Separation-Based Distance Measures for Causal Graphs

Jonas Wahl

German Research Centre for
Artificial Intelligence (DFKI)

Jakob Runge

Center for Scalable Data Analytics and
Artificial Intelligence (ScaDS.AI) Dresden/Leipzig,
TU Dresden
and
Institute of Computer Engineering
and Microelectronics, TU Berlin

Abstract

Assessing the accuracy of the output of causal discovery algorithms is crucial in developing and comparing novel methods. Common evaluation metrics such as the structural Hamming distance are useful for assessing individual links of causal graphs. However, many state-of-the-art causal discovery methods do not output single causal graphs, but rather their Markov equivalence classes (MECs) which encode all of the graph's separation and connection statements. In this work, we propose additional measures of distance that capture the difference in separations of two causal graphs which link-based distances are not fit to assess. The proposed distances have low polynomial time complexity and are applicable to directed acyclic graphs (DAGs) as well as to maximal ancestral graph (MAGs) that may contain bidirected edges. We complement our theoretical analysis with toy examples and empirical experiments that highlight the differences to existing comparison metrics.

1 INTRODUCTION

Inferring causal relations from observational data in the form of a *causal graph* is a highly challenging task for which causality researchers have proposed numerous algorithms. While the assumptions on which these algorithms rely differ widely, many existing approaches, including the PC algorithm (Spirtes et al., 1993), and its descendants stablePC (Colombo and Maathuis, 2014), consistentPC (Li et al., 2019), PCMCI+ (Runge, 2020), LPMCI (Gerhardus and Runge, 2020), FCI (Spirtes et al., 1995), as well as GES (Chickering, 2003) have in common that they are provably able to infer the correct causal graph up to *Markov equivalence* in the infinite sample limit if their respective assumptions are fulfilled. More precisely, each of these methods assumes that the process that generated the data under investigation can be represented by a causal graph that belongs to a predefined class $\mathbb{G}$, most commonly the class of *directed acyclic graphs (DAGs)* for the PC algorithm, GES, and many others. The class of graphs $\mathbb{G}$ is equipped with a notion of *separation*, for instance, *d*-separation for DAGs. Separation is a graphical criterion that specifies whether two nodes X, Y in a graph are either *separated* or *connected* by another subset of nodes $\mathcal{S}$. In practice, separation is used to translate conditional independence statements regarding the node variables of causal graphs into graphical language. Two graphs belonging to the same class $\mathbb{G}$ are called *Markov equivalent* if they share exactly the same separation/connection statements.

Even if their inferences are correct, the methods mentioned above do not output the full 'ground truth' graph underlying the data-generating process but only its Markov equivalence class (MEC), or in other words, the totality of separation/connection statements implied by the 'true' graph. Some methods such as LiNGaM (Shimizu et al., 2006)) utilize further assumptions about functional dependencies and noise distributions to go beyond the MEC, but without such more restrictive assumptions, the MEC is all one can infer.

Newly developed causal discovery algorithms are typically validated empirically through a range of sim-

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

ulations and experiments that compare their output graphs to a known ground truth. The difference of the output and the true graph is most commonly measured with the structural Hamming distance (SHD) (Acid and Campos, 2003; Tsamardinos et al., 2006), or false positive/negative edge detection rates. Such metrics can be used to assess the correctness of the presence and absence of individual adjacencies and their orientations which represent direct causal relationships. To assess the quality of an output graph with respect to the downstream task of quantifying total causal effects, further comparison metrics have been developed. The most notable example is the structural intervention distance (SID) (Peters and Bühlmann, 2015), which was recently reframed by Henckel et al. (2024) as a special case of a so-called *adjustment identification distance (AID)* alongside a sped-up algorithm for its computation and numerous other improvements and generalizations. Measures of predictive performance that have been fine-tuned to causal graphs exist as well, see e.g. (Liu et al., 2010; Biza et al., 2020), but so far, causal discovery researchers do not seem to have incorporated them into their performance analyses on a broad scale.

Contributions In this work, we propose to add another family of comparison metrics to the existing canon that compares the implied separation/connection statements of two graphs. Surprisingly, even though separation statements are what is actually being inferred by many algorithms, they are rarely considered in evaluations, with a few notable exceptions, e.g. (Hyttinen et al., 2014). To fill this gap, we discuss two types of separation-based graph distances:

- **s/c-metrics** formalize the simple idea to count all possible separation statements, potentially with a prescribed maximal size of the separating set, and compare their validity in both graphs under investigation. This type of comparison arguably yields the most complete picture, and produces a distance measure that is a proper metric in the mathematical sense, no matter the type of separation under discussion. However, it is of limited practical applicability due to the exponential increase of separation statements that need to be checked when the graph size grows. Even though this issue can be alleviated to a degree by working with Monte-Carlo style random approximations of the metric (see Appendix H), a heavy computational overhead remains. In summary, this type of separation-based evaluation yields the arguably most complete picture but lacks desirable scaling properties.
- **separation distances** provide scalable alternatives that transfer the logic of adjustment-based distances (Henckel et al., 2024) to separations. Instead of considering all possible separation statements in two graphs $\mathcal{G}, \mathcal{H}$, separation distances require the user to specify a *separation strategy* that chooses a single separation set $\mathcal{S}$ for every pair of separable nodes in $\mathcal{H}$, and validate whether $\mathcal{S}$ remains a separating set in the graph $\mathcal{G}$. We present different possible separation strategies for DAGs as well as for maximal ancestral graphs (MAGs) and their MECs represented by CPDAGs and PAGs. Using the toolset developed in (Henckel et al., 2024), we show that the computational complexity of separation distances is of low polynomial order in the number of nodes, making them employable even for large graphs. For MAGs, to our knowledge, these are the first examples of causal comparison metrics beyond the SHD.

We illustrate the differences between separation-based and existing graph distances with a number of toy examples and empirical simulations. We also discuss potential pitfalls in the typical link-based evaluation of causal discovery algorithms, by showing that adding or changing the orientations of few edges on a graph can lead to significant changes in the separations encoded in its MEC. Our findings highlight the importance of utilizing a broad range of evaluation metrics that take into account different causal characteristics of the output of causal discovery algorithms, from direct and total effects to the separation statements we consider here.

2 NOTATION

Throughout this work, we consider causal graphs, generically denoted by $\mathcal{G}$ or $\mathcal{H}$, over a set $\mathcal{V}$ of N nodes. More specifically, we will focus on directed acyclic graphs (DAGs), maximal ancestral graphs (MAGs) and their MECs which can be represented by complete partially directed acyclic graphs (CPDAGs) for DAGs and by partial ancestral graphs (PAGs) for MAGs. We recall that a mixed graph (MG) is a graph that may contain directed $\rightarrow$ or bidirected edges $\leftrightarrow$ and that a path $\pi = (\pi(1), \pi(2), \dots, \pi(n))$ in an MG is a sequence of adjacent edges in which no non-endpoint appears twice. A non-endpoint node C on a path π in a MG is a *collider* if its preceding and succeeding edge both point towards C, and a *non-collider* if it is not a collider. A path π is m-blocked by a subset of nodes $\mathcal{S}$ if $\mathcal{S}$ contains a non-collider on π or if there is a collider on π that does not have any descendants in $\mathcal{S}$. A set $\mathcal{S}$ is said to m-separate[1] two nodes (X, Y) on $\mathcal{G}$ if it

[1] When $\mathcal{G}$ is a DAGs, it is more common to speak of d-separation rather than m-separation.

m-blocks all paths between them, and we denote this by $X \bowtie_{\mathcal{G}} Y|\mathcal{S}$. In this case, $\mathcal{S}$ will also be called a separator for (X, Y). A path that is not m-blocked by the empty set is called m-open. A mixed graph contains an *almost directed cycle* if there are is a directed path $X \to \cdots \to Y$ and a bidirected edge $X \leftrightarrow Y$. A MAG is a mixed graph that (a) does not contain any directed or almost directed cycle (ancestral) and (b) in which any two non-adjacent nodes can be m-separated by some set $\mathcal{S}$ (maximal). Other types of separation, such as σ-separation for cyclic graphs (Bongers et al., 2021) exist as well, and many results of this work, particularly those pertaining to s/c-metrics, generalize to any class of graphs equipped with a notion of separation. The symbol $\mathbb{G}$ will be our generic symbol for a class of graphs on the node set $\mathcal{V}$, which we hide in the notation because it stays fixed throughout this article. For example, we just write $\mathbb{G} = \{\text{DAGs}\}$ for the set of DAGs over $\mathcal{V}$.

We also fix the following conventions:

- $\mathcal{C}$ denotes the set of all triples $(X, Y, \mathcal{S})$ with $X, Y \in \mathcal{V},\ X \neq Y$ and $\mathcal{S} \subset \mathcal{V} \backslash \{X, Y\}$. We implicitly identify triples $(X, Y, \mathcal{S})$ and $(Y, X, \mathcal{S})$. $\mathcal{C}^k$ is the set of triples $(X, Y, \mathcal{S}) \in \mathcal{C}$ with $|\mathcal{S}| = k$. In particular, $\mathcal{C} = \bigsqcup_{k=0}^{N-2} \mathcal{C}^k$.
- We write $\mathcal{C}_{con}(\mathcal{G}) = \{(X, Y, \mathcal{S}) \in \mathcal{C} \mid X \not\bowtie_{\mathcal{G}} Y|\mathcal{S}\}$ for the set of connection statements in a graph $\mathcal{G}$ and $\mathcal{C}_{sep}(\mathcal{G}) = \{(X, Y, \mathcal{S}) \in \mathcal{C} \mid X \bowtie_{\mathcal{G}} Y|\mathcal{S}\}$ for the set of its separations. We also write $\mathcal{C}^k_{con}(\mathcal{G})$ $(\mathcal{C}^k_{sep}(\mathcal{G}))$, for the set of connection (separation) statements of order k, i.e. with $|\mathcal{S}| = k$.

Two causal graphs $\mathcal{G}, \mathcal{H} \in \mathbb{G}$ are called *Markov equivalent* if they share the same separations, i.e. if $\mathcal{C}_{sep}(\mathcal{G}) = \mathcal{C}_{sep}(\mathcal{H})$. This equivalence relation partitions $\mathbb{G}$ into disjoint Markov equivalence classes, and we write $\text{MEC}(\mathcal{G}) = \{\mathcal{G}' | \mathcal{G}' \sim \mathcal{G}\}$ for the MEC of the graph $\mathcal{G}$. If $\mathbb{G} = \{\text{DAGs}\}$, any MEC can be represented uniquely as a CPDAG which may contain directed as well as undirected edges. The CPDAG representing $\text{MEC}(\mathcal{G})$ is defined as the graph with the same skeleton as $\mathcal{G}$ in which an edge is oriented $X \to Y$ if and only if it is oriented in this way for every class member $\mathcal{G}' \in \text{MEC}(\mathcal{G})$. The MECs of MAGs can be represented as PAGs which may contain directed, bidirected, undirected and semidirected ($\circ\!\!\to$) edges. The PAG representing $\text{MEC}(\mathcal{G})$ is the graph with the same skeleton as $\mathcal{G}$ in which an edge mark is drawn if and only if the edge mark appears in every class member $\mathcal{G}' \in \text{MEC}(\mathcal{G})$. If we start with a PAG or CPDAG $\mathcal{H}$, we write $\text{MAG}(\mathcal{H})$, respectively $\text{DAG}(\mathcal{H})$ for the set of all MAGs/DAGs in the respective MEC, i.e. $\text{MAG}(\mathcal{H}) = \text{MEC}(\mathcal{G})$ for any $\mathcal{G} \in \text{MAG}(\mathcal{H})$.

We also adopt the following terminology for graphical reasoning. A node Y is a parent of a node X in a mixed graph if there is a directed edge $Y \to X$ and a child if there is a directed edge $Y \leftarrow X$. If there is a bidirected edge $Y \leftrightarrow X$, then Y is called a *sibling*[2] of X. Y is an *ancestor* of X if there exists a directed path from Y to X and a *descendant* of X if there is a directed path from X to Y. Two nodes that share a common child are called *spouses*. We write $\text{ch}_{\mathcal{G}}(X), \text{pa}_{\mathcal{G}}(X), \text{an}_{\mathcal{G}}(X), \text{des}_{\mathcal{G}}(X), \text{sib}_{\mathcal{G}}(X), \text{sp}_{\mathcal{G}}(X)$ for the children, parents, ancestors, descendants, siblings or spouses of X in $\mathcal{G}$ respectively. Y is a *possible parent* of X in a MAG $\mathcal{G}$ if $Y \in \text{ppa}_{\mathcal{G}}(X) := \bigcup_{\mathcal{G}' \in \text{MEC}(\mathcal{G})} \text{pa}_{\mathcal{G}'}(X)$ and a *possible ancestor* if $Y \in \text{pan}_{\mathcal{G}}(X) := \bigcup_{\mathcal{G}' \in \text{MEC}(\mathcal{G})} \text{anc}_{\mathcal{G}'}(X)$. Clearly possible parents/ancestors only depend on the MEC of a graph and can be read off its graphical representation. For instance, Y is a possible parent of X in a DAG $\mathcal{G}$ if it is connected to X by an edge $Y \to X$ or $Y - X$ in $\text{CPDAG}(\mathcal{G})$. Hence we can also write $\text{ppa}_{\mathcal{G}}(X)$ whenever $\mathcal{G}$ is a CPDAG or a PAG. We use the shortcut notation $\text{pa}_{\mathcal{G}}(X \cup Y) := \text{pa}_{\mathcal{G}}(X) \cup \text{pa}_{\mathcal{G}}(Y)$, and similarly for children, spouses etc. If $\mathcal{T} \subset \mathcal{V}$ is a subset of nodes of a graph $\mathcal{G}$, then $\mathcal{G}_{\mathcal{T}}$ is the subgraph of $\mathcal{G}$ over $\mathcal{T}$.

3 EXISTING DISTANCE MEASURES

The Structural Hamming Distance The SHD (Acid and Campos, 2003; Tsamardinos et al., 2006) is the most commonly used metric to compare two causal graphs. If $\mathcal{G}, \mathcal{H}$ are two causal graphs, $\text{SHD}(\mathcal{G}, \mathcal{H})$ is defined as the number of node pairs that do not have the same type of edge between them in both graphs. For instance, the node pair (X, Y) contributes to $\text{SHD}(\mathcal{G}, \mathcal{H})$ if $X \to Y$ in $\mathcal{G}$ but there is either no edge between X and Y in $\mathcal{H}$ or an edge that is oriented differently. The SHD is popular as it can be computed fast with only a few lines of code and it can be adapted to graphs of arbitrary types including those representing MECs. However, the SHD has also been criticized for not properly capturing the *full* causal implications of the differences in the graphs $\mathcal{G}$ and $\mathcal{H}$, as it only assesses the presence/absence of direct effects. We illustrate this critique with the following toy example which shows that small structural changes to a graph can change its causal implications considerably.

Example 1 The DAGs $\mathcal{G}$ and $\mathcal{H}$ in Figure 1 differ in the orientation of one edge only, meaning that their

[2] These conventions unfortunately differ across the literature. We decided to employ the family tree inspired conventions in which sibling means hidden common parent and spouse means common child.

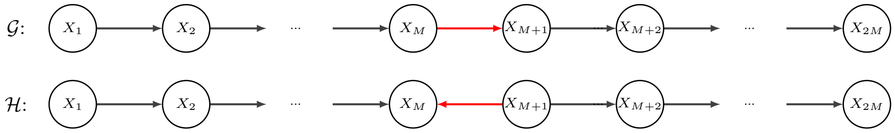

Figure 1: Two DAGs $\mathcal{G}$ and $\mathcal{H}$.

SHD is low relative to the total number of node pairs to the point of being negligible for a large number of nodes. On the other hand, any two nodes are connected by an open causal path in $\mathcal{G}$, while in $\mathcal{H}$ any node on the left of the collider X_M is d-separated from every node on the right. In particular, no causal effects are permeated from the left to the right of the graph, and one could therefore argue that causally these graphs are not very close. In Section 5 we also show empirically that edge removal or reversal as in this example, has a much more pronounced effect on separation- and adjustment-based distance measures than on the SHD.

Adjustment Identification Distances AIDs (Henckel et al., 2024) are designed to compare DAGs with a view on causal effect estimation downstream tasks, and they include the structural intervention distance (SID) of Peters and Bühlmann (2015) as a special case. The core idea of AIDs is to choose an adjustment strategy that produces identification formulas for causal effects implied by the graph $\mathcal{H}$. In a second step, one then needs to verify whether the inferred identification formulas are valid in the base graph $\mathcal{G}$. Whenever this is not the case, a penalty of 1 is added to the distance score. We summarize their exact definitions in Appendix E. Henckel et al. (2024) provide an implementation of AIDs that achieves a computational complexity of only $\mathcal{O}(N^2)$ in sparse graphs. In addition, they propose an adaptation of AIDs to CPDAGs. However, there is no obvious relationship between the AID of two MECs represented by CPDAGs and their class members as the following example illustrates.

Example 2 We first consider the two DAGs $\mathcal{G}', \mathcal{H}'$ in Figure 2 which are both causal chains, but with a different variable ordering. Clearly, these graphs are very different structurally as they do not share a single edge. This difference is witnessed by the parent-AID (= SID) when applying it to the DAGs $\mathcal{G}', \mathcal{H}'$ directly. However, in either of the associated CPDAGs, there are no identifiable causal effects. Therefore, the CPDAG-parent-AID as defined in (Henckel et al., 2024) is zero. As a consequence, an MEC inferred by a causal discovery method may score perfectly when comparing it to the CPDAG of a known ground truth DAG, while *any* of the class members get a non-zero score. In other words, there is no guarantee that the AID between two MECs is bounded below (above) by the minimal (maximal) value obtained by their class members.

Our main motivation for introducing separation distances is to define metrics that measure causal claims that are invariant under Markov equivalence. While this is desirable when the output of the discovery task are MECs, it comes at the cost that such distances can never distinguish members of the same MEC by construction. Therefore, when the output of a discovery method is more fine-grained than the MEC, separation distances should be combined with other metrics such as AIDs or the SHD. More generally, we are of the opinion that an evaluation across a broad range of metrics is desirable in most cases anyway. Nevertheless, even when more than the MEC can be inferred, a separation-based comparison is still a valid endeavor. Inferring separations is a valuable aspect of a correct discovery *even if* the method aims for more.

4 SEPARATION-BASED DISTANCES

s/c-Metric Arguably, the most straightforward way to compare the implied separations/connections of two causal graphs is to define a graded sum over all possible separation/connection statements. If $\mathcal{G}$ is a causal graph with an appropriate notion of separation, we use the separation indicator function $\iota_{\mathcal{G}} : \mathcal{C} \to \{0, 1\}$,

$$\iota_{\mathcal{G}}(X, Y, \mathcal{S}) = \begin{cases} 1 & \text{if } X \not\bowtie_{\mathcal{G}} Y | \mathcal{S} \\ 0 & \text{if } X \bowtie_{\mathcal{G}} Y | \mathcal{S}. \end{cases}$$

Definition 4.1. Consider two causal graphs $\mathcal{G}, \mathcal{H}$ over the same set of N nodes, equipped with an appropriate notion of separation. Define

$$d^k_{s/c}(\mathcal{G}, \mathcal{H}) := \frac{1}{|\mathcal{C}^k|} \sum_{(X,Y,\mathcal{S}) \in \mathcal{C}^k} |\iota_{\mathcal{G}}(X, Y, \mathcal{S}) - \iota_{\mathcal{H}}(X, Y, \mathcal{S})|$$

for $k = 0, \ldots, N-2$. The **s/c-metric up to order K** is defined as

$$d^{\leq K}_{s/c}(\mathcal{G}, \mathcal{H}) = \frac{1}{K+1} \sum_{k=0}^{K} d^k_{s/c}(\mathcal{G}, \mathcal{H}).$$

If $K = N - 2$, we speak of the **full s/c-metric** and write $d_{s/c}$ instead of $d^{\leq N-2}_{s/c}$.

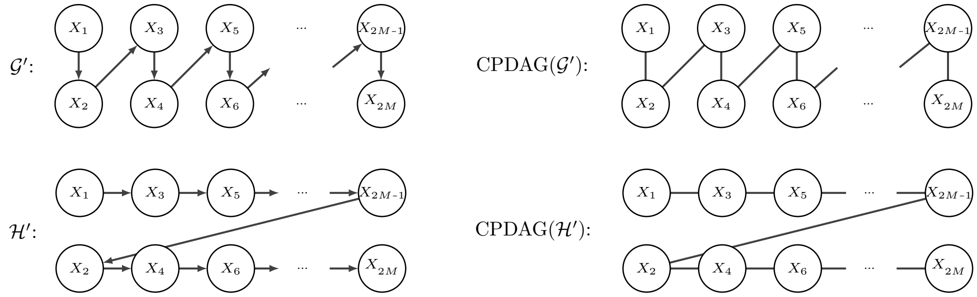

Figure 2: Two DAGs $\mathcal{G}', \mathcal{H}'$ and their CPDAGs.

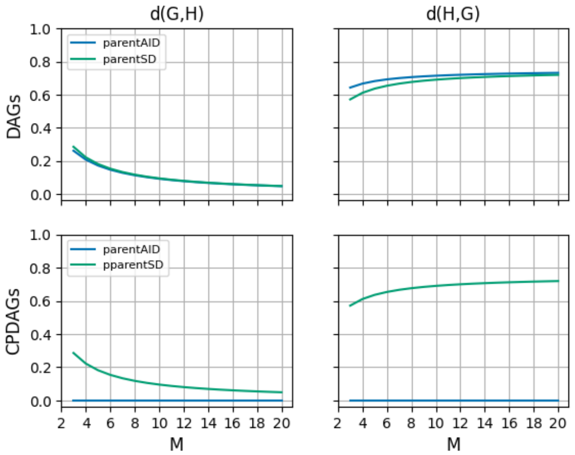

Figure 3: Empirical comparison between the parent-AID and the separation-based metrics *parent-SD* and *pparent-SD* to be defined in Section 4 on the DAGs $\mathcal{G}', \mathcal{H}'$ and their CPDAGs. The label M refers to the index in Figure 2. The qualitative difference between the adjustment-based metric and its separation-based analog is negligible for DAGs but clearly visible for CPDAGs.

In words, the s/c-metric is a weighted count of the disagreement in separation/connection statements between the two graphs. It defines a mathematical metric on MECs, see Lemma D.2 in Appendix D. If a graph $\mathcal{G}$ is fixed as a reference point, for instance because it is the ground-truth in a simulated experiment, we can also compare the separation and connection statements implied by $\mathcal{G}$ separately to those implied by $\mathcal{H}$. This yields notions of *false positive and false negative rate for separations*, see Appendix D for exact definitions and a more detailed discussion. Informal versions of such error rates were already used in the simulation section of (Hyttinen et al., 2014) but have not been adopted more broadly.

Separation-based measures of this kind have an obvious drawback. The number of separation statements $(X, Y, \mathcal{S})$ grows exponentially in the number of nodes so that they are slow to compute and do not scale to large graphs. Therefore, in practice, a manageable subset of separation statements needs to be selected to keep computations manageable. For this, there are two possible approaches. The first one is to choose statements randomly to compute a Monte-Carlo style approximation of the full metric. The second approach is to select only few separating statements according to a predefined deterministic strategy. We provide a plot of the quality of Monte-Carlo style approximation in Appendix H. In the main text, we focus on the second approach, since it delivers a greater computational speed-up and requires a deeper theoretical investigation.

Separation Strategies Separation strategies are the analog of adjustment strategies Henckel et al. (2024) for separation. Roughly speaking, the fundamental idea of Henckel et al. (2024) is to associate to every pair of nodes (X, Y) in a DAG $\mathcal{H}$ a so-called adjustment set $\mathrm{ad}^{\mathcal{H}}(X, Y)$ according to a fixed strategy, and to validate whether the induced adjustment formula for computing the causal effect of X on Y remains valid in a second DAG $\mathcal{G}$. This second DAG $\mathcal{G}$ is typically (but not necessarily) the ground truth in a simulated experiment. (Henckel et al., 2024) explicitly consider parent adjustment, ancestor adjustment, and optimal adjustment (Runge, 2021) as possible adjustment strategies. We will propose separation analogs of these strategies for DAGs, MAGs and their MECs (CPDAGs and PAGs). This task is non-trivial along two dimensions: first, the direct analogs of parent and ancestor adjustment, parent separation and ancestor separation, while valid for DAGs, are not valid separation strategies for general mixed graphs, see Figure 6 in Appendix B. Based on the work (van der Zander and Liśkiewicz, 2020), we will therefore also introduce ZL-separation for MAGs, named after that paper's authors. Secondly, while being a separating set is invariant under Markov equivalence, the specific *choice* *of separating set*, if done naively, may depend on the MEC member (more on this below).

We fix a class of graphs $\mathbb{G}$ with an appropriate notion of separation in which any pair of non-adjacent nodes can be separated. We primarily have the classes $\mathbb{G} = \{\text{DAGs}\}, \{\text{CPDAGs}\}, \{\text{MAGs}\}, \{\text{PAGs}\}$ with m-separation in mind, but we will briefly discuss extensions to cyclic graphs towards the end of this section.

Definition 4.2. A *separation strategy (sep-strategy)* for $\mathcal{G} \in \mathbb{G}$ is a map $\mathfrak{S}_{\mathcal{G}}(\cdot, \cdot)$ that associates to every pair of non-adjacent nodes (X, Y) of $\mathcal{G}$ a set $\mathfrak{S}_{\mathcal{G}}(X, Y)$ such that $X \bowtie_{\mathcal{G}} Y | \mathfrak{S}_{\mathcal{G}}(X, Y)$. A *universal sep-strategy for* $\mathbb{G}$ is a map $\mathfrak{S} : \mathcal{G} \mapsto \mathfrak{S}_{\mathcal{G}}$ that associates a sep-strategy to every $\mathcal{G} \in \mathbb{G}$.

Given two graphs $\mathcal{G}, \mathcal{H} \in \mathbb{G}$ and a sep-strategy $\mathfrak{S}_{\mathcal{H}}(\cdot, \cdot)$ for the latter, we define the indicator function

$$\iota^{\mathfrak{S}}_{\mathcal{G},\mathcal{H}}(X, Y) = \begin{cases} 1 & \text{if } X \not\bowtie_{\mathcal{G}} Y | \mathfrak{S}_{\mathcal{H}}(X, Y) \\ 0 & \text{if } X \bowtie_{\mathcal{G}} Y | \mathfrak{S}_{\mathcal{H}}(X, Y). \end{cases}$$

for any pair of $\mathcal{G}$-non-adjacent nodes $(X, Y) \notin \mathcal{E}_{\mathcal{H}}$.

Definition 4.3 ($\mathfrak{S}$-separation distance)**.** Let $\mathfrak{S}$ be a universal sep-strategy for $\mathbb{G}$. The (normalized) $\mathfrak{S}$-*separation distance* ($\mathfrak{S}$-SD) of $\mathcal{G}$ and $\mathcal{H} \in \mathbb{G}$ is defined as

$$d^{\mathfrak{S}}(\mathcal{G}, \mathcal{H}) := \frac{1}{N(N-1)} \sum_{(X,Y) \in \text{nadj}(\mathcal{H})} \iota^{\mathfrak{S}}_{\mathcal{G},\mathcal{H}}(X, Y),$$

where the sum runs over non-adjacent node pairs in $\mathcal{H}$.

Thus, the $\mathfrak{S}$-SD measures whether the separating set for (X, Y) on $\mathcal{H}$ that has been selected according to the strategy $\mathfrak{S}$ remains a separating set on $\mathcal{G}$, and if this is not the case, a penalty is incurred.

Remark 4.4. Like the adjustment identification distances of Henckel et al. (2024), separation distances are not symmetric, i.e. $d^{\mathfrak{S}}(\mathcal{G}, \mathcal{H}) \neq d^{\mathfrak{S}}(\mathcal{H}, \mathcal{G})$, as the two graphs play different roles in the distance's computation. We also define a symmetric $\mathfrak{S}$-SD of $\mathcal{G}$ and $\mathcal{H}$ by taking the geometric mean of both directions

$$d^{\mathfrak{S}}_{\text{sym}}(\mathcal{G}, \mathcal{H}) := \tfrac{1}{2} \left(d^{\mathfrak{S}}(\mathcal{G}, \mathcal{H}) + d^{\mathfrak{S}}(\mathcal{H}, \mathcal{G}) \right).$$

Other symmetrizations, e.g. taking the harmonic mean instead of the geometric one, are also possible.

Since any DAG is a MAG, any universal sep-strategy for $\{\text{MAGs}\}$ is also a universal sep-strategy for $\{\text{DAGs}\}$. Similarly, any universal sep-strategy $\mathfrak{S}$ for $\{\text{CPDAGs}\}$ defines a universal sep-strategy $\mathfrak{S}'$ for $\{\text{DAGs}\}$ by $\mathfrak{S}'_{\mathcal{G}} = \mathfrak{S}_{\text{MEC}(\mathcal{G})}$ as any separating set for a MEC is a separating set for all its members. We note the subtlety that the converse of this statement need not be true because a sep-strategy $\mathfrak{S}$ defined on DAGs *may not be well-defined* on CPDAGs, in the sense that we may have $\mathfrak{S}_{\mathcal{G}}(X, Y) \neq \mathfrak{S}_{\mathcal{G}'}(X, Y)$ for some nodes X, Y even if $\mathcal{G} \sim \mathcal{G}'$ are Markov equivalent.

Before discussing further theoretical properties of SDs, we present several possible sep-strategies in increasing order of generality. For proofs of all results of this section, we refer to Appendix B.

Parent Separation on DAGs As for adjustment, the most straightforward idea to separate non-adjacent nodes in a DAG is by using their parents. This yields the symmetric sep-strategy $\mathfrak{S}_{\mathcal{G}}(X, Y) = \mathfrak{S}_{\mathcal{G}}(Y, X) = \text{pa}_{\mathcal{G}}(X \cup Y),\ (X, Y) \notin \mathcal{E}_{\mathcal{G}}$. However, parent separation is only a universal sep-strategy on DAGs. On MAGs, two non-adjacent nodes might not be separable through their parents, see Figure 6 in Appendix B for a counterexample. In addition, parent separation does not directly extend to CPDAGs, as generically $\text{pa}_{\mathcal{G}}(X \cup Y) \neq \text{pa}_{\mathcal{G}'}(X \cup Y)$ for Markov equivalent $\mathcal{G} \sim \mathcal{G}'$. We call the corresponding separation distance for DAGs the *parent-SD.*

Ancestor Separation on DAGs Another universal sep-strategy for DAGs that parallels adjustment is ancestor separation, i.e. $\mathfrak{S}_{\mathcal{G}}(X, Y) = (\text{an}_{\mathcal{G}}(X \cup Y)) \setminus \{X, Y\},\ (X, Y) \notin \mathcal{E}_{\mathcal{G}}$. Again, this is not a valid sep-strategy for MAGs, see Figure 6, nor does it extend to CPDAGs. We call the corresponding separation distance the *ancestor-SD.*

p-Parent Separation on CPDAGs Replacing parents by possible parents yields the sep-strategy for $\mathfrak{S}_{\mathcal{G}}(X, Y) = \text{ppa}_{\mathcal{G}}(X \cup Y),\ (X, Y) \notin \mathcal{E}_{\mathcal{G}}$ whenever $\mathcal{G}$ is a CPDAG. We prove that possible parents are valid separators in Appendix B, Lemma B.1.

ZL-Separation on MAGs and their MECs To define a universal sep-strategy that remains valid on MAGs, we first recall that a set of nodes $\mathcal{S}$ is a minimal separator for (X, Y) on a graph $\mathcal{G}$ if $X \bowtie_{\mathcal{G}} Y | \mathcal{S}$ and $X \not\bowtie_{\mathcal{G}} Y | \mathcal{S}'$ for any proper subset $\mathcal{S}' \subsetneq \mathcal{S}$. (van der Zander and Liśkiewicz, 2020) describe a fast algorithm that returns a minimal separator for non-adjacent nodes X, Y in any MAG $\mathcal{G}$. We will call this separator the *ZL*-separator $\text{ZL}_{\mathcal{G}}(X, Y)$, and $\mathfrak{S}_{\mathcal{G}}(X, Y) = \text{ZL}_{\mathcal{G}}(X, Y)$ defines a universal sep-strategy for MAGs. Since defining a separator as the output of an algorithm is inconvenient, we will now characterize it differently. The algorithm of (van der Zander and Liśkiewicz, 2020) computes the ZL-separator by first computing a so-called *nearest separator* to (X, Y) and the refining this nearest separator to a minimal one. A set of nodes $\mathcal{S} \subset \text{pan}_{\mathcal{G}}(X \cup Y)$ is called a *nearest separator* relative to (X, Y) on a MAG $\mathcal{G}$ if (i) $X \bowtie_{\mathcal{G}} Y | \mathcal{S}$; and (ii) for every $W \in \text{pan}_{\mathcal{G}}(X \cup Y) \setminus \{X, Y\}$ and every path π connecting W and Y on the moralized graph[3] $(\mathcal{G}_{\mathrm{pan}_{\mathcal{G}}(X \cup Y)})^m$ such that $\mathrm{nodes}(\pi) \cap \mathcal{S} \neq \emptyset$, any other set $\mathcal{S}' \subset \mathrm{pan}_{\mathcal{G}}(X \cup Y)$ that separates (X, Y) must also satisfy $\mathrm{nodes}(\pi) \cap \mathcal{S}' \neq \emptyset$.

Lemma 4.5. *Let X, Y be non-adjacent nodes on a MAG $\mathcal{G}$, and let $\mathcal{S}$ be a nearest separator for (X, Y). Then any separator $\mathcal{S}' \subset \mathcal{S}$ is also nearest for (X, Y).*

Theorem 4.6. *Let X, Y be non-adjacent nodes on a MAG $\mathcal{G}$. The ZL-separator $\mathrm{ZL}_{\mathcal{G}}(X, Y)$ is the unique separator that is both minimal and nearest for (X, Y). Moreover, $\mathrm{ZL}_{\mathcal{G}}(X, Y) = \mathrm{ZL}_{\mathcal{G}'}(X, Y)$ for Markov equivalent MAGs $\mathcal{G} \sim \mathcal{G}'$.*

We call the separation distance based on ZL-separation the *ZL-SD*. Due to Theorem 4.6, the ZL-SD is a well-defined separation distance on MAGs and their MECs.

MB-enhanced Sep-Strategies In this paragraph, we present a way to adapt a universal sep-strategy $\mathfrak{S}$ using the *Markov blanket (MB)* to produce a new universal sep-strategy $\mathrm{MB}(\mathfrak{S})$. This MB-enhancement can be carried out on any of the considered classes of graphs but its primary advantage is for DAGs/CPDAGs where the MB-enhanced strategy can be computed at lower computational cost than the original strategy, at least for sparse graphs. To define MB-enhancement, we recall that the *Markov blanket* $\mathrm{MB}_{\mathcal{G}}(X)$ of a node X in a MAG $\mathcal{G}$ is defined as the smallest set of nodes $\mathcal{S}$ with the property that $X \bowtie_{\mathcal{G}} Y | \mathcal{S}$ for all $Y \notin \mathcal{S}$. If $\mathcal{G}$ is a DAG, the Markov blanket of X can be conveniently characterized as the union of its parents $\mathrm{pa}_{\mathcal{G}}(X)$, its children $\mathrm{ch}_{\mathcal{G}}(X)$ and its spouses $\mathrm{sp}_{\mathcal{G}}(X)$, i.e. the nodes that share a common child with X. For general MAGs, the Markov blanket can be characterized graphically as well but this characterization is slightly more involved, see Appendix C. We observe that $\mathrm{MB}_{\mathcal{G}}(X) = \mathrm{MB}_{\mathcal{G}'}(X)$ for Markov equivalent $\mathcal{G} \sim \mathcal{G}'$, see Lemma C.1, so that the Markov blanket is well-defined on MECs.

Now, given a universal sep-strategy $\mathfrak{S}$ on a class of graphs $\mathbb{G}$, we define the *MB-enhanced sep-strategy* as

$$\mathfrak{S}_{\mathcal{G}}(X, Y) = \begin{cases} \mathrm{MB}_{\mathcal{G}}(X) & \text{if } Y \notin \mathrm{MB}(X) \\ \mathfrak{S}_{\mathcal{G}}(X, Y) & \text{else.} \end{cases}$$

for two non-adjacent nodes X, Y. Thus, we use the Markov blanket of X as a separator for all nodes except for those that are non-adjacent to X but part of $\mathrm{MB}_{\mathcal{G}}(X)$ themselves. For these nodes, we employ the initial sep-strategy. The advantage of the MB-enhanced strategy over the original one is then that the MB-separator only depends on X and not on Y for 'most' node pairs (X, Y). In DAGs/CPDAGs of bounded node degree, the number of exceptional cases is small compared to the number of nodes N which allows us to leverage the results of Henckel et al. (2024) to achieve a faster implementation of complexity $\mathcal{O}(N^2)$, see below. On MAGs, this speed-up is only retained under stronger assumptions, see Appendix C.

Extensions to cyclic graphs The s/c-metric can be defined for any class of graphs under consideration as long as this class is accompanied by a fitting notion of separation. To extend separation distances to other types of graphs, it is necessary to specify a fitting sep-strategy. For mixed graphs with cycles and σ-separation, a sep-strategy can be defined by noting that two nodes in such a graph $\mathcal{G}$ can be σ-separated by $\mathcal{S}$ if and only if they can be m-separated by $\mathcal{S}$ in the *acyclification* $\mathrm{ac}(\mathcal{G})$ of $\mathcal{G}$ (Bongers et al., 2021). Therefore, the problem of defining a sep-strategy can be reduced to defining a sep-strategy in mixed graphs. Moreover, any mixed graph can be projected onto a MAG over the same nodes with equivalent m-separations. Therefore, a valid sep-strategy can be defined by projecting $\mathrm{ac}(\mathcal{G})$ onto its MAG projection and employing a sep-strategy, such as ZL-separation, for MAGs.

Further properties of $\mathfrak{S}$-SDs In this section, we focus on the theoretical properties of separation distances of MAGs which are also inherited by DAGs. We denote the undirected skeleton of a MAG $\mathcal{G}$ by $\mathrm{sk}(\mathcal{G})$. We also recall that a triple of nodes (X, Y, Z) is called an *unshielded triple* if Y is adjacent to both X and Z but X and Z are non-adjacent. The unshielded triple (X, Y, Z) is an *unshielded collider* if both edges have arrowheads pointing into Y. We denote the set of all unshielded colliders on a $\mathcal{G}$ by $\mathcal{U}_c(\mathcal{G})$. Finally, in a MAG $\mathcal{G}$, a path π between nodes X and Y is a *discriminating path* for node V if (i) π includes at least three edges; (ii)V is a non-endpoint of π and adjacent to Y on π; and (iii) X and Y are non-adjacent in $\mathcal{G}$ and every node between X and V on π is both a collider on π and a parent of Y. Richardson and Spirtes (2002) showed that two MAGs $\mathcal{G}, \mathcal{H}$ over the same nodes are Markov equivalent if and only if (a) they share the same skeleton; (b) they share the same unshielded colliders and (c) if π is a discriminating path for node V in both graphs, then V is a collider on π in $\mathcal{G}$ if and only if it is a collider on π in $\mathcal{H}$.

We have already discussed that two Markov equivalent MAGs satisfy $d^{\mathfrak{S}}(\mathcal{G}, \mathcal{H}) = 0$. The next result shows to what extent graphs at zero distance are similar.

Theorem 4.7. *Let $\mathfrak{S}$ be a universal sep-strategy for MAGs. If $d^{\mathfrak{S}}(\mathcal{G}, \mathcal{H}) = 0$, then*

(i) $\mathrm{sk}(\mathcal{G}) \subset \mathrm{sk}(\mathcal{H})$;

[3] the moralized graph of a MAG $\mathcal{H}$ is the undirected graph in which nodes A and B share an edge if and only if on $\mathcal{H}$ there is a path between them on which all nodes except A and B are colliders. We denote it by $\mathcal{H}^m$.

(ii) If (X, Y, Z) is an adjacent triple[4] *on both graphs, and an unshielded collider on $\mathcal{H}$, then it is an unshielded collider on $\mathcal{G}$.*

(iii) if π is a discriminating path for node V in both graphs, and V is a collider on π in $\mathcal{H}$, then it is a collider on π in $\mathcal{G}$.

In particular, $d^{\mathfrak{S}}_{sym}(\mathcal{G}, \mathcal{H}) = 0$ if and only if $\mathcal{G}$ and $\mathcal{H}$ are Markov equivalent.

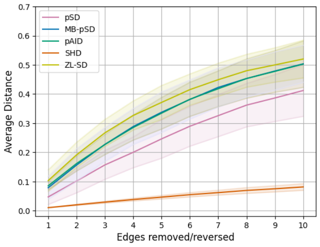

Figure 4: Effect of edge removal and reversal on different distance metrics. Parent-AID and SD are affected much more strongly by such local operations than the SHD.

Complexity of Implementations The computational complexity of our algorithms may depend on the sep-strategy and we provide detailed computations in Appendix F. In general, the distance computation consists of two steps: a separator computation step in the graph $\mathcal{H}$ and a verification step on the graph $\mathcal{G}$ that checks whether the separators $\mathfrak{S}_{\mathcal{H}}(X, Y)$ found on $\mathcal{H}$ remain separators on $\mathcal{G}$. For non-enhanced sep-strategies excluding ZL-separation, the separator computation is $\mathcal{O}(N^3)$ or lower, for ZL-separation in a MAG this step is $\mathcal{O}(N^4)$ ($\mathcal{O}(N^3)$ on sparse graphs) (van der Zander and Liśkiewicz, 2020). The verification step can be executed with worst-case complexity $\mathcal{O}(N^2)$ ($\mathcal{O}(N)$ on sparse graphs) (van der Zander et al., 2014) per node pair (X, Y) and thus $\mathcal{O}(N^4)$ in total (sparse: $\mathcal{O}(N^3)$). This is one degree slower than the complexity $\mathcal{O}(N^3)$ (sparse: $\mathcal{O}(N^2)$) of the algorithm that computes the parent- or ancestor-AID in (Henckel et al., 2024). For MB-enhanced strategies, however, we can mimic the ideas of Henckel et al. (2024) to achieve a lowered complexity of $\mathcal{O}(N^2)$ on sparse graphs. This is because, thanks to the *Bayes-Ball algorithm* (Geiger et al., 1990; Shachter, 1998), see also (Henckel et al., 2024, Appendix D), we can verify many separation statements while looping over only one node X instead over pairs of nodes (X, Y). The loop over the second node Y only has to be entered in exceptional cases. The number of exceptional cases is bounded by the size of the Markov blanket which is in turn bounded by d^2, where d is the maximal node degree of the graph. Therefore, on sparse graphs with $d = \text{const.}$, we reach a complexity of $\mathcal{O}(N^2)$ similar to the one for AIDs. A similar trick can be applied in the separator computation step to achieve a complexity of $\mathcal{O}(N^2)$. Empirically, computing the MB-enhanced parent-SD on 100 pairs of Erdös-Renyi graphs with $N = 1000$ nodes and an expected number of $10N$ edges, took $0.59s$ on average on an Apple M1max 64GB chip as opposed to $4.35s$ for the non-enhanced parent-SD.

5 EMPIRICAL RESULTS

In Figure 5, we draw an Erdös-Renyi DAG $G \sim G(N, p)$ from which we then randomly remove one edge and reverse the orientation of another edge to obtain a second graph $\mathcal{H}$. We then compare the normalized distances of these two graphs as well as the distances of their CPDAGs. We conduct a similar experiment for mixed graphs which we generate in an Erdös-Renyi fashion with edge density p and a probabilty of 0.25 that a given edge is bidirected. Among SDs, the ZL-SD reacts the most strongly to the edge deletion/reversal. Parent-SD, MB-enhanced parent-SD and parent-AID both also witness the difference in graphs much more clearly than the SHD. On average, the parent- and pparent-SD seem to behave more conservative than their adjustment analog and somewhat interpolate between the SHD and the parent-AID. For mixed graphs/MAGs the only available distances are the SHD and the ZL-SD which exhibit a clear difference in values for sparser graphs. Interestingly, for very dense graphs the ZL-SD even drops below the difference in SHD which is likely due to the fact that in such graphs barely any nodes can be separated anymore. In Figure 4, we plot the change in distance metrics to a base graph after we have removed an edge and reversed another one k-times, $k = 1, \dots, 10$. The base graph is a randomly drawn Erdös-Renyi DAG $G \sim G(25, 0.25)$ and the removed and reversed edges are drawn uniformly random from existing edges. We have repeated this experiment 100 times and Figure 4 shows the averages across these runs. Parent-AID and all SDs are more strongly affected by these removal and reversal operations than the SHD. Remarkably, the Markov enhanced parent-SD behaves very similar to the parent-AID. In Appendix H, we provide further experiments on the correlation of different distances and on s/c-metrics.

In summary, our experiments and theoretical considerations demonstrate that AID, SDs and the SHD measure different notions of similarity of causal graphs, and the choice of metric should depend on the ultimate goal of the causal discovery effort.

[4] that is to say, Y is adjacent to both X and Z

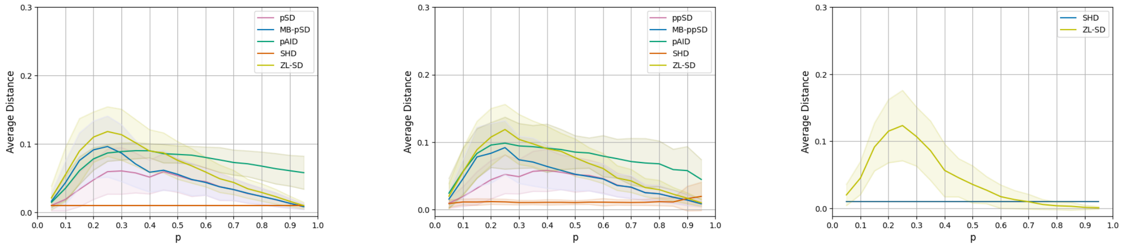

Figure 5: Average values of different graph distances applied to an Erdös-Renyi DAG $\mathcal{G} \sim G(N = 25, p)$ and another DAG $\mathcal{H}$ which is obtained from $\mathcal{G}$ by randomly deleting one edge and reversing one edge orientation. The plot on the left shows the distance values for the DAGs themselves while the plot in the middle compares their CPDAGs for increasing values of p. For mixed graphs, the only available distances are the ZL-SD and the SHD which are depicted in the rightmost plot. We ran 100 experiments per parameter.

	separation-based		adjustment-based	structure-based
	s/c-metric	SD	AID	SHD
compares?	all separation statements	selected separation statements	selected adjustment sets	edges (presence and orientation)
local?	no	no	no	yes

variants	FP-/FN-rates;
	randomly chosen
	sep-statements;
		symmetrization;
		different
		sep-strategies;
			symmetrization;
			different
			adjustment
			strategies;
				edge or orientation
				FP-/FN-rates;

applicable to	DAGs, MAGs, CPDAGS, PAGs cyclic graphs;	DAGs, MAGs, CPDAGS, PAGs cyclic graphs;	DAGs, CPDAGS, further extensions possible;	all graphs;
scales to large graphs?	no	yes	yes	yes

particularly
recommended
for
	small graphs;	MECs,
		graph classes
		beyond DAGs;
			evaluation of
			causal discovery
			for downstream
			effect estimation;
				all graphs;

Table 1: Summary of measures of comparison for causal graphs.

6 CONCLUSION

We have introduced new separation-based graph distances that allow to quantify how similar two causal graphs are in terms of their implied separation statements. Many of these metrics are fast to compute, scalable, and applicable to DAGs, their Markov equivalence classes (CPDAGs) as well as mixed graphs that incorporate bidirected edges for hidden confounding. We have compared them with other metrics through toy examples and empirical experiments. We summarize the properties of the available distance measures in Table 1. Our work is in line with other recent attempts to provide more comprehensive tools to evaluate causal discovery algorithms such as (Henckel et al., 2024; Faller et al., 2024; Ramsey et al., 2024). Considering these recent developments regarding evaluation together with new proposals for more realistic data simulation (Gamella et al., 2025; Andrews and Kummerfeld, 2024; Ormaniec et al., 2024) and the use of more real-world data , we believe that it would be worthwhile to conduct an extensive re-evaluation of popular causal discovery methods across a broad range of measures and data sources in future work.

Code availability An implementation of the distance measures introduced in this work is available in the repository https://github.com/JonasChoice/sep_distances.

Acknowledgements JW and JR received funding from the European Research Council (ERC) Starting Grant CausalEarth under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 948112). JW also received support from the German Federal Ministry of Education and Research (BMBF) as part of the project MAC-MERLin (Grant Agreement No. 01IW24007). We thank Leonard Henckel for valuable discussions and his suggestion to use Markov blankets as separators. We also thank Simon Bing for proof-reading a previous version of this manuscript.

References

Acid, S. and Campos, L. M. d. (2003). Searching for Bayesian Network Structures in the Space of Restricted Acyclic Partially Directed Graphs. *Journal of Artificial Intelligence Research*, 18:445–490.

Andrews, B. and Kummerfeld, E. (2024). Better simulations for validating causal discovery with the dag-adaptation of the onion method. *arXiv preprint arXiv:2405.13100*.

Biza, K., Tsamardinos, I., and Triantafillou, S. (2020). Tuning Causal Discovery Algorithms. In *Proceedings of the 10th International Conference on Probabilistic Graphical Models*, pages 17–28. PMLR.

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

Chickering, D. M. (2003). Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, 3:507–554.

Claassen, T. and Heskes, T. (2012). A Bayesian approach to constraint based causal inference. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, pages 207–216, Arlington, Virginia, USA. AUAI Press.

Colombo, D. and Maathuis, M. H. (2014). Order-Independent Constraint-Based Causal Structure Learning. *Journal of Machine Learning Research*, 15(116):3921–3962.

Eulig, E., Mastakouri, A. A., Blöbaum, P., Hardt, M., and Janzing, D. (2023). Toward Falsifying Causal Graphs Using a Permutation-Based Test. arXiv:2305.09565.

Faller, P. M. and Janzing, D. (2025). On different notions of redundancy in conditional-independence-based discovery of graphical models. *arXiv preprint arXiv:2502.08531*.

Faller, P. M., Vankadara, L. C., Mastakouri, A. A., Locatello, F., and Janzing, D. (2024). Self-compatibility: Evaluating causal discovery without ground truth. In *International Conference on Artificial Intelligence and Statistics*, pages 4132–4140. PMLR.

Faltenbacher, S., Wahl, J., Herman, R., and Runge, J. (2025). Internal incoherency scores for constraint-based causal discovery algorithms. *arXiv preprint arXiv:2502.14719*.

Gamella, J. L., Peters, J., and Bühlmann, P. (2025). Causal chambers as a real-world physical testbed for ai methodology. *Nature Machine Intelligence*, pages 1–12.

Geiger, D., Verma, T., and Pearl, J. (1990). d-separation: From theorems to algorithms. In *Machine intelligence and pattern recognition*, volume 10, pages 139–148. Elsevier.

Gerhardus, A. and Runge, J. (2020). High-recall causal discovery for autocorrelated time series with latent confounders. In *Advances in Neural Information Processing Systems*, volume 33, pages 12615–12625. Curran Associates, Inc.

Heisterkamp, S. H. (2009). Directed acyclic graphs and the use of linear mixed models. *Technical Report*.

Henckel, L., Perković, E., and Maathuis, M. H. (2022). Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):579–599.

Henckel, L., Würtzen, T., and Weichwald, S. (2024). Adjustment identification distance: A gadjid for causal structure learning. In *The 40th Conference on Uncertainty in Artificial Intelligence*.

Hyttinen, A., Eberhardt, F., and Järvisalo, M. (2014). Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, pages 340–349.

Kocaoglu, M. (2023). Characterization and learning of causal graphs with small conditioning sets. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Li, H., Cabeli, V., Sella, N., and Isambert, H. (2019). Constraint-based causal structure learning with consistent separating sets. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Machlanski, D., Samothrakis, S., and Clarke, P. S. (2024). Robustness of algorithms for causal structure learning to hyperparameter choice. In *Causal Learning and Reasoning*, pages 703–739. PMLR.

Mian, O. A., Marx, A., and Vreeken, J. (2021). Discovering Fully Oriented Causal Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8975–8982.

Ormaniec, W., Sussex, S., Lorch, L., Schölkopf, B., and Krause, A. (2024). Standardizing structural causal models. *arXiv preprint arXiv:2406.11601.*

Pellet, J.-P. and Elisseeff, A. (2008). Finding latent causes in causal networks: an efficient approach based on markov blankets. *Advances in Neural Information Processing Systems*, 21.

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

Pitchforth, J. and Mengersen, K. (2013). A proposed validation framework for expert elicited Bayesian Networks. *Expert Systems with Applications*, 40(1):162–167.

Ramsey, J., Spirtes, P., and Zhang, J. (2006). Adjacency-faithfulness and conservative causal inference. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, pages 401–408. AUAI Press.

Ramsey, J. D., Andrews, B., and Spirtes, P. (2024). Choosing dag models using markov and minimal edge count in the absence of ground truth. *arXiv preprint arXiv:2409.20187.*

Richardson, T. and Spirtes, P. (2002). Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962 – 1030.

Runge, J. (2020). Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1388–1397. PMLR.

Runge, J. (2021). Necessary and sufficient graphical conditions for optimal adjustment sets in causal graphical models with hidden variables. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems 34 (NeurIPS 2021).*

Shachter, R. D. (1998). Bayes-ball: Rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 480–487.

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

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*, volume 81 of *Lecture Notes in Statistics*. Springer, New York, NY.

Spirtes, P., Meek, C., and Richardson, T. (1995). Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 499–506.

Textor, J., van der Zander, B., Gilthorpe, M. S., Liskiewicz, M., and Ellison, G. T. (2016). Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology*, 45(6):1887–1894.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78.

van der Zander, B. and Liśkiewicz, M. (2020). Finding minimal d-separators in linear time and applications. In Adams, R. P. and Gogate, V., editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 637–647. PMLR.

van der Zander, B., Liskiewicz, M., and Textor, J. (2014). Constructing separators and adjustment sets in ancestral graphs. In *UAI 2014 Workshop Causal Inference: Learning and Prediction*, page 11.

Wienöbst, M. (2023). On the computational complexity of graph moralization. *Blog post at mwien.github.io.*

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

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes

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

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

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

Separation-based Distance Measures for Causal Graphs: Supplementary Material

A RELATED WORK

The articles (Acid and Campos, 2003; Tsamardinos et al., 2006) introduce the SHD, (Peters and Bühlmann, 2015) introduce the SID. The latter is generalized to a broader class of adjustment identification distances in (Henckel et al., 2024). Liu et al. (2010) and Biza et al. (2020) define algorithms that judge causal graphs based on their predictive abilities for the purpose of hyperparameter selection. In addition, several works address method evaluation and causal graph falsification. Faller et al. (2024) propose a criterion for evaluating causal discovery algorithms in the absence of a ground truth which quantifies the compatibility of causal graphs that were inferred over different subsets of variables. Pitchforth and Mengersen (2013) review methods to validate expert-elicited Bayesian networks. Machlanski et al. (2024) discuss causal model evaluation in the context of causal effect estimation, and Eulig et al. (2023) develop a permutation-based test for causal graph falsification. Ramsey et al. (2024) propose a statistical test for the validity of the Markov property of a distribution on a causal graph. Faltenbacher et al. (2025) investigate how coherently the separations of the output graph of a constraint-based causal discovery method reflect the conditional independencies that where measured during its run. Faller and Janzing (2025) investigate how to use redundant test results to correct errors in the learned graph.

B PROOFS

Lemma B.1. *Let $\mathcal{G}$ be a DAG. For any pair of non-adjacent nodes (X, Y), the set of possible parents $\mathrm{ppa}_{\mathcal{G}}(X \cup Y)$ is a d-separating set.*

To prove Lemma B.1 we recall that separations on CPDAGs can be nicely handled with *definite status paths.* Fix a CPDAG $\mathcal{H}$ and a path $\pi = (\pi(1), \ldots, \pi(n))$ on $\mathcal{H}$. A non-collider $\pi(i)$ is called a *definite non-collider* on π if the triple $(\pi(i-1), \pi(i), \pi(i+1))$ is unshielded or if one of its adjacent edges is oriented away from $\pi(i)$ in $\mathcal{H}$. An arbitrary node on $\pi(j)$ on π is said to be of *definite status on* π if it is a definite non-collider or a collider. Finally the path π is called a *definite status path* if all of its non-endpoint nodes are of definite status. A definite status path π from X to Y is called d-blocked by a set of nodes $\mathcal{S} \subset \mathcal{V} \backslash \{X, Y\}$ if one of its non-colliders is in $\mathcal{S}$ or if one of its colliders does not have any descendants (including itself) in $\mathcal{S}$. Zhang (2008) proved that X and Y are d-separated by a set $\mathcal{S}$ in one (every) DAG in $\mathrm{DAG}(\mathcal{H})$ if and only if every definite status path on $\mathcal{H}$ is d-blocked by $\mathcal{S}$.

Proof of Lemma B.1. We consider $\mathcal{H} := \mathrm{CPDAG}(\mathcal{G})$, consider the pair of non-adjacent nodes (X, Y) and the set of possible parents $\mathcal{S} := \mathrm{ppa}_{\mathcal{H}}(X \cup Y)$. By the previous considerations, we need to show that $\mathcal{S}$ d-blocks any definite status path between X and Y. Let $\pi = (\pi(1), \pi(2), \ldots, \pi(n)), n \geq 3$ be such a definite status path with $\pi(1) = X$ and $\pi(n) = Y$. We will need to consider several different cases. First, if $\pi(1) \leftarrow \pi(2)$ or $\pi(1) \circ\!\!-\!\!\circ \pi(2)$ and similarly if $\pi(n-1) \rightarrow \pi(n)$ or $\pi(n-1) \circ\!\!-\!\!\circ \pi(n)$, then $\mathcal{S}$ contains a non-collider of π and hence π is d-blocked. Therefore, we now assume that $\pi(1) \rightarrow \pi(2)$ and $\pi(n-1) \leftarrow \pi(n)$. Then, there must be a collider on π. Next, we assume that $\mathcal{S}$ d-unblocks π and derive a contradiction.

Case 1: We first consider the case, where there is exactly one collider $C = \pi(i), 1 < i < n$ on π. Then every edge $(\pi(k-1), \pi(k))$ must be directed towards C, i.e. $\pi(k-1) \rightarrow \pi(k)$ for $k \leq i$ and $\pi(k) \leftarrow \pi(k+1)$ for $k \geq i$ as any other orientation would generate an additional collider on π in any DAG $\mathcal{G}' \in \mathrm{DAG}(\mathcal{H})$. Since $\mathcal{S}$ d-unblocks π, there must be a possible parent $P \in \mathrm{ppa}_{\mathcal{H}}(X) \cap \mathrm{des}(C)$. Then the path $X = \pi(1) \rightarrow \pi(2) \rightarrow \ldots C \rightarrow \ldots P \circ\!\!-\!\!\circ X$ would induce a cycle in some DAG $\mathcal{G}' \in \mathrm{DAG}(\mathcal{H})$. The case $P \in \mathrm{ppa}_{\mathcal{H}}(Y) \cap \mathrm{des}(C)$ is analogues.

Case 2: Now we assume that there is more than one collider on π. Let $C = \pi(i_C)$ be the first and $C' = \pi(i_{C'})$ be the last collider on the path , so that in particular $i_C < i_{C'}$. By the same reasoning as in Case 1, the subpath $(\pi(1), \ldots, \pi(i_C)$ must be right-directed and the subpath $(\pi(i_{C'}), \ldots, \pi(n))$ must be left-directed. If $\mathcal{S}$ d-unblocks π, then there must exist $P \in \mathcal{S} \cap (\text{des}(C) \cup \{C\})$ and $P' \in \mathcal{S} \cap (\text{des}(C') \cup \{C'\})$ with descending paths ξ and ξ' from C to P and C' to P' respectively. If $P \in \text{ppa}_{\mathcal{H}}(X)$, then the concatenation of $(\pi(1), \ldots, \pi(i_C))$, ξ and the edge $P \circ\!\!-\!\!\circ X$ (or $P \to X$) would yield a cycle in some $\mathcal{G}' \in \text{DAG}(\mathcal{H})$. By the same argument, $P' \in \text{ppa}_{\mathcal{H}}(Y)$ would induce a cycle in some $\mathcal{G}' \in \text{DAG}(\mathcal{H})$. Therefore, we must have $P \in \text{ppa}_{\mathcal{H}}(Y)$ and $P' \in \text{ppa}_{\mathcal{H}}(X)$. We now assume that P is the first node on the descending path ξ that is in $\text{ppa}_{\mathcal{H}}(Y)$ (otherwise, we replace P by the first node P^* on ξ with this property). Similarly we assume that P' is the first node on the descending path ξ' that is in $\text{ppa}_{\mathcal{H}}(X)$. Consider the path Π obtained by concatenating in this order the path $(\pi(1), \ldots, \pi(i_C))$, the path $\xi = (\xi(1), \ldots, \xi(k))$, the edge (P, Y), the inverse of the path $(\pi(i_{C'}), \ldots, \pi(n))$, the path $\xi' = (\xi'(1), \ldots, \xi'(l))$ and the edge (P', X). We will show the contradictory conclusion that Π must be a cycle.

If the edge (P, Y) was left-directed $P \leftarrow Y$ in *some* $\mathcal{G}' \in \text{DAG}(\mathcal{H})$, then $\xi'(l-1) \to P = \xi(l) \leftarrow Y$ would be a collider and we show that this is contradictory. Indeed, if this collider was unshielded, then P would be a child of Y in every $\mathcal{G}' \in \text{DAG}(\mathcal{H})$ which contradicts the assumption $P \in \text{ppa}_{\mathcal{H}}(Y)$. However, if the collider was shielded, there would have to be an edge $(\xi(l-1), Y)$ and since P was the first node on ξ that is a possible parent of Y, this edge would have to be directed as $\xi(l-1) \leftarrow Y$. But then the edge (P, Y) must be directed as $P \leftarrow Y$ in *every* $\mathcal{G}'' \in \text{DAG}(\mathcal{H})$ to avoid the cycle $Y \to \xi(l-1) \to P \to Y$. But this again contradicts the assumption $P \in \text{ppa}_{\mathcal{H}}(Y)$. Consequently, the edge (P, Y) must be right-directed $P \to Y$ in *every* $\mathcal{G}' \in \text{DAG}(\mathcal{H})$. We can repeat the same line of reasoning for the edge (P', X) to see that this edge needs to be right-directed as well. But then we have achieved our final contradiction that the path Π must be circular.

In both cases we have derived the desired contradiction and hence $\mathcal{S}$ must d-block π.

□

Lemma B.2. *Let X, Y be non-adjacent nodes on a MAG $\mathcal{G}$, and let $\mathcal{S}$ be a nearest separator for (X, Y). Then any separator $\mathcal{S}' \subset \mathcal{S}$ is also nearest for (X, Y).*

Proof. It suffices to show that $\mathcal{S}'$ satisfies condition (ii) in the definition of nearest separator. Let $W \in \text{pan}_{\mathcal{G}}(X \cup Y) \backslash \{X, Y\}$ and let π be a path connecting W and Y on the moralized graph $(\mathcal{G}_{\text{pan}_{\mathcal{G}}(X \cup Y)})^m$ with $\text{nodes}(\pi) \cap \mathcal{S}' \neq \emptyset$. Since $\mathcal{S}' \subset \mathcal{S}$, we also have $\text{nodes}(\pi) \cap \mathcal{S} \neq \emptyset$, and since $\mathcal{S}$ is nearest, we must have $\text{nodes}(\pi) \cap \mathcal{T} \neq \emptyset$ for any separating set $\mathcal{T} \subset \text{pan}_{\mathcal{G}}(X \cup Y)$, establishing property (ii) for $\mathcal{S}'$. □

Theorem B.3. *Let X, Y be non-adjacent nodes on a MAG $\mathcal{G}$. The ZL-separator $\text{ZL}_{\mathcal{G}}(X, Y)$ is the unique separator that is both minimal and nearest for (X, Y). Moreover, $\text{ZL}_{\mathcal{G}}(X, Y) = \text{ZL}_{\mathcal{G}'}(X, Y)$ for Markov equivalent MAGs $\mathcal{G} \sim \mathcal{G}'$.*

Proof. We first show that $\text{ZL}_{\mathcal{G}}(X, Y)$ is both minimal and nearest for (X, Y). Minimality was already established in (van der Zander and Liśkiewicz, 2020), and by construction of the algorithm that outputs $\text{ZL}_{\mathcal{G}}(X, Y)$ in (van der Zander and Liśkiewicz, 2020), $\text{ZL}_{\mathcal{G}}(X, Y)$ is a subset of a nearest separator. By Lemma B.2, $\text{ZL}_{\mathcal{G}}(X, Y)$ is also nearest. To prove uniqueness, let $\mathcal{S}, \mathcal{S}'$ be two minimal nearest separators for (X, Y). By symmetry we need only show $\mathcal{S} \subset \mathcal{S}'$. Let $W \in \mathcal{S}$. We want to show that $W \in \mathcal{S}'$. Since $\mathcal{S}$ is a minimal separator, there must exist a path Π from X to Y on $\mathcal{G}$ that is not closed by $\mathcal{S} \backslash \{W\}$ but closed by $\mathcal{S}$. In particular, W is the unique element of $\mathcal{S}$ that is a non-collider on Π. By (Richardson and Spirtes, 2002, Lemma 3.17) the sequence of non-colliders on Π forms a undirected path π on $(\mathcal{G}_{\text{pan}_{\mathcal{G}}(X \cup Y)})^m$ from X to Y. The subpath π' starting from W and ending at Y has the property that $\text{nodes}(\pi') \cap \mathcal{S} = \{W\}$. Now, we use the assumption that $\mathcal{S}$ is a nearest separator for (X, Y). Condition (ii) of the definition of nearest separators namely implies that $\text{nodes}(\pi') \cap \mathcal{S}' \neq \emptyset$. If $W \in \text{nodes}(\pi') \cap \mathcal{S}'$, then in particular $W \in \mathcal{S}'$ and we are done. To finish the proof, we show that $W \notin \text{nodes}(\pi') \cap \mathcal{S}'$ leads to a contradiction. If $W \notin \text{nodes}(\pi') \cap \mathcal{S}'$, we choose $W' \in \text{nodes}(\pi') \cap \mathcal{S}'$ and consider the subpath π'' starting from W'. Since $\mathcal{S}'$ was also assumed to be nearest, $\text{nodes}(\pi'') \cap \mathcal{S}' \neq \emptyset$ implies that also $\text{nodes}(\pi'') \cap \mathcal{S} \neq \emptyset$. But π'' was a subpath of $\pi' \backslash \{W\}$ so it follows that $\text{nodes}(\pi') \backslash \{W\} \cap \mathcal{S} \neq \emptyset$ which contradicts $\text{nodes}(\pi') \cap \mathcal{S} = \{W\}$. This concludes the proof.

Finally, we note that minimality and being nearest are invariant under Markov equivalent which proves that $\text{ZL}_{\mathcal{G}}(X, Y) = \text{ZL}_{\mathcal{G}'}(X, Y)$ for Markov equivalent MAGs $\mathcal{G} \sim \mathcal{G}'$. □

Theorem B.4. *Let $\mathfrak{S}$ be a universal sep-strategy for MAGs. If $d^{\mathfrak{S}}(\mathcal{G},\mathcal{H})=0$, then*

1. $\mathrm{sk}(\mathcal{G})\subset\mathrm{sk}(\mathcal{H})$*;*
2. *If (X,Y,Z) is an adjacent triple on both graphs, and an unshielded collider on $\mathcal{H}$, then it is an unshielded collider on $\mathcal{G}$.*
3. *if π is a discriminating path for node V in both graphs, and V is a collider on π in $\mathcal{H}$, then it is a collider on π in $\mathcal{G}$.*

In particular, $d^{\mathfrak{S}}_{sym}(\mathcal{G},\mathcal{H})=0$ if and only if $\mathcal{G}$ and $\mathcal{H}$ are Markov equivalent.

Proof. If X and Y are non-adjacent in $\mathcal{H}$, they are separated by $\mathfrak{S}_{\mathcal{H}}(X,Y)$, and since $d^{\mathfrak{S}}(\mathcal{G},\mathcal{H})=0$, $\mathfrak{S}_{\mathcal{H}}(X,Y)$ also separates them in $\mathcal{G}$. Hence X and Y are also non-adjacent in $\mathcal{G}$, proving the first claim. To establish the second claim, consider an unshielded collider $(X,Y,Z)\in\mathcal{U}_c(\mathcal{H})$ that is also an adjacent triple on $\mathcal{G}$. By the first claim (X,Y,Z) is also unshielded in $\mathcal{G}$. $(X,Y,Z)\in\mathcal{U}_c(\mathcal{H})$ implies that $Y\notin\mathfrak{S}_{\mathcal{H}}(X,Z)$. Since $d^{\mathfrak{S}}(\mathcal{G},\mathcal{H})=0$, $\mathfrak{S}_{\mathcal{H}}(X,Z)$ is therefore a separating set not containing Y on $\mathcal{G}$ as well, so that (X,Y,Z) must be a collider on $\mathcal{G}$. Now, let $\pi=(X,W_1,\dots,W_s,V,Y)$ be a discriminating path for node V between nodes X and Y on both graphs such that V is a collider on π in $\mathcal{H}$. By induction over i, $W_i\in\mathfrak{S}_{\mathcal{H}}(X,Y)$ for $i=1,\dots,s$ and moreover because of the collider property $V\notin\mathfrak{S}_{\mathcal{H}}(X,Y)$. If V was not a collider on π in $\mathcal{G}$, then $\mathfrak{S}_{\mathcal{H}}(X,Y)$ would open the path π in $\mathcal{G}$ which would contradict $d^{\mathfrak{S}}(\mathcal{G},\mathcal{H})=0$. Hence V must be a collider on π in $\mathcal{G}$.

Finally, let us consider the symmetrized distance $d^{\mathfrak{S}}_{\mathrm{sym}}(\mathcal{G},\mathcal{H})$ which is zero if and only if $d^{\mathfrak{S}}(\mathcal{G},\mathcal{H})=0$ and $d^{\mathfrak{S}}(\mathcal{H},\mathcal{G})=0$. So, by the first part of the theorem, both graphs have the same skeleton, the same unshielded triples and the same colliders on shared discriminating paths. Hence they are Markov equivalent. Conversely, if $\mathcal{G}$ and $\mathcal{H}$ are Markov equivalent, then any separating strategy for one graph is a separating strategy for the other, so $d^{\mathfrak{S}}_{\mathrm{sym}}(\mathcal{G},\mathcal{H})=0$. □

The following figure proves that neither parent nor ancestor separation are valid sep-strategies for MAGs.

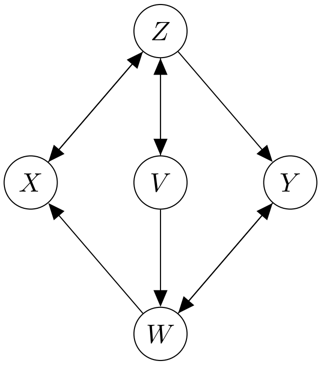

Figure 6: In this MAG $\mathcal{G}$, $\mathrm{pa}_{\mathcal{G}}(X\cup Y)=\{W,Z\}$ does not m-separate X and Y as it unblocks the path $X\leftrightarrow Z\leftrightarrow V\rightarrow W\leftrightarrow Y$. The same is true for the ancestral set $\mathrm{anc}_{\mathcal{G}}(X\cup Y)\backslash\{X,Y\}$ as well as for the set of potential parents $\mathrm{ppa}_{\mathcal{G}}(X\cup Y)$ as both coincide with $\mathrm{pa}_{\mathcal{G}}(X\cup Y)$ in this example.

C MB-SEPARATION IN MAGS

In this section, we extend the Markov blanket separation distance to MAGs. Recall that the Markov blanket of X is the smallest set of nodes of $\mathcal{G}$ with the property that $X\bowtie_{\mathcal{G}}Y|\mathrm{MB}_{\mathcal{G}}(X)$ for all $Y\notin\mathrm{MB}_{\mathcal{G}}(X)$. The following characterization of the Markov blanket for MAGs has been considered in various places, see for instance (Pellet and Elisseeff, 2008). It is based on the notion of collider paths: A path $\pi=(\pi(1),\pi(2),\dots,\pi(n))$ from $X=\pi(1)$ to $Y=\pi(n)$ is a *collider path* if $n>2$ and all middle nodes $\pi(2),\dots,\pi(n-1)$ are colliders on π. The Markov blanket $\mathrm{MB}_{\mathcal{G}}(X)$ of a node X on a MAG $\mathcal{G}$ then consists of all nodes Y such that

(i) Y is adjacent to X or,

(ii) there exists a collider path from Y to X.

We record that the Markov blanket is invariant under Markov equivalence.

Lemma C.1. *If two MAGs $\mathcal{G} \sim \mathcal{G}'$ are Markov equivalent, then for any node X, we have* $\mathrm{MB}_{\mathcal{G}}(X) = \mathrm{MB}_{\mathcal{G}'}(X)$.

Proof. Since m-separation is invariant under Markov equivalence $\mathrm{MB}_{\mathcal{G}}(X)$ has the property that $X \bowtie_{\mathcal{G}'} Y|\mathrm{MB}_{\mathcal{G}}(X)$ for all $Y \notin \mathrm{MB}_{\mathcal{G}}(X)$. Since $\mathrm{MB}_{\mathcal{G}'}(X)$ is the smallest subset with this property, we must have $\mathrm{MB}_{\mathcal{G}'}(X) \subset \mathrm{MB}_{\mathcal{G}}(X)$. Repeating the argument with reversed roles of $\mathcal{G}$ and $\mathcal{G}'$ finishes the proof. □

Therefore if $\mathfrak{S}$ defines a sep-strategy on Markov equivalence classes, then so does its MB-enhanced strategy. For PAGs (and consequently also for MAGs), we can therefore define the MB-enhancement of ZL-separation

$$\mathfrak{S}_{\mathcal{G}}(X,Y) = \begin{cases} \mathrm{MB}_{\mathcal{G}}(X) & \text{if } Y \notin \mathrm{MB}_{\mathcal{G}}(X) \\ \mathrm{ZL}_{\mathcal{G}}(X,Y) & \text{else.} \end{cases}$$

for all non-adjacent pairs (X, Y). In DAGs, the computational advantage of Markov blanket separation over other sep-strategies is that its implementation can avoid a double loop over nodes X and Y for all but the exceptional cases, leading to lower computional complexity if the number of exceptions is small, see Appendix F below. This is the case if the maximal node degree d, i.e. the maximal number of neighbors per node is small compared to the total number of nodes: any $Y \in \mathrm{MB}_{\mathcal{G}}(X)$ that is non-adjacent to X can be reached in exactly two steps, so the number of exceptional cases is d^2 at most. In MAGs, this guarantee can no longer be upheld, since collider paths can be of arbitrary lengths and thus a small node degree is no longer sufficient to bound the size of the set of exceptions. Such bounds can only be retained if in addition to the node degree, the maximal lengths of collider paths is assumed to be small compared to the number of nodes.

D FURTHER RESULTS ON S/C-METRICS

Two causal graphs $\mathcal{G}, \mathcal{H}$ of the same type are called *K-th order Markov equivalent* if their separations/connections coincide up to order K (Kocaoglu, 2023), that is if $d^{\leq K}_{s/c}(\mathcal{G}, \mathcal{H}) = 0$.

Lemma D.1. *Let $\mathbb{G}$ be a class of graphs with an appropriate notion of separation, e.g. $\mathbb{G} = \{MAGs\}$ and m-separation. $d^{\leq K}_{s/c}$ is a metric on the set ${}^{\mathbb{G}}\!/\!_{\sim}$ of K-th order Markov equivalence classes of $\mathbb{G}$.*

Corollary D.2. *Consider a class of graphs $\mathbb{G}$ with an appropriate notion of separation.*

Then, $d_{s/c}(\mathcal{G}, \mathcal{H}) = 0$ if and only if $\mathcal{G}$ and $\mathcal{H}$ are Markov equivalent. Moreover, $d_{s/c}$ defines a metric (in the mathematical sense of the word) on the set of Markov equivalence classes over $\mathbb{G}$.

Proof of Lemma D.1. Consider the real vector space V_k of maps $\mathcal{C}_k \to \mathbb{R}$ on which $\|f\|_k = \frac{1}{\mathcal{C}_k} \sum_{(X,Y,\mathcal{S}) \in \mathcal{C}_k} |f_k(X, Y, \mathcal{S})|$ defines a norm. Then $\|(f_0, \ldots, f_K)\| := \sum_{k=0}^{K} \|f_k\|_k$ defines a norm on the graded vector space $V_K = \bigoplus_{k=0}^{K} V_k$. We now define the mapping

$$g_K : \mathbb{G} \to V_K, \qquad \mathcal{G} \mapsto (1 - i_{\mathcal{G}}|_{\mathcal{C}_0}, \ldots, 1 - i_{\mathcal{G}}|_{\mathcal{C}_K}),$$

where $i_{\mathcal{G}}|_{\mathcal{C}_k}$ is the separation indication function restricted to the set of k-th order separation/connection statements. We observe that $g_K(\mathcal{G}_0) = g_K(\mathcal{G})$ if and only if $\mathcal{G}$ and $\mathcal{H}$ have the same separations up to order K. In other words, the mapping g_K is well-defined on the quotient ${}^{\mathbb{G}}\!/\!_{\sim}$ of K-th order Markov equivalence classes and becomes an embedding. Since $d^{\leq K}_{s/c}(\mathcal{G}, \mathcal{H})$ is nothing but

$$d^{\leq K}_{s/c}(\mathcal{G}, \mathcal{H}) = \|g_K(\mathcal{G}) - g_K(\mathcal{H})\|,$$

it follows directly that $d^{\leq K}_{s/c}$ is a metric on ${}^{\mathbb{G}}\!/\!_{\sim}$. Finally, since the usual notion of Markov equivalence means that two graphs share exactly the same separations, $\mathcal{G}$ and $\mathcal{H}$ are Markov equivalent if and only if $d_{s/c}(\mathcal{G}, \mathcal{H}) = d^{\leq N-2}_{s/c}(\mathcal{G}, \mathcal{H}) = 0$, so this is indeed the special case where $K = N - 2$ by which point all separation statements have been exhausted.

□

Remark D.3. (i) Since $d^k_{s/c}(\mathcal{G},\mathcal{H})$ is bounded by 1, $d^{\leq K}_{s/c}(\mathcal{G},\mathcal{H})$ is also bounded by 1 due to the normalization constant $\frac{1}{K+1}$. This value is taken if $\mathcal{G}$ is the fully disconnected and $\mathcal{H}$ is a fully connected graph.

(ii) If one prefers to assign more importance to differences in low order statements than to differences in higher order statements, this can be reflected by introducing a weight $w_k,\ k = 0,\dots N-2$ and replace the s/c-metric with a weighted version

$$d^w_{s/c}(\mathcal{G},\mathcal{H}) = \frac{1}{\sum_k w_k}\sum_{k=0}^{K} w_k \cdot d^k_{s/c}(\mathcal{G},\mathcal{H}).$$

D.1 Markov and Faithfulness metric

The s/c-metric introduced in the previous subsection is a symmetric notion of distance for two causal graphs to which differences in connections and differences in separations between the two graphs contribute equally. We can also consider separations and connections separately at the price of losing symmetry: we have to specify one graph as a reference point for the separations or connections that we would like to compare. A version of the distance measures that we are about to introduced (without the grading by order) has been used implicitly in (Hyttinen et al., 2014) to evaluate their causal discovery method.

Definition D.4. Consider two causal graphs $\mathcal{G},\mathcal{H}$ over the same set of nodes, equipped with an appropriate notion of separation. We first define

$$d^k_c(\mathcal{G},\mathcal{H}) := \frac{1}{|\mathcal{C}^k_{con}(\mathcal{G})|}\sum_{(X,Y,\mathcal{S})\in\mathcal{C}^k_{con}(\mathcal{G})} (1-\iota_{\mathcal{H}}(X,Y,\mathcal{S})).$$

for $k = 0,\dots,N-2$. Then we call

$$d^{\leq K}_c = \frac{1}{K+1}\sum_{k=0}^{K} d^k_c(\mathcal{G},\mathcal{H}).$$

the **c-metric** or **Markov metric** of $\mathcal{H}$ to $\mathcal{G}$ **of order K**. If $K = N-2$, we just speak of the **c-metric** or **Markov metric** and write d_c instead of d^{N-2}_c.

We recall that a distribution $P_{\mathcal{V}}$ over the node variables $\mathcal{V}$ is called *Markovian* on a causal graph $\mathcal{G}$, if the graphical separation $X \bowtie_{\mathcal{G}} Y|\mathcal{S}$ implies the conditional independence $X \perp\!\!\!\perp_{P_{\mathcal{V}}} Y|\mathcal{S}$, and is called *faithful* on $\mathcal{G}$ if the graphical connection $X \not\bowtie_{\mathcal{G}} Y|\mathcal{S}$ implies the conditional dependence $X \not\perp\!\!\!\perp_{P_{\mathcal{V}}} Y|\mathcal{S}$

The name 'Markov metric' is justified by the following result.

Lemma D.5. *Consider two causal graphs $\mathcal{G},\mathcal{H}$ over the same set of nodes* $\mathbf{X}$*, equipped with an appropriate notion of separation. Suppose that $P_{\mathcal{V}}$ is a distribution on* $\mathbf{X}$ *that is Markovian and faithful on the graph $\mathcal{G}$. Then $d_c(\mathcal{G},\mathcal{H}) \neq 0$, if and only if $P_{\mathcal{V}}$ is not Markovian on $\mathcal{H}$.*

Proof. If $d_c(\mathcal{G},\mathcal{H}) \neq 0$, then there must be a triple $(X,Y,\mathcal{S}) \in \mathcal{C}_{con}(\mathcal{G}) \cap \mathcal{C}_{sep}(\mathcal{H})$. Since $(X,Y,\mathcal{S}) \in \mathcal{C}_{con}(\mathcal{G})$ and $P_{\mathcal{V}}$ is faithful on $\mathcal{G}$, we must have that $X \not\perp\!\!\!\perp_{P_{\mathcal{V}}} Y|\mathcal{S}$. If $P_{\mathcal{V}}$ would be Markovian on $\mathcal{H}$, $(X,Y,\mathcal{S}) \in \mathcal{C}_{sep}(\mathcal{H})$ would imply $X \perp\!\!\!\perp_{P_{\mathcal{V}}} Y|\mathcal{S}$, which is a contradiction. Conversely, if $P_{\mathcal{V}}$ is not Markovian on $\mathcal{H}$, there must be a triple $(X,Y,\mathcal{S}) \in \mathcal{C}_{sep}(\mathcal{H})$ with $X \not\perp\!\!\!\perp_{P_{\mathcal{V}}} Y|\mathcal{S}$. Since $P_{\mathcal{V}}$ is Markovian on $\mathcal{G}$, it follows that $(X,Y,\mathcal{S}) \in \mathcal{C}_{con}(\mathcal{G})$ and thus $(X,Y,\mathcal{S}) \in \mathcal{C}_{con}(\mathcal{G}) \cap \mathcal{C}_{sep}(\mathcal{H})$. Hence $d_c(\mathcal{G},\mathcal{H}) \neq 0$. □

Consider a simulated experiment to evaluate a causal discovery method $\mathcal{M}$ which outputs the graph $\mathcal{H}$. If the parameters of the data-generation process are chosen in such a way that the Markov property and causal Faithfulness w.r.t to the ground truth graph $\mathcal{G}$ is guaranteed, according to the previous lemma, $d_c(\mathcal{G},\mathcal{H}) \neq 0$ means that $P_{\mathcal{V}}$ does not have the Markov property on the output graph. In other words, $d_c(\mathcal{G},\mathcal{H})$ measures how far the pair $(\mathcal{H}, P_{\mathcal{V}})$ is from being Markovian. As the Markov property is the most fundamental link between the distribution and the graph, a large Markov metric should be a clear warning sign that the algorithm is not performing well or that the data-generation process is far from being faithful on $\mathcal{G}$. Markovianity on $\mathcal{G}$ is usually a given in a data simulation.

We can define an analogous notion for separations instead of connections.

Definition D.6. Consider two causal graphs $\mathcal{G}, \mathcal{H}$ over the same set of nodes $\mathbf{X}$, equipped with an appropriate notion of separation. We first define

$$d_s^k(\mathcal{G}, \mathcal{H}) := \frac{1}{|\mathcal{C}_{sep}^k(\mathcal{G})|} \sum_{(X,Y,\mathcal{S}) \in \mathcal{C}_{sep}^k(\mathcal{G})} \iota_{\mathcal{H}}(X, Y, \mathcal{S}).$$

for $k = 0, \dots, N-2$. Then we call

$$d_s^{\leq K} = \frac{1}{K+1} \sum_{k=0}^{K} d_s^k(\mathcal{G}, \mathcal{H}).$$

the **s-metric** or **Faithfulness metric** of $\mathcal{H}$ to $\mathcal{G}$ **of order K**. If $K = N-2$, we just speak of the **c-** or **Faithfulness metric** and write d_c instead of d_c^{N-2}.

The following result is the analogue of Lemma D.5 for the Faithfulness metric.

Lemma D.7. *Consider two causal graphs $\mathcal{G}, \mathcal{H}$ over the same set of nodes $\mathbf{X}$, equipped with an appropriate notion of separation. Suppose that $P_{\mathcal{V}}$ is a distribution on $\mathbf{X}$ that is Markovian and faithful on the graph $\mathcal{G}$. Then $d_s(\mathcal{G}, \mathcal{H}) \neq 0$ if and only if $P_{\mathcal{V}}$ is not faithful on $\mathcal{H}$.*

Proof. If $d_s(\mathcal{G}, \mathcal{H}) \neq 0$, then there must be a triple $(X, Y, \mathcal{S}) \in \mathcal{C}_{sep}(\mathcal{G}) \cap \mathcal{C}_{con}(\mathcal{H})$. Since $(X, Y, \mathcal{S}) \in \mathcal{C}_{sep}(\mathcal{G})$ and $P_{\mathcal{V}}$ is Markovian on $\mathcal{G}$, we must have that $X \perp\!\!\!\perp_{P_{\mathcal{V}}} Y | \mathcal{S}$. If $P_{\mathcal{V}}$ would be faithful on $\mathcal{H}$, $(X, Y, \mathcal{S}) \in \mathcal{C}_{con}(\mathcal{H})$ would imply $X \not\!\perp\!\!\!\perp_{P_{\mathcal{V}}} Y | \mathcal{S}$, which is a contradiction. Conversely, if $P_{\mathcal{V}}$ is not faithful on $\mathcal{H}$, there must a triple $(X, Y, \mathcal{S}) \in \mathcal{C}_{con}(\mathcal{H})$ with $X \perp\!\!\!\perp_{P_{\mathcal{V}}} Y | \mathcal{S}$. Since $P_{\mathcal{V}}$ is faithful on $\mathcal{G}$, it follows that $(X, Y, \mathcal{S}) \in \mathcal{C}_{sep}(\mathcal{G})$ and thus $(X, Y, \mathcal{S}) \in \mathcal{C}_{sep}(\mathcal{G}) \cap \mathcal{C}_{con}(\mathcal{H})$. But this means that $d_s(\mathcal{G}, \mathcal{H}) \neq 0$. □

The name 'Faithfulness metric' is thus motivated by the fact that if $\mathcal{G}$ is a ground truth graph in a simulated experiment, the Faithfulness metric measures the amount of Faithfulness violations in a method's output graph.

Remark D.8.

(i) Like the s/c-metric before, both the Markov and the Faithfulness metric only depend on the Markov equivalence classes of the two graphs in play.

(ii) At first glance, it might seem like $d_s(\mathcal{G}, \mathcal{H}) = d_c(\mathcal{H}, \mathcal{G})$ but this is not the case. The crucial difference lies in the normalization constants of the k-th order contributions which need not coincide in $d_s(\mathcal{G}, \mathcal{H})$ and $d_c(\mathcal{H}, \mathcal{G})$.

Analogy to False Positive and False Negative Rate The term d_c^k in the Markov metric measures the number of connections of the ground truth graph $\mathcal{G}$ with $|\mathcal{S}| = k$ that the graph $\mathcal{H}$ misses (*"false negatives"*) relative to the total number of connections of $\mathcal{G}$ (*"positives"*) with $|\mathcal{S}| = k$. Thus, since we scale the Markov metric by $\frac{1}{N-1}$ we compute the average *false negative rate* across all orders for separation/connection statements. Similarly, the Faithfulness metric measures the number of separations of the ground truth graph $\mathcal{G}$ with $|\mathcal{S}| = k$ that the graph $\mathcal{H}$ mistakes for connections (*"false positives"*), relative to the total number of separations of $\mathcal{G}$ (*"negatives"*) with $|\mathcal{S}| = k$. The Faithfulness metric, can thus be interpreted as the average *false positive rate* across all orders for separation/connection statements. Like usual FPRs and FNRs, we can also combine Markov and Faithful distance into a ROC-curve to obtain an additional quality metric, see (Hyttinen et al., 2014).

E A SUMMARY OF ADJUSTMENT IDENTIFICATION DISTANCES

Adjustment Identification Distances (AIDs) were introduced in (Henckel et al., 2024). As we often refer to these distances in the main document, we repeat their definition here for the convenience of the reader. This section does not contain any original results, and the reader is encouraged to consult (Henckel et al., 2024) for more details.

AIDs are based on the notion of identifying formulas in causal graphs. An identifying formula for the effect of variable X on variable Y in a DAG $\mathcal{G}$ is an equation that expresses the interventional distribution $P(Y|\text{do}(X))$ purely in terms of the observational distribution of the graphical nodes for any distribution P compatible with $\mathcal{G}$. An effect is called identifiable if there is at least one identifying formula for it. A (sound and complete) identification strategy is then defined as an algorithm that inputs a tuple $(\mathcal{G}, X, Y)$ and that returns a correct identifying formula if there is one, and NONE otherwise.

More specifically Henckel et al. (2024) focus on identification through adjustment. If $X, Y \in \mathcal{V}$ are nodes and $\mathcal{S} \subset \mathcal{V} \backslash \{X, Y\}$ is a subset of nodes, then $\mathcal{S}$ is a valid adjustment set for the effect of X on Y if $P(Y|\mathrm{do}(X)) = \int P(y|x, \mathbf{s})P(\mathbf{s})d\mathbf{s}$ for any distribution P compatible with $\mathcal{G}$, now assuming that P has a density with respect to the Lebesgue measure. Since $\mathrm{pa}_{\mathcal{G}}(X)$ is a valid adjustment set for (X, Y) whenever Y is not itself a parent of X, the first option to define an identification strategy is to use the identification formula obtained through parent adjustment; this is what Henckel et al. (2024) call the parent adjustment strategy. Other identification strategies used in (Henckel et al., 2024) are ancestor adjustment, which employs the identifying formula obtained by conditioning on all ancestors of X, and optimal adjustment (Henckel et al., 2022) which makes use of the adjustment set of minimal variance. To build a distance measure for DAGs from identification strategies, the final missing ingredient is a verifier, i.e. an algorithm that inputs a tuple $(\mathcal{G}, X, Y)$ and an identifying formula and that outputs whether this formula is, in fact, a correct identifying formula for the effect of X on Y on $\mathcal{G}$.

With these tools at hand, the adjustment identification distance $\mathrm{AID}(\mathcal{G}, \mathcal{H})$ for a given adjustment strategy is defined by (1) computing the identifying formula prescribed by the chosen strategy in $\mathcal{H}$ for each pair of nodes $(X, Y),\ X \neq Y$; (2) verifying for each pair of nodes whether this formula is correct in $\mathcal{G}$; and (3) incurring a penalty of 1 whenever the formula is false in $\mathcal{G}$.

To generalize AIDs to CPDAGs, it is first necessary to observe that in CPDAGs causal effects may longer be identifiable. Therefore, in CPDAG-AIDs, a penalty is incurred not only when the identifying formula computed in $\mathcal{H}$ is incorrect in $\mathcal{G}$, but also if an effect is identifiable in one graph but not in the other.

F ALGORITHMS

F.1 Algorithms to compute separation distances

In this section, we will present pseudocode to compute the parent-, the MB- and the ZL-separation distance, including a more detailed discussion of their computational complexity. Each algorithm consists of two steps, a separator computation step in one graph and a separator verification in the other. The computational bottleneck is the verification step which will therefore be our main focus.

Separator Computation Separators can be computed as follows. We record the algorithmic complexity of each step, and our usage of 'sparse' refers to a bounded node degree d independent of the number of nodes N. Recall that a graph is of bounded node degree d if each node is adjacent to at most d other nodes. The number of edges of a graph will be denoted by M.

Computing the parent separators $\mathrm{pa}_{\mathcal{H}}(X \cup Y)$ for all node pairs (X, Y) is of computational complexity $\mathcal{O}(N^3)$ in general: one needs to compute $\mathrm{pa}_{\mathcal{H}}(X)$ for all X in a first loop ($\mathcal{O}(N^2)$) and then take the union $\mathrm{pa}_{\mathcal{H}}(X \cup Y)$ ($\mathcal{O}(N)$) for all node pairs (X, Y) ($\mathcal{O}(N^2)$ times), so $\mathcal{O}(N^3)$ in total. In the sparse case, taking the union is only $\mathcal{O}(1)$ so that the complexity reduces to $\mathcal{O}(N^2)$. Similar arguments also yield the same general complexity $\mathcal{O}(N^3)$ for computing ancestor and potential parent separators. Computing the ZL-separator of (X, Y) in a MAG $\mathcal{H}$ is of complexity $\mathcal{O}(N + M)$ (van der Zander and Liśkiewicz, 2020) and this needs to be executed $\mathcal{O}(N^2)$ times, so that the complexity in terms of N is $\mathcal{O}(N^4)$ in general and $\mathcal{O}(N^3)$ in the sparse case. Computing the Markov blanket $MB_{\mathcal{H}}(X)$ in a DAG $\mathcal{H}$ for all X can be done by computing the moralized graph ($\mathcal{O}(N^3)$[5] and $\mathcal{O}(N^2)$ in the sparse case) and by subsequently computing the adjacencies of X in this graph for all X $\mathcal{O}(N^2)$, yielding an upper bound of $\mathcal{O}(N^3)$ in general and $\mathcal{O}(N^2)$ in the sparse case. Computing the MB-enhanced (possible) parent separator is therefore of complexity $\mathcal{O}(N^3)$ in general and $\mathcal{O}(N^2)$ in the sparse case as well.

To compute the ZL-separator, we use the algorithm introduced in (van der Zander and Liśkiewicz, 2020) which is of computational complexity $\mathcal{O}(N + M)$. The ZL-separator $\mathrm{ZL}_{\mathcal{G}}(X, Y)$ depends on both arguments X, Y and hence we need to loop through both, leading to a complexity of $\mathcal{O}(N^4)$ or $\mathcal{O}(N^3)$ if the input graph is sparse.

Separation Verifier Since verifying whether the separators computed in a graph $\mathcal{H}$ are in fact separators in another graph $\mathcal{G}$ is the computationally most costly step, we will provide more details here. In general,

[5] this can actually be reduced to $\approx \mathcal{O}(N^{2.37})$, see (Heisterkamp, 2009; Wienöbst, 2023)

the complexity of this part of the distance computation is $\mathcal{O}(N^2 \cdot (N+M))$ or $\mathcal{O}(N^4)$ in terms of N only, as Algorithm 1 illustrates. In the sparse case where $M \in \mathcal{O}(N)$, this becomes $\mathcal{O}(N^2)$. We write $\mathrm{nadj}(\mathcal{H})$ for the set of non-adjacent node pairs in a graph $\mathcal{H}$ (computable in $\mathcal{O}(N^2)$).

Algorithm 1 Pseudocode to verify separators and compute the corresponding non-normalized SD.

```
Require: causal graph $\mathcal{G}$, list $\mathfrak{S}_{\mathcal{H}}(X,Y), (X,Y) \in \mathrm{nadj}(\mathcal{H})$ of proposed separator found in a previous step based
  on another graph $\mathcal{H}$.
  Initialize distance $d \leftarrow 0$;
  for $(X,Y) \in \mathrm{nadj}(\mathcal{H})$                                   {loop of length $\mathcal{O}(N^2)$}
   do
    check $X \bowtie_{\mathcal{G}} Y | \mathfrak{S}_{\mathcal{H}}(X,Y)$;                       {complexity $\mathcal{O}(N+M)$}
    if check returns false then
      $d$ += 1;
    end if
  end for
.
```

On a DAG or CPDAG $\mathcal{G}$, MB-enhancement yields an improvement thanks to the *Bayes-Ball algorithm* (Geiger et al., 1990; Shachter, 1998), see also (Henckel et al., 2024, Appendix D). For a given node X and a set $\mathcal{S}$, the Bayes-Ball algorithm is able to compute the set $\mathrm{nsep}_{\mathcal{G}}(X, \mathcal{S}) = \{Y \mid Y \not\bowtie_{\mathcal{G}} X | \mathcal{S}\}$ in time $\mathcal{O}(N+M)$. We apply it in Algorithm 2 to achieve the desired reduction in complexity.

Algorithm 2 Pseudocode to verify MB-enhanced separators and to compute the corresponding non-normalized SD. The worst case computational complexity is $\mathcal{O}(N^2 \cdot (N+M))$. If $\mathcal{G}$ and $\mathcal{H}$ are sparse, the computational complexity reduces to $\mathcal{O}(N^2)$.

```
Require: DAG $\mathcal{G}$, list of Markov blankets $\mathrm{MB}_{\mathcal{H}}(X), X \in \mathcal{V}$, list of separators for exceptional cases $\mathfrak{S}_{\mathcal{H}}(X,Y), Y \in$
  $\mathrm{MB}_{\mathcal{H}}(X) \setminus (\mathrm{pa}_{\mathcal{H}}(X) \cup \mathrm{ch}_{\mathcal{H}}(X))$.
  Initialize distance $d \leftarrow 0$;
  for $X \in \mathcal{V}$                                                     {loop of length $N$}
   do
    get $\mathrm{nsep}_{\mathcal{G}}(X, \mathcal{S})$ with Bayes-Ball;        {complexity $\mathcal{O}(N+M)$ and $\mathcal{O}(N)$ if $\mathcal{G}$ is sparse}
    $d \leftarrow |\mathrm{nsep}_{\mathcal{G}}(X, \mathcal{S}) \cap \mathcal{V} \setminus \mathrm{MB}_{\mathcal{H}}(X)|$;                           {complexity $\mathcal{O}(N)$}
    for $Y \in \mathrm{MB}_{\mathcal{H}}(X) \setminus (\mathrm{pa}_{\mathcal{H}}(X) \cup \mathrm{ch}_{\mathcal{H}}(X))$;        {loop of length $\mathcal{O}(N), \mathcal{O}(1)$ if $\mathcal{H}$ is sparse}
     do
      check $X \bowtie_{\mathcal{G}} Y | \mathfrak{S}_{\mathcal{H}}(X,Y)$;          {complexity $\mathcal{O}(N+M)$, $\mathcal{O}(N)$ if $\mathcal{G}$ is sparse}
      if check returns false then
        $d$ += 1;
      end if
    end for
  end for
.
```

F.2 Algorithms to compute s/c-metrics

For the sake of completion, even though the computation is straightforward, we also provide pseudocode to compute the s/c-metric. Once again, we rely on an *oracle function* $\iota_{\mathcal{G}}(X, Y, \mathcal{S})$ that takes in a causal graph and a triple $(X, Y, \mathcal{S}) \in \mathcal{C}$ and outputs 0 if X and Y are separated and 1 if they are connected by $\mathcal{S}$ in $\mathcal{G}$. Practical implementations of such an oracle for DAGs (d-separation) are available in the R package *Dagitty* (Textor et al., 2016) and the Python packages *networkx* and *Tigramite* (`https://github.com/jakobrunge/tigramite`). The oracle implemented in Tigramite is also applicable to MAGs (m-separation), tsMAGs and tsDAGs.

G The CHALLENGE OF INVALID OUTPUT GRAPHS

In this section, we would like to draw attention to an additional challenge that occurs when applying evaluation metrics to causal discovery algorithms based on the PC-algorithm. Separation-based (or adjustment-based)

Algorithm 3 Pseudocode to compute the s/c-metric.

```
Require: $\mathcal{G}$, $\mathcal{H}$ causal graphs (e.g. DAGs) or Markov equivalence classes (e.g. CPDAGs) with $N$ nodes, Oracle
  for fitting notion of separation. Sets $\mathcal{L}_k$ of order $k$-triples to be tested for $k = 0, \ldots, N-2$.
  if $\mathcal{G}$ (or $\mathcal{H}$) is Markov equivalence class then
    $\mathcal{G}$ (or $\mathcal{H}$) $\leftarrow$ member of $\mathcal{G}$ (or $\mathcal{H}$);
  end if
  dist = 0.0;
  for $k = 0, \ldots, N-2$ do
    $\text{dist}_k = 0.0$;
    count = 0;
    for triples $(X, Y, \mathcal{S}) \in \mathcal{L}_k$ do
      count += 1;
      $\text{dist}_k$ += $|\iota_{\mathcal{G}}(X, Y, \mathcal{S}) - \iota_{\mathcal{H}}(X, Y, \mathcal{S})|$;
    end for
    dist += $\frac{\text{dist}_k}{\text{count}}$;
  end for
  return $\frac{\text{dist}}{N-1}$.
.
```

metrics that compare two graphs $\mathcal{G}$ and $\mathcal{H}$ require that both of these graphs are able to decide whether $(X, Y, \mathcal{S})$ is a separation or a connection statement (or whether $\mathcal{S}$ is adjustment set for (X, Y)) in the respective graph. For instance, if the applied notion of separation is d-separation, then $\mathcal{G}$ and $\mathcal{H}$ should be DAGs or CPDAGs. While score-based causal discovery methods like GES (Chickering, 2003), NOTEARS (Zheng et al., 2018), GLOBE (Mian et al., 2021), BCCD (Claassen and Heskes, 2012), parametric methods like LinGaM (Shimizu et al., 2006) or logic-based methods like (Hyttinen et al., 2014) guarantee that their output is either a proper causal graph or a MEC, this is no longer true for all PC-based algorithms. Due to assumption violations or contradictory independence test results during their execution, the order independent version of PC (Colombo and Maathuis, 2014) might run into logical conflicts between different orientation rules. or it might label unshielded triples $X - Y - Z$ in the graph as ambiguous. The conservative PC algorithm of (Ramsey et al., 2006) labels an unshielded triples $X - Y - Z$ ambiguous if the middle node Y is in some but not all separating sets that the method found for X and Z. The majority decision rule (Colombo and Maathuis, 2014) labels a triple ambiguous if Y belongs to exactly half of all separating sets found for X and Z. If conflicts or ambiguities occur, the output graph is *invalid* in the sense that it does no longer imply separation/connection statements for arbitrary triples $(X, Y, \mathcal{S})$. Hence computing separation-based or adjustment-based metrics is no longer straightforward.

Dealing with Conflicting Orientations Since conflicts between different orientation rules are serious errors that flag assumption violations of some form, the most conservative way to treat them in a performance evaluation of a PC-like method is to record the proportion of their occurrence and then disregard graphs with conflicts for further steps. A well-performing method should lead to (a) a low proportion of graphs with conflicts and (b) good results w.r.t. the chosen evaluation metric on its valid output graphs.

Dealing with Ambiguities In contrast to conflicting orientations, the appearance of ambiguities is a sign that a method is conservative about the inferences it is willing to make. Discarding graphs with ambiguities entirely, therefore seems like an overly harsh punishment for a method not being willing to make strong claims. We will now describe a procedure for computing best-case and worst-case distances to a ground truth for graphs with ambiguities by considering all possible ways in which ambiguities may be interpreted. The proposed strategy is similar to how the structural intervention distance is computed for a CPDAG by iterating through all DAGs that are represented by a given CPDAG, see (Peters and Bühlmann, 2015). The procedure will start with the following input data:

- A (ground truth) DAG or CPDAG $\mathcal{G}$;
- An undirected graph $\mathcal{H}$ that will serve as the skeleton of a DAG or tsDAG;
- A partition of the set of unshielded triples $\mathcal{U} = \mathcal{U}_c \cup \mathcal{U}_{nc} \cup \mathcal{U}_a$ of $\mathcal{H}$ into colliders ($\mathcal{U}_c$), non-colliders ($\mathcal{U}_{nc}$) and

Algorithm 4 Pseudocode to compute distance measures for graphical outputs with ambiguities.

```
Require: $\mathcal{G}$, $\mathcal{H}$, $\mathcal{U}_c$, $\mathcal{U}_a$ as defined above. A distance metric $d(\cdot, \cdot)$.

Initialize empty list $\mathcal{L}$.
for all $\mathcal{B} \subset \mathcal{U}_a$ do
    $\mathcal{H}_\mathcal{B} \leftarrow \mathcal{H}$;
    COL $\leftarrow \mathcal{U}_c \cup \mathcal{B}$;
    Collider phase: try: orient all $(X, Y, Z) \in$ COL as colliders on $\mathcal{H}_\mathcal{B}$;
    if No conflicting orientations and no cycle in $\mathcal{H}_\mathcal{B}$ then
        Orientation phase: try: Apply Meek's orientation rules to $\mathcal{H}_\mathcal{B}$;
        if No conflicting orientations and no cycle in $\mathcal{H}_\mathcal{B}$ then
            append $\mathcal{H}_\mathcal{B}$ to $\mathcal{L}$;
        end if
    end if
end for
if $\mathcal{L} = \emptyset$ then
    max-dist, mean-dist, min-dist $\leftarrow 1$;
else
    max-dist $\leftarrow \max_{\mathcal{B} \in \mathcal{L}} d(\mathcal{G}, \mathcal{H}_\mathcal{B})$;
    mean-dist $\leftarrow \text{mean}_{\mathcal{B} \in \mathcal{L}} d(\mathcal{G}, \mathcal{H}_\mathcal{B})$;
    min-dist $\leftarrow \min_{\mathcal{B} \in \mathcal{L}} d(\mathcal{G}, \mathcal{H}_\mathcal{B})$;
end if
return max-dist, mean-dist, min-dist.
```

ambiguous triples ($\mathcal{U}_a$).

This information is part of the output of the conservative PC algorithm (Ramsey et al., 2006), the stable PC algorithm with the majority rule (Colombo and Maathuis, 2014) or the Tigramite implementation of PCMCI+ with the conservative or majority contemporary collider rule (Runge, 2020).

From this input, we now generate a list of CPDAGs ($\mathcal{H}_\mathcal{B}$) where $\mathcal{B} \subset \mathcal{U}_a$. For each CPDAG $\mathcal{H}_\mathcal{B}$ in this list, we then compute the desired distance $d(\mathcal{G}, \mathcal{H}_\mathcal{B})$ and, finally, we arrive at the best-case, average and worst-case estimate

$$\min_{\mathcal{B} \subset \mathcal{U}_a} d_x(\mathcal{G}, \mathcal{H}_\mathcal{B}), \qquad \text{mean}(d_x(\mathcal{G}, \mathcal{H}_\mathcal{B})) \qquad \max_{\mathcal{B} \subset \mathcal{U}_a} d_x(\mathcal{G}, \mathcal{H}_\mathcal{B}).$$

The details are outlined as pseudocode in Algorithm 4.

H ADDITIONAL EXPERIMENTS

In this section, we provide additional empirical experiments on separation-based distance measures.

H.1 Correlations of symmetrized distance measures

In Figure 7, we plot the correlation coefficients of the parent-SD with other distance measures for Erdös-Renyi DAGs $G(N, p)$ with $N = 25$ nodes for different values of p (500 runs per parameter). We decided to use symmetrized versions of the included distance measures as this is a bit more similar to the SHD for a fair comparison. Parent-SD and its MB-enhanced variant are strongly correlated for all values of p, suggesting that it might generically be advantageous to employ the latter due to its faster runtime. All other metrics are strongly correlated for very sparse graphs, but this correlation drops off rapidly and then increases again. Interestingly, for $0.25 \leq p \leq 0.85$, parent-SD and SHD even become negatively correlated. These correlations do not differ substantially between DAGs and CPDAGs. We include a similar plot for the correlation of the ZL-SD and the SHD on mixed graphs further below in Appendix H.2. In addition, we compute the correlation of separation distances with the full separation metric that computes all separation statements on $N = 10$ nodes in Appendix H.3. While this correlation is close to 1 for sparse DAGs, it decreases when the DAGs become more dense.

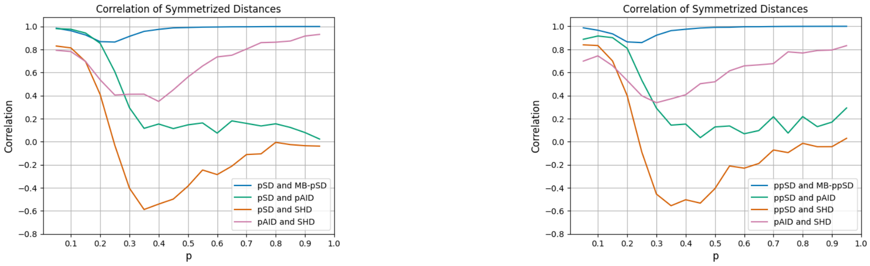

Figure 7: Correlation coefficients of symmetrized distance measures applied to Erdös-Renyi graphs $\mathcal{G}, \mathcal{H} \sim G(N, p)$ (left) and their associated CPDAGs (right) for increasing values of p. We ran 500 experiments per parameter.

H.2 Correlation of ZL-SD and SHD on mixed graphs

First, we repeat the experiments on the correlation of distance measures on mixed graphs. In this context, the available comparison metrics are the ZL-SD and the SHD, so these are the only ones we consider. We generate 500 random mixed graphs by the following scheme. We first generate a causal order with a random permutation. Then for every node pair (X, Y) for which $X < Y$ in the causal order, we generate an edge between them with probability p. If an edge is drawn, we orient the edge as $X \to Y$ with probability $q = 0.2, 0.7, 0.9$ and as $X \leftrightarrow Y$ with probability $1 - q$. The resulting graphs are acyclic but not necessarily MAGs as they might not be ancestral or might have almost cycles. The lack of ancestrality is unproblematic for the computation of the ZL-SD as the separator search will simply return *None* for two non-adjacent nodes that cannot be separated. We did not apply any checks for almost-cycles as this would have introduced a significant computational overhead and would not have changed the general form of these plots.

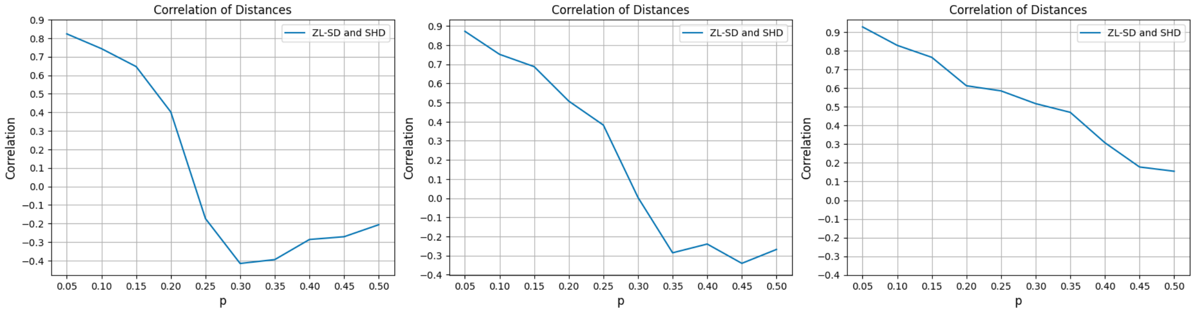

Figure 8: Correlation of ZL-SD and SHD across 500 random mixed graphs with probability $q = 0.2$ (left), $q = 0.7$ (middle) and $q = 0.9$ (right) that an existing edge is directed.

H.3 Correlation of SDs and the s-metric

In this section, we present a plot on the correlation of separation distances with the full s-metric across 100 pairs of Erdös-Renyi DAGs $\mathcal{G}, \mathcal{H} \sim G(N, p)$ on 10 nodes for different values of p. More precisely, we compare the values of $d^{\mathfrak{S}}(\mathcal{G}, \mathcal{H})$ to the values of $d^s(\mathcal{H}, \mathcal{G})$ as these metrics are the most similar conceptually: in both cases separations found in $\mathcal{H}$ are verified in $\mathcal{G}$. We see that all SDs are strongly correlated with $d^s(\mathcal{H}, \mathcal{G})$ for sparse graphs but the correlation drops for more dense ones. This illustrates that the choice of only specific separators according to a sep-strategies approximates a full comparison well on sparse graphs but the tradeoff of this selection becomes more apparent, the denser the graphs become.

H.4 Monte-Carlo approximation of the full s/c-metric

Due to the exponential growth of the number of separation statements in the number of nodes N, computing the full s/c-distance as a weighted average across all separation statements is very time-consuming and only feasible for a small number of nodes. A possible way to alleviate this issue is to specify a fixed number L of separation statements to be tested per order and to then draw these L statements randomly. In this case the normalization

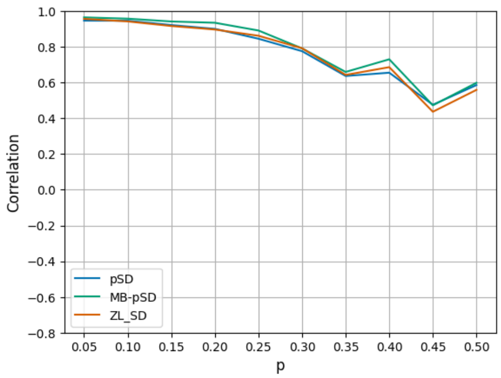

Figure 9: Correlation of SDs with the full s-metric $d^s(\mathcal{H}, \mathcal{G})$.

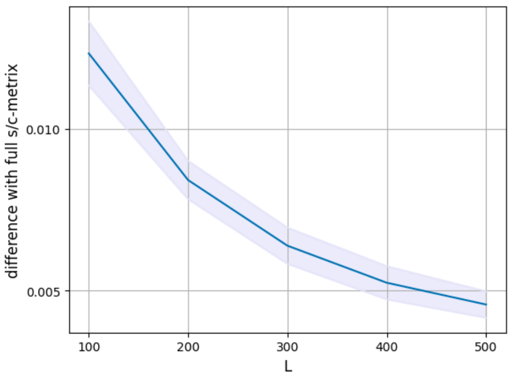

Figure 10: Difference to the full s/c-metric when only $L = 100, 200, 300, 400, 500$ separation statements are drawn randomly per order on Erdös-Renyi DAGs with $N = 10$ nodes and edge probability $p = 0.4$.

constant in the order terms of the s/c-metric has to be set to L^{-1}. Figure 10 below shows that at least for small graphs, for which the full s/c-distance can still be checked, this approximation yields very good results. We note however that there is a trade-off between the quality of the approximation and the computational speed-up as high values of L will give better approximation while low values will reduce the computational effort. Nevertheless, even with low values of L the number of separation statements to be checked, and hence the computational cost, is still significantly higher than for the deterministic sep-strategies.

Reproducibility The Python code used for the experiments in this work is available in the repository `https://github.com/JonasChoice/SDs_experiments`.