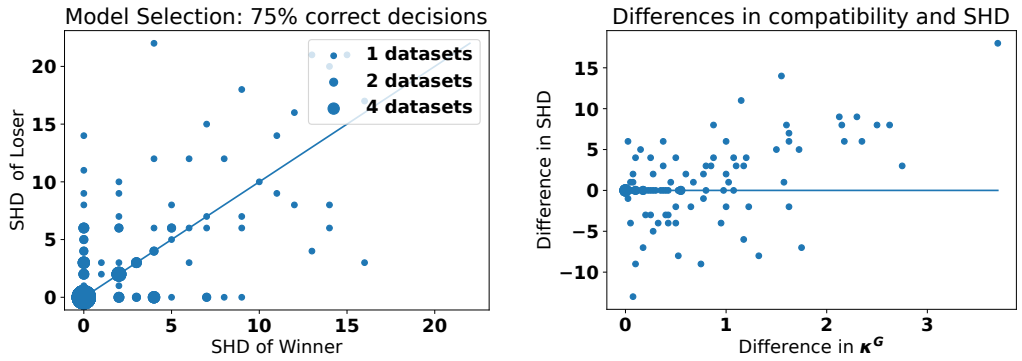


Figure 29: The incompatibility score κ^G as metric for model selection between PC and GES analogously to the main paper. We picked the model with the better κ^G score as winner and report the SHD of the winner and of the loser on x-axis and y-axis respectively. We picked the strictly better parameter w.r.t. SHD in 43% of datasets and the strictly worse in 25%. The dot size indicates that several points overlap.

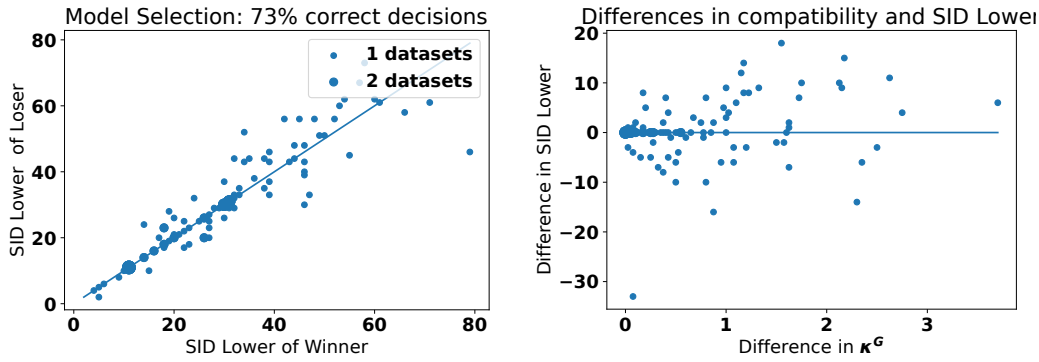


Figure 30: The incompatibility score κ^G as metric for model selection between PC and GES analogously to the main paper. We picked the model with the better κ^G score as winner and report the lower bound for the SID of the winner CPDAG and of the loser on x-axis and y-axis respectively. We picked the strictly better parameter w.r.t. the lower bound on the SID in 34% of datasets and the strictly worse in 27%. The dot size indicates that several points overlap.

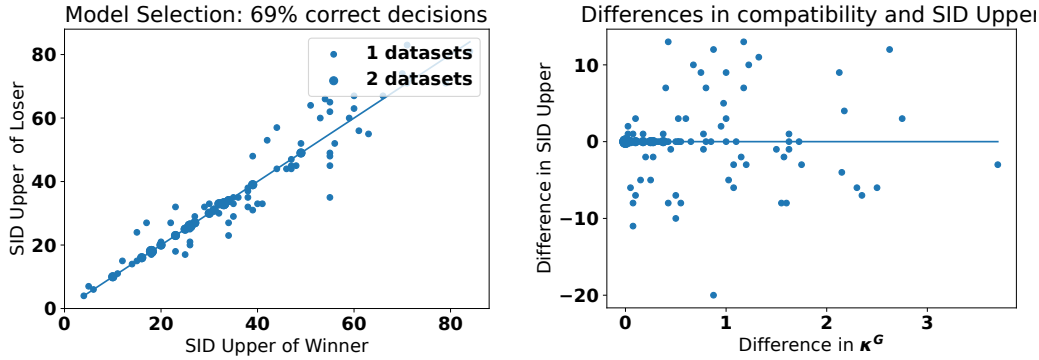


Figure 31: The incompatibility score κ^G as metric for model selection between PC and GES analogously to the main paper. We picked the model with the better κ^G score as winner and report the upper bound for the SID of the winner CPDAG and of the loser on x-axis and y-axis respectively. We picked the strictly better parameter w.r.t. the upper bound on the SID in 28% of datasets and the strictly worse in 31%. The dot size indicates that several points overlap.

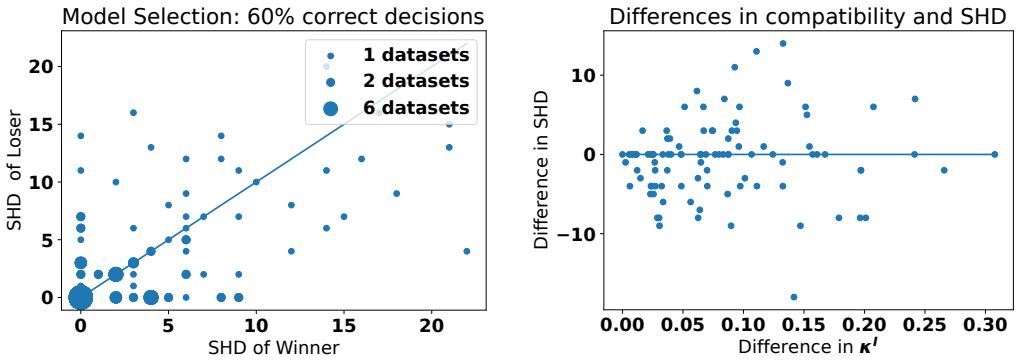


Figure 32: The incompatibility score κ^I as metric for model selection between PC and GES analogously to the main paper. We picked the strictly better parameter w.r.t. to SHD in 28% of datasets and the strictly worse in 40%. The dot size indicates that several points overlap.

A16.4 Additional Real Data

We applied our incompatibility scores to the dataset presented by Sachs et al. (2005).

We compared the SHD of $\mathcal{A}(\mathbf{X})$ for all algorithms $\mathcal{A}$ and also the F_1 score with respect to the existence of an edge in the skeleton of the resulting graph. It is worth noting that the SHD (and also the incompatibility scores) of different model types like ADMGs, PAGs and CPDAGs are not directly comparable and can only give an intuition about the relative performance of the algorithms. For the algorithms using the kernel independence test (KCI), we also randomly subsampled 1000 datapoints to speed up the computation. For all algorithms we picked $\alpha = 0.01$.

We suspected the real dataset might contain confounding. Therefore we started by using an algorithm that does not assume causal sufficiency. As the LiNGAM-based method does not merely return a Markov equivalence class, we picked the RCD algorithm first. We used both incompatibility scores κ^I and κ^G. As we can see in table 1, the graphical incompatibility score is in a medium to high range, compared to the results in figure 14 (i.e. compared to the setting where we know the true SHD). This lead us to the conclusion that possibly either linearity or the non-Gaussian additive noise assumption are violated. (Note, that the interventional score is 0, i.e. the best score possible. Looking at figure 33, this is probably as in the joint graph, no interventional probability is identifiable. This shows that compatibility alone does not suffice as criterion, but one also needs to account for how much an algorithm "commits" to falsifiable statements.) We therefore tried FCI with the correlation-based Fisher Z test and used again the interventional score κ^I and the graphical score κ^G. Again, the graphical score is not really low and now additionally the interventional score seems to be quite high, compared to the histograms in figures 10 and 15. So as a third attempt, we tried FCI with the kernel-based independence test proposed by Zhang et al. (2011) and (as we cannot use κ^I in a non-linear setting[21]) we only report κ^G. This yielded an incompatibility score of zero. We additionally wanted to try the PC algorithm, despite the fact that it assumes causal sufficiency. This again led to a good score, although not as good as the result of FCI with KCI. So in this case, the incompatibility scores would have directed us towards the models with the best F_1 score and the one with the second best SHD. But as the best SHD of 20 (which is still comparably high) shows, the good incompatibility score is not enough to *guarantee* a good performance.

[21] The goal of our self-compatibility is to find out if the assumptions of a causal discovery algorithm are violated to the extent that the output of the algorithm is changed non-negligibly. So if we already use an algorithm that does not assume linear dependencies, it is not clear what information we would gain from conducting a test that relies on linearity.

Table 1: Comparison of causal discovery algorithms on cell dataset

	RCD	**FCI + Fisher Z**	**FCI + KCI**	**PC + KCI**
κ^I	0.0	0.82	-	-
κ^G	6.65	6.5	0.0	0.68
SHD	84	62	24	**20**
Skeleton F_1	0.5	0.49	**0.62**	**0.62**

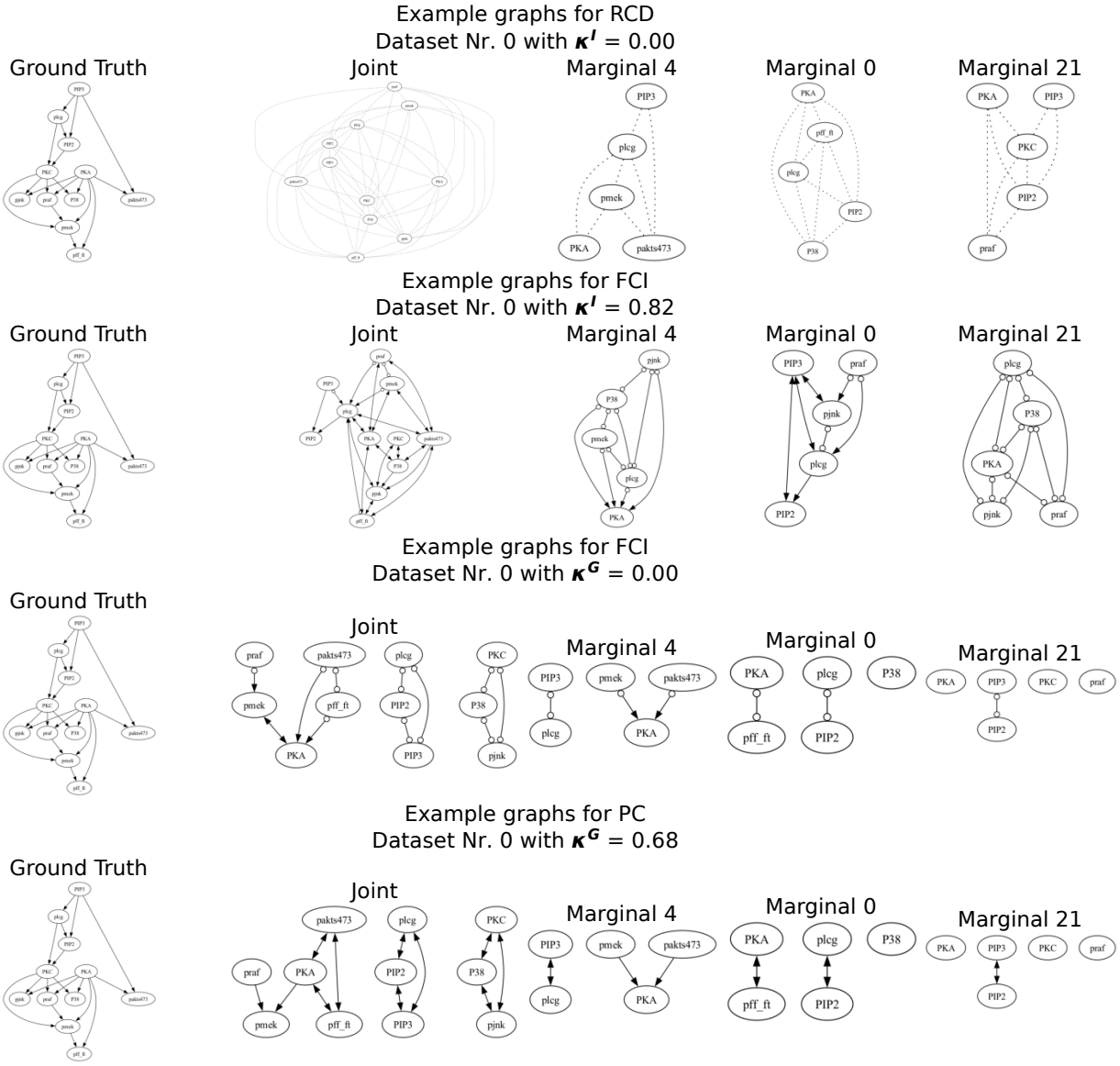


Figure 33: Randomly drawn example graphs from the experiment on the cell dataset. The figure shows the ground truth graph, the joint graph found by the algorithm on all variables and some randomly drawn marginal graphs, i.e. graphs that the algorithm found on subsets of variables. The algorithms are (from top to bottom) RCD, FCI with Fisher Z test, FCI with kernel independence test and PC with kernel independence test.