Proceedings of Machine Learning Research 275:1–33, 2025 4th Conference on Causal Learning and Reasoning

Cross-validating causal discovery via Leave-One-Variable-Out

Daniela Schkoda DANIELA.SCHKODA@TUM.DE
Technical University of Munich, Germany

Philipp M. Faller PHILIPP.FALLER@PARTNER.KIT.EDU
Karlsruhe Institute of Technology, Germany
Amazon Research Tübingen, Germany

Patrick Blöbaum BLOEBP@AMAZON.COM
Amazon Research Tübingen, Germany

Dominik Janzing JANZIND@AMAZON.COM
Amazon Research Tübingen, Germany

Editors: Biwei Huang and Mathias Drton

Abstract

We propose a new approach to falsify causal discovery algorithms without ground truth, which is based on testing the causal model on a variable pair excluded during learning the causal model. Specifically, given data on $X, Y, \boldsymbol{Z} = X, Y, Z_1, \ldots, Z_k$, we apply the causal discovery algorithm separately to the 'leave-one-out' data sets $X, \boldsymbol{Z}$ and $Y, \boldsymbol{Z}$. We demonstrate that the two resulting causal models, in the form DAGs, ADMGs, CPDAGs or PAGs, often entail conclusions on the dependencies between X and Y and allow to estimate $\mathbb{E}(Y \mid X = x)$ without any joint observations of X and Y, given only the leave-one-out datasets. This estimation is called "Leave-One-Variable-Out (LOVO)" prediction. Its error can be estimated since the joint distribution $P(X, Y)$ is available, and X and Y have only been omitted for the purpose of falsification.

We present two variants of LOVO prediction: One graphical method, which is applicable to general causal discovery algorithms, and one version tailored towards algorithms relying on specific a priori assumptions, such as linear additive noise models. Simulations indicate that the LOVO prediction error is indeed correlated with the accuracy of the causal outputs, affirming the method's effectiveness.

Keywords: Out of variable generalization, Benchmarking causal models, Benchmarking without ground truth.

1. Introduction

Causal discovery (Spirtes et al., 1993), the inference of (typically acyclic) causal graphs from observational data, has attained substantial research interest since the development of the PC algorithm (Spirtes et al., 1993; Glymour et al., 2019), which leverages the causal Markov condition and faithfulness assumption. Research gained further momentum after it was observed that additional assumptions render identification solvable also within Markov equivalence classes, see, e.g., Kano and Shimizu (2003); Shimizu et al. (2006); Sun et al. (2006); Hoyer et al. (2008); Zhang and Hyvärinen (2009); Peters et al. (2011); Kocaoglu et al. (2017); Gnecco et al. (2021); Rolland et al. (2022); Montagna et al. (2023). These approaches are meanwhile complemented by supervised learning methods: Lopez-Paz et al. (2015) treats cause-effect inference as a binary classification problem, Nauta et al. (2019); Lachapelle et al. (2020); Zheng et al. (2020); Ke et al. (2023) use

© 2025 D. Schkoda, P.M. Faller, P. Blöbaum & D. Janzing.

techniques from deep learning to learn multivariate causal graphs using architectures tailored for learning properties of probability distributions. Further, it has been shown that data from changing environments helps in identification of causal models (Tian and Pearl, 2001; Peters et al., 2016; Zhang et al., 2017; Mooij et al., 2020; Rothenhäusler et al., 2021). However, even after decades of creative contributions, it is fair to say that causal discovery did not experience any widely celebrated breakthroughs in practical applications despite interesting success stories, e.g., Shen et al. (2020); Lagemann et al. (2023). One reason, if not the main one, is that researchers working on practical applications find it hard to judge which method works best or if any works reasonably well for their use case. Extensive evaluations are mainly performed on simulated data, a practice about which serious doubts are in order (Reisach et al., 2021). Despite the existence of instructive examples for interventional data (Lagemann et al., 2023), such as knockout experiments in genetics (Hamilton et al., 1989), it is a serious limitation to entirely rely on interventions. This is not only because interventional experiments are expensive but also because interventions can not necessarily be attributed to single nodes (Zhang et al., 2023). Thus, some datasets may require a separate discussion about what node is intervened on, which motivated research on learning causal graphs from unknown intervention targets (Jaber et al., 2020). Further, "ground truth" reported in the literature (Sachs et al., 2005) has later been challenged elsewhere (Mooij et al., 2020, Section 5.8). In other words, despite all the interesting experimental data sets, automatic retrieval of a vast number of interventional data for trustworthy benchmarking seems currently out of reach. In practical applications, researchers often solve causal inference tasks like treatment effect estimation in directed acyclic graphs (DAGs) with comparable low complexity and raise doubts about whether more complex DAGs can be trusted (Imbens, 2020).

Motivated by the lack of benchmarking data, Faller et al. (2024) suggests a "self-compatibility" check, which applies causal discovery algorithms to subsets of variables and quantifies to what extent the algorithm's outputs on subsets contradict the output on the entire set of variables. The paper discusses different measures of disagreement. However, *some* disagreement is not unlikely, and the paper provides no guidance on which level of disagreement is large enough to reject the algorithm's outcome. Given data on variables $\boldsymbol{W} = (X, Y, \boldsymbol{Z}) = (X, Y, Z_1, \ldots, Z_k)$, similar to Faller et al. (2024), our method begins with applying the causal discovery algorithm in question to subsets of variables, namely $X, \boldsymbol{Z}$ and $Y, \boldsymbol{Z}$. The novelty lies in introducing a new measurement: We demonstrate that the causal models on the two subsets often allow us to infer $\mathbb{E}[Y|X = x]$ only based on the two datasets $X, \boldsymbol{Z}$ and $Y, \boldsymbol{Z}$. Comparing this to an estimate obtained from the joint distribution $P(X, Y)$, we obtain the LOVO prediction error. Unlike the discrete metrics in Faller et al. (2024)), this prediction error is continuous and, therefore, constitutes an optimizable loss function. Moreover, by focusing on a bivariate relation, we ensure a statistically well-behaved target. Most importantly, we establish a rigorous criterion to decide when to reject the algorithm's output by developing a baseline to which we compare the prediction error.

While our task can be seen as a missing data problem (for which causal structure is known to enable better imputation, see e.g. Mohan and Pearl (2021)), here, missingness comes from dropping on purpose for testing causal hypotheses. We call our scenario "Leave-One-Variable-Out (LOVO)" cross-validation in analogy to leave-one-out (LOO) cross-validation in statistical learning (Stone, 2018). While LOO cross-validation evaluates a model prediction for a *datapoint* that has not been used for learning, we test the prediction of the causal model at a *variable pair* (X, Y) whose relation has not been used for learning. We will discuss conditions under which causal models render this task solvable. In other words, while statistical learning is based on an inductive bias that allows

to interpolate a function at a point that has not been seen before, causal learning may "interpolate" dependencies between variable pairs that have not been seen together. This way, we further elaborate on the view of previous works (Tsamardinos et al., 2012; Dhir and Lee, 2019; Gresele et al., 2022; Guo et al., 2024) that causal models help for generalization across variables, called "out-of-variable generalization" in Guo et al. (2024), with the difference that we predict a statistical relation that is actually known and only temporarily dropped for the purpose of testing. One of the early works that observed that causal models enable predicting relations between variables not observed together is Tsamardinos et al. (2012), where Maximal Ancestral Graphs (MAGs) on overlapping subsets of variables are used to infer dependencies between variables, each of which only occurs in one of the subsets.

Structure of the paper: After Section 2 formally defines LOVO prediction and lays out the general structure of LOVO cross-validation, in Section 3, we examine under which conditions causal models enable LOVO prediction and propose a practical estimation method. Section 4 defines the "non-causal baseline," i.e., a simple LOVO prediction rule to be used when nothing is known about the causal structure. Section 5 reports experiments where we apply LOVO prediction to the causal discovery algorithms DirectLiNGAM (Shimizu et al., 2011) and Repetitive Causal Discovery (Maeda and Shimizu, 2020). All proofs are given in the appendix.

Notation and technical assumptions: To simplify mathematical discussions and notation, we will always assume that the joint distribution $P(X, Y, \boldsymbol{Z})$ has a density (the probability mass function for the case of discrete variables) with respect to a product measure. Further, except for results that explicitly refer to continuous variables, we use discrete sum over probabilities without being explicit about replacing them with integrals over densities otherwise. For standard concepts of causal discovery like the causal Markov condition, d-separation, and Markov equivalence, we refer to the literature, e.g., Spirtes et al. (1993); Pearl (2000).

2. Building blocks for LOVO cross-validation

We interpret LOVO prediction as the task of inferring $P(Y|X)$, or the reduced problem of inferring the regression function $f(x) = \mathbb{E}[Y|X = x]$, or the correlation ρ_{XY}, from $P(X, \boldsymbol{Z})$ and $P(Y, \boldsymbol{Z})$. This prediction is used to assess the reliability of a causal discovery algorithm as follows: Given data on the whole set of variables $\boldsymbol{W}$, we select one pair of variables (X, Y) and create the two leave-one-out datasets $(X, \boldsymbol{Z})$ and $(Y, \boldsymbol{Z})$. Next, we run the causal discovery method separately on both sets. As we show later, the outputs G_X, G_Y often provide information about the relation between X and Y, aiding in LOVO prediction. Correspondingly, we call this specific way of LOVO prediction based on G_X, G_Y *causally informed LOVO prediction*. Its prediction error is then estimated by comparing it to an estimate obtained from the joint distribution $P(Y, X)$. To obtain the overall LOVO cross-validation error, we repeat the procedure for all choices of pairs (X, Y) from $\boldsymbol{W}$ and take the average. Finally, to decide whether the error is still acceptable or so large that we should reject the outcome of the causal discovery algorithm, we compare the cross-validation error to the error of a baseline LOVO predictor, which estimates $P(Y|X)$ from $P(X, Z)$ and $P(Y, Z)$ without using any causal information.

Since we perform causal inference on subsets of the whole dataset, we need a class of causal graphs that enables marginalizations. Following Faller et al. (2024), we use acyclic directed mixed graphs (ADMGs) (Richardson, 2003) which contain the usual causal edges $\rightarrow$ as well as edges

$\leftrightarrow$ (indicating a confounding path that can not be blocked by any observed variable). There exist slightly different definitions of ADMGs across the literature, varying in whether to allow co-occurrence of both types of edges as confounded causal links $A \overset{\leftrightarrow}{\to} B$. We allow these confounded causal links[1] and say that a child B of A is a confounded child, if $A \overset{\leftrightarrow}{\to} B$, and an unconfounded child if $A \to B$ is the sole edge between the two nodes, similarly for parents. If $A \leftrightarrow B$, we say that A and B are siblings. We denote the children, parents, and siblings of a node A by $\mathrm{ch}(A), \mathrm{pa}(A)$, and $\mathrm{sib}(A)$, respectively. Moreover, we use the symbol $-$ to denote any type of directed or bidirected edge, and $A \not- B$ to indicate that A and B are not connected by an edge. For details of marginalization in ADMGs, we refer to Richardson (2003), but it is rather intuitive: directed paths $A \to * \to B$ turn into edges $A \to B$ when marginalizing over the mediators, confounding paths $A \leftarrow * \to B$ or $A \leftarrow * \leftrightarrow B$ into bidirected edges $A \leftrightarrow B$ when marginalizing over the common cause. By G, we refer to the DAG or ADMG for the entire set of nodes $\boldsymbol{W}$[2], and G_X, G_Y are its marginalizations when leaving out Y, X.

3. Constructing LOVO predictors via causal discovery

3.1. Connection of causality and LOVO prediction

We first consider a small toy scenario on three variables (X, Y, Z), whose joint graph is assumed to be a DAG, to provide an intuition under which conditions causally informed LOVO prediction is feasible. Alongside, we illustrate that it makes sense to proceed by the following steps:
1. Infer the joint DAG (or later ADMG) G from the two marginal graphs G_X, G_Y.
2. Use the joint graph G to reconstruct $P(X, Y)$ from $P(X, Z)$ and $P(Y, Z)$.
Suppose a causal discovery algorithm yields the outputs

$$G_X = (X \to Z), \quad G_Y = (Z \to Y)$$

when applied to the leave-one-out subsets $(X, Z), (Y, Z)$. First, we want to explore if these outputs allow us to draw conclusions about the entire graph. Employing the marginalization rules mentioned earlier, we find that the edge $X \to Z \in G_X$ can arise if and only if $X \to Z$, $X \to Y \to Z$, or both structures are contained in G. Since the second option contradicts $G_Y = (Z \to Y)$, we conclude $X \to Z \in G$. Similarly, we obtain that $Z \to Y \in G$. Lastly, we need to check if G could contain other edges: By acyclicity, the only potential additional edge is $X \to Y$. However, if $X \to Y \in G$, then G_Y would be $Z \overset{\rightarrow}{\leftrightarrow} Y$. Knowing G, we can conclude $X \perp\!\!\!\perp Y \mid Z$, and therefore,

$$P(X, Z, Y) = P(X, Z)P(Y|Z),$$

which determines $\mathbb{E}[Y|X = x]$. Thus, if the true underlying graph is $X \to Z \to Y$, we are able to construct a LOVO predictor. While we use a conditional independence statement entailed by G here, we want to stress that not every LOVO predictor is based on conditional independencies. Theorem 5 or Table 1 in the supplement contain cases where it can be inferred despite $X \not\perp\!\!\!\perp Y \mid \boldsymbol{Z}$, using other statistical properties entailed by the joint causal model. If, for instance, the joint DAG is $X \to Y \to Z$, linear models enable the identification of $\mathbb{E}[Y|X = x]$. Moreover, in the preceding step to infer G from G_X, G_Y, we heavily make use of the arrows' directions. If, in the example above, we had only gained knowledge on the skeletons of G_X, G_Y, the joint graph could be any

1. Nonetheless, our approach can be adapted to the other definition, see Supplement B.1.
2. This implicitly assumes that the joint data is Markov to some ADMG.

graph in which Z is not an isolated node. This is not special to this example; instead inferring $\mathbb{E}[Y|X = x]$ can not follow from the conditional independencies observed in $P(X, \boldsymbol{Z})$ and $P(Y, \boldsymbol{Z})$ via any *mathematical laws* (except for degenerate cases, e.g, when $\boldsymbol{Z}$ uniquely determines X, Y), as formalized in the following lemma:

Lemma 1 (No probabilistic law enables LOVO prediction) *Let X, Y be real-valued variables whose conditional distributions $P(X|\boldsymbol{Z} = \boldsymbol{z})$, $P(Y|\boldsymbol{Z} = \boldsymbol{z})$ have densities $p(x|\boldsymbol{z})$ and $p(y|\boldsymbol{z})$ with respect to the Lebesgue measure. Let $\boldsymbol{Z} = \{Z_1, \ldots, Z_k\}$ be variables with arbitrary range. Then $P(X, \boldsymbol{Z})$ and $P(Y, \boldsymbol{Z})$ can never uniquely determine $P(X, Y)$. In particular, even the sign of their correlation is ambiguous.*

The proof is quite explicit about the remaining ambiguity: when generating $P(X, \boldsymbol{Z})$ and $P(Y, \boldsymbol{Z})$ via structural equation models with noise variables $N^X_{\boldsymbol{Z}}$ and $N^Y_{\boldsymbol{Z}}$, respectively, the dependencies between $N^X_{\boldsymbol{Z}}$ and $N^Y_{\boldsymbol{Z}}$ only influence the joint distribution, but not the marginals. Note that $\boldsymbol{Z}$ can consist of multiple variables here; thus, the lemma is a general result.

However, also causal models do not always enable LOVO prediction. For example, $G = (Z \to X \to Y)$ can not be uniquely reconstructed from its marginal graphs $G_X = (Z \to X)$ and $G_Y = (Z \to Y)$ since, e.g., the graphs $X \leftarrow Z \to Y$ and $Z \to Y \to X$ have the same marginal graphs. While the graph $G = (X \to Y \leftarrow Z)$ is uniquely determined from its marginal graphs, $G_X = (X \neq Z)$ and $G_Y = (Z \to Y)$, here, the second step fails. Because $X \perp\!\!\!\perp Z$, we can not combine $P(X, Z)$ and $P(Y, Z)$ to extract information on the connection strength from X to Y. Appendix A presents an overview of the realizability of LOVO for all possible graphs consisting of three nodes and two edges, indicating that in many cases, one of the two steps fails. However, in practice, it suffices if we can construct a LOVO predictor for a few pairs of nodes in the graph; then, we simply compute the cross-validation error as an average over those pairs for which we can construct the LOVO predictor. Furthermore, small graphs are particularly challenging for LOVO prediction because the overlap between (X, Z) and (Y, Z) is small. The next section and the simulations reveal that for larger graphs, we typically discover at least one (and often several) pairs that can be handled.

3.2. LOVO prediction via parent adjustment

This section discusses the general case in which $\boldsymbol{Z}$ may contain multiple variables, and the joint graph may be a DAG or an ADMG. Analogously to our first example, whenever there is a set $\boldsymbol{Z}_S$ that renders X and Y conditionally independent, we can define a LOVO predictor via the equation

$$P(y|x) = \sum_{\boldsymbol{z}_S} P(y|\boldsymbol{z}_S)P(\boldsymbol{z}_S|x).$$

Since graphs with more than three nodes often contain at least some pairs of conditionally independent nodes, we mainly rely on this LOVO predictor. Specifically, in a DAG, X and Y are conditionally independent with respect to an appropriate conditioning set if and only if they are not connected by an edge. In this case, the union of parents of X and Y is a d-separating set. In contrast, in ADMGs, the absence of a direct link does not guarantee the existence of an m-separating set (e.g., for $X \to Z \leftrightarrow Y$). Here, the union of parents is m-separating if there is no link and all the parents are unconfounded. Hence, the question arises of how to identify these unlinked pairs with only the marginal graphs available. For example, Figure 1 shows how if we would add the edge

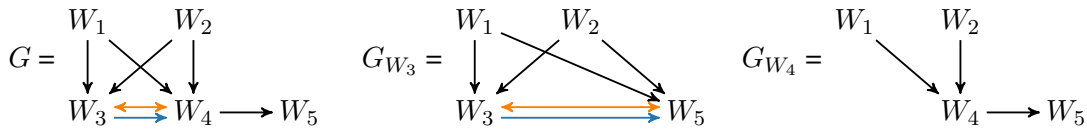

Figure 1: Adding edges between $X = W_3$ and $Y = W_4$ introduces changes in the marginal graphs.

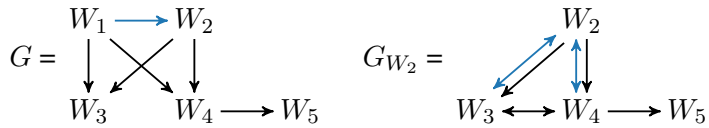

Figure 2: When G is known to be causally sufficient, $W_1 \not\to W_2$ can be inferred based on just G_{W_2}.

$W_3 \to W_4$ or $W_3 \leftrightarrow W_4$ to G, this introduces additional edges in the marginal graphs, represented by the blue and orange edges. Therefore, the marginal graphs allow us to conclude that these edges are not in G. Similarly, one can find that $W_3 \leftarrow W_4 \notin G$. In general, even though X does not appear in G_Y as a node, it indirectly shapes the graph. As a confounder, that is $C \leftarrow X \to D \in G$, it creates the edge $C \leftrightarrow D \in G_Y$, and as a mediator $P \to X \to C \in G$, it leads to $P \to C \in G_Y$. These relations allow us to deduce the absence of an edge, as formalized in the following lemma.

Lemma 2 (excluding links in ADMGs) *Let G an ADMG whose marginalizations are G_X and G_Y. If X has a child in G_X that is neither a sibling nor a child of Y in G_Y, or the same holds with reversed the roles of X and Y, then $X \not\to Y$ in G.*

For the graph G depicted in Figure 1, the lemma allows to deduce $W_3 \not\to W_4$ as W_4 has the child W_5 in G_{W_4}, which is not a child or sibling of W_3 in G_{W_3}. Similarly, one can conclude $W_1 \not\to W_5$ and $W_2 \not\to W_5$. In contrast, the lemma can neither certify $W_1 \not\to W_2$ nor $W_3 \not\to W_5$.

If we know that G is a DAG, we can decide whether an edge exists and, if so, determine its type in even more cases. Knowing that G contains no bidirected edges, we can trace each bidirected edge $C \leftrightarrow D \in G_X$ back to C and D being children of Y in the joint model. Continuing the previous example, but now looking at the pair (W_1, W_2), the bidirected edge $W_3 \leftrightarrow W_4 \in G_{W_2}$ implies that W_3, W_4 are both children of W_1, compare Figure 2. Therefore, if there would be an edge $W_1 \to W_2 \in G$, the edges $W_3 \leftrightarrow W_2, W_4 \leftrightarrow W_2$ would appear in G_{W_2}, which they do not. Thus, $W_1 \not\to W_2 \in G$. The next lemma summarizes all conclusions we can make.

Lemma 3 (determining edge types in DAGs) *Assume G is a DAG with marginalizations G_X, G_Y. Then X has at least two children in G if and only if G_Y contains a bidirected edge. Moreover,*

(1) if X has at least two children in G, then $X \to Y \in G$ if and only if Y has a sibling in G_Y.

(2) If Y has at least two children in G, and X has fewer, then $X \to Y \in G$ if and only if X has multiple children in G_X.

(3) In the case that both have fewer than two children in G,

- *(a) if $ch^{G_X}(X) \neq ch^{G_Y}(Y)$, then $X \not\to Y \in G$;*
- *(b) if neither $pa^{G_X}(X) \subseteq pa^{G_Y}(Y)$, nor vice versa, then $X \not\to Y \in G$;*
- *(c) if $ch^{G_X}(X) = ch^{G_Y}(Y) = \{C\}$, and neither $pa^{G_Y}(Y) \subseteq pa^{G_X}(X) \subseteq pa^{G_Y}(Y) \cup pa^{G_Y}(C)$ nor vice versa, then $X \not\to Y \in G$;*

(d) if $ch^{G_X}(X) = ch^{G_Y}(Y) = \{C\}$, and neither $pa^{G_Y}(Y) \subseteq pa^{G_X}(X) \subseteq pa^{G_Y}(Y) \cup pa^{G_Y}(C)$ nor at the same time $pa^{G_X}(X) \subseteq pa^{G_Y}(C)$ and $pa^{G_Y}(Y) \subseteq pa^{G_X}(C)$, then $X \to Y \in G$.

(4) All the above criteria hold for reversed roles of X and Y.

This lemma leaves only a few exceptions where it does not make a statement about the edge type; in particular, it determines the edge type whenever X or Y has at least two children in G or when they do not have precisely the same children in the marginal graphs. Moreover, the lemma is exhaustive in the sense that if none of the conditions apply, it is impossible to determine whether X and Y are linked. In the example, it finds all five unlinked pairs.

One limitation of both lemmas is that they are restricted to causal discovery algorithms outputting ADMGs. In contrast, many algorithms learn DAGs or produce partially oriented graphs such as CPDAGs or PAGs. However, given the straightforward nature of the proof of the lemmas, it is unsurprising that analogous results can readily be derived for other graph types, see Appendix B.

Once a pair is identified as unlinked, the next step is to derive a d- or m-separating set by inferring the union of parents and verifying their unconfoundedness with X, Y. If there is no edge $X - Y \in G$, the parents and the siblings of X are the same in both G_X and G, as are those of Y. This allows us to directly assess the parents and their unconfoundedness directly from G_X, G_Y, and leads to the following LOVO predictor.

Theorem 4 (LOVO via parent adjustment) *Let all parents of X be unconfounded. Likewise, let all parents of Y be unconfounded. If $\boldsymbol{Z}_S$ denotes the union of the parents of X and Y, then we have*

$$P(y|x) = \sum_{\boldsymbol{z}_S} P(y|\boldsymbol{z}_S)P(\boldsymbol{z}_S|x). \tag{1}$$

Overall, our method first employs Lemma 2 or 3 to determine whether X and Y are linked. Lemma 3 is used when assuming causal sufficiency in the joint model; otherwise, Lemma 2 is used. If the variables are unlinked, their parents are determined using $\text{pa}_G(X) = \text{pa}_{G_X}(X)$ and $\text{pa}_G(Y) = \text{pa}_{G_Y}(Y)$. To implement Theorem 4 in practice, we resort to the following simple procedure:

Three-step LOVO predictor

(1) Learn a predictor $\hat{P}(Y|\boldsymbol{Z})$ or a regression function $\hat{f}$ with $\hat{f}(\boldsymbol{z}) := \hat{\mathbb{E}}[Y|\boldsymbol{Z} = \boldsymbol{z}]$.

(2) Apply the predictor to the $\boldsymbol{z}$-values of the pairs $(x, \boldsymbol{z})$ sampled from $P(X, \boldsymbol{Z})$ to generate artificial pairs $(x, \hat{y})$, with $\hat{y}$ sampled from $\hat{P}(y|\boldsymbol{z})$ or chosen as $\hat{y} := \hat{f}(\boldsymbol{z})$, respectively.

(3) Use these pairs to learn the predictor $\hat{P}(Y|X)$, $\hat{\mathbb{E}}[Y|X]$ or $\hat{\rho}_{XY}$.

As in the example on three nodes, Theorem 4 relies on conditional independence statements postulated by G. In contrast, the preceding steps to infer the non-existence of the edge and the joint parents from the marginal distributions employ the built-in inductive bias of causal models, particularly the faithfulness of the joint model. Although we do not claim that LOVO prediction necessitates causal models, the results in this section suggest that they are a natural way to solve this task and conditional independence information alone is insufficient for LOVO prediction. This is a key distinction to the falsification approach in Biza et al. (2020), where the estimated Markov blanket of a target variable Y is used for predicting Y from as few variables as possible, and the predictions are good whenever Y is conditionally independent of the remaining variables with respect to the estimated Markov blanket.

3.3. LOVO tailored to Linear non-Gaussian Acyclic Models (LiNGAM)

Some causal discovery algorithms are based on structural equation models, such as the linear additive noise model (LiNGAM). The LiNGAM for the DAG G postulates that

$$W_i = \sum_{W_j \in \mathrm{pa}(i)} \lambda_{ij} W_j + N_j, \quad i = 1, \ldots, k+2, \tag{2}$$

where the λ_{ij} are real coefficients and the N_j are independent centered non-Gaussian variables. Since the structure matrix $\Lambda = (\lambda_{ij})$ collecting all coefficients is sparse according to the acyclic graph G, it can be transformed into a strictly lower triangular matrix through simultaneous row and column permutations. We assume faithfulness, that is, for all edges $(i, j) \in G$, the total causal effect

$$m_{ij} = \sum_{\substack{\pi \text{ directed path} \\ \text{from } j \text{ to } i}} \prod_{k \to l \text{ edge on } \pi} \lambda_{lk} \tag{3}$$

is not zero. This assumption is fulfilled for Lebesgue almost all structure matrices Λ compatible with a fixed graph G. Reflecting the model, the corresponding algorithms commonly output not only a DAG but also the matrix Λ. Thus, when developing a LOVO predictor, it appears natural to incorporate the learned matrix in order to falsify the algorithm's entire output. Additionally, this enables LOVO prediction even if a direct link exists, as long as we can determine its type using Lemma 3.

Theorem 5 (LOVO via LiNGAM) *Suppose $P^{\boldsymbol{W}}$ follows a linear additive noise model for some DAG G. If Lemma 3 allows us to identify the edge type between X and Y, ($X \to Y, Y \to X$, or $X \neq Y$), and the children sets are not $ch(Y) = \{X, Z_j\}$ and $ch(X) = \{Z_j\}$ or vice versa, then*[3]

(1) the structure matrix Λ can be uniquely identified from $P(X, \boldsymbol{Z})$ and $P(Y, \boldsymbol{Z})$.

(2) If, in addition, all second and higher order moments of $\boldsymbol{N}$ are finite, $P(X, \boldsymbol{Z})$ and $P(Y, \boldsymbol{Z})$ uniquely determines $P(X, Y, \boldsymbol{Z})$, except for a measure zero set of moments of $\boldsymbol{N}$.

As in the previous theorem, all graphical assumptions in the theorem can be verified from the marginal graphs. Knowing the structure matrix, we can construct the LOVO predictor as follows. Denoting $\boldsymbol{Z}_S = \mathrm{pa}^G(Y) \cap \boldsymbol{Z}$, and using that under the model assumptions $Y = \lambda_{Y\boldsymbol{Z}_S} \boldsymbol{Z}_S + \lambda_{YX} X + N_Y$, and N_Y is centered and independent of $(X, \boldsymbol{Z}_S)$, we obtain

$$\mathbb{E}(Y \mid X = x) = \lambda_{Y\boldsymbol{Z}_S} \mathbb{E}(\boldsymbol{Z}_S \mid X = x) + \lambda_{YX} x,$$

which can be estimated from $P(X, \boldsymbol{Z})$.

4. Baseline: LOVO prediction in absence of causal information

Although the two proposed LOVO predictors provide a reasonable approximation of $\mathbb{E}(Y \mid X = x)$ whenever the marginal graphs G_X, G_Y are accurate, some error will persist. To decide which level of deviation is still acceptable, we ask whether the causal information helped the prediction by

3. Lemma 3 assumes G_X, G_Y to be ADMGs, whereas most LiNGAM-based causal discovery algorithms produce DAGs entailing explicit latent nodes. However, such a DAG can be easily transformed into an equivalent ADMG by replacing each structure $W_1 \leftarrow L \to W_2$, with L a latent node, by $W_1 \leftrightarrow W_2$.

comparing the prediction error to the error of the best LOVO predictor *without causal information*, also called baseline predictor. Note that one may consider $P(Y|X) = P(Y)$ (that is, assuming independence of X and Y) as the best predictor in the absence of any causal knowledge. We reject this idea for two reasons: First, the dependencies between $X, \boldsymbol{Z}$ and between $Y, \boldsymbol{Z}$ may be so strong that it is impossible that X and Y are independent, see No 4. in Table 1 in the supplement, last column. Second, the predictor $P(Y|X) = P(Y)$ is unlikely to be the right one in graphs with several nodes unless one assume relatively sparse graphs.

Instead of assuming independent X, Y as the best "causally agnostic" predictor, we use the "MaxEnt prediction" (Jaynes, 2003), which is the joint distribution that maximizes entropy subject to the given marginal distributions $P(X, \boldsymbol{Z})$ and $P(Y, \boldsymbol{Z})$ (Garrido Mejia et al., 2022). It is given by the unique joint distribution with $X \perp\!\!\!\perp Y \,|Z$. To see this, note that the joint entropy reads (Cover and Thomas, 1991)

$$H(X, Y, Z) = H(X, Z) + H(Y|X, Z) = H(X, Z) + H(Y|Z) - I(Y : X\,|Z),$$

which is maximal when the conditional mutual information $I(Y : X\,|Z)$ vanishes.

To justify MaxEnt as a reasonable approach for our purpose, we remind the reader of the intuition that the MaxEnt distribution is the "most mixed" distribution within the set of distributions satisfying the given bivariate constraints, which seems like a better compromise rather than choosing distributions closer to the boundary. Grünwald and Dawid (2004) provides a game-theoretic view on MaxEnt and shows that it minimizes the worst-case logarithmic cross-entropy loss.

Definition 6 (MaxEnt Baseline predictor) *Given $P(X, \boldsymbol{Z})$ and $P(Y, \boldsymbol{Z})$, the MaxEnt baseline predictor is defined by*

$$P^{\text{MaxEnt}}(y|x) = \sum_{\boldsymbol{z}} P(y|\boldsymbol{z})P(\boldsymbol{z}|x). \tag{4}$$

One can easily show, see Lemma 8 in the appendix, that the MaxEnt predictor is correct for all DAGs whose "moral graph" [4] does not have an edge $X - Y$. Since the overall shape of the MaxEnt predictor aligns with the one of the parent adjustment LOVO predictor, we can again use the three-step procedure to estimate it from finite data. An important difference between the two predictors is that the MaxEnt baseline generally adjusts for more variables. However, comparison of regression models with an equal number of features is "fairer" with respect to statistical inaccuracies entailed by finite data. Therefore, when using the MaxEnt predictor as a baseline against the parent adjustment predictor, we recommend comparing against *random adjustment sets* $\boldsymbol{Z}_R$ of equal size. In this slightly modified version, the baseline is generically worse than parent adjustment whenever $P(X, Y, \boldsymbol{Z})$ is Markov to a graph G, in which $\boldsymbol{Z}_R$ does not d-separate X and Y.

5. Experiments

The code to reproduce our experiments is available at GitHub[5].

4. The moral graph (Lauritzen, 1996) is the undirected graph obtained by removing orientations and connecting parents of a common child.
5. https://github.com/DanielaSchkoda/LOVO.

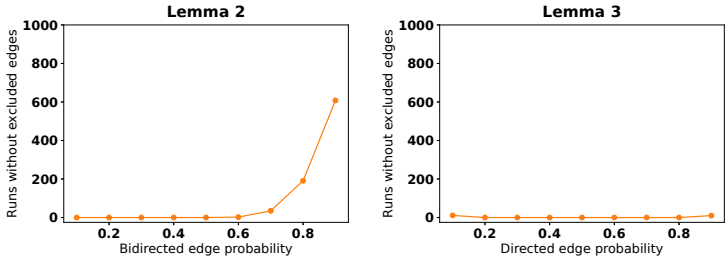

Figure 3: For Lemma 2 and small values of q, and Lemma 3 regardless of p, only in a few graphs no single unlinked pair can be detected, so LOVO is realizable in most cases.

5.1. LOVO prediction given the true marginal graphs G_X, G_Y

First, we want to shed light on the question of how frequently Lemmas 2 - 3 succeed in excluding edges. To this end, we randomly generate 1000 Erdős–Rényi DAGs on 10 nodes by choosing a random ordering and then inserting each edge with probability p varying between 0.1 and 0.9. For each generated graph, we check if, for at least one pair of nodes (X, Y), we can rule out that they are linked based on the marginal graphs G_X, G_Y. For Lemma 2, we use ADMGs instead of DAGs; they are generated following the same procedure, except that we fix $p = 0.3$ and additionally include bidirected edges with a probability $q \in [0.1, 0.9]$. Figure 3 shows in how many runs no single pair without edge can be found, and therefore, LOVO prediction would not be possible. In Appendix E.1, we illustrate the average number of pairs identified as unlinked in each graph.

Next, we assess the parent adjustment LOVO predictor, where we use the correlation ρ between X and Y as the estimation target. We generate graphs as above with $p = 0.3$, and, for ADMGs, $q = 0.1$. To obtain data in accordance with the graphs, we employ a linear additive noise model, with noise uniformly distributed on $[-1, 1]$ and coefficients drawn uniformly from $[-1, -0.5] \cup [0.5, 1]$. We set the sample size to $n = 5000$. Again, based on the true marginal graphs G_X, G_Y, for each pair (X, Y), we evaluate whether they might be linked. If not, we compute the three-step LOVO predictor $\hat{\rho}^{\text{LOVO}}$ according to Theorem 4, as well as the baseline predictor $\hat{\rho}^{\text{Base}}$, for which we calculate a MaxEnt predictor with a random adjustment set of the same size as the union of parents multiple times, and then take the average. Moreover, we directly calculate the sample correlation coefficient $\hat{\rho}$ from $P(X, Y)$ in order to estimate the prediction errors $\hat{\rho}^{\text{Base}} - \hat{\rho}$, $\hat{\rho}^{\text{LOVO}} - \hat{\rho}$. For a more accurate error assessment, in the above steps, we never use all samples; instead, we split the data into three parts of sample size $n/3$ each. The first two parts are used for $P(X, \boldsymbol{Z})$ and $P(Y, \boldsymbol{Z})$, respectively, required in the three-step procedure, while the third part is reserved to estimate $\hat{\rho}$. Finally, we average the results across all pairs to derive the cross-validation errors $CV^{\text{LOVO}}, CV^{\text{Base}}$, which are compared in Figure 4 (left and middle). As before, we use ADMGs combined with Lemma 2 and DAGs with Lemma 3. In the ADMG setting, the LOVO predictor abstains in 0.5% of the replications, and in the DAG setting, never. The supplement includes similar experiments for other graph types.

To analyze the LOVO via LiNGAM predictor, we sample DAGs and data as before and again use the correlation as the estimation target. The right plot in Figure 4 compares the prediction error of LOVO to the MaxEnt baseline predictor with all variables $\boldsymbol{Z}$ as the adjustment set. Again, the LOVO predictor never abstains.

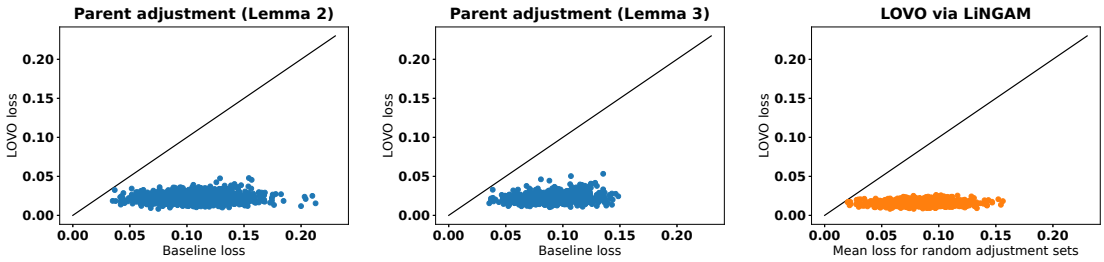

Figure 4: When provided with the true marginal graphs G_X and G_Y, the parent adjustment LOVO predictor and the LiNGAM LOVO predictor outperform the baseline.

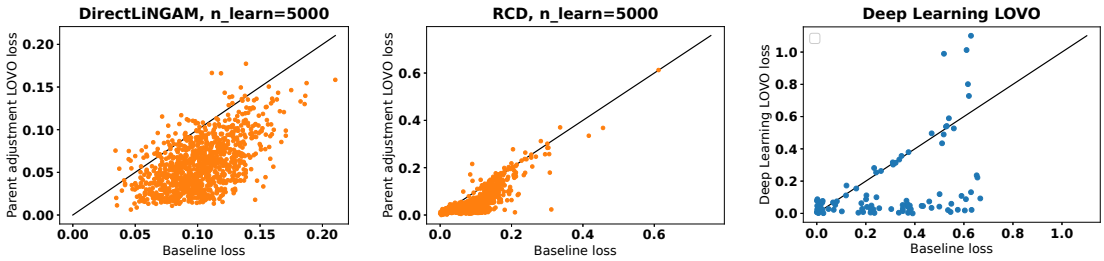

Figure 5: The scatter plots show LOVO versus baseline loss for parent adjustment LOVO applied to graphs estimated with DirectLiNGAM and RCD; and for DL LOVO prediction.

5.2. LOVO applied to DirectLiNGAM and RCD

Next, we apply the LOVO predictor to two causal discovery algorithms, namely DirectLiNGAM (Shimizu et al., 2011) and Repetitive Causal Discovery (RCD) (Maeda and Shimizu, 2020). The first method produces DAGs, and the second one uses the alternative definition of ADMGs, which forbids the co-occurrence of a directed and a bidirected edge. Correspondingly, we use the slightly adjusted LOVO predictors described in Appendix B.1-B.2. We sample DAGs and the data as before, but double the sample size and use the first half to learn G_X, G_Y. For DirectLiNGAM, we use $p = 0.5$ to ensure that Lemma 7 often applies. For RCD, we adhere to $p = 0.3$ but decrease the number of nodes to 5 because of its slower execution time. Figure 5 (left and middle) compares the LOVO cross-validation error to the baseline. For DirectLiNGAM, LOVO abstained in 23% of the cases, and for RCD in 3%. To examine whether the LOVO cross-validation error indeed increases with the number of mistakes in the learned graphs, we repeat the above experiment with varying sample sizes for learning the graphs, specifically, $n_{\text{learn}} = 100, 500, 1000, 5000$. In Appendix E.2, we plot the LOVO loss for each value of n_{learn}. Moreover, we concatenate all the results to calculate the Spearman correlation coefficient of the LOVO cross-validation error and

1. whether an edge $X - Y$ exists in G, averaged over all (X, Y) used in the cross-validation.
2. the sum of the Structural Hamming Distances (SHDs) of $\hat{G}_X$ to G_X and of $\hat{G}_Y$ to G_Y, averaged over all (X, Y) used in the cross-validation.

Including the first measurement is motivated by the fact that the parent adjustment LOVO predictor relies on the absence of an edge and, therefore, mistakes here are expected to lead to inaccurate

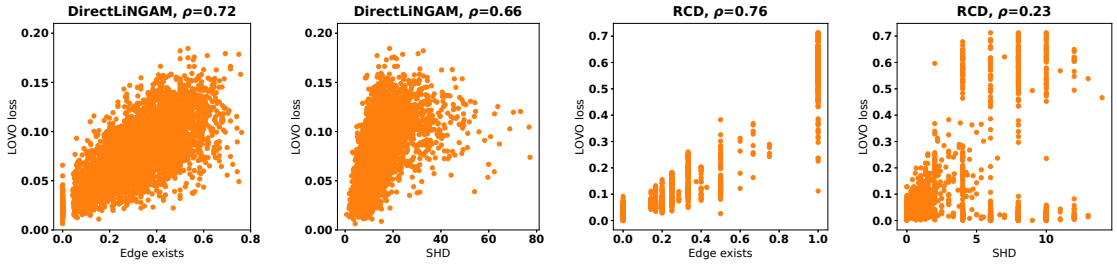

Figure 6: The scatter plots show how LOVO performance correlates with causal discovery performance. The LOVO error increases with the number of pairs misidentified as unlinked and with the SHD. The corresponding Spearman correlation coefficients included in the titles all significantly deviate from zero, with p-values $0.0, 0.0, 0.0$, and $4 \cdot 10^{-44}$.

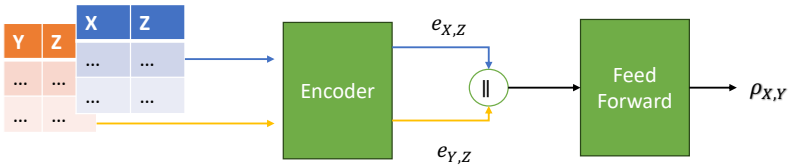

Figure 7: Architecture of our DL LOVO predictor: the encoder learns appropriate features of the two marginal distributions from which the correlation of X and Y is inferred.

LOVO prediction. The second measurement more straightforwardly evaluates the accuracy of the learned graphs. We obtain significant positive correlations in all cases, as presented in Figure 6.

Appendix E.4 includes experiments on real-world data.

5.3. Training DL for trivariate LOVO

enecccfgidufkrkkcddtvncefhjucgtirvkfkufrvnlu So far, we have constructed LOVO predictors either by restricting to unlinked pairs or by assuming LiNGAM. To support the hypothesis that also causal models without such restrictive assumptions help for LOVO predictions, we now show that a deep learning architecture that has been proposed for causal discovery can be modified to a LOVO predictor. To this end, we built on Ke et al. (2023), who use a transformer-based architecture to directly infer the adjacency matrix of the causal graph from a given dataset. We apply the encoder part from their architecture to each marginal dataset to get representation vectors $e_{X,Z}$ and $e_{Y,Z}$. These representations are concatenated and used as input to a feed-forward layer (see Figure 7 for an overview and Section E.5 in the appendix for more details). This way, the network can be trained to predict the correlation $\rho_{X,Y}$ from given marginal datasets. We want to stress that we do not input any parametric model assumptions in the architecture and, therefore, also do not input any preconceptions about how to construct a LOVO predictor. Figure 5 (right) shows that it outperforms our baseline in most cases. To shed light on the challenging question of whether our DL LOVO predictor implicitly learns a causal representation, we try to predict the causal structure between X and Z from the learned representation $e_{X,Z}$. If a second model could learn to map the representation to the causal structure, this suggests that the learned features are suitable for both tasks. To this end, we train a simple feed-forward network to predict the underlying causal structure, encoded as categories

$\{\rightarrow, \leftarrow, \leftrightarrow, \not\sim\}$ (see again Section E.5 for more details). As a na "ive baseline, we consider the average training label (where the categories are represented via one-hot-encoding). Indeed, we can predict the causal structure better than our baseline (see Figure 16 in the appendix).

6. Conclusions

We have shown that causal hypotheses built via applying causal discovery to two Leave-One-Variable-Out datasets can, in principle, enable the prediction of the statistical relations between the two variables X, Y that were dropped. As a concrete LOVO predictor, we first propose prediction via adjusting for parents, which relies on the absence of edges and is applicable to general causal discovery algorithms that produce DAGs or ADMGs. We further demonstrate how LOVO prediction can be customized for specific structural equation models, such as LiNGAM, enabling prediction even when a direct link is present. In simulation experiments, we observe a significant correlation between the LOVO prediction error and the accuracy of the estimated causal graphs. This reinforces our conjecture that the goodness of LOVO predictions can be utilized to evaluate (inferred) causal relationships.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 883818). DS acknowledges support by the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. PMF was supported by a doctoral scholarship of the Studienstiftung des deutschen Volkes (German Academic Scholarship Foundation).

References

Konstantina Biza, Ioannis Tsamardinos, and Sofia Triantafillou. Tuning causal discovery algorithms. In Proc. 10th International Conference on Probabilistic Graphical Models, volume 138, pages 23–25. PMLR, 2020.

Thomes Cover and Joy Thomas. Elements of Information Theory. Wileys Series in Telecommunications, New York, 1991.

Anish Dhir and Ciarán M. Lee. Integrating overlapping datasets using bivariate causal discovery. In Proc. 34th AAAI Conference on Artificial Intelligence. AAAI Press, 2019.

Philipp M. Faller, Leena Chennuru Vankadara, Atalanti A. Mastakouri, Francesco Locatello, and Dominik Janzing. Self-compatibility: evaluating causal discovery without ground truth. In Proc. 27th International Conference on Artificial Intelligence and Statistics, volume 238, pages 4132–4140. PMLR, 2024.

Sergio H. Garrido Mejia, Elke Kirschbaum, and Dominik Janzing. Obtaining causal information by merging datasets with MAXENT. In Proc. 25th International Conference on Artificial Intelligence and Statistics, volume 151, pages 581–603. PMLR, 2022.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. Frontiers in Genetics, 10, 2019.

Nicola Gnecco, Nicolai Meinshausen, Jonas Peters, and Sebastian Engelke. Causal discovery in heavy-tailed models. The Annals of Statistics, 49(3):1755–1778, 2021.

J. B. Grace, T. M. Anderson, E. W. Seabloom, E. T. Borer, Peter B. Adler, W. Stanley Harpole, Yann Hautier, Helmut Hillebrand, Eric M. Lind, Meelis Pärtel, Jonathan D. Bakker, Yvonne M. Buckley, Michael J. Crawley, Ellen I. Damschen, Kendi F. Davies, Philip A. Fay, Jennifer Firn, Daniel S. Gruner, Andy Hector, Johannes M. H. Knops, Andrew S. MacDougall, Brett A. Melbourne, John W. Morgan, John L. Orrock, Suzanne M. Prober, and Melinda D. Smith. Integrative modelling reveals mechanisms linking productivity and plant species richness. Nature, 529 (7586):390–393, 2016.

Luigi Gresele, Julius Von Kügelgen, Jonas Kübler, Elke Kirschbaum, Bernhard Schölkopf, and Dominik Janzing. Causal inference through the structural causal marginal problem. In Proc. 39th International Conference on Machine Learning, volume 162, pages 7793–7824. PMLR, 2022.

Peter Grünwald and Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. The Annals of Statistics, 32(4):1367–433, 2004.

Siyuan Guo, Jonas Wildberger, and Bernhard Schölkopf. Out-of-variable generalization for discriminative models. In Proc. 12th International Conference on Learning Representations, 2024.

Carol M. Hamilton, Martí Aldea, Brian K. Washburn, Paul Babitzke, and Sidney R Kushner. New method for generating deletions and gene replacements in escherichia coli. Journal of Bacteriology, 171:4617 – 4622, 1989.

Patrik Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In Advances in Neural Information Processing Systems, volume 21. Curran Associates, Inc., 2008.

Guido 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, 2020.

Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: characterization and learning. In Advances in Neural Information Processing Systems, volume 33, pages 9551–9561. Curran Associates, Inc., 2020.

Edwin T. Jaynes. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, MA, 2003.

Yutaka Kano and Shohei Shimizu. Causal inference using nonnormality. In Proc. International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion, pages 261–270, 2003.

Nan Rosemary Ke, Silvia Chiappa, Jane X. Wang, Jorg Bornschein, Anirudh Goyal, Melanie Rey, Matthew Botvinick, Theophane Weber, Michael Curtis Mozer, and Danilo Jimenez Rezende.

Learning to induce causal structure. In International Conference on Learning Representations, 2023.

Murat Kocaoglu, Alexandros G. Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causal inference. In Proc. 31st AAAI Conference on Artificial Intelligence, pages 1156–1162. AAAI Press, 2017.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. In International Conference on Learning Representations, 2020.

Kai Lagemann, Christian Lagemann, Bernd Taschler, and Sach Mukherjee. Deep learning of causal structures in high dimensions under data limitations. Nature Machine Intelligence, 5(11):1306–1316, 2023.

Steffen Lauritzen. Graphical Models. Clarendon Press, 1996.

David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Ilya Tolstikhin. Towards a learning theory of cause-effect inference. In Proc. 32nd International Conference on International Conference on Machine Learning, volume 37, pages 1452–1461. PMLR, 2015.

Takashi Nicholas Maeda and Shohei 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.

Karthika Mohan and Judea Pearl. Graphical models for processing missing data. Journal of the American Statistical Association, 116(534):1023–1037, 2021.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Scalable causal discovery with score matching. In Proc. 2nd Conference on Causal Learning and Reasoning, volume 213, pages 752–771. PMLR, 2023.

Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. Journal of Machine Learning Research, 21(99):1–108, 2020.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. Machine Learning and Knowledge Extraction, 1(1):312–340, 2019.

Judea Pearl. Causality. Cambridge University Press, 2000.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. In Proc. 27th Conference on Uncertainty in Artificial Intelligence, page 589–598. AUAI Press, 2011.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society Series B: Statistical Methodology, 78(5):947–1012, 2016.

Alexander G. Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! Causal discovery benchmarks may be easy to game. In Advances in Neural Information Processing Systems, volume 34, pages 27772–27784. Curran Associates, Inc., 2021.

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

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In Proc. 39th International Conference on Machine Learning, volume 162. PMLR, 2022.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: heterogeneous data meets causality. Journal Royal Statistical Society Series B, 83:215–246, 2021.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. Science, 308(5721): 523–529, 2005.

Saber Salehkaleybar, AmirEmad Ghassami, Negar Kiyavash, and Kun Zhang. Learning linear non-Gaussian causal models in the presence of latent variables. Journal of Machine Learning Research, 21(39):1–24, 2020.

Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, Gyorgy Simon, Michael Weiner, Paul Aisen, Ronald Petersen, Clifford Jack, Andrew Saykin, William Jagust, John Trojanowki, Arthur Toga, Laurel Beckett, Robert Green, John Morris, Leslie Shaw, Zaven Khachaturian, Greg Sorensen, Maria Carroll, and Kristin Fargher. Challenges and opportunities with causal discovery algorithms: application to Alzheimer's pathophysiology. Scientific Reports, 10:2975, 02 2020.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7:2003–2030, 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. Journal of Machine Learning Research, 12:1225–1248, 2011.

Peter Spirtes, Clark Glymour, and Richard Scheines. Causation, Prediction, and Search. Springer-Verlag, New York, NY, 1993.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological), 36(2):111–133, 2018.

Xiaohai Sun, Dominik Janzing, and Bernhard Schölkopf. Causal inference by choosing graphs with most plausible Markov kernels. In Proc. 9th International Symposium on Artificial Intelligence and Mathematics, pages 1–11, 2006.

Jin Tian and Judea Pearl. Causal discovery from changes. In Proc. 17th Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2001.

Ioannis Tsamardinos, Sofia Triantafillou, and Vincenzo Lagani. Towards integrative causal analysis of heterogeneous data sets and studies. Journal of Machine Learning Research, 13:1097–1157, 2012.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In Advances in Neural Information Processing Systems, volume 36, pages 50254–50292. Curran Associates, Inc., 2023.

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

Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In Proc. 26th International Joint Conference on Artificial Intelligence, pages 1347–1353, 2017.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric DAGs. In Proc. 23rd International Conference on Artificial Intelligence and Statistics, volume 108, pages 3414–3425. PMLR, 2020.

Appendix A. LOVO predictors for DAGs with two arrows

As mentioned earlier, the case of three nodes is particularly challenging, and our approach presented in Section 3.2 may not always succeed. Therefore, we present alternative LOVO predictors that can be beneficial in these cases. Specifically, we consider the "promise"-scenario of three variable (X, Y, Z), where we are given the information that the joint distribution $P(X, Y, Z)$ has been generated by a causal directed acyclic graph (DAG) with *two arrows* only. In Table 1, we group the 12 possible DAGs according to the 3 possible skeletons $X - Z - Y$, $X - Y - Z$, $Y - X - Z$, with each skeleton allowing for 4 different DAGs. We will see that in each of these groups, the collider is special, but the three other Markov equivalent DAGs entail the same LOVO predictor.

No. 1-3: DAGs with $X \perp\!\!\!\perp Y \,|Z$ This is the simplest case where the conditional independence directly entails the solution

$$P(X, Z, Y) = P(X, Z)P(Y|Z), \tag{5}$$

without any parametric assumptions. The solution is most intuitive for the DAGs $X \to Z \to Y$ (No.1) and $X \leftarrow Z \to Y$ (No.2), where the algebraic structure of (5) resembles the data generating process by applying the stochastic map $P(Y|Z)$ to the joint distribution of X, Z. While 5 is certainly also valid for $X \leftarrow Z \leftarrow Y$, now $P(Y|Z)$ turns into an "anticausal" conditional. When parametric assumptions are imposed for causal conditionals (e.g. linear non-Gaussian models (Kano and Shimizu, 2003) or non-linear additive noise models (Hoyer et al., 2008)), $P(Y|Z)$ now results from Bayesian inversion of those models. For linear models, the Pearson correlation between X and Y is easily obtained via[6]

$$\rho_{XY} = \rho_{XZ} \cdot \rho_{ZY}. \tag{6}$$

If X, Y have zero mean and unit variance, the best linear predictor for Y from X then reads $\mathbb{E}[Y|X = x] = \rho_{XY} \cdot x$. By slight abuse of terminology, we therefore call ρ_{XY} the "linear LOVO predictor," which implicitly refers to this convention.

Note that this LOVO predictor coincides with the MaxEnt baseline predictor, and therefore, the cases where (6) does *not* hold are the interesting ones for us.

6. This follows from zero partial correlation, which is defined by $\rho_{X,Y|Z} = \frac{\rho_{XY} - \rho_{X,Z}\rho_{Z,Y}}{\sqrt{1-\rho^2_{X,Z}}\sqrt{1-\rho^2_{Y,Z}}}$.

No.	DAG	linear predictor	stochastic matrix predictor	bivariate causality	necessary conditions
1	$X \to Z \to Y$	$\rho_{XY} = \rho_{XZ} \cdot \rho_{YZ}$	$P_{Y\|X} = P_{Y\|Z} P_{Z\|X}$	$X \to Z$ $Y \leftarrow Z$	
2	$X \leftarrow Z \to Y$	$\rho_{XY} = \rho_{XZ} \cdot \rho_{YZ}$	$P_{Y\|X} = P_{Y\|Z} P_{Z\|X}$	$X \leftarrow Z$ $Y \leftarrow Z$	
3	$X \leftarrow Z \leftarrow Y$	$\rho_{XY} = \rho_{XZ} \cdot \rho_{YZ}$	$P_{Y\|X} = P_{Y\|Z} P_{Z\|X}$	$X \leftarrow Z$ $Y \to Z$	
4	$X \to Z \leftarrow Y$	$\rho_{XY} = 0$	$P_{Y\|X} = P_Y$	$X \to Z$ $Y \to Z$	$\rho_{XZ}^2 + \rho_{YZ}^2 \leq 1$
5	$X \to Y \to Z$	$\rho_{XY} = \rho_{XZ}/\rho_{YZ}$	$P_{Y\|X} = P_{Z\|Y}^{-1} P_{Z\|X}$	$X \to Z$ $Y \to Z$	$I(X:Z)$ $\leq I(Y:Z)$
6	$X \leftarrow Y \to Z$	$\rho_{XY} = \rho_{XZ}/\rho_{YZ}$	$P_{Y\|X} = P_{Z\|Y}^{-1} P_{Z\|X}$	$X \leftrightarrow Z$ $Y \to Z$	$I(X:Z)$ $\leq I(Y:Z)$
7	$X \leftarrow Y \leftarrow Z$	$\rho_{XY} = \rho_{XZ}/\rho_{YZ}$	$P_{Y\|X} = P_{Z\|Y}^{-1} P_{Z\|X}$	$X \leftarrow Z$ $Y \leftarrow Z$	$I(X:Z)$ $\leq I(Y:Z)$
8	$X \to Y \leftarrow Z$	?	?	$X \not\leftrightarrow Z$ $Y \leftarrow Z$	
9	$Y \to X \to Z$	$\rho_{XY} = \rho_{YZ}/\rho_{XZ}$	$P_{Y\|X} = P_{Y\|Z} P_{X\|Z}^{-1}$	$X \to Z$ $Y \to Z$	$I(X:Z)$ $\geq I(Y:Z)$
10	$Y \leftarrow X \to Z$	$\rho_{XY} = \rho_{YZ}/\rho_{XZ}$	$P_{Y\|X} = P_{Y\|Z} P_{X\|Z}^{-1}$	$X \to Z$ $Y \leftrightarrow Z$	$I(X:Z)$ $\geq I(Y:Z)$
11	$Y \leftarrow X \leftarrow Z$	$\rho_{XY} = \rho_{YZ}/\rho_{XZ}$	$P_{Y\|X} = P_{Y\|Z} P_{X\|Z}^{-1}$	$X \leftarrow Z$ $Y \leftarrow Z$	$I(X:Z)$ $\geq I(Y:Z)$
12	$Y \to X \leftarrow Z$	?	?	$X \leftarrow Z$ $Y \not\leftrightarrow Z$	

Table 1: All possible DAGs on X, Y, Z with two arrows, together with their LOVO predictors.

No. 4: variable Z as collider Due to $X \perp\!\!\!\perp Y$, we ignore X and take $P(Y)$ as the correct LOVO predictor for Y. We will later see, however, that this case is hard to recognize from the bivariate distributions because the bivariate causal models $X \to Z$ and $Y \to Z$ can also originate from the joint models $X \to Y \to Z$ and $Y \to X \to Z$.

No. 5-7: DAGs with $X \perp\!\!\!\perp Z \,|Y$ Now, the conditional distribution of X given Z is a concatenation of Markov kernels

$$P(Z|X) = P(Z|Y) \cdot P(Y|X). \tag{7}$$

In linear models we conclude $\rho_{XY} \cdot \rho_{YZ} = \rho_{XZ}$, from which we can directly construct the linear LOVO predictor.

In the case where X, Y, Z are variables with finite ranges $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, we introduce the stochastic matrices $P_{X|Z} := (p(x|z))_{x \in \mathcal{X}, z \in \mathcal{Z}}$ and obtain the matrix equation $P_{Z|X} = P_{Z|Y} \cdot P_{Y|X}$. Whenever

the matrix $P_{Z|Y}$ is invertible[7], we thus obtain

$$P_{Y|X}^{\text{mediator Y}} := P_{Z|Y}^{-1} \cdot P_{Z|X}. \tag{8}$$

No. 8: variable Y as collider This is a negative case: as explained in Section 3.1, it is unclear how to construct a good LOVO predictor.

No. 9-11: DAGs with $Y \perp\!\!\!\perp Z \,|X$ Here we obtain

$$P_{Y|Z} = P_{Y|X} P_{X|Z}, \tag{9}$$

which implies the predictor

$$P_{Y|X}^{\text{mediator X}} := P_{Y|Z} P_{X|Z}^{-1}, \tag{10}$$

if $P_{X|Z}$ is invertible.

No. 12: variable X as collider As for No. 8, we can not infer the strength of the influence of Y on X and abstain from constructing a LOVO predictor.

Inspecting the column with the bivariate causal graphs in Table 1, we find only 7 cases where the joint DAG can be uniquely identified from the bivariate graphs, namely the numbers No. 1, 2, 3, 6, 8, 10, 12. Unfortunately, recognizing 8 and 12 is not helpful for our purpose because we can not offer a LOVO predictor there. Moreover, the predictor of No. 1-3 coincides with our baseline. Overall, we are left with only two cases where a LOVO predictor is realizable and simultaneously beats the baseline. We can improve upon that by considering additional conditions on the marginals from which the respective DAG can be excluded, as listed in the last column of the table.

For instance, $X \perp\!\!\!\perp Z \,|Y$ implies that the dependence between X and Z can not be larger than the dependence between Y and Z. In the non-parametric case, this can be formalized via the Shannon mutual information, for which we have the data processing inequality $I(X : Z) \leq I(Y : Z)$. For linear models, we have $\rho_{XZ} \leq \rho_{YZ}$. Further, the collider $X \to Z \leftarrow Y$ is only possible if $\rho_{XZ}^2 + \rho_{YZ}^1 \leq 1$, otherwise the correlation matrix

$$\begin{pmatrix} 1 & \rho_{XZ} & 0 \\ \rho_{XZ} & 1 & \rho_{YZ} \\ \rho_{YZ} & 0 & 1 \end{pmatrix},$$

would not be positive semi-definite.

Appendix B. LOVO extended to further graph types

Three graph types, which have so far not been considered and yet are the outputs of many causal discovery algorithms, are ADMGs without confounded causal links, Directed Acyclic Graphs (DAGs), Completed Partially Directed Acyclic Graphs (CPDAGs) and Partial Ancestral Graphs (PAGs). This section extends parent adjustment to these graph types. As before, we identify unlinked pairs and then find valid adjustment sets. To highlight the differences between the graph types, we revisit the example used in Subsection 3.2, but this time with the edge $W_1 \to W_3$ omitted, resulting in the graph shown in Figure 8. For this slightly modified graph, Lemma 2 identifies all unlinked pairs except (W_3, W_5) and Lemma 3 finds all six unlinked pairs. When referring to this example, we may use the shorthand i to denote the node W_i.

7. Note that the inverse is not a stochastic matrix except for the trivial case of determinism.

B.1. ADMGs without confounded causal links

The alternative ADMG definition prohibiting confounded causal links uses single bidirected edges in place of our confounded causal links $\overset{\leftrightarrow}{\rightarrow}$ or $\overset{\leftrightarrow}{\leftarrow}$. This substitution leads to a loss of information. For instance, when leaving out X in each of the following three graphs,

$$Y \leftarrow X \rightarrow Z\,, \qquad Y \overset{\frown}{\leftarrow} X \overset{\frown}{\rightarrow} Z\,, \quad \text{and } Y \underset{\frown}{\leftarrow} X \underset{\frown}{\rightarrow} Z,$$

according to the ADMG definition used so far, one obtains three distinct marginal graphs, namely

$$Y \leftrightarrow Z, \quad Y \overset{\leftrightarrow}{\rightarrow} Z, \quad \text{and } Y \overset{\leftrightarrow}{\leftarrow} Z.$$

In contrast, the other definition always yields $Y \leftrightarrow Z$. Despite these differences, Lemma 2 remains valid. Recall that it asserts that whenever X has a child in G_X that is not a sibling of Y in G_Y, then X and Y are not linked in G. We write G_X, G_Y for the ADMGs with the previously used definition, and $\overline{G}_X, \overline{G}_Y$ for the ADMGs where confounded causal links are substituted for bidirected edges. Because every child of X in $\overline{G}_X$ is also a child of X in G_X, and every sibling of Y in G_Y is also a sibling of Y in $\overline{G}_Y$, lemma remains true when considering $\overline{G}_X$ and $\overline{G}_Y$. Similarly, one can show that Lemma 3 still holds. However, the next step, that is, reading off the parents of X, Y from the marginal graphs, becomes more involved. Specifically, a bidirected edge between $Y \leftrightarrow Z_i \in \overline{G}_Y$ precludes their parent-child relationship: In G, we could have $Y \rightarrow Z_i, Y \leftarrow Z_i$, or $Y \not\rightarrow Z_i$. So, each sibling of Y in $\overline{G}_Y$ could be a parent in G or not, and therefore, we can not infer the parents whenever Y has siblings in $\overline{G}_Y$. Accordingly, whenever $\text{sib}^{\overline{G}_X}(X) \neq \varnothing$ or $\text{sib}^{\overline{G}_Y}(Y) \neq \varnothing$, we refrain from a prediction for all pairs with siblings. Note that when Lemma 3 is employed, we anyways exclude all these pairs since an edge might exist according to condition (1). However, this does not necessarily apply when Lemma 2 is used. As stated above, in the running example, Lemma 2 detects five pairs as unlinked. For all of them, X has no siblings in $\overline{G}_X$ and likewise Y none in $\overline{G}_Y$, so we do not need to abstain.

B.2. Directed Acyclic Graphs

Turning to DAGs, it is evident that they are not able to model latent confounding, which, however, is typically present in the marginal graphs G_X and G_Y. To address this mismatch, we assume to be given DAGs $\tilde{G}_X, \tilde{G}_Y$, which are derived from G_X, G_Y by orienting each bidirected edge into a directed one. Often, only one direction avoids the creation of cycles. Otherwise, we opt for a random direction. We are aware that it might be unrealistic to expect an algorithm assuming causal sufficiency to generate precisely such a DAG when latent confounding is actually present. Nonetheless, we argue that this assumption is still reasonable since we do not necessitate the LOVO predictor based on this assumption to yield a perfect prediction but only require it to outperform the non-causal predictor based on a random adjustment set.

Within this scenario, inferring the edge type is possible in fewer cases. Figure 8 illustrates why in the example, the pair $(1, 5)$ can not be certified as unlinked anymore, but $(1, 2)$ still can.

Lemma 7 (excluding links based on DAGs) *Let G be a DAG, and $\tilde{G}_X, \tilde{G}_Y$ DAGs obtained from G_X, G_Y by exchanging each bidirected edge for a directed one. If X has a child in $\tilde{G}_X$ that is not a neighbor of Y in $\tilde{G}_Y$, then G does not contain any edge $X - Y$.*

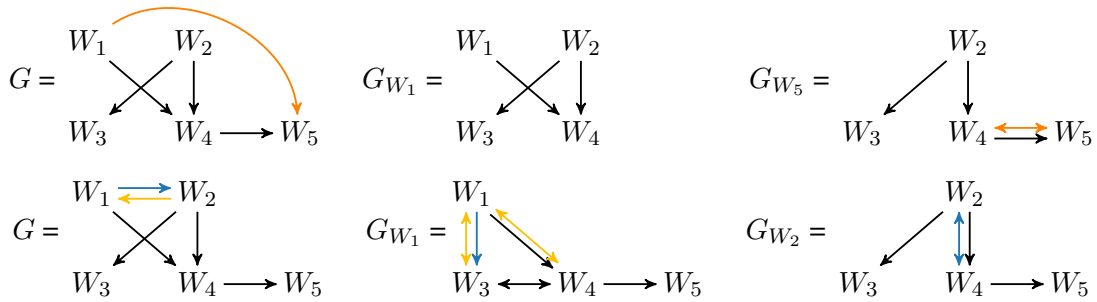

Figure 8: Adding the edge $1 \to 5$ to G, introduces a bidirected edge in G_{W_5}. However, this additional edge is not visible in $\tilde{G}_{W_5}$ since it turns into $4 \to 5$, which was already present before. Thus, $(1, 5)$ can not be confirmed as unlinked. In contrast, we can conclude $1 \not\perp 2$. Lemma 7 states that if 2 and 1 were connected in G, then 1 would be linked to 3 in G_{W_1} since 3 is a child of 2 in G_{W_2}. Indeed, adding either $1 \to 2$ or $2 \to 1$ would yield a connection between 1 and 3 as visualized by the blue and yellow edges.

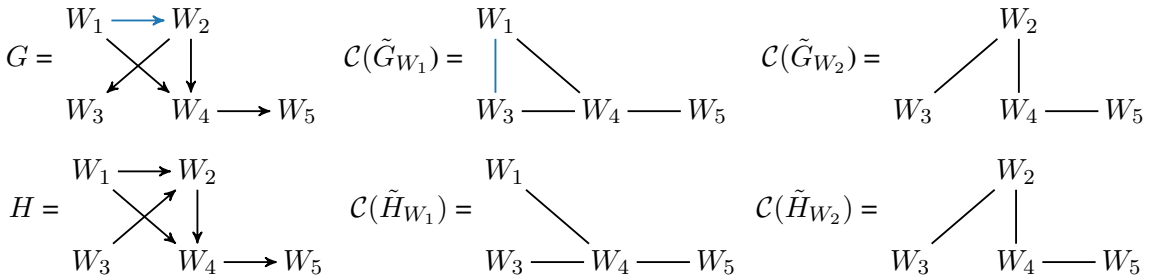

Figure 9: Altough adding the edge $1 \to 2$ to G introduces changes in the marginal CPDAG $\mathcal{C}(\tilde{G}_{W_1})$, we can not identify this pair as unlinked. In contrast to the scenario in Figure 8, we can not be certain that 3 is a child of 2 and do not find any child to utilize in Lemma 7. Indeed, there is another graph H, with the edge between 2 and 3 reversed, an edge $1 \to 2$, and the same marginal CPDAGs as G.

Proof Since X has a child C in $\tilde{G}_X$, either $X \leftrightarrow C$ or $X \to C$ in G_X. In the first case, $X \leftarrow Y \to C \in G$, and thus $Y \to C \in G_Y$ contrary to the assumption of the lemma. In the second case, if G would feature the edge $X \to Y$, then combining both edges yields $Y \leftarrow X \to C$ and therefore Y and C are connected in G_Y, contrary to the assumption of the lemma. Similarly, if the reverse edge $Y \to X$ would be part of G, then $Y \leftarrow X \to C \in G$, which results in $Y \to C \in G_Y$ and again contradicts the assumption. ■

In the example, the lemma detects four of the six unlinked pairs, namely $(1, 2), (1, 3), (2, 5)$, and $(3, 4)$.

B.3. Partially Directed Acyclic Graphs and Partial Ancestral Graphs

A CPDAG is an equivalence class of DAGs, which collects all DAGs entailing the same conditional independence statements. Consequently, CPDAGs also do not capture latent confounding. Because algorithms learning CPDAGs often rely on conditional independence testing, we again expect latent

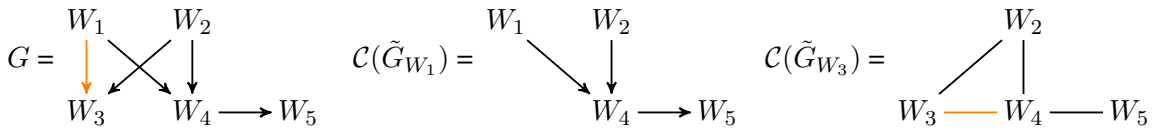

Figure 10: Unlike the pair $(1, 2)$, the pair $(1, 3)$ can be confirmed to not be connected. The graph $\mathcal{C}(\tilde{G}_{W_3})$ features directed edges, which allow to conclude that 4 is definitively a child of 1 in $\tilde{G}_{W_1}$, while not being a neighbor of 3 in $\tilde{G}_{W_3}$. As before, the orange edges show that, if $1 \to 3 \in G$, nodes 3 and 4 were neighbors in $\tilde{G}_{W_3}$.

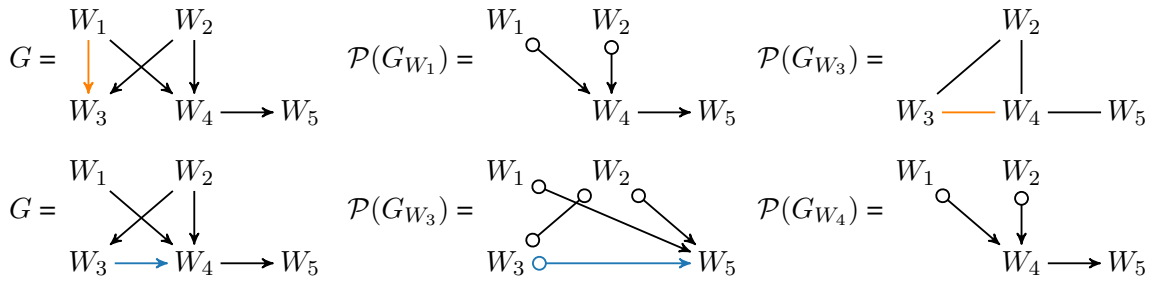

Figure 11: The edge $1 \to 3$ can not be excluded since neither 1 nor 3 has definite children in their marginal PAGs . In contrast, $3 \to 4$ can be excluded due to the child 5 of 4 in $\mathcal{P}(G_{W_4})$.

confounding to manifest as directed or undirected edge in the learned CPDAG. More precisely, we assume we are given the CDPAGs $\mathcal{C}(\tilde{G}_X)$ and $\mathcal{C}(\tilde{G}_Y)$ corresponding to $\tilde{G}_X$ and $\tilde{G}_Y$ as defined in the previous subsection. Since a CPDAG represents an equivalence class of DAGs, or in other words, a DAG except for uncertainty regarding some of the edge directions, Lemma 7 remains valid. Unsurprisingly, we can exclude fewer pairs because edges might be unoriented and, therefore, we can less frequently find a node that is *definitely* a child of X in $\tilde{G}_X$ and at the same time not a neighbor of Y in $\tilde{G}_Y$. Figure 9 and 10 demonstrate this for our running example. Overall, the lemma can certify two pairs as unlinked, namely $(1, 3)$ and $(3, 4)$.

PAGs are similar to CPDAGs insofar as they both represent an equivalence class of graphs encoding the same conditional independence statements. The key difference is that PAGs can represent latent confounding. Specifically, PAGs feature directed edges ($X \to Y$) and bidirected edges ($X \leftrightarrow Y$) with the same meaning as in ADMGs. As in the alternative definition, confounded causal links are not allowed and are replaced by single bidirected edges. Additionally, circle edges ($X \circ\!\!-\!\!\circ Y$, $X \circ\!\!\to Y$) can be present, where a circle at an endpoint indicates uncertainty about the presence of an arrowhead or tail, similar to the undirected edges in CPDAGs. Sometimes, undirected edges ($X - Y$) are also included to represent selection bias; however, we do not consider them here. Since the circles represent uncertainty, we can directly read off the *definite* as well as the *possible* children, parents, neighbors, and siblings of each node. Therefore, Lemma 2 is still applicable if only the PAGs $\mathcal{P}(G_X), \mathcal{P}(G_Y)$ corresponding to G_X, G_Y are available.[8] Again, due to the uncertainty, we can find fewer unlinked pairs. In the running example, only the pair $(3, 4)$ is found to be unlinked, as illustrated in Figure 11.

For CPDAGs as well as PAGs, having found unlinked pairs, the next step is determining a valid adjustment set. With many edges being unoriented, it is often impossible to identify the parents of

8. Note that these PAGs may not exist as explained above. In such cases, the lemma is not applicable.

X, Y. Therefore, we rely on a different adjustment set: Assuming a causally sufficient joint model, no path between X and Y with more than two edges can consist solely of colliders. Thus, these paths can be blocked by including all interior nodes in the adjustment set. For paths of length two ($X - Z - Y$), a case distinction must be made:

- If Z could be either a collider or a non-collider, we can not determine a valid adjustment set and must abstain.
- If Z is definitely a non-collider, it is included in the adjustment set.
- If Z is a definite collider and was not already added to the the adjustment set in any previous step, the adjustment set remains valid. Otherwise, we abstain.

If this procedure succeeds in finding a valid set, the LOVO predictor can be formed via the three-step predictor defined in Section 3.2.

Appendix C. When is MaxEnt LOVO correct?

The following simple criterion tells us when the MaxEnt predictor is right:

Lemma 8 (MaxEnt baseline) *Let G be a causal DAG connecting $\boldsymbol{Z}, X, Y$ and G^m be the corresponding moral graph.*[9] *If G_m does not contain the edge $X - Y$, then the MaxEnt predictor is correct.*

Proof Due to the Markov condition for undirected graphs (Lauritzen, 1996), $X \perp\!\!\!\perp Y \,|\boldsymbol{Z}$ if there is no link $X - Y$ in G^m, which implies $P(y|x, \boldsymbol{z}) = P(y|\boldsymbol{z})$. □

If $\boldsymbol{Z}$ consists of just one variable Z, the number of DAGs for which Lemma 8 holds can be counted as follows: obviously, it only holds for DAGs with less than 3 arrows. For the one with 2 arrows, the skeleton must read $X - Z - Y$. To ensure that G^m does not contain $X - Y$, there can not be a collider at Z, thus only the Markov equivalence class of $X \to Z \to Y$ is remaining (with 3 elements). For the 6 DAGs with one arrow, $X \perp\!\!\!\perp Y \,|Z$ is always satisfied. Hence, we obtain 9 DAGs for which our MaxEnt LOVO is optimal, and the total number of DAGs with 3 nodes reads 25.

Appendix D. Proofs

D.1. Proof of Lemma 1

Define the conditional cumulative distribution functions $F_X(x|\boldsymbol{z}) := P(X \leq x|\boldsymbol{Z} = \boldsymbol{z})$ and $F_Y(y|\boldsymbol{z}) := P(Y \leq y|\boldsymbol{Z} = \boldsymbol{z})$. We then define structural equation models for $P(Y|\boldsymbol{Z} = \boldsymbol{z})$ $P(Y|\boldsymbol{Z} = \boldsymbol{z})$ with uniformly distributed noise variables: $X = f_X(\boldsymbol{Z}, N_X)$ and $Y = f_Y(\boldsymbol{Z}, N_Y)$, where $f_X(\boldsymbol{z}, N_X) = F_X^{-1}(N_X|\boldsymbol{z})$ and $f_Y(\boldsymbol{z}, N_Y) = F_Y^{-1}(N_Y|\boldsymbol{z})$. Whenever we generate $\boldsymbol{z}$-values with distribution $P(\boldsymbol{Z})$, we obtain the right marginal distributions $P(X, \boldsymbol{Z})$ and $P(Y, \boldsymbol{Z})$. Note that this holds even for dependent noise with arbitrary $P(N_X, N_Y)$ with the only constraint that their marginals need to be uniform (in other words, $P(N_X, N_Y)$ is a copula) since the dependences between N_X and N_Y do not affect the marginals. When we choose $P(N_X, N_Y) = P(N_X)P(N_Y)$,

9. Following Lauritzen (1996), page 7, the moral graph corresponding to a DAG G is the undirected graph that contains an edge $a - b$ if and only if a and b are directly connected in G or if they have a common child.

we obtain the MaxEnt solution $P^{\mathrm{MaxEnt}}(X, Y, \boldsymbol{Z}) = P(X, \boldsymbol{Z})P(Y|\boldsymbol{Z})$. However, when we choose $N_Y = N_X$, the variables X and Y are positively correlated when conditioned on $\boldsymbol{Z}$. When we choose $N_Y = (1 - N_X)$ instead, X and Y are negatively correlated when conditioned on $\boldsymbol{Z}$. Let $\mathrm{Cov}^{\mathrm{MaxEnt}}(X, Y)$, $\mathrm{Cov}^{\mathrm{pos}}(X, Y)$, and $\mathrm{Cov}^{\mathrm{neg}}(X, Y)$ denote the covariances of X, Y with respect to the three different choices of the dependences of N_X, N_Y. We then have

$$\mathrm{Cov}^{\mathrm{pos}}(X, Y) > \mathrm{Cov}^{\mathrm{MaxEnt}}(X, Y) > \mathrm{Cov}^{\mathrm{neg}}(X, Y).$$

This follows because

$$\mathrm{Cov}^{\mathrm{pos}}(X, Y|\boldsymbol{Z} = \boldsymbol{z}) > \mathrm{Cov}^{\mathrm{MaxEnt}}(X, Y|\boldsymbol{Z} = \boldsymbol{z}) > \mathrm{Cov}^{\mathrm{neg}}(X, Y|\boldsymbol{Z} = \boldsymbol{z}),$$

holds for any $\boldsymbol{z}$ from the law of total covariance:

$$\mathrm{Cov}(X, Y) = \mathbb{E}[\mathrm{Cov}(X, Y|\boldsymbol{Z})] + \mathrm{Cov}(\mathbb{E}[X|\boldsymbol{Z}], \mathbb{E}[Y|\boldsymbol{Z}]),$$

since the conditional expectations $\mathbb{E}[X|\boldsymbol{Z}]$ the $\mathbb{E}[Y|\boldsymbol{Z}]$ are both functions of $\boldsymbol{Z}$, which only depend on the respective marginal distribution and are therefore unaffected by the dependences of the noise variables. □

D.2. Proof of Lemma 2

Going slightly beyond the statement in the lemma, we show the following criteria for excluding each possible edge type.

1) We can exclude the existence of a direct edge $X \to Y$ in G if
   a) X occurs causally after Y in the sense that in G_X there exists an ancestor of X that is at the same time a descendant of Y in G_Y; or
   b) at least one of the following implications is violated
      i) $P \to X \in G_X \implies P \to Y \in G_Y$,
      ii) $X \to C \in G_X \implies Y \to C \in G_Y$ or $Y \leftrightarrow C \in G_Y$,
      iii) $X \leftrightarrow S \in G_X \implies Y \leftrightarrow S \in G_Y$,
      iv) $Y \to C \in G_Y \implies X \to C \in G_X$.
2) Likewise, the arrow $Y \to X$ can be excluded by swapping the roles of X and Y.
3) The bidirected edge $X \leftrightarrow Y$ can be excluded if one of the following implications is violated
   a) $X \to C \in G_X \implies Y \to C \in G_Y$ or $Y \leftrightarrow C \in G_Y$,
   b) $Y \to C \in G_Y \implies X \to C$ or $X \leftrightarrow C \in G_X$.

Since the criteria in point 3) already entail the criteria in 1) and 2), all three points taken together yield the lemma.

We prove statement 1), and the rest works similarly. Condition a) excludes $X \to Y$ because we had a causal cycle otherwise. We show b) by showing its contrapositive, that is, if $X \to Y \in G$, then all four implications hold.

i) If $P \to X \in G_X$, then $P \to X \in G$ since all directed edges in $G_X \setminus G$ are of the form $P \to C$ for $P \in \mathrm{pa}(Y)$, $C \in \mathrm{ch}(Y)$ but X is not a child of Y. Combined with $X \to Y$ this yields $P \to Y \in E_Y$.

ii) If $X \to C \in G_X$, then $X \to Y \to C \in G$ or $X \to C \in G$. In the first case, it directly follows that $Y \to C \in G_Y$. In the second case, $Y \leftrightarrow C$ is added in the marginalization G_Y since $X \to Y \in G$.

iii) If $X \leftrightarrow S \in G_X$, then $X \leftrightarrow S \in G$, which combined with $X \to Y \in G$ implies $Y \leftrightarrow S \in G_Y$.
iv) If $Y \to C \in G_Y$, then $X \to C \in G$ or $Y \to C \in G$, which both imply $X \to C \in G_X$ since $X \to Y \in G$. □

D.3. Proof of Lemma 3

The criterion for X having at least two children or not, as well as condition (1), can be derived from the marginalization rule that $C_1 \leftarrow X \to C_2 \in G$ turns into an edge $C_1 \leftrightarrow C_2$ in G_Y and this is the only way that bidirected edges can arise.

Conditions (2) and (3a) follow from rule 1b)iv) specified in the proof of Lemma 2, and (3b) follows from rule 1b)i).

Turning to conditions (3c) and (3d), if X and Y have the same child C in G_X, G_Y, the structure of $X, Y, C \in G$ can in principle be either of the following:

$$i)\ X \to Y \to C, \quad ii)\ Y \to X \to C, \quad \text{or} \quad iii)\ X \to C \leftarrow Y.$$

To differentiate between them, we include the parents in our consideration. Denoting $P = \text{pa}^G(X) \cap \boldsymbol{Z}, Q = \text{pa}^G(Y) \cap \boldsymbol{Z}$ and $R = \text{pa}^G(C) \cap \boldsymbol{Z}$, we obtain the following differences in the marginal graphs.

	i)	ii)	iii)
G_X	$\begin{tikzcd} P \arrow[d] & Q \arrow[d] & R \arrow[d] \\ X \arrow[r] & C \end{tikzcd}$	$\begin{tikzcd} P \arrow[d] & Q \arrow[d] & R \arrow[d] \\ X \arrow[r] & C \end{tikzcd}$	$\begin{tikzcd} P \arrow[d] & Q \arrow[d] & R \arrow[d] \\ X \arrow[r] & C \end{tikzcd}$
G_Y	$\begin{tikzcd} P \arrow[d] & Q \arrow[d] & R \arrow[d] \\ Y \arrow[r] & C \end{tikzcd}$	$\begin{tikzcd} Q \arrow[d] & P \arrow[d] & R \arrow[d] \\ Y \arrow[r] & C \end{tikzcd}$	$\begin{tikzcd} Q \arrow[d] & P \arrow[d] & R \arrow[d] \\ Y \arrow[r] & C \end{tikzcd}$

These differences further imply different relations for the sets of parents in the marginal graphs:
i) $\text{pa}^{G_X}(X) \subseteq \text{pa}^{G_Y}(Y) \subseteq \text{pa}^{G_X}(X) \cup \text{pa}^{G_X}(C)$,
ii) $\text{pa}^{G_Y}(Y) \subseteq \text{pa}^{G_X}(X) \subseteq \text{pa}^{G_Y}(Y) \cup \text{pa}^{G_Y}(C)$,
iii) $\text{pa}^{G_X}(X) \subseteq \text{pa}^{G_Y}(C),\ \text{pa}^{G_Y}(Y) \subseteq \text{pa}^{G_X}(C)$.
Combining them yields the conditions in the lemma.

Finally, we prove that if neither of the conditions in the lemma apply, we can not identify the edge type. First, note that if no condition in the lemma is satisfied, then either $\text{ch}^{G_X}(X) = \text{ch}^{G_Y}(Y) = \{C\}$ and two of the conditions i) - iii) apply at the same time, or both have no child in the marginal graphs and $\text{pa}^{G_X}(X) \subseteq \text{pa}^{G_Y}(Y)$ or vice versa. For all these cases, we need to find two DAGs G_1, G_2 on the entire set of nodes with different edge types between X and Y but with the same marginalizations. In the case of one common child and i) as well as ii), we can define G_1, G_2 by

$$\text{pa}^{G_i}(X) = \text{pa}^{G_X}(X), \text{pa}^{G_i}(W) = \text{pa}^{G_Y}(W) \text{ for all } W \in \boldsymbol{W} \setminus \{X\}, i = 1, 2.$$

Additionally, in G_1 we include $X \to Y \to C$, and in G_2 we add $Y \to X \to C$. If i) and iii) hold, we define G_1, G_2 by

$$\text{pa}^{G_i}(X) = \text{pa}^{G_X}(X), \text{pa}^{G_i}(W) = \text{pa}^{G_Y}(W) \text{ for all } W \in \boldsymbol{W} \setminus \{X\}, i = 1, 2,$$

and include $X \to Y \to C$ in G_1 , as well as $X \to C \leftarrow Y$ in G_2. All other cases work similarly. □

D.4. Proof of Theorem 5

(1) For a matrix $M \in \mathbb{R}^{l\times m}$, and an index $i \in [l]$, $M_{i:,:}$ denotes the submatrix of all rows starting from the ith one. Similarly, for a subset $A \subseteq [l]$, $M_{A,:}$ is the submatrix consisting of all rows with indices in A, and $M_{\widehat{A},:}$ the submatrix that arises by omitting all A rows. Throughout the proof we assume that $\boldsymbol{W}$ is enumerated as $\boldsymbol{W} = (X, Y, Z_1, \dots, Z_k)$. Following Salehkaleybar et al. (2020), we can rewrite (2) as $\boldsymbol{W} = M\boldsymbol{N}$ with the "mixing matrix" $M := (I - \Lambda)^{-1}$ which linearly combines the independent "sources" $N_1, \dots, N_{k+2}$. Note that $I - \Lambda$ is invertible since Λ is strictly lower triangular after applying simultaneous row and column permutations, and the entries of M coincide with the total causal effects defined via (3). By observing only the variables $(X, \boldsymbol{Z})$, we have a (slightly) over-complete ICA with $k + 1$ observed variables and $k + 2$ sources, namely

$$\begin{pmatrix} X \\ \boldsymbol{Z} \end{pmatrix} = M_{\widehat{2},:} N \tag{11}$$

where $M_{\widehat{2},:}$ is the submatrix with the row for $W_2 = Y$ missing, and, similarly

$$\begin{pmatrix} Y \\ \boldsymbol{Z} \end{pmatrix} = M_{\widehat{1},:} N. \tag{12}$$

The main idea of the proof is to identify $M_{\widehat{2},:}$ from $P(X, \boldsymbol{Z})$, as well as $M_{\widehat{1},:}$ from $(Y, \boldsymbol{Z})$ and then combine them to reconstruct M. From Theorem 15 in Salehkaleybar et al. (2020), if Y has at least two children, then,

a) if Y has a unique (with respect to the topological order) oldest child W_j, then $M_{\widehat{2},:}$ can be identified up to swapping the columns corresponding to Y and W_j and up to rescaling of the column corresponding to Y.
b) Otherwise, $M_{\widehat{2},:}$ can be identified uniquely up to rescaling the Y column.

If Y has exactly one child W_j, then the column in $M_{\widehat{2},:}$ corresponding to Y is a multiple of the column for W_j, in formulas,

$$M_{\widehat{2},2} = \lambda_{j,2} M_{\widehat{2},j}.$$

If Y has no children at all, $M_{\widehat{2},2} = 0$. In both cases, obtaining $\boldsymbol{N}'$ from $\boldsymbol{N}$ by omitting N_2, and, in the case of one child, additionally replacing N_j by $N'_j = N_j + \lambda_{j,2} N_2$, the vector $(X, \boldsymbol{Z})$ fulfills the complete, and therefore identifiable, ICA model

$$\begin{pmatrix} X \\ \boldsymbol{Z} \end{pmatrix} = M_{\widehat{2},\widehat{2}} N'.$$

Thus,

c) if Y has at most one child, the submatrix $M_{\widehat{2},\widehat{2}}$ can be identified uniquely.

Relating back to M, in all cases, $P(X, \boldsymbol{Z})$ uniquely determines $M_{\widehat{2},\widehat{2}}$ or $M_{\widehat{2,\{2,j\}}}$, with W_j being the oldest child of Y. In scenario a), additionally, we know two candidate columns A, B, where either $M_{\widehat{2},2}, M_{\widehat{2},j} = A, B$ up to rescaling or vice versa. To find the correct assignment, we use the information obtained from $P(Y, \boldsymbol{Z})$; that is, we identified all columns of $M_{\widehat{1},:}$ except for column 1, and at most one other column. In particular, we determined $M_{3:,2}$ or $M_{3:,j}$. Therefore, comparing whether $A_{2:}$ or $B_{2:}$ coincides with $M_{3:,2}$ or $M_{3:,j}$, yields correct assignment as well as correct scale. However, this fails in one exceptional case, specifically when $A_{2:} = B_{2:}$ up to scaling. Writing out

the entries in $M_{2:,1}$ $M_{2:,j}$ in terms of the λ_{ij}, and using faithfulness, we obtain that this can occur only if $\text{ch}(Y) = \{X, Z_j\}$ and $\text{ch}(X) = \{Z_j\}$, which is excluded in the assumptions of the theorem.

In case b), the only ambiguity in $M_{\widehat{2},:}$ concerns the scaling of the Y column, which again can derived from the information we have on $M_{\widehat{1},:}$.

The same holds for reversed roles. So, if for both (11), and (12) identifiability cases a) or b) apply, we can infer $M_{\widehat{2},:}$, and $M_{\widehat{1},:}$ and combine them to M.

If case c) applies in one of the ICAs, assume the one related to $(X, \boldsymbol{Z})$, we are still missing the value of m_{21}. Since Y has multiple, and X at most one child, according to Lemma 3, $X \to Y \in G$ if and only if X has multiple children in G_X. If so, we can choose one of these children C. Employing that in the joint model, all directed paths from X to C go through Y, we obtain

$$m_{21} = m_{C1}/m_{C2}.$$

If $X \not\to Y$, $\lambda_{21} = 0$, which determines m_{21} via $\Lambda = I - M^{-1}$.

If case c) applies in both ICAs, then both X, Y have at most one child, which corresponds to conditions (3a)-(3d) in Lemma 3. In cases (3a)-(3c), we know $X \not\to Y$, and therefore $\lambda_{21} = \lambda_{12} = 0$, which gives m_{12}, m_{21} via $\Lambda = I - M^{-1}$. In case (3d), we know that X and Y have the same child C in the marginal graphs, whereas in the joint graph, $X \to Y \to C$ without a direct connection between X and C, since this would contradict the fact that X has only one child. Therefore, as above

$$m_{21} = m_{C1}/m_{C2}.$$

Moreover, we that find $m_{12} = 0$ due to acyclicity. Again, the same holds for reversed roles. Finally, we can compute Λ as $\Lambda = I - M^{-1}$.

(2) To prove the identifiability of $P(X, Y, \boldsymbol{Z})$, we use the fact that once the projected mixing matrices $M_{\widehat{1},:}$, $M_{\widehat{2},:}$ in (11), (12) are known, under the genericity assumption on the moments, all cumulants of the exogenous sources $\boldsymbol{N}$ can be identified (citation removed for anonymous submission). These cumulants uniquely determine $P(\boldsymbol{N})$, which, combined with the overall mixing matrix M, yields $P(X, Y, \boldsymbol{Z})$. □

Appendix E. Additional experiments

E.1. How often do Lemmas 2 - 3 succeed in excluding edges?

While the measurements depicted in Figure 3 give insight into how often the Lemmas find at least one pair without edge per graph, which is the crucial factor for the realizability of LOVO, another interesting question is what proportion of unlinked pairs are recognized by the lemmas as such. To address this, Figure 12 compares the average number of detected unlinked pairs (blue) to the number of pairs in the graph that are actually not connected, which is expected to be $(1-p) \cdot \binom{10}{2}$ in an Erdős–Rényi DAG with edge probability p, and $(1-p)(1-q) \cdot \binom{10}{2}$ for ADMGs with directed edge probability p and bidirected edge probability q (grey).

E.2. LOVO applied to DirectLiNGAM and RCD with varying sample size

This section provides more details on the experiment described in 5.2. Specifically, we analyze the LOVO prediction loss for varying sample sizes, which is expected to relate to the accuracy of the learned graphs. Indeed, the mean SHD of a marginal graph learned with DirectLiNGAM

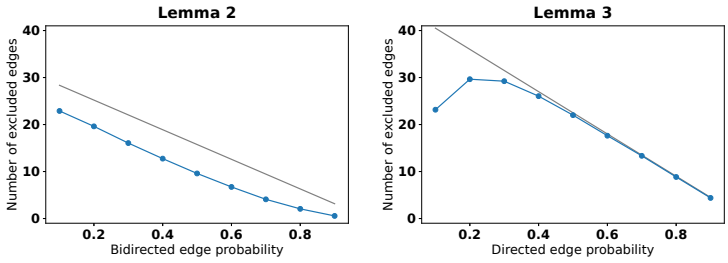

Figure 12: For Lemma 3 and $p \geq 0.3$, the average number of detected absent edges (blue) is close to the true number of absent edges (grey), whereas Lemma 2 does not find all absent edges.

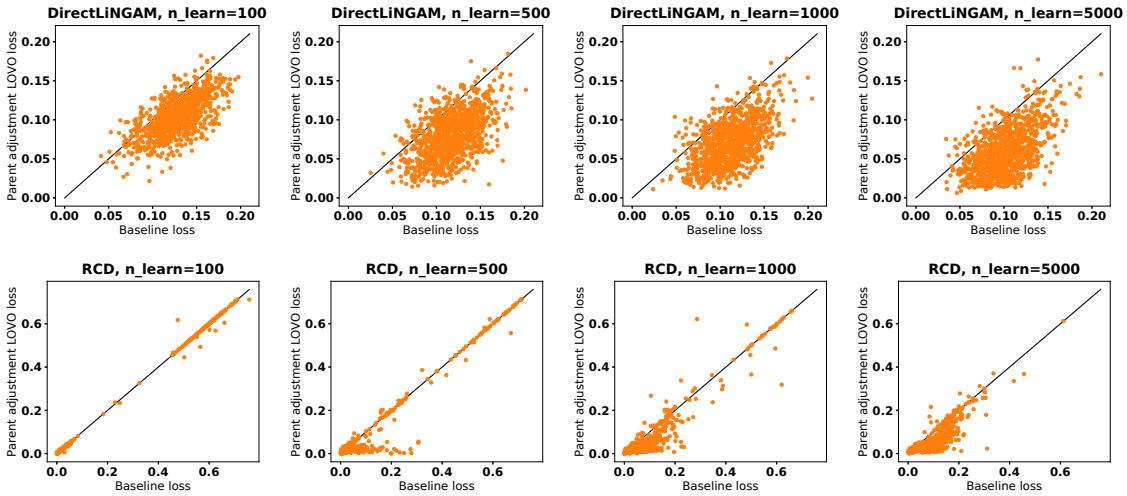

Figure 13: Evolution of LOVO prediction loss when learning sample sizes increases, and by that, the accuracy of the learned graphs increases.

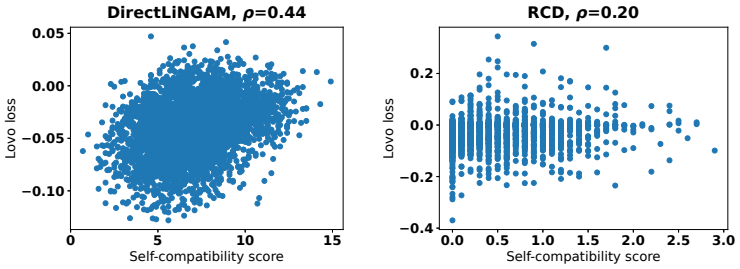

Figure 14: While the self-compatibility score and LOVO loss are significantly correlated, the two metrics show distinct patterns, reflecting their different objectives of incompatibility counting and LOVO prediction.

is $12.3, 6.9, 6.5, 6.5$ for $n_{\text{learn}} = 100, 500, 1000, 5000$, respectively, and $4.0, 2.4, 1.5, 0.3$ for RCD. Figure 13 shows that the LOVO prediction loss tends to decrease with increased learning sample size. A notable observation is that, for RCD with smaller sample sizes, the LOVO loss is very close to the baseline loss and often abstains from making predictions, doing so in about 52% of the replications. This is due to RCD almost always only learning bidirected and no directed edges, meaning it does not commit to any causal directions, which makes it harder to challenge its output; compare Section A11 in Faller et al. (2024). In the cases where predictions are made, the learned union of parents, which is the adjustment set in LOVO, is almost always empty. Consequently, $\hat{\rho}^{\text{LOVO}} = 0$, and both prediction errors are close to the absolute value of true correlation ρ_{XY}. Moreover, the scatter plot reveals two clusters cluster characterized by error values below and above 0.3. The cluster with lower errors corresponds to pairs correctly identified as unlinked, while the other cluster contains pairs where an edge exists. Note that we can relate the points in the scatter plot to pairs, even though each point represents the cross-validation error averaged over all pairs, since the cross-validation error was often only computed using one pair and the LOVO predictor abstained for all other pairs. The cluster related to higher values gets smaller for increased sample size and eventually disappears. Also the number of how often LOVO abstains decreases; to $37\%, 29\%, 3\%$ for $n_{\text{learn}} = 500, 1000, 5000$. In contrast, for DirectLiNGAM, LOVO abstained more rarely for lower sample sizes, specifically in $0.3\%, 2\%, 1\%, 23\%$ of the replications for $n_{\text{learn}} = 100, 500, 1000, 5000$.

E.3. Comparison to Self-Compatibility

To compare with the self-compatibility score from Faller et al. (2024), we repeat the same experiment. Recall that the self-compatibility score runs the algorithm in question on various subsets and counts the incompatibilities the resulting graphs entail. Figure 14 plots the LOVO loss against the self-compatibility score evaluated on RCD and DirectLiNGAM. For the data-generation, the same parameters as in the preceding subsection are used, and we collect the results for all the four different values of n_{learn} into one plot.

E.4. LOVO applied to DirectLiNGAM and RCD on real-world data

We evaluate RCD and DirectLiNGAM on two real-world data sets: a protein-signaling data set from Sachs et al. (2005) with 7466 samples across 11 features and a biodiversity and productivity data

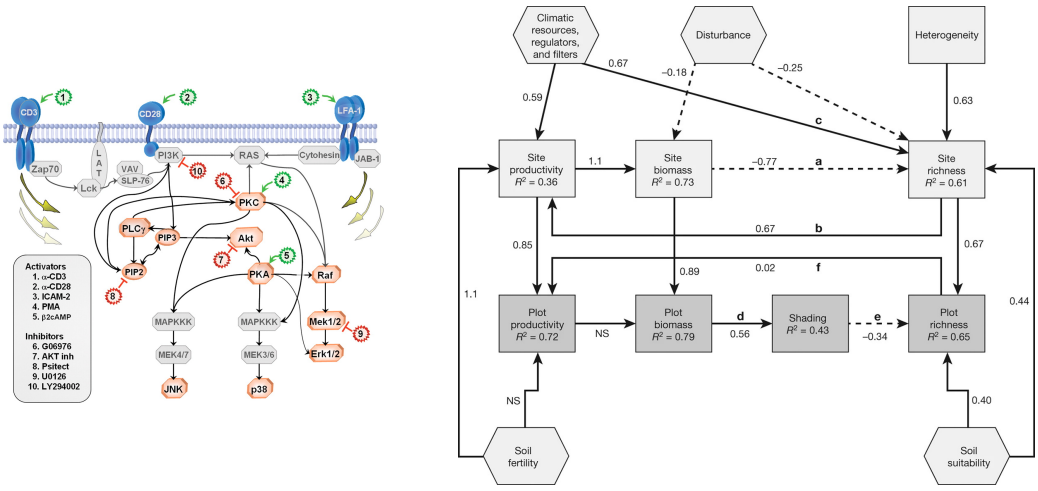

Figure 15: For both data sets, the original papers postulate ground-truth causal graphs, which are visualized here. The left plot shows the Sachs data; all the new nodes refer to features in the data set, and the remaining ones are unobserved variables.

Table 2: The LOVO and baseline errors for the real-world data sets suggest that neither DirectLiNGAM nor RCD is a suitable algorithm for these data.

		DirectLiNGAM	RCD	"Ground truth"
Sachs et al. (2005)	LOVO error	0.152	-	0.187
	Baseline error	0.146	-	0.187
Grace et al. (2016)	LOVO error	0.403	-	0.354
	Baseline error	0.376	-	0.436

set from Grace et al. (2016) comprising 1126 samples and 12 variables (Figure 15). For the second data set, the original study estimated effects using two separate models: one for the upper seven variables and another for the remaining variables in the lower three rows, except *soil fertility*. Given the larger sample size in the second model, where the variables are measured for each of the 1126 *plots* versus the coarser-scale *sites* used for the upper variables, we focus our analysis on the second model.

On both datasets, RCD very rarely commits to a non-trivial output in the sense that it almost always returns a complete ADMG, which indicates that the datasets do not satisfy the method's assumptions. Given these outputs not entailing any statements, LOVO always abstains. In contrast, DirectLiNGAM always returns a DAG, and we can compute the LOVO cross-validation error, which is worse than the baseline in both cases; see Table 2. As a comparison, we also apply LOVO to the hypothesized ground truth graphs depicted in Figure 15. The required marginal graphs G_X, G_Y are derived via the marginalization rules. For the protein dataset, the LOVO prediction error is comparable to the baseline error, while for the ecology dataset, the LOVO error is smaller.

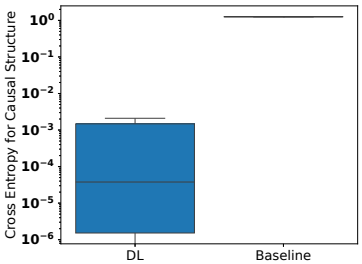

Figure 16: Prediction error for recovering the causal structure from the learned representation of the deep learning model on unseen examples (as described in Section 5.3). Our prediction model has a lower loss than the baseline in all examples.

E.5. Further details for Subsection 5.3

Architecture The main component of the deep learning model used in Section 5.3 is the encoder from Ke et al. (2023). This encoder gets as input a data matrix $\boldsymbol{W} \in \mathbb{R}^{n \times d}$, where $n \in \mathbb{N}$ is the number of samples and $d \in \mathbb{N}$ is the number of variables (in our case $d = 2$). Initially, each entry of this matrix is embedded into $\mathbb{R}^{h/2}$ with a linear transformation. We also embed the column index (i.e. the node identity) of every entry into $\mathbb{R}^{h/2}$. Unlike Ke et al. (2023), we use another linear transformation for this. We concatenate this input embedding and identity embedding to get an h-dimensional representation of every entry. Eventually, we add a row of zeros to the initial data matrix $\boldsymbol{W}$, which will be used later for the encoder summary. This gives us an initial embedding $e^0 \in \mathbb{R}^{n+1 \times d \times h}$.

Ke et al. (2023) propose to alternatingly use attention blocks that calculate attention weights between samples for every node and between nodes for every sample. At attention layer i we expect an input tensor $e^{i-1} \in \mathbb{R}^{n+1 \times d \times h}$ and start by applying a classical self-attention block to all matrices $e^{i-1}_j \in \mathbb{R}^{d \times h}, j = 1, \dots, n$ that result from indexing the sample dimension in e^{i-1}. Implementation-wise, this amounts to passing e^{i-1} to a standard attention layer and considering the first dimension as batch dimension. We then apply a feed-forward layer. This results in a tensor $\hat{e}^i \in \mathbb{R}^{n+1 \times d \times h}$. The second attention block is then applied to the matrices $\hat{e}^i_{:,j}, j = 1, \dots, d$, that result from indexing the nodes, i.e., we reshape the tensor to have the second dimension as batch-dimension (and reshape it back after the attention block). After every attention block, we add a feed-forward layer, and we add a pre-layer norm and a residual connection to every attention block and feed-forward layer.

The final encoder summary is obtained by another attention block, where we consider the column dimension the batch dimension again, and we only use the $(n+1)$-th row as key and all other rows as queries. This gives us a final embedding $e \in \mathbb{R}^{d \times h}$, which we flatten to be in $\mathbb{R}^{d \cdot h}$.

Instead of the decoder proposed by Ke et al. (2023), we add another feed-forward layer that receives the concatenated embeddings of each marginal dataset as input and outputs a scalar.

For the second model that predicts the causal structure from the hidden representations $e_{X,Z}$ of the first model, we simply used a feed forwards layer with four output dimensions and a softmax layer to encode the four possible causal structures $\{\rightarrow, \leftarrow, \leftrightarrow, \text{no edge}\}$ as categories.

Training Note that Ke et al. (2023) propose to train the model on a dataset, where each "datapoint" consists of a synthetically generated adjacency matrix as target with a matrix containing multiple

Table 3: Hyperparameters for training the deep learning models.

Hyperparameter	Value
Batch size	1
Learning rate	1e-4
Gradient clipping value	10
Epochs	2
Encoder layers	3
Feed forward hidden layers	1
Feed forward widening	4
Activation	GELU
Attention heads	8
Hidden dimension h	64
Samples per dataset	3000
Test examples	100
Loss Transitive Prediction	MSE
Training examples Transitive Prediction	100000
Loss Structure from Embedding	Cross-entropy
Training examples Structure from Embedding	1000

samples drawn from this graph as input features. For every such datapoint, we generate one of the 12 DAGs consisting of two nodes and three edges (compare Table 1) with equal probability, and data with sample size $3n$ as in Section 5.1. Again, we split the data into three equal-sized subsets. Let $m \in \mathbb{N}$ be the number of datapoints of the training set and denote the sample matrices for the j-th adjacency matrix with

$$\begin{aligned} M^j_{X,Z} &:= \{x^j_i, z^j_i\}_{i=1,\dots,n} \\ M^j_{Y,Z} &:= \{y^j_i, z^j_i\}_{i=n+1,\dots,2n} \\ M^j_{X,Y} &:= \{x^j_i, y^j_i\}_{i=2n+1,\dots,3n}. \end{aligned}$$

We then solve the minimization problem

$$\arg\min_{f\in\mathcal{F}} \sum_{j=1}^{m} \left(f(M^j_{X,Z}, M^j_{Y,Z}) - \hat{\rho}^j_{X,Y} \right)^2,$$

where $\mathcal{F}$ is the function class defined by the model architecture and $\hat{\rho}^j_{X,Y}$ is the correlation coefficient computed from the third part of the samples $M^j_{X,Y}$. The parameter settings of the main model training are summarized in Table 3. Most notably, we generated 100000 pairs of marginal data matrices and respective ground truth correlation coefficient $\rho_{X,Y}$.

The second model (that is trained to predict the causal structure) is simply a feed-forward network with a single hidden layer. To train it, we generate $k \in \mathbb{N}$ more marginal samples $\{x^j_i, z^j_i\}_{i=1,\dots,M}$ and apply the encoder from the pre-trained model above to get a dataset

$$\{(e^j_{X,Z}, s_j)\}_{j=1,\dots,k},$$

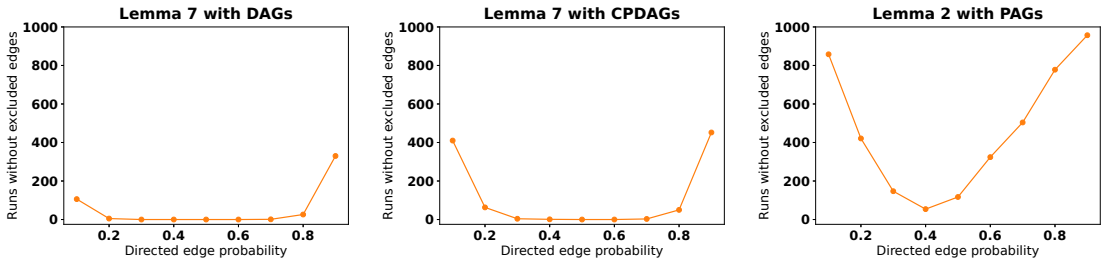

Figure 17: Lemma 7 detects at least one feasible pair in almost all runs for $p \in [0.2, 0.8]$ and DAGs as well as CPDAGs. Lemma 2 employed with PAGs finds feasible pairs in most runs when $p \in [0.3, 0.5]$, but often does not find any when p is close to 0 or 1.

where $s_j \in \{\rightarrow, \leftarrow, \leftrightarrow, \not\sim\}$ denotes the true underlying causal structure that generated the j-datapoint. We use a cross-entropy loss to train this model. The second model is trained on 1000 pairs of embeddings $e_{X,Z}$ and (one-hot encoded) underlying structure. We trained the LOVO prediction model using a squared loss and the second model using the cross-entropy loss. Unless stated otherwise, we used the same parameters for the main model and the second model.

Computational resources The main deep learning model from Section 5.3 was trained on an AWS EC2 instance of type `p3.2xlarge`. These machines contain Intel Xeon E5-2686-v4 processors with 2.3 GHz and 8 virtual cores as well as 61 GB RAM. The training ran less than an hour. All inference and further experiments were run on a MacBook Pro with Apple M1 processor and 32 GB RAM and can be run in less than an hour.

E.6. Evaluating the LOVO variants for DAGs, CPDAGs, and PAGs

This section sheds light on our LOVO version for DAGs, CPDAGs, and PAGs as proposed in Appendix B. As in Section 5.1, we first explore how often Lemmas 7 for DAGs and CPDAGs and Lemma 2 using the limited information entailed by PAGs succeed in finding unlinked pairs. Again, we report the mean number of excluded edges in each run (Figure 18) as well as the number of runs where no single unlinked pair can be found (Figure 17). Next, we employ the LOVO prediction based on the true marginal graphs and simulated data and plot the LOVO loss in comparison to the baseline loss in Figure 19. For DAGs, we use parent adjustment as before, whereas for CPDAGs and PAGs, we adjust using the set of all nodes on paths between X and Y excluding colliders as described in Section B.3.

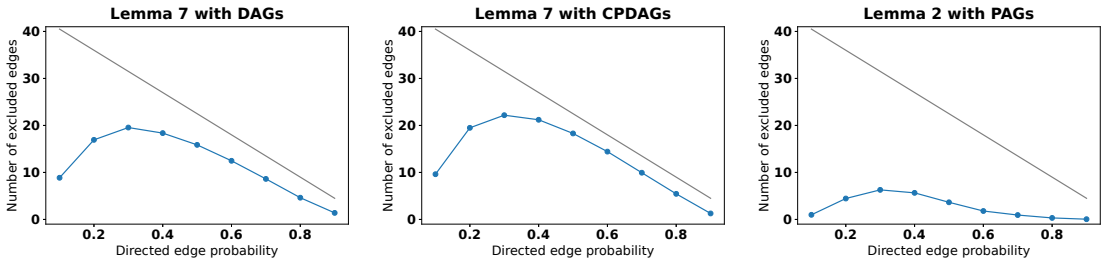

Figure 18: For Lemma 7 and $p \geq 0.3$, the average number of detected absent edges (blue) is close to the true number of absent edges (grey), whereas Lemma 2 does not find all absent edges.

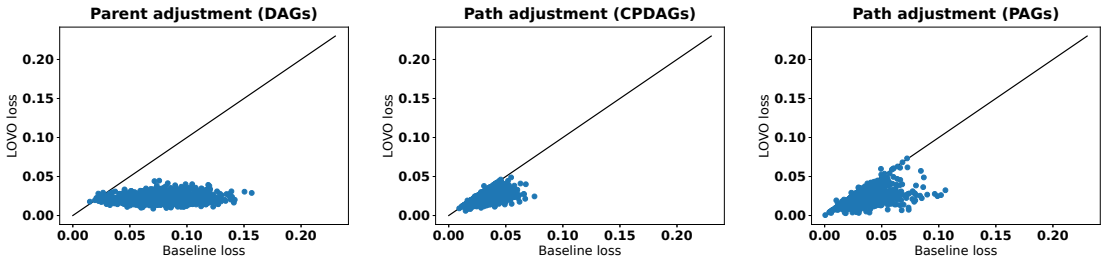

Figure 19: In all three cases, the LOVO loss rarely surpasses the baseline loss. For DAGs, we observe that the LOVO error is considerably smaller in most cases, while for CPDAGs and PAGs, the errors are often identical. This can be attributed to the same phenomenon encountered when applying LOVO to RCD in 5.2: we frequently need to abstain, resulting in only one or a few pairs contributing to the cross-validation average. If all these pairs have the empty set as their adjustment set, the baseline loss aligns with the LOVO loss.