Foundations of Computational Mathematics
https://doi.org/10.1007/s10208-022-09581-9

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics


# Causal Structure Learning: A Combinatorial Perspective

**Chandler Squires**[1] · **Caroline Uhler**[2]

Received: 28 May 2022 / Accepted: 8 June 2022


**Abstract**
In this review, we discuss approaches for learning causal structure from data, also called *causal discovery*. In particular, we focus on approaches for learning directed acyclic graphs and various generalizations which allow for some variables to be unobserved in the available data. We devote special attention to two fundamental combinatorial aspects of causal structure learning. First, we discuss the structure of the search space over causal graphs. Second, we discuss the structure of *equivalence classes* over causal graphs, i.e., sets of graphs which represent what can be learned from observational data alone, and how these equivalence classes can be refined by adding *interventional* data.

**Keywords** Causal inference · Causal structure discovery · Markov equivalence

**Mathematics Subject Classification** 62H22 · 68R05 · 05C75

## 1 Introduction

Many important scientific, sociological, and technological questions are fundamentally *causal*: "which genes regulate one another?," "how would raising minimum wage affect unemployment rate?," "which treatment most effectively prolongs the lifespan

Communicated by Teresa Krick and Hans Munthe-Kaas.

Invited paper based on the FoCM 2021 Online Seminar lecture *Causal Inference and Overparameterized Autoencoders in the Light of Drug Repurposing for COVID-19* presented by Caroline Uhler in January 2021.

✉ Caroline Uhler
cuhler@mit.edu

Chandler Squires
csquires@mit.edu

[1] Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2] Broad Institute and Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Published online: 01 August 2022

of breast cancer patients?." In each case, answering the question requires predicting how a system, e.g., a cell, economy, or human body, will react to external manipulation. *Structural causal models* can be used to formalize such questions, to create algorithms that determine whether such questions can be answered from available data sources, and to develop general-purpose methods for learning the answers to such questions. In the framework of structural causal models, a *directed graph* is used to reflect how the variables in these models depend causally on one another. Each node $i$ of the directed graph is associated with a variable $X_i$, and an edge $i \to j$ indicates that the variable $X_i$ is a direct cause of the variable $X_j$. In some special, well-studied settings, background knowledge and human reasoning can be used to propose plausible directed graph models. However, in large systems such as gene regulatory networks, the directed graph is *not* known a priori, making it necessary to develop methods for *learning* the graph from data. Once this graph is learned, it can be used to predict the effects of interventions or distributional shifts, in contrast to traditional machine learning methods which can only make predictions on inputs that come from the same distribution as the training data.

The problem of learning such a causal graph from data, known as *causal structure learning* (or *causal discovery*), has been the focus of much recent work in computer science, statistics, and bioinformatics , covered in a number of recent reviews [40, 41, 64, 74, 113]. Compared to these reviews, we here emphasize the *combinatorial* aspects of causal structure learning, including characterizations of equivalence classes of graphs, computing the size and number of these equivalence classes, and how the characterization and properties are influenced by the presence of latent variables or interventional data. After discussing these topics, we will cover methods for causal structure learning which are based heavily on the combinatorial structure over the space of directed graphs. Focusing on this combinatorial structure has three significant advantages:

1. Causal structure learning can be dramatically simplified when fixing some combinatorial aspect of the problem, such as the ordering of the variables.
2. Understanding the combinatorial aspects of structure learning allows a number of different methods to be synthesized into a single framework and eases future methodological development.
3. Insights into the combinatorial aspects of structure learning are also useful for other tasks, such as experimental design.

The framework provided by the combinatorial viewpoint encompasses methods for learning causal models with unobserved variables, as well as methods for learning from a combination of observational and interventional data. The second point is especially important, since interventional data are often crucial for identifying the true causal model and subsequently using the causal model for predicting the effects of interventions or distributional shifts.

## 2 Structural Causal Models

A structural causal model defines causal relationships over a set of random variables $\{X_i\}_{i=1}^p$. These relationships are summarized by a directed acyclic graph (DAG) $\mathcal{G}$ over nodes $i = 1, \ldots, p$, where the node $i$ in $\mathcal{G}$ is associated with the variable $X_i$. Given a DAG $\mathcal{G}$, we let $\mathrm{pa}_{\mathcal{G}}(i)$ denote the *parents* of the node $i$, i.e., $\mathrm{pa}_{\mathcal{G}}(i) = \{j \mid j \to i \text{ in } \mathcal{G}\}$. Then, a (Markovian) *structural causal model* (SCM) [124] with *causal graph* $\mathcal{G}$ consists of a set of *endogenous variables* $\{X_i\}_{i=1}^p$, a set of *exogenous variables* $\{\epsilon_i\}_{i=1}^p$, a product distribution $\mathbb{P}_\epsilon$ over the exogenous variables, and a set of *structural assignments* $\{f_i\}_{i=1}^p$. In particular, the structural assignment $f_i$ asserts the relation $X_i = f_i(X_{\mathrm{pa}_{\mathcal{G}}(i)}, \epsilon_i)$. Via these structural assignments, the distribution $\mathbb{P}_\epsilon$ over the exogenous variables induces a distribution $\mathbb{P}_X$ over the endogenous variables, called the *entailed distribution* [124]. In particular, we have $\mathbb{P}_X(X_i \mid X_{\mathrm{pa}_{\mathcal{G}}(i)}) = \mathbb{E}_{\epsilon_i}[\mathbb{1}_{X_i = f_i(X_{\mathrm{pa}_{\mathcal{G}}(i)}, \epsilon_i)} \mid X_{\mathrm{pa}_{\mathcal{G}}(i)}]$ and

$$\mathbb{P}_X(X) = \prod_{i=1}^{p} \mathbb{P}_X(X_i \mid X_{\mathrm{pa}_{\mathcal{G}}(i)}). \tag{1}$$

***Example 1*** (*A simple structural causal model of genetic inheritance*) As a running example, we will consider a simplified model of genetic inheritance of weight among a family of mice. Let $X_2$ and $X_3$ represent the weights, in grams, of an unrelated male and female mouse, respectively. Let $X_4$ represent the weight of their offspring, and $X_5$ represent the weight of the offspring's offspring. Finally, let $X_1$ be a binary variable representing whether the two parent mice are genetically modified for increased weight. Assume that these variables are related via the following set of assignments:

$$\begin{aligned}
X_1 &= \epsilon_1 & \epsilon_1 &\sim \mathsf{Ber}(0.5)\\
X_2 &= \epsilon_2 + 2X_1 & \epsilon_2 &\sim \mathcal{N}(25, 1)\\
X_3 &= \epsilon_3 + 2X_1 & \epsilon_2 &\sim \mathcal{N}(20, 1)\\
X_4 &= \frac{1}{2}(X_2 + X_3) + \epsilon_4 & \epsilon_4 &\sim \mathcal{N}(0, 1)\\
X_5 &= X_5 + \epsilon_5 & \epsilon_4 &\sim \mathcal{N}(0, 2)
\end{aligned}$$

where the set of $\epsilon$ are mutually independent. The parent sets are $\mathrm{pa}_{\mathcal{G}}(1) = \varnothing$, $\mathrm{pa}_{\mathcal{G}}(2) = \{1\}$ $\mathrm{pa}_{\mathcal{G}}(3) = \{1\}$, $\mathrm{pa}_{\mathcal{G}}(4) = \{2, 3\}$, and $\mathrm{pa}_{\mathcal{G}}(5) = \{4\}$. The causal graph is given in Fig. 1, and

$$\begin{aligned}
\mathbb{P}_X(X) = \mathsf{Ber}(X_1; 0.5) &\times \mathcal{N}(X_2; 25 + 2X_1, 1) \times \mathcal{N}(X_3; 20 + 2X_1, 1)\\
&\times \mathcal{N}(X_4; \frac{1}{2}(X_2 + X_3), 1) \times \mathcal{N}(X_5; X_4, 1)
\end{aligned}$$

is the entailed distribution. □

The above definition of structural causal models can be generalized in at least two ways. First, one may remove the assumption that the distribution over the exogenous

variables is a product distribution, i.e., one may allow dependence between $\epsilon_i$ and $\epsilon_j$ for $i \neq j$. Such SCMs are called *semi-Markovian* and are taken as the basic definition of SCMs by some authors [122]. Instead of allowing for dependencies between exogenous variables, we use Markovian SCMs as the basic definition and assume that any unmodeled dependence between endogenous variables is due to some other *unobserved* endogenous variables, which we will cover in 2.3. Second, one may remove the assumption that $\mathcal{G}$ is acyclic. The assumption of acyclicity is natural when considering endogenous variables which are defined at certain time points, since the intuitive notion of causality dictates that a cause precedes any of its effects. However, if the endogenous variables are not well defined in time, e.g., if they represent the average state of a system in equilibrium, then feedback loops may occur. We will briefly discuss recent progress on causal structure learning for cyclic causal models in 5.

### 2.1 Markov Properties and Markov Equivalence in DAGs

Given a DAG $\mathcal{G}$, the set of distributions $\mathbb{P}_X$ that factorize according to 1 are said to follow the *Markov factorization property* with respect to $\mathcal{G}$. Depending on assumptions on the structural equations $\{f_i\}_{i=1}^p$ and the exogenous variables $\{\epsilon_i\}_{i=1}^p$, the Markov factorization property implies many other testable properties of the distribution $\mathbb{P}_X$. For instance, the *entire* set of conditional independence statements entailed by the Markov factorization property can be characterized simply in terms of a graphical criterion, known as *d-separation*, that can be read off from the DAG $\mathcal{G}$. The definition of d-separation relies on the notion of a *collider* along a path from $i$ to $j$. Given a path $\gamma = \langle \gamma_1 = i, \gamma_2, \ldots, \gamma_M = j \rangle$ from $i$ to $j$, the node $\gamma_m$ is a *collider* if $\gamma_{m-1} \rightarrow \gamma_m \leftarrow \gamma_{m+1}$, i.e., two arrowheads "collide" at $\gamma_m$. Then, a path $\gamma$ *d-connects* $i$ and $j$ given the set $C \subseteq [p] \backslash \{i, j\}$ if:

1. All non-colliders on the path do not belong to $C$.
2. All colliders on the path either belong to $C$ or have a descendant which belongs to $C$.

Finally, $i$ and $j$ are *d-connecting* given $C$ if there exists any d-connecting path given $C$; otherwise, they are *d-separated*. We denote that $i$ and $j$ are d-separated in $\mathcal{G}$ given $C$ via $i \perp\!\!\!\perp_{\mathcal{G}} j \mid C$. We denote the complete set of d-separation statements in a DAG $\mathcal{G}$ as $\mathcal{I}_{\perp\!\!\!\perp}(\mathcal{G})$; i.e.,

$$\mathcal{I}_{\perp\!\!\!\perp}(\mathcal{G}) = \{(i, j, C) \mid i, j \in [p], C \subseteq [p] \backslash \{i, j\}, i \perp\!\!\!\perp_{\mathcal{G}} j \mid C\}.$$

***Example 2*** (*d-connection and d-separation*) In $\mathcal{G}^*$ from Fig. 1a, there are two paths between 2 and 3, the path $\gamma_1 = 2 \leftarrow 1 \rightarrow 3$, and the path $\gamma_2 = 2 \rightarrow 4 \leftarrow 3$. For $C = \varnothing$, $\gamma_1$ is a d-connecting path between 2 and 3, since 1 is a non-collider and does not belong to $C$, while $\gamma_2$ is *not* a d-connecting path, since 4 is a collider but neither 4 nor 5 is in $C$. Thus, 2 and 3 are d-connected given $C = \varnothing$. For $C = \{1\}$, neither $\gamma_1$ nor $\gamma_2$ are d-connecting paths, so 2 and 3 are d-separated given $C = \{1\}$. Finally, for any $C$ containing 4 or 5, $\gamma_2$ is a d-connecting path between 2 and 3. Thus, 2 and 3 are d-connected given $C = \{4\}$, $C = \{5\}$, $C = \{1, 4\}$, etc. □

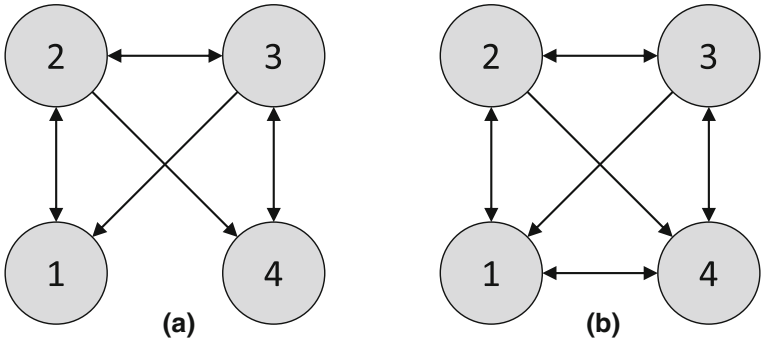


**Fig. 1** **a** Causal graph $\mathcal{G}^*$ for the structural causal model in 1. **b** A minimal I-MAP $\mathcal{G}_2$ for $\mathcal{G}^*$, described in 3

Given a distribution $\mathbb{P}_X$, we call $X_i$ and $X_j$ *conditionally independent* given $X_C$ if $\mathbb{P}_X(X_i, X_j \mid X_C) = \mathbb{P}_X(X_i \mid X_C)\mathbb{P}_X(X_j \mid X_C)$.

This is denoted by $i \perp\!\!\!\perp_{\mathbb{P}_X} j \mid C$.

We denote the set of all conditional independence statements in $\mathbb{P}_X$ as

$$\mathcal{I}_{\perp\!\!\!\perp}(\mathbb{P}_X) = \{(i, j, C) \mid i, j \in [p], C \subseteq [p] \backslash \{i, j\}, i \perp\!\!\!\perp_{\mathbb{P}_X} j \mid C\}.$$

If all d-separation statements in the DAG $\mathcal{G}$ hold as conditional independence statements in $\mathbb{P}_X$, i.e., $\mathcal{I}_{\perp\!\!\!\perp}(\mathcal{G}) \subseteq \mathcal{I}_{\perp\!\!\!\perp}(\mathbb{P}_X)$, then $\mathbb{P}_X$ is said to satisfy the *global Markov property* with respect to $\mathcal{G}$. Suppose that $\mathbb{P}_X$ has a density with respect to some product measure. Then, without any additional assumptions on the structural equations or the distributions of exogenous variables, the Markov factorization property and the global Markov property are equivalent [109].

Conversely, a given distribution $\mathbb{P}_X$ may satisfy the global Markov property with respect to many different DAGs. These DAGs are called *independence maps (I-MAPs)* of the distribution $\mathbb{P}_X$. As an extreme example, the complete graph implies *no* conditional independencies in $\mathbb{P}_X$, so it is an I-MAP of all distributions. However, the complete graph does not capture any of the independence structure in $\mathbb{P}_X$. For a variety of purposes, including computational and statistical efficiency in inference and estimation, it is preferable to find a DAG $\mathcal{G}$ that captures as many of the independences of $\mathbb{P}_X$ as possible. This intuition is captured in the definition of a *minimal* I-MAP for $\mathbb{P}_X$, which is an I-MAP $\mathcal{G}$ of $\mathbb{P}_X$, such that the deletion of *any* edge will result in a new DAG $\mathcal{G}'$ which is no longer an I-MAP for $\mathbb{P}_X$. The following example shows that a distribution $\mathbb{P}_X$ can have several minimal I-MAPs.

***Example 3*** (A distribution $\mathbb{P}_X$ can have multiple minimal I-MAPs) Let $\mathbb{P}_X$ be the distribution in 1. Then the DAG $\mathcal{G}^*$ in Fig. 1a is a minimal I-MAP for $\mathbb{P}_X$. To see this, we consider the deletion of each edge. Deleting $1 \to 2$ or $1 \to 3$ implies that $X_1 \perp\!\!\!\perp X_2$, or $X_1 \perp\!\!\!\perp X_3$, respectively, both of which are false. Similarly, deleting $2 \to 4$ or $3 \to 4$ implies that $X_2 \perp\!\!\!\perp X_4$, or $X_3 \perp\!\!\!\perp X_4$, respectively, but both are false. Finally, deleting $4 \to 5$ implies that $X_4 \perp\!\!\!\perp X_5$, which is again false.

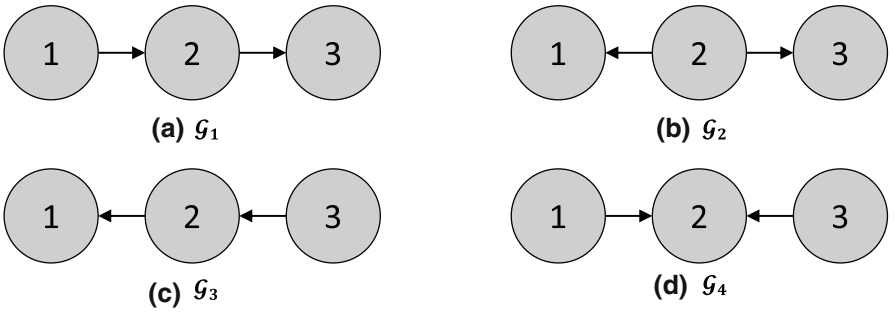


**Fig. 2** **a–c** Three Markov equivalent graphs from 4. **d** A fourth graph that is not Markov equivalent to the other three

$\mathbb{P}_X$ has other minimal I-MAPs, including the DAG $\mathcal{G}_2$ in Fig. 1b. Deleting $2 \to 1$ and $3 \to 1$ implies $X_2 \perp\!\!\!\perp X_1 \mid X_4, X_3$ and $X_3 \perp\!\!\!\perp X_1 \mid X_2$, respectively, both of which are false. Deleting $2 \to 3$ implies that $X_2 \perp\!\!\!\perp X_3 \mid X_4$, deleting $4 \to 2$ implies $X_4 \perp\!\!\!\perp X_2$, deleting $4 \to 3$ implies $X_4 \perp\!\!\!\perp X_3 \mid X_2$, and deleting $4 \to 5$ implies $X_4 \perp\!\!\!\perp X_5$, showing that $\mathcal{G}_2$ is indeed minimal. □

Suppose $\mathbb{P}_X$ is entailed by an SCM with causal graph $\mathcal{G}^*$. Since $\mathbb{P}_X$ may have multiple minimal I-MAPs, it is natural to ask, under some set of assumptions, whether $\mathcal{G}^*$ can be distinguished from the other minimal I-MAPs, and if not, whether a small *subset* of the minimal I-MAPs can be distinguished as candidates for $\mathcal{G}^*$. As we will discuss in 4, without assumptions on the functional forms of the structural assignments $f_i$, one cannot in general distinguish $\mathcal{G}^*$ from all other graphs using only $\mathbb{P}_X$. In particular, two DAGs $\mathcal{G}$ and $\mathcal{G}'$ with the same set of d-separation statements (i.e., $\mathcal{I}_{\perp\!\!\!\perp}(\mathcal{G}) = \mathcal{I}_{\perp\!\!\!\perp}(\mathcal{G}')$) are called *Markov equivalent*, and we denote this by $\mathcal{G} \approx_{\mathcal{M}} \mathcal{G}'$. The set of all DAGs that are Markov equivalent to $\mathcal{G}^*$ is called the *Markov equivalence class* (*MEC*) of $\mathcal{G}^*$, denoted $\mathcal{M}(\mathcal{G}^*)$, and $\mathbb{P}_X$ can in general only identify $\mathcal{G}^*$ up to $\mathcal{M}(\mathcal{G}^*)$.

***Example 4*** (Markov equivalence) The three DAGs in 2a–c are all Markov equivalent to one another, since for all three graphs, the only d-separation statement is that 1 and 3 are d-separated given 2. However, the DAG in 2d is *not* a member of the same MEC, since in $\mathcal{G}_4$, 1 and 3 are (unconditionally) d-separated, but are d-connected given 2. □

However, under certain assumptions, it is possible to distinguish the set $\mathcal{M}(\mathcal{G}^*)$ from all other minimal I-MAPs of $\mathbb{P}_X$. This is the case under the *sparsest Markov representation (SMR)* assumption [131], which states that, for any minimal I-MAP $\mathcal{G}'$ of $\mathbb{P}_X$ such that $\mathcal{G}' \notin \mathcal{M}(\mathcal{G}^*)$, we have $|\mathcal{G}'| > |\mathcal{G}^*|$, where $|\mathcal{G}|$ denotes the number of edges in $\mathcal{G}$. Under this assumption, $\mathcal{M}(\mathcal{G}^*)$ can be identified by enumerating over minimal I-MAPs of $\mathbb{P}_X$ and picking the sparsest minimal I-MAP.

More generally, to identify $\mathcal{M}(\mathcal{G}^*)$, structure learning algorithms require some form of *faithfulness* assumption. The strongest such assumption, referred to simply as *the faithfulness assumption*, is exactly the converse to the global Markov property: all conditional independence statements in $\mathbb{P}_X$ must hold as d-separation statements in $\mathcal{G}^*$, i.e., $\mathcal{I}_{\perp\!\!\!\perp}(\mathbb{P}_X) = \mathcal{I}_{\perp\!\!\!\perp}(\mathcal{G}^*)$. The faithfulness assumption is a "genericity" assumption

in the sense that for parametric models, such as linear Gaussian models, the set of parameters which violate the faithfulness assumption is of Lebesgue measure zero [149]. This is demonstrated by the following example.

***Example 5*** Consider the distribution $\mathbb{P}_X$ entailed by the following SCM:

$$\begin{aligned} X_1 &= \epsilon_1 & \epsilon_1 &\sim \mathcal{N}(0,1) \\ X_2 &= \epsilon_2 + \beta_{12} X_1 & \epsilon_2 &\sim \mathcal{N}(0,1) \\ X_3 &= \epsilon_3 + \beta_{13} X_1 & \epsilon_2 &\sim \mathcal{N}(0,1) \\ X_4 &= \beta_{24} X_2 + \beta_{34} X_3 + \epsilon_4 & \epsilon_4 &\sim \mathcal{N}(0,1) \end{aligned}$$

Denoting the corresponding causal graph by $\mathcal{G}$, then the d-separation statements are given by $\mathcal{I}_{\perp\!\!\!\perp}(\mathcal{G}) = \{(1, 4, \{2, 3\}), (2, 3, \{1\})\}$. However, if $\beta_{12}\beta_{24} + \beta_{13}\beta_{34} = 0$, then $\mathrm{Cov}(X_1, X_2) = 0$, so by Gaussianity, we have that $1 \perp\!\!\!\perp_{\mathbb{P}_X} 4$, i.e., $(1, 4, \varnothing) \in \mathcal{I}_{\perp\!\!\!\perp}(\mathbb{P}_X)$ but $(1, 4, \varnothing) \notin \mathcal{I}_{\perp\!\!\!\perp}(\mathcal{G}^*)$. The set of parameters $(\beta_{12}, \beta_{13}, \beta_{24}, \beta_{34})$ satisfying this equality is of Lebesgue measure zero. □

In this example, the effect of $X_1$ on $X_4$ along the paths $1 \to 2 \to 4$ and $1 \to 3 \to 4$ perfectly "cancels out." While perfect cancelation may only occur for very specific parameters, structure learning algorithms do not have direct access to $\mathbb{P}_X$, and must *test* for conditional independence using samples from $\mathbb{P}_X$. Thus, *near* cancelations, e.g., if $\beta_{12}\beta_{24} + \beta_{13}\beta_{34} = 0.0015$, may be indistinguishable from cancelations at small sample sizes. To overcome noise and provide finite sample or high-dimensional guarantees for structure learning algorithms, it is necessary to make stronger assumption, such as *strong* faithfulness [181], which assumes that the (conditional) mutual information between d-connected variables is bounded away from zero. However, the set of parameters which violate the strong faithfulness assumption can have large Lebesgue measure [168]. This has motivated the development of structure learning algorithms under assumptions that only require some *subset* of the missing d-separation statements in $\mathcal{I}_{\perp\!\!\!\perp}(\mathcal{G})$ to hold "strongly" in $\mathbb{P}_X$, thus reducing the size of the set of violating parameters. Such assumptions, including a strong version of the SMR assumption, are reviewed and compared in [131, 183].

Since in general $\mathcal{G}^*$ can only be identified up to its MEC, the natural search space for causal structure learning algorithms is over MECs, rather than DAGs. Consequently, characterizing the structure *within* and *between* MECs has been an important problem for developing structure learning algorithms. We will discuss useful characterizations of the MEC in 3. One way to overcome the limitations on learning from observational data is by using data from *interventions*, which we now formalize.

### 2.2 Interventions and Interventional Markov Equivalence

To formalize the effect of an *intervention* $I$ in an SCM, we consider a new *interventional SCM* where we modify some subset of the structural assignments and/or the distributions of exogenous noise variables, without introducing new nodes into any of the parent sets. If a node $i$ has either its structural assignment $f_i$ or the distribution

of its exogenous noise $\epsilon_i$ modified by intervention $I$, it is called a *target* of the intervention, and we write $i \in I$. The new SCM induces a different distribution $\mathbb{P}_X^I$ on $X$, called the *interventional distribution*, which takes the form

$$\mathbb{P}_X^I(X) = \prod_{i \notin I} \mathbb{P}_X(X_i \mid X_{\mathrm{pa}_\mathcal{G}(i)}) \prod_{i \in I} \mathbb{P}_X^I(X_i \mid X_{\mathrm{pa}_\mathcal{G}(i)}). \tag{2}$$

In general, an intervention consists of any modification of the structural assignment or exogenous noise. To distinguish this most general form of intervention from more stringent definitions of intervention, we will follow [124] and call these *soft* interventions (also referred to as *mechanism changes* in [163]). Particular subclasses of interventions have generated special interest. Most significantly, a *hard* intervention, also called a *perfect*, *surgical* [26], or *structural* [44] intervention, is one which completely removes the dependence of a target $X_i$ on its parents. However, perfect interventions allow for the target to depend on $\epsilon_i$, so that the target's value may still be random, i.e., the interventional distribution is

$$\mathbb{P}_X^I(X) = \prod_{i \notin I} \mathbb{P}_X(X_i \mid X_{\mathrm{pa}_\mathcal{G}(i)}) \prod_{i \in I} \mathbb{P}_X^I(X_i). \tag{3}$$

More extremely, if the structural assignment of $X_i$ is changed to a constant $a_i$, then there is no randomness left in $X_i$. Such a perfect intervention is called a *do-intervention* [113]. In this case, the interventional distribution is

$$\mathbb{P}_X^I(X) = \prod_{i \notin I} \mathbb{P}_X(X_i \mid X_{\mathrm{pa}_\mathcal{G}(i)}) \prod_{i \in I} \mathbb{1}_{X_i = a_i}. \tag{4}$$

***Example 6*** (The interventional SCM for mouse genetic modification) Suppose we implement an intervention on the model in 1, where we edit the genome of the offspring mouse to reduce its weight. In particular, the effect of this intervention is to change the distribution of $\epsilon_4$ to $\mathcal{N}(-10, 0.1)$. The interventional distribution is

$$\begin{aligned}\mathbb{P}_X(X) = \mathsf{Ber}(X_1; .5) \times \mathcal{N}(X_2; 25 + 2X_1, 1) \times \mathcal{N}(X_3; 20 + 2X_1, 1) \\ \times \mathcal{N}(X_4; \frac{1}{2}(X_2 + X_3) - 10, 1) \times \mathcal{N}(X_5; X_4, 1).\end{aligned}$$

This intervention is *not* a perfect intervention, since $X_4$ still depends on its parent $X_2$ and $X_3$. If instead the genetic modification perfectly ensures that the offspring weights 15 grams, i.e. $X_4 = 15$ always, then the intervention would be a perfect intervention—in particular, a *do*-intervention. In this case, the interventional distribution becomes

$$\begin{aligned}\mathbb{P}_X(X) = \mathsf{Ber}(X_1; .5) \times \mathcal{N}(X_2; 25 + 2X_1, 1) \times \mathcal{N}(X_3; 20 + 2X_1, 1) \\ \times \mathbb{1}_{X_4 = 15} \times \mathcal{N}(X_5; X_4, 1),\end{aligned}$$

where $X_4$ does not depend on its parents anymore. □

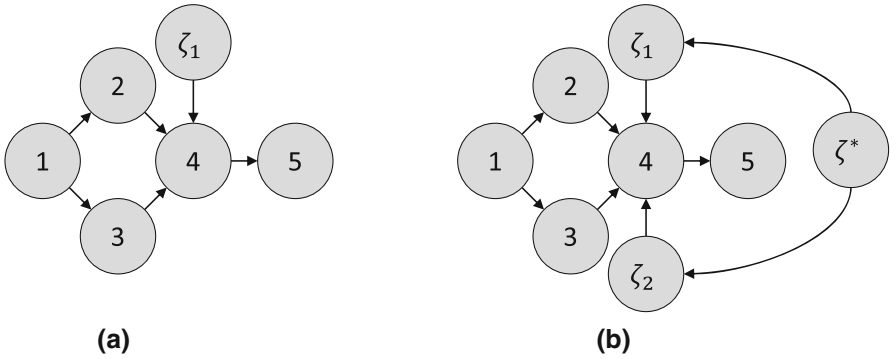


**Fig. 3** **a** $\mathcal{I}$-DAG from 7. **b** $\mathcal{I}$-DAG from 8.

The causal DAG also implies relationships between the observational and interventional distributions. The simplest approach to deriving these relationships is to *extend* the DAG to include variables which represent different interventions, as proposed in [177] and used by [153]. This approach can be seen as an important special case of the *Joint Causal Inference (JCI)* framework [115]. For a single intervention $I$ with targets $T$, this can be achieved by adding a node $\zeta$ with children $T$. $\zeta$ represents a binary variable, where $\zeta = 1$ denotes that a sample comes from the intervention $I$, and $\zeta = 0$ denotes otherwise.

***Example 7*** (Binary encoding of an intervention) Consider the intervention $I_1$ in 6, where the intervention is applied with probability 0.5. Then the joint distribution over $X$ and $\zeta$ is

$$\begin{aligned}\mathbb{P}_{X,\zeta}(X,\zeta) = \mathsf{Ber}(\zeta;.5) \quad &\times \mathsf{Ber}(X_1;.5) \times \mathcal{N}(X_2; 25 + 2X_1, 1) \times \mathcal{N}(X_3; 20 + 2X_1, 1) \\ &\times \mathcal{N}(X_4; \frac{1}{2}(X_2 + X_3) - 10\zeta, 1) \times \mathcal{N}(X_5; X_4, 1).\end{aligned}$$

The causal DAG for $\zeta, X_1, X_2, X_3, X_4, X_5$ is shown in 3a. The node 5 is d-separated from $\zeta$ given 4. Therefore, $\mathbb{P}(X_5 \mid X_4, \zeta = 1) = \mathbb{P}(X_5 \mid X_4, \zeta = 0)$, i.e., $\mathbb{P}_X(X_5 \mid X_4) = \mathbb{P}_X^{I_1}(X_5 \mid X_4)$. □

To generalize to *multiple* interventions, we add a node for each intervention. In particular, consider a set of interventions $\mathcal{I} = \{I_1, \ldots, I_M\}$. For the intervention $I_m$ with targets $T_m$, we introduce a node $\zeta_m$ with children $T_m$. Again, $\zeta_m = 1$ denotes that the sample comes from the intervention $I_m$, and $\zeta_m = 0$ otherwise. However, each sample can only be generated from a single intervention, i.e., $\zeta_m = 1$ for at most one $m$. To reflect this constraint, we include a final node $\zeta^*$, which takes values in $0, 1, \ldots, M$, to indicate which intervention the sample comes from, i.e., $\zeta_m = 1$ if and only if $\zeta^* = m$. Thus, if $\zeta^* = 0$, the sample comes from the observational distribution. The resulting DAG is called the *interventional DAG* ($\mathcal{I}$-DAG) [177].

***Example 8*** (Binary encoding of a set of interventions) Let $I_1$ be the intervention in 6, and let $I_2$ be an intervention which changes the distribution of $\epsilon_4$ to $\mathcal{N}(-5, 0.1)$.

Suppose each intervention has a 40% chance of being applied. In the remaining 20% of the time, no intervention takes place. Then the joint distribution over $X$, $\zeta_1$, $\zeta_2$, and $\zeta^*$ is

$$\begin{aligned}\mathbb{P}_{X,\zeta}(X,\zeta) =& \mathsf{Cat}(\zeta^*; (0,1,2), (0.2, 0.4, 0.4)) \times \left(1 - \mathbb{1}_{\zeta_1=1,\zeta^*\neq 1}\right) \times \left(1 - \mathbb{1}_{\zeta_2=1,\zeta^*\neq 2}\right) \\ &\times \mathsf{Ber}(X_1; .5) \times \mathcal{N}(X_2; 25 + 2X_1, 1) \times \mathcal{N}(X_3; 20 + 2X_1, 1) \\ &\times \mathcal{N}(X_4; \frac{1}{2}(X_2 + X_3) - 10\zeta_1 - 5\zeta_2, 1) \times \mathcal{N}(X_5; X_4, 1).\end{aligned}$$

The causal DAG for $\zeta_1$, $\zeta_2$, $\zeta^*$, $X_1$, $X_2$, $X_3$, $X_4$, $X_5$ is shown in 3b. □

Following [177], we define a *conditional invariance* statement to be a conditional independence statement where the conditioning set includes intervention variables, e.g., $\mathbb{P}_{X,\xi}(X_i \mid X_C, \xi^* = m) = \mathbb{P}_{X,\xi}(X_i \mid X_C, \xi^* = 0)$. This statements posits that a conditional distribution in the $m$th interventional setting is the same as it is in the observational setting, i.e., the conditional distribution is *invariant* under the intervention. A set of observational and interventional distributions satisfies the *$\mathcal{I}$-Markov property* with respect to a DAG $\mathcal{G}$ and a set of interventions $\mathcal{I}$ if it satisfies the global Markov property with respect to $\mathcal{G}$, and satisfies all conditional invariance statements entailed by the $\mathcal{I}$-DAG. Similarly to the observational case, given a set $\mathcal{I}$ of interventions, if two DAGs $\mathcal{G}$ and $\mathcal{G}'$ entail the same set of conditional independence and conditional invariance statements, we call them *$\mathcal{I}$-Markov equivalent*, denoted $\mathcal{G} \approx_{\mathcal{M}_\mathcal{I}} \mathcal{G}'$. The resulting $\mathcal{I}$-Markov equivalence class ($\mathcal{I}$-MEC) is thus a (not necessarily strict) subset of the MEC, as demonstrated by the following example.

***Example 9*** (Interventional Markov equivalence) Given the intervention set $\mathcal{I} = \{I_1\}$ for $I_1$ with target 1, the graphs $\mathcal{G}_2$ and $\mathcal{G}_3$ in 2 are $\mathcal{I}$-Markov equivalent, since they both entail the invariance statements $\mathbb{P}^{I_1}(X_2) = \mathbb{P}(X_2)$ and $\mathbb{P}^{I_1}(X_3) = \mathbb{P}(X_3)$. However, $\mathcal{G}_1$ does *not* entail these invariance statements, so it is not $\mathcal{I}$-Markov equivalent to $\mathcal{G}_2$ and $\mathcal{G}_3$. □

### 2.3 Graphical Representations for Latent Confounding

Thus far, we have discussed how a structural causal model defines a *data generating process* for a particular system and interventions on that system. In the simplest case, called the *causally sufficient* setting, one directly observes the generated data. However, it is often the case that observations are subject to additional processing, in which case we call the setting *causally insufficient*. Two forms of causal insufficiency are commonly considered. First, under *latent confounding*, some of the endogenous variables are simply unobserved, and we call these variables *latent confounders*. Thus, instead of observing samples from the distribution $\mathbb{P}_X$, one observes samples from a *marginal* distribution $\mathbb{P}_{X'}$ for $X' \subset X$. For instance, suppose that in 1, the experimentalist does not record the variable $X_1$ indicating whether the mice were genetically modified. Then, an observer looking at their data would see samples from the distribution $\mathbb{P}_{X_2,X_3,X_4,X_5}$. Second, under *selection bias*, the probability that a sample is observed may depend on the values of some of the variables in the sample. Thus, if

we introduce a binary variable $S$ to indicate whether a sample is observed, and we have $\mathbb{P}(S = 1 \mid X)$ describe the selection process, then one observes samples from the *conditional* distribution $\mathbb{P}(X \mid S = 1)$. For instance, suppose that in 1, the experimentalist only records those experiments for which the mouse in the final generation weighs more than 20 grams. Then, someone looking at their data would see samples from the distribution $\mathbb{P}_X(\cdot \mid X_5 \geq 20)$.

In this section, we will focus on the first type of causal insufficiency, latent confounding. We postpone discussion of selection bias to 5. Without causal sufficiency, one must somehow account for latent confounders to perform accurate causal structure learning. When the latent confounders have special structure, it may be possible to explicitly *recover* the relationship of the latent confounders and the observed variables. One such case is when each latent confounder is a parent of a large portion of the observed variables, which is termed *pervasive confounding*. In such settings, the observed data may be "deconfounded" by removing its top principal components [51, 141], even when the causal relations are nonlinear [3]. A large range of assumptions on the structure between the unobserved and observed variables may be suitable for different applications. A thorough summary of methods using such assumptions is outside of the scope of the current review. Instead, we focus on a different approach for accounting for latent confounders, which acknowledges their presence but does *not* attempt to explicitly recover their relationships with the observed variables.

Structural assumptions on latent confounders can leave a wide range of signatures on the distribution of the observed variables. These signatures include not only conditional independence constraints, which can be expressed in the form $\mathbb{P}_X(X_i, X_j \mid X_C) = \mathbb{P}_X(X_i \mid X_C)\mathbb{P}_X(X_j \mid X_C)$, but also more complex constraints. This includes both equality constraints on the distribution $\mathbb{P}_X$, commonly called *Verma constraints*, as well as inequality constraints. The full set of constraints is referred to as a *marginal DAG model* [46], and can be graphically modeled using a hypergraph. Indeed, [46] show that ordinary mixed graphs are incapable of representing marginal DAG models. Nevertheless, ordinary mixed graphs are capable of encoding a rich subset of the constraints implied by a marginal DAG model. For example, an *acyclic directed mixed graph (ADMG)* encodes a subset of the equality constraints of the marginal DAG model via the associated *nested Markov model* [133, 145]; in fact, the nested Markov model is known to encode *all* equality constraints in the case of discrete variables [47]. It is outside the scope of this review to provide a full overview of the different types of graphs used to capture the constraints of marginal DAG models, instead see [46] and [103] for more thorough overviews.

In our review, we focus on (directed) *ancestral graphs*, which encode only conditional independencies, are closed under marginalization, and have at most one edge between each pair of vertices. Directed ancestral graphs are *mixed* graphs, consisting of both directed and *bidirected* edges. A bidirected edge between two nodes indicates the possibility that they are both children of the same unobserved variable(s). Similarly to directed graphs in the causally sufficient setting, the mixed graphs in the causally insufficient case are required to obey a form of acyclicity condition. In particular, a mixed graph with directed and bidirected edges is called "ancestral" if there are no directed cycles, and if any two nodes that are connected by a bidirected edge (called *spouses*) are not ancestors of one another [132].

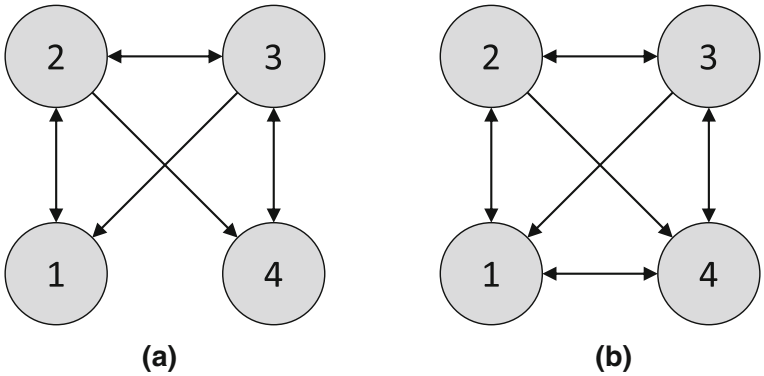


**Fig. 4** **a** An ancestral graph that is not maximal; **b** its maximal completion

Similarly to DAG models, there is a notion of separation in directed ancestral graphs called *m-separation*. The same definition works as for d-separation: two nodes are *m-connected* by a path $\gamma$ given a set of nodes $C$ if (1) every non-collider on the path is not in $C$, and (2) every collider on the path is either in $C$ or has a descendant in $C$. Unfortunately, this notion of separation has the property that two non-adjacent nodes may have *no* m-separating set. Fortunately, adding a bidirected edge between two such nodes does not affect the set of m-separation statements which hold in the directed ancestral graph ( [132], Theorem 5.1). The operation of adding bidirected edges between all such nodes is called taking the *maximal completion* of a directed ancestral graph, and a directed ancestral graph is called *maximal* if it is its own maximal completion. It is natural in structure learning to restrict the search space to directed maximal ancestral graphs (*DMAGs*), so that each adjacency between nodes corresponds exactly to the lack of an m-separating set.

***Example 10*** (*Maximal completion*) 4 shows a graph (left) which is not maximal, since 1 and 4 are m-connected given any of the sets $\{\varnothing, \{2\}, \{3\}, \{2, 3\}\}$, but they are not adjacent. The graph on the right is its maximal completion. □

## 3 Identifiability

As alluded to in the previous section, two Markov equivalent DAGs cannot be distinguished from observational data alone. In particular, given a DAG $\mathcal{G}$, consider the collection of distributions $\mathbb{M}(\mathcal{G})$ which factorize according to $\mathcal{G}$, i.e., can be written in the form 1. This collection depends on the allowed set of conditional distributions $\mathbb{P}_X(X_i \mid X_{\mathrm{pa}(i)})$. If the set of conditional distributions is unrestricted, then we have that $\mathbb{M}(\mathcal{G}) = \mathbb{M}(\mathcal{G}')$ if and only if $\mathcal{I}_{\perp\!\!\!\perp}(\mathcal{G}) = \mathcal{I}_{\perp\!\!\!\perp}(\mathcal{G}')$, i.e., Markov equivalent DAGs give rise to the exact same set of distributions. If the conditional distributions are restricted to specific classes, such as Gaussians or discrete measures, then this equivalence remains [109, 159].

Broadly speaking, there are two approaches to distinguishing between Markov equivalent DAGs. The first approach, which we call the *functional form* approach, considers restricting the class of conditional distributions in such a way that identifiability is possible from only observational data. The second approach, which we call the *equivalence class approach*, does not restrict the class of conditional distributions, but instead uses interventional data to refine the level of identifiability from the MEC to the $\mathcal{I}$-MEC. Given enough interventions, the equivalence class approach is sufficient for completely identifying a DAG or an ADMG [43].

### 3.1 Functional Form Approaches to Identifiability

Suppose the true causal graph is $X_1 \to X_2$. The core idea in this class of approaches is to find asymmetries between models learned in the "causal" ($X_1 \to X_2$) and "anti-causal" ($X_2 \to X_1$) directions. The asymmetries in this bivariate case are often easy to subsequently extend to the multivariate case.

As a canonical example, assume that noise is *additive*, i.e., $X_2 = f_2(X_1) + \epsilon_2$, with $\epsilon_2 \perp\!\!\!\perp X_1$. By making assumptions about the functional form of $f_2$ and the distribution of $\epsilon_2$, it is often possible to show that the induced distribution $\mathbb{P}_X$ cannot be induced by a model of the form $X_1 = f_1(X_2) + \epsilon_1$, $\epsilon_1 \perp\!\!\!\perp X_2$, under the same assumptions on $f_1$ and $\epsilon_1$. For example, [87, 143, 144] assume that each function $f_i$ is linear, and each $\epsilon_i$ is non-Gaussian. Indeed, [76] shows that in linear models, symmetry is only possible in the Gaussian case, and gives more general results for the case where $f_i$ is nonlinear, which form the basis for structure learning methods such as the *Causal Additive Model* (CAM) algorithm [23]. Even in the linear Gaussian case, it is possible to achieve identifiability by imposing additional assumptions, such as equal error variances for each $\epsilon_i$ [123]. It is also possible to move beyond the additive noise case, e.g., by allowing for further nonlinearities after the addition of noise [185].

Thus far, we have discussed identification strategies designed for continuous random variables. Similar results are achievable in the discrete case, e.g., by assuming that the exogenous noise terms have low entropy [93] or by assuming the existence of a (hidden) low cardinality representation of the cause variable that mediates its effects [24].

### 3.2 The Equivalence Class Approach to Identifiability

When no assumptions are made on the functional form, and only observational data is available, the true graph $\mathcal{G}^*$ can only be identified up to the MEC, i.e., the set of DAGs $\mathcal{G}'$ such that $\mathcal{G}' \approx_{\mathcal{M}} \mathcal{G}^*$. Thus, for the purposes of algorithm design, it becomes interesting to characterize when two DAGs are Markov equivalent.

#### 3.2.1 Characterizations of Markov Equivalence Classes

**Characterizations of Markov equivalence in DAGs** There are numerous ways to characterize Markov equivalence in DAGs, and we will cover three main characterizations: a *graphical* characterization, a *transformational* characterization, and a

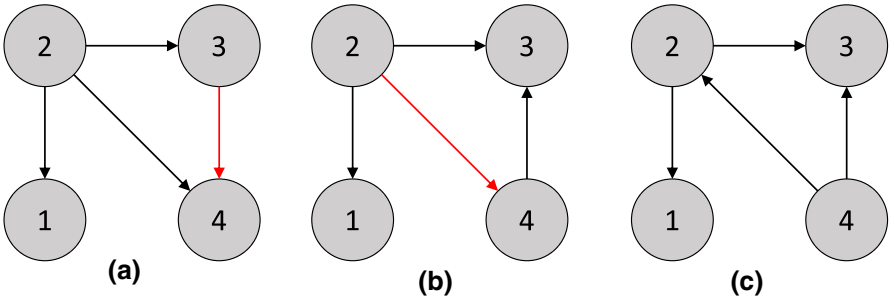


**Fig. 5** Transformational characterization of equivalence in a DAG. Starting with the DAG on the left, we proceed to the right by performing covered edge reversals on the red edges (Color figure online)

*geometric* characterization. The graphical characterization is based on two notions. The *skeleton* of a DAG $\mathcal{G}$ is defined as the set $\text{skel}(\mathcal{G}) = \{(i, j) \mid i \to j \text{ or } j \to i \text{ in } \mathcal{G}\}$. The v-structures (also called *immoralities*) are defined as $\text{vstruct}(\mathcal{G}) = \{(i, j, k) \mid i \to j \leftarrow k \text{ in } \mathcal{G}, (i, k) \notin \text{skel}(\mathcal{G})\}$. Verma and Pearl [170] show that two DAGs $\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent if and only if they have the same skeleton and v-structures, i.e., $\mathcal{G} \approx_{\mathcal{M}} \mathcal{G}'$ if and only if $\text{skel}(\mathcal{G}) = \text{skel}(\mathcal{G}')$ and $\text{vstruct}(\mathcal{G}) = \text{vstruct}(\mathcal{G}')$. Given this graphical notion, it is natural to represent an MEC via an *essential graph*, which is a mixed graph with the same adjacencies as all DAGs in the equivalence class, and with the edge $i \to j$ directed only if $i \to j$ in all DAGs in the equivalence class. Meanwhile, the transformational characterization is based on a single notion: a *covered edge* is an edge $i \to j$ in $\mathcal{G}$ such that $\text{pa}_{\mathcal{G}}(i) = \text{pa}_{\mathcal{G}}(j) \backslash \{i\}$. $\mathcal{G}'$ and $\mathcal{G}$ are related by a *covered edge flip* if $\mathcal{G}'$ has all of the same edges as $\mathcal{G}$, except that the covered edge $i \to j$ in $\mathcal{G}$ is oriented as $j \to i$ in $\mathcal{G}'$. From the graphical characterization, one can deduce that if $\mathcal{G}$ and $\mathcal{G}'$ are related by a series of covered edge flips, then $\mathcal{G} \approx_{\mathcal{M}} \mathcal{G}'$. The transformational characterization states that the converse is also true: If $\mathcal{G} \approx_{\mathcal{M}} \mathcal{G}'$, then $\mathcal{G}$ can be transformed into $\mathcal{G}'$ by a series of covered edge flips [29]. This transformation is illustrated in 5. Finally, the geometric characterization encodes each graph as an integer-valued vector in the space $\mathbb{Z}^{2^{[p]}}$. First, we introduce a set of basis vectors $\delta_A$ for all subsets $A \subset [p]$. Then, the *standard imset* for a DAG $\mathcal{G}$ is given by $u_{\mathcal{G}} = \delta_{[p]} - \delta_{\varnothing} + \sum_{i=1}^{p} \left(\delta_{\text{pa}_{\mathcal{G}}(i)} - \delta_{\{i\} \cup \text{pa}_{\mathcal{G}}(i)}\right)$. Alternatively, [160] introduces the *characteristic imset* $c_{\mathcal{G}}$, with $c_{\mathcal{G}}(A) = 1$ if and only if there exists some $i \in [p]$ such that $A \backslash \{i\} \subseteq \text{pa}_{\mathcal{G}}(i)$. Two DAGs $\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent if and only if $u_{\mathcal{G}} = u_{\mathcal{G}'}$, or equivalently, $c_{\mathcal{G}} = c_{\mathcal{G}'}$.

**Characterizations of interventional Markov equivalence in DAGs** As discussed in 2.2, the effect of an intervention can be formalized by introducing new binary variables to represent each intervention [115]. Therefore, the same characterizations of Markov equivalence that apply in the observational case just discussed also apply in the interventional case. However, it is still instructive to directly characterize the interventional Markov equivalence class. Consider a set of interventions $\mathcal{I}$ such that $\varnothing \in \mathcal{I}$ (i.e., observational data is available). Extending a result for perfect interventions [70, 177] shows that two DAGs $\mathcal{G}$ and $\mathcal{G}'$ are $\mathcal{I}$-Markov equivalent if and only if they (1) have the same skeleton and v-structures, as in the case of a DAG and (2) for all

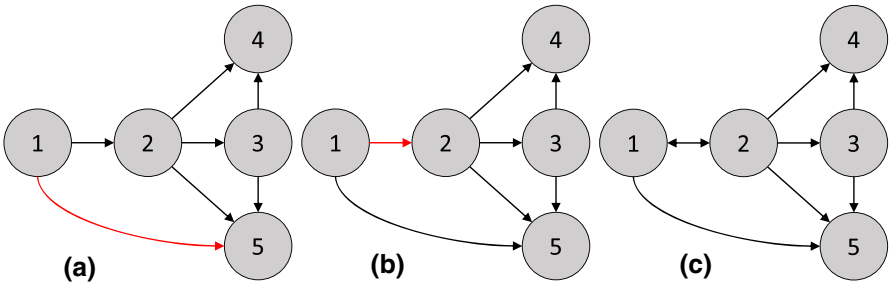


**Fig. 6** Transformational characterization of equivalence in a DMAG. Starting with the DMAG on the left, we proceed to the right by performing legitimate mark changes on the red edges (Color figure online)

$I \in \mathcal{I}$ and $i \in I$, $j \notin I$, we have $j \to i$ in $\mathcal{G}$ if and only if $j \to i$ in $\mathcal{G}'$. Note that this is equivalent to stating that the two $\mathcal{I}$-DAGs do not differ in v-structures of the form $\zeta_I \to i \leftarrow j$, confirming the equivalence with the observational characterization applied to $\mathcal{I}$-DAGs. As an example, under the set of interventions $\mathcal{I} = \{\varnothing, \{1\}\}$, the graphs $\mathcal{G}_2$ and $\mathcal{G}_3$ in 2 are $\mathcal{I}$-Markov equivalent, but $\mathcal{G}_1$ is not, since its $\mathcal{I}$-DAG would not have the v-structure $\zeta_{\{1\}} \to 1 \leftarrow 2$.

**Characterizations of Markov equivalence in DMAGs** As in the case of DAGs, equivalence between DMAG models can be characterized in multiple ways, and we will cover the graphical and transformational characterizations. For both characterizations, we must define the notion of a *discriminating path* for a vertex $k$. A path $\gamma = \langle i, \ldots, k, j \rangle$ is a discriminating path for $k$ if (i) there is at least one node on the path between $i$ and $k$, (ii) every node between $i$ and $k$ is a collider on the path, and (iii) every node between $i$ and $k$ is a parent of $j$. We denote the set of discriminating paths for node $k$ in $\mathcal{G}$ as $\text{discr}_k(\mathcal{G})$. A fundamental result [151] states that two DMAGs $\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent if and only if (i) they have the same skeleton and v-structures, and (ii) for all $k$, for all $\gamma \in \text{discr}_k(\mathcal{G}) \cap \text{discr}_k(\mathcal{G}')$, $k$ is a collider on $\gamma$ in $\mathcal{G}$ if and only if $k$ is a collider on $\gamma$ in $\mathcal{G}'$. Checking this graphical condition for Markov equivalence can be computationally expensive, motivating recent work [77] which provides a new graphical characterization of Markov equivalence in DMAGs that can be checked more efficiently. We next describe the transformational characterization of Markov equivalence in DMAGs. As in the case of DAGs, the transformational characterization requires us to define a local structural modification. In particular, the modification of the edge $i \to j$ in $\mathcal{G}$ to the edge $i \leftrightarrow j$ in $\mathcal{G}'$, or vice versa, is called a *legitimate mark change* [182] if (i) $\text{pa}_{\mathcal{G}}(i) \subseteq \text{pa}_{\mathcal{G}}(j)$, (ii) $\text{sp}_{\mathcal{G}}(i) \backslash \{j\} \subseteq \text{sp}_{\mathcal{G}}(j) \cup \text{pa}_{\mathcal{G}}(j)$, and (iii) there is no $\gamma \in \text{discr}_i(\mathcal{G})$ for which $j$ is the endpoint adjacent to $i$. The authors in [182] show that $\mathcal{G} \approx_{\mathcal{M}} \mathcal{G}'$ if and only if $\mathcal{G}$ and $\mathcal{G}'$ are connected by a series of legitimate mark changes. This transformation is illustrated in 6.

### 3.2.2 Combinatorial Aspects of Markov Equivalence

Since DAGs in general can only be identified up to ($\mathcal{I}$-)Markov equivalence, it has been of significant interest to study the size of a given MEC, the number of MECs

**Table 1** Number of MECs (first row), the ratio of the number of MECs to the number of DAGs (second row), and the ratio of the number of MECs of size 1 compared to the total number of MECs (third row), up to 10 nodes

| $p$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| # MEC | 11 | 185 | 8.78e4 | 1.06e6 | 3.13e8 | 2.12e11 | 3.26e14 | 1.12e18 |
| $\frac{\#\text{MEC}}{\#\text{DAG}}$ | 0.44 | 0.34 | 0.30 | 0.28 | 0.27 | 0.27 | 0.27 | 0.27 |
| $\frac{\#\text{MEC}-1}{\#\text{MEC}}$ | 0.36 | 0.32 | 0.30 | 0.29 | 0.28 | 0.28 | 0.28 | 0.28 |

over a given number of variables, and the minimum number of interventions required to identify a DAG (i.e., obtain a $\mathcal{I}$-MEC of size 1).

The first problem—computing the number of DAGs within a given MEC, or computationally equivalently, sampling uniformly from the MEC—is important for a number of experimental design algorithms [56], which use Monte Carlo approximations to compute expectations over the MEC and pick interventions with good average-case behavior. A recent advance [175] provides a polynomial-time algorithm for this task based on a representation of the equivalence class via clique trees, improving over previous algorithms with exponential worst-case runtime [6, 14, 52, 57, 72, 162].

To address the second problem, [61] develops a program for enumerating all MECs on graphs with a given number of nodes, and obtained results for graphs of up to 10 nodes, shown in 1. Further theoretical works [126, 127] study the problem of enumerating all MECs for a fixed skeleton using the idea of generating functions from combinatorics. The computational results in [61] suggest that, asymptotically, the average MEC contains approximately 4 DAGs, and that roughly one quarter of all MECs are comprised of only a single DAG, in which case no interventional data is needed to identify the causal DAG. However, proving these conjectured limits, as well as efficiently enumerating the number of MECs on a given number of nodes, remain open combinatorial and computational problems. Less work has been done to characterize the average number of interventions required to identify a DAG. For a given DAG, [152] characterizes the minimum-size set of single-node interventions needed to identify the underlying causal DAG, using a representation based on clique trees. However, this work does not address the average of this quantity over all DAGs on a given number of nodes. Meanwhile, [88] conducts a computational study of the average number of *greedily* selected interventions to identify a graph, where the average is with respect to a directed Erdös-Rényi graph model. In this model, the results suggest that the number of interventions necessary is typically less than 4, but further work is necessary to characterize the average with respect to the uniform distribution over graphs and to address the case where interventions are picked optimally.

## 4 Methods for Causal Structure Learning

Thus far, we have discussed what is in principle identifiable about the underlying causal DAG with observational and interventional data. Now, we present algorithms which carry these principles of identifiability into practice. In particular, we will dis-

cuss a number of algorithms which are *consistent*, i.e., in the limit of infinite data, they provably learn all identifiable causal structures. We will also highlight some *heuristic* algorithms, which do not have consistency guarantees but often perform well in practice. We begin with a broad overview of the different paradigms for causal structure learning, before diving into methods which explicitly leverage the combinatorial structures already discussed. At the highest level, methods for estimating causal models from data fall into two broad categories: *constraint-based* methods and *score-based* methods. Constraint-based methods are natural when viewing causal structure learning as a constraint satisfaction problem, where conditional independences or other constraints that can be inferred from data are used to iteratively prune the space of possible graphs. In contrast, score-based methods arise from viewing causal structure learning as a *combinatorial optimization* problem. These methods assign a score to each graph (or equivalence class) which quantifies how well it fits the data, then search the space of graphs (or equivalence classes) to find a model which optimizes the score. To highlight the general principles of these two paradigms, we will first concentrate on the causally sufficient case with only observational data. Then, in 4.3, we discuss algorithms that can make use of interventional data, and in 4.4, we briefly discuss algorithms for learning in the presence of latent confounding.

**Constraint-based approaches** The most prominent constraint-based approach to causal structure learning is the PC algorithm [86, 149]. The PC algorithm begins with a complete undirected graph and iteratively deletes edges by testing conditional independences involving conditioning sets of increasing cardinality. Then, the second phase of the PC algorithm orients v-structures by reusing the conditional independences found in the first phase. Additional orientations can be inferred via the *Meek orientation rules* [111].

The method for testing conditional independence (CI) depends on modeling assumptions as well as practical considerations such as computational complexity. For example, in a multivariate Gaussian distribution, two variables $X_i$ and $X_j$ are conditionally independent given the variables $X_C$ if and only if the partial correlation $\rho_{ij|C}$ is zero. Since the distribution of sample partial correlation coefficients is well known (see, e.g., [86]), hypothesis testing for CI in the Gaussian setting is straightforward and computationally efficient. On the other hand, in nonparametric settings, hypothesis tests for conditional independence can often be performed based on more complicated test statistics [75, 158, 186]. Unfortunately, impossibility results [142] state that any *uniformly* valid conditional independence test (i.e., one whose false positive rate tends to at most the significance level $\alpha$, over all possible distributions $\mathbb{P}$ where $X \perp\!\!\!\perp_{\mathbb{P}} Y \mid Z$) will have no statistical power (i.e., the probability of a true positive will also be at most $\alpha$). Thus, testing conditional independence requires additional assumptions on the set of possible distributions, such as complexity restrictions on the function space of $\mathbb{E}_{\mathbb{P}}[X \mid Z]$.

Under such complexity assumptions, conditional independence tests allow constraint-based approaches to directly be applied to nonparametric settings, even permitting high-dimensional consistency bounds in these settings [69]. Furthermore, because conditional independences also characterize DMAG models, constraint-based approaches can be easily extended to settings with latent variables [35]. Pushing further, one may encode conditional independences as logical constraints, allowing them to be used

in answer set programming (ASP) solvers. These solvers can search over more general model classes and easily incorporate background knowledge [80, 180]. However, ASP-based causal structure learning methods are widely viewed as being difficult to scale for many practical applications.

**Score-based approaches** Score-based methods for causal structure learning originated in parametric settings, such as in discrete or linear Gaussian models. In parametric settings, the score $S(\mathcal{G})$ of a graph $\mathcal{G}$ is often based on the *marginal likelihood* $\mathbb{P}(\mathbb{X} \mid \mathcal{G})$ of the data $\mathbb{X}$ given the graph $\mathcal{G}$, with respect to some prior $\mathbb{P}(\theta)$ over the parameters $\theta$. In some cases, e.g., when choosing a *conjugate* prior for the likelihood function, $\mathbb{P}(\mathbb{X} \mid \mathcal{G})$ can be computed in closed form [55]. Alternatively, it is common to use a consistent approximation of the marginal likelihood, in the form of the *Bayesian information criterion (BIC)* score [31, 32]. Such likelihood-based scores can be extended to nonparametric settings, e.g., by using Gaussian process priors [50] or non-paranormal distributions [117]. The BIC score and related scores are also a natural starting point from which to develop more sophisticated scores with better statistical and computational properties, see, e.g., [21].

Finding the highest scoring DAG model is generally NP-hard [30], imposing a trade-off between computational efficiency and algorithmic consistency guarantees. Score-based methods can generally be subdivided into three categories based on how they address this trade-off. On the one end of the spectrum, *exact* score-based approaches find some $\widehat{\mathcal{G}}$ that exactly optimizes the score $S$. Exact approaches address computational issues using a variety of combinatorial optimization techniques and heuristics, e.g., dynamic programming [95, 121], A*-style state-space search [179], or methods from integer linear programming [11, 38, 39, 82]. For example, the GOBNILP algorithm [39] uses the geometric characterization of Markov equivalence classes to reduce structure learning to an integer linear programming problem. This reduction allows the use of techniques such as cutting planes and pricing to handle the exponential number of decision variables and constraints.

*Greedy* score-based approaches trade off to achieve better computational efficiency over exact approaches by relaxing the requirement that $\widehat{\mathcal{G}}$ optimizes $S$. Most prominently, greedy equivalence search (GES) [31] and its variants [32] perform a search over equivalence classes of graphs that greedily optimizes $S$. While greedy algorithms are not *exact*, they are still *consistent*, placing them in a middle ground on the computational–statistical trade-off. Notably, [106] shows that GES and a number of other greedy approaches can also be viewed geometrically. In particular, these methods can be seen as *edge walks* between vertices of the *characteristic imset polytope*, i.e., the convex hull of all characteristic imsets on $p$ variables. Finally, at the other extreme of this trade-off, gradient-based methods [101, 178, 187, 190] relax the discrete search space over DAGs to a *continuous* search space, allowing gradient descent and other techniques from continuous optimization to be applied to causal structure learning. However, the search space of these problems is highly non-convex, so that the optimization procedure may become stuck in a local minima. Thus, consistency guarantees for these methods will depend on theoretical advances in global minimization of such non-convex optimization problems.

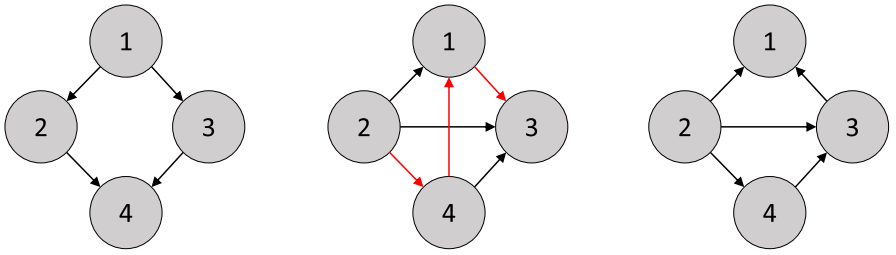


**Fig. 7** A greedy step over minimal I-MAPs performed by GSP. **a** True graph $\mathcal{G}^*$, to which the distribution $\mathbb{P}(X)$ is faithful. **b** Minimal I-MAP associated with the permutation $\pi^{(0)} = [2, 4, 1, 3]$, with covered edges shown in red. **c** Minimal I-MAP associated with the permutation $\pi^{(1)} = [2, 4, 3, 1]$, obtained after flipping the covered edge $1 \to 3$ (Color figure online)

### 4.1 Learning DAGs Using Permutation-Based Algorithms

Beyond the constraint-based and score-based paradigms for causal structure learning already discussed, there are a variety of *hybrid* methods [7, 117, 138, 140, 166], which generally use constraints to reduce the search space, and scores to optimize over this reduced search space. In this section, we discuss the *greedy sparsest permutation* (GSP) algorithm, a hybrid method that constrains the search space to the set of (estimated) minimal I-MAPs of $\mathbb{P}_X$. By focusing on this method, we highlight the combinatorial nature of the problem of causal structure learning.

As discussed in 2, a distribution $\mathbb{P}_X$ may permit several different minimal I-MAPs. Since the minimal I-MAPs of $\mathbb{P}_X$ are the (locally) sparsest DAGs which can correctly model $\mathbb{P}_X$, they form a natural space over which to search for the true DAG $\mathcal{G}^*$. Furthermore, the space of minimal I-MAPs of $\mathbb{P}_X$ can be described as the image of a $\mathbb{P}_X$-dependent map, with the $\mathbb{P}_X$-independent domain of $S_p$ of *permutations* of $[p]$. We denote by $i <_\pi j$ that $i$ is earlier in the permutation $\pi$ than $j$, and we call a graph $\mathcal{G}$ *consistent* with a permutation $\pi$ if and only if $i <_\pi j$ implies that $j \not\to i$ in $\mathcal{G}$. The following result establishes the existence of a unique map from permutations to minimal I-MAPs.

**Theorem 1** (from [169]) *Given a permutation $\pi$ and a distribution $\mathbb{P}_X$, there exists a unique graph $\mathcal{G}_{\mathbb{P}_X}(\pi)$ that is consistent with $\pi$ and is a minimal I-MAP for $\mathbb{P}_X$. This graph has edges*

$$\{i \to j \mid X_i \not\perp\!\!\!\perp X_j \mid X_{\mathrm{pre}_\pi(j)\setminus\{i\}}\} \qquad \textit{where } \mathrm{pre}_\pi(j) = \{k \mid k <_\pi j\}.$$

Given a graph $\mathcal{G}$, let $|\mathcal{G}|$ be the number of edges in the graph. The *sparsest I-MAP theorem* [131] establishes that, under a mild condition, the sparsest minimal I-MAPs of $\mathbb{P}_X$—i.e, those such that $|\mathcal{G}_{\mathbb{P}_X}(\pi)|$ is minimized—are Markov equivalent to the underlying causal graph $\mathcal{G}^*$. In particular, the required condition for this result is strictly weaker than the *restricted faithfulness assumption* [129], which only requires that $\mathcal{I}_{\perp\!\!\!\perp}(\mathbb{P}_X)$ and $\mathcal{I}_{\perp\!\!\!\perp}(\mathcal{G}^*)$ agree on conditional independences/d-separations involving nodes connected by paths of lengths one or two. The sparsest I-MAP theorem directly

suggests the *sparsest permutation* (SP) algorithm: enumerate over all permutations $\pi \in S_p$, estimating the minimal I-MAP $\mathcal{G}_{\mathbb{P}_X}(\pi)$ for each of these permutations using conditional independence testing, and return the sparsest graphs.

However, the SP algorithm is clearly computationally prohibitive, since the size of $S_p$ is super-exponential in $p$. To address this issue, [148] proposed the *greedy sparsest permutation* (GSP) algorithm. GSP searches greedily over the space of permutations, and hence, minimal I-MAPs. In particular, at each step $i$ of the algorithm, GSP maintains a permutation $\pi^{(i)}$ and its corresponding minimal I-MAP $\mathcal{G}_{\mathbb{P}_X}(\pi^{(i)})$. At this step, GSP searches over the Markov equivalence class of $\mathcal{G}_{\mathbb{P}_X}(\pi^{(i)})$ for some DAG $\mathcal{G}'$ which is *not* a minimal I-MAP of $\mathbb{P}_X$. This search can be executed by repeatedly flipping covered edges to generate new permutations. Upon finding $\mathcal{G}'$ which is not a minimal I-MAP of $\mathbb{P}_X$, there must be some strict sub-DAG $\mathcal{G}''$ of $\mathcal{G}'$ which is a minimal I-MAP of $\mathbb{P}_X$. GSP then takes the topological ordering of this sub-DAG as the new permutation $\pi^{(i+1)}$, with $\mathcal{G}''$ as its corresponding minimal I-MAP $\mathcal{G}_{\mathbb{P}_X}(\pi^{(i+1)})$. One greedy step of GSP is demonstrated in 7.

As in the case for other greedy approaches, GSP has an interpretation as an edge walk over a convex polytope. In particular, starting from the *permutahedron*, i.e., the convex hull of all permutations on $p$ nodes, we may define the *DAG associahedron* by contracting all edges $\pi^{(i)} - \pi^{(j)}$ of the permutahedron for which $\mathcal{G}_{\mathbb{P}_X}(\pi^{(i)}) = \mathcal{G}_{\mathbb{P}_X}(\pi^{(j)})$. As shown in [148], this contraction results in a convex polytope, GSP is equivalent to an edge walk along this polytope, and, under conditions that are strictly weaker than the faithfulness assumption, this edge walk terminates in the Markov equivalence class of the causal graph $\mathcal{G}^*$ underlying $\mathbb{P}_X$. The central technical ingredient in this proof is the existence of *Chickering sequences*. In particular, [31] proves the *Meek conjecture* for DAGs [112]: if $\mathcal{G}_M$ is an I-MAP of $\mathcal{G}_0 = \mathcal{G}^*$, then there exists a sequence $(\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_M)$ composed only of edge additions and covered edge reversals. This sequence is called a Chickering sequence [148] and its existence guarantees the consistency of GSP.

In addition to the consistency of GSP and the algorithms discussed previously, which provides guarantees as the sample size goes to infinity, it is important to understand the performance of different algorithms for finite sample size. Simulation results suggest that score-based and hybrid approaches perform better for fixed sample sizes [8, 74, 117]. However, a theoretical characterization of the trade-offs between these algorithms on finite samples is not well understood and is an important area for future research, as also briefly described in 5.

### 4.2 Bayesian Methods for Causal Structure Learning

Thus far, we have only discussed causal structure learning methods which return a *point estimate*—i.e., a single DAG that (approximately or locally) maximizes a score, and/or satisfies inferred conditional independences. However, when the amount of data is small, there may be substantial *uncertainty* about the underlying graph (or equivalence class). A common framework for quantifying this uncertainty is *Bayesian inference*. Given some dataset $\mathbb{D}$, instead of returning a point estimate, Bayesian methods return (an approximation to) the posterior $\mathbb{P}(\mathcal{G} \mid \mathbb{D})$ over graphs. This posterior allows one

to compute marginal probabilities of any feature of interest, such as the posterior probability of some edge $i \to j$.

Bayesian methods for causal structure learning can be divided into three types of approaches: *exact* approaches (e.g., [42]) and two types of approximate approaches: *variational* and *sampling-based* approaches. Similarly to the gradient-based approaches discussed before, variational approaches do not necessarily return a consistent estimate of the posterior; rather, they project the posterior onto a *variational family* $\{Q(\cdot \mid \theta)\}_{\theta \in \Theta}$, which is more computationally convenient. However, traditional variational families, such as multivariate Gaussians, are continuous and thus do not apply to the discrete setting of DAGs. Thus, until recently, variational methods for Bayesian causal structure learning have not been widely studied. For a recent work in this space, see [107], which uses relaxations of DAGs to a continuous search space and neural networks to parameterize a flexible variational family.

On the other hand, sampling-based approaches to Bayesian causal structure learning have been much more widely studied. Markov chain Monte Carlo (MCMC) methods have been especially popular, beginning with the *structure MCMC* algorithm [110], which runs a Metropolis–Hastings algorithm over the space of DAG models, using edge additions and deletions to move in this space. However, this approach suffers from slow *mixing times* due to regions of high-probability DAG models being separated by large regions of low-probability DAG models, i.e., if structure MCMC finds some high-probability DAG $\mathcal{G}_0$, some other high-probability DAG $\mathcal{G}_M$ may only be reachable from $\mathcal{G}_0$ by a sequence of DAGs $\mathcal{G}_1, \ldots \mathcal{G}_{M-1}$ which have very low probability. Thus, the probability that structure MCMC traverses this path becomes incredibly low, so that $\mathcal{G}_M$ will not be sampled without running the algorithm for many steps.

This difficulty has motivated a search for "smoother" sampling spaces, either by adding moves to structure MCMC [62, 67], or by changing the search space, as was done in *order MCMC* [45, 49], *partial order MCMC* [118], and *partition MCMC* [99]. These methods run a Markov chain over some "coarser" space (permutations, partial orders, or ordered partitions), then sample DAGs conditionally based on their consistency with the coarser structure. The *minimal I-MAP MCMC* algorithm [5] also runs a Markov chain over the coarser space of permutations. However, instead of conditionally sampling a DAG based on each permutation, it estimates the minimal I-MAP associated to each sampled permutation.

Since the space of permutations is much smaller than the space of DAGs or MECs, the minimal I-MAP MCMC algorithm can mix more quickly than previous algorithms. But this comes at a price: Minimal I-MAP MCMC does not sample over the entire posterior distribution of DAG models, but only a restricted subset. Luckily, this price is small: intuitively, conditional on an order, the minimal I-MAP asymptotically has the highest posterior probability, so a point mass on the minimal I-MAP is a good approximation of the true conditional distribution. Indeed, [5] shows that the posterior approximation error for any bounded function of the graph decreases exponentially with the number of samples. By highlighting this algorithm, we once again see the computational benefits that are possible when considering the combinatorial nature of the causal structure learning problem.

### 4.3 Causal Structure Learning Using Interventional Data

As discussed in 3, interventional data can significantly improve the identifiability of causal models. Several approaches have been proposed for learning from a combination of observational and experimental data, going back at least to the Bayesian approaches of [36] and [42]. As in the case of learning from purely observational data, these approaches can be divided into constraint-based approaches, such as the COmbINE [165] algorithm, and score-based approaches. Score-based approaches include greedy algorithms, such as *Greedy Interventional Equivalence Search (GIES)* [70], and gradient-based algorithms, such as meta-learning approaches [89] and DCDI [22]. Note that, unlike in the case of GES for observational data, GIES is known to not be consistent for interventional data [171].

The Joint Causal Inference framework [115] discussed in 2.2 suggests a natural way to extend causal structure learning algorithms for observational data to settings with interventional data. In particular, an algorithm for the observational setting can be used to learn the $\mathcal{I}$-DAG by appending indicator variables to the dataset for each intervention $I \in \mathcal{I}$, as long as the algorithm can incorporate appropriate forms of background knowledge. This background knowledge includes *exogeneity*—i.e., intervention variables are not caused by the original "system" variables, *randomized context*—i.e., lack of confounding between the intervention and system variables, and *generic context*—i.e., that the intervention variables are deterministically related to one another. As an example, [153] shows that the GSP algorithm can be adapted to include these assumptions, along with any assumptions about known targets of each intervention, while maintaining consistency of the algorithm. They call the resulting algorithm the *Unknown Target Intervention GSP (UT-IGSP)* algorithm to emphasize its ability to handle interventions with unknown targets, extending previous works where targets were assumed to be known [171, 177]. Finally, it is also natural to develop Bayesian variants of causal structure learning algorithms for interventional data, e.g., [27] shows how to compute posteriors over DAGs in the setting when the data is multivariate Gaussian.

### 4.4 Causal Structure Learning in the Presence of Latent Confounding

The approaches to causal structure learning in the causally insufficient setting follow the same broad categorization as approaches in the causally sufficient setting. In particular, the *Fast Causal Inference (FCI)* algorithm [150] is a constraint-based algorithm for learning DMAGs, similar in spirit to the PC algorithm. The FCI algorithm has inspired several variants, including *Really Fast Causal Inference (RFCI)* [35], and *FCI+* [34]. Score-based methods include both greedy search strategies, such as *Greedy FCI (GFCI)* [119], *MAG Max–Min Hill Climbing* ($M^3HC$) [167], and *Conservative rule and Causal effect Hill Climbing (CCHM)* [33], exact score-based approaches, such as AGIP [28], and gradient-based approaches [16].

As in the case of learning DAGs, we will discuss a hybrid method for learning DMAGs, which combines elements of both score-based and constraint-based approaches, and elucidates the combinatorial aspects of learning DMAGs. This

method, called the *Greedy Sparsest Poset* (GSPo) method, restricts the search space of DMAGs to minimal I-MAPs of the distribution $\mathbb{P}_X$. This space can be realized as the image of a map $\mathcal{G}_{\mathbb{P}_X}$ from *partially ordered sets (posets)* to graphs. A partially ordered set $\pi$ defines a relation $\preceq_\pi$ that captures the notion of an ordering via three requirements: reflexivity ($i \preceq_\pi i$ for all $i$), antisymmetry ($i \preceq_\pi j$ and $j \preceq_\pi i$ implies $i = j$), and transitivity ($i \preceq_\pi j$ and $j \preceq k$ implies $i \preceq_\pi k$). Because of the definition of the ancestrality condition, the set of complete DMAGs can be put in bijection to the set of posets, so that posets form a natural domain for the map $\mathcal{G}_{\mathbb{P}_X}$.

The authors in [13] show that $\mathcal{G}_{\mathbb{P}_X}(\pi)$ can be constructed using a procedure similar to the procedure defined for DAGs in 1, although the construction requires two iterations of conditional independence testing between pairs of variables instead of one. They also provided a version of the sparsest I-MAP theorem for DMAGs, i.e., under a restricted faithfulness assumption, the sparsest minimal I-MAPs of $\mathbb{P}_X$ are all Markov equivalent to the underlying DMAG $\mathcal{G}^*$. Motivated by the GSP algorithm for learning DAGs, [13] introduce the *greedy sparsest poset (GSPo)* algorithm for learning DMAGs, which uses legitimate mark changes to search over posets and iteratively find sparser I-MAPs. Over 100,000 synthetic examples suggest that the GSPo algorithm is consistent, but proof of its consistency is an important open problem, and closely tied to the open problem of generalizing Meek's conjecture [31, 112] to DMAGs.

## 5 Discussion and Open Problems

In this review article, we sought to cover both classical and recent approaches to causal structure learning, emphasizing the combinatorial nature of this problem. We end by discussing several related areas of work that were not covered in depth and remain under active development.

***Learning with both interventions and latent confounding*** While we separately discussed learning with interventional data and learning under confounding, it is natural to combine these two settings. Recent work [83] considers this combination for DMAGs, introducing the new notion of $\Psi$-Markov equivalence to capture pairs of graphs and interventions which induce the same set of conditional independencies and conditional invariances. This work allows for both *soft* and *unknown-target* interventions. Furthermore, [83] provides a graphical characterization of $\Psi$-Markov equivalence, and introduces a constraint-based algorithm, called $\Psi$-FCI, for learning the $\Psi$-Markov equivalence class from data. As a next step it is natural to consider score-based algorithms, both exact and greedy, for learning DMAGs, ADMGs, and other subclasses of marginal DAG models, using a combination of observational and interventional data.

***Learning with assumptions on the latent structure*** As indicated in 2.3, in some cases with unobserved confounding, it is desirable to *recover* the unobserved variables and their relationship to the observed variables. Naturally, recovery of these details requires assumptions on their structure. A common assumption, called the *exogeneity* or *measurement* assumption, is that all unobserved variables are upstream of the observed variables, i.e., none of the unobserved variables are caused by any of the observed variables.

With the exogeneity assumption as a starting point, additional assumptions may be made to (approximately) recover the latent variables, and possibly, the structure between them. For example, several works [51, 141] consider recovering the unobserved variables in settings with *pervasive confounding*, i.e., when each unobserved variable has a direct effect on a large number of observed variables. As an important special case of this setting, some works have considered recovering a *mixture* of DAG models [65, 137, 156, 157], where there is a single unobserved variable that is a parent of all variables in the graph. Alternatively, many works [25, 68, 84, 91, 100, 136, 176, 184] consider recovering unobserved variables under the measurement assumption and a form of *purity* or *anchor* assumption, where each unobserved variable must have some number of observed variables which are only their children. Few works consider recovering unobserved variables *without* the assumption of exogeneity, with [154] being a recent exception.

***Learning in the presence of selection bias*** As suggested in 2.3, considerable effort has gone into characterizing the distributional constraints imposed by marginalization of DAG models. However, in many applications, the observed distribution is the result of both marginalization and *conditioning* of an underlying distribution. In particular, such observed distributions are induced by *selection bias*, where the probability that a sample is observed is dependent on the value of some of the variables in the sample. General maximal ancestral graphs (see 2.3), which allow for undirected edges in addition to directed and bidirected edges, are conditional independence models which are closed under marginalization and conditioning. As in the case of marginalization, several graphical representations, including MC graphs [96] and summary graphs [174], have been introduced to capture constraints induced by such conditional models. However, to the best of our knowledge, there is no graphical representation which exactly captures all equality and inequality constraints induced by conditioning a DAG model, in contrast to the case for marginal models [46]. Thus, important next steps include (1) developing a graphical representation which fully captures both marginalization and conditioning, (2) developing notions of Markov equivalence in this setting, including with interventional data, and (3) developing structure learning algorithms in this general setting.

***Learning cyclic causal models*** As indicated in 2, a widespread assumption in causal modeling and causal structure learning is that the structural causal model (SCM) induces an acyclic graph. However, this may not be the case if the SCM models a system that involves feedback loops. While the underlying dynamics of the system are necessarily acyclic over time, feedback loops can arise when modeling the equilibrium states of such systems [19]. For example, in gene regulatory networks, we may have that gene A regulates gene B, and gene B also regulates gene A, so that intervening on either gene will affect the value of the other gene. Recent work [20] has investigated the semantics of cyclic causal models, showing that Markov properties and other desirable properties hold in the case of certain solvability conditions. Despite the technical difficulties associated with cyclic models, several approaches have been proposed for learning their structure from data. These approaches include many algorithms designed for the linear case, including LLC [78], score-based approaches [58], and BackShift [134]. Algorithms for the general case include SAT-based approaches [81], exact score-based approaches [130], and constraint-based approaches [48, 114, 155].

***Statistical and computational complexity of causal structure learning*** In conjunction with methodological developments for settings with cycles, latent confounding, selection bias, and interventional data, it is important to understand the fundamental statistical and computational limits of causal structure learning, and any trade-offs between these. The analysis of existing causal structure learning algorithms gives *upper bounds* on what is statistically and computationally achievable. Recent work derives upper bounds for a wide range of settings, including the linear equal-variance setting [60], the linear non-Gaussian setting [173], other parametric settings [120, 128], as well as nonparametric settings [53]. On the other hand, it is important to understand the fundamental *lower bounds* on the sample complexity needed by *any* causal structure learning algorithm. Such lower bounds have been established for the exponential family setting [59] and the linear equal-variance setting [54], but the lower bounds for a wide range of settings and assumptions remain uncharacterized.

Furthermore, since consistency of causal structure learning algorithms always requires some form of "faithfulness" or genericity assumption (see 2), there are likely trade-offs between the strength of faithfulness assumption imposed and computational and statistical complexity. Indeed, an interesting open question is to characterize the weakest assumption needed for causal structure learning, with the *sparsest Markov representation* assumption [131] being one candidate. Finally, the works discussed above are all in causally sufficient settings with only observational data. Incorporating interventional data into these analyses would open the possibility for a reduction in overall sample complexity, and may introduce a landscape of trade-offs between interventional and observational sample complexities. Indeed, interventional data has been considered in recent works [1, 17] on the statistical and computational complexity of *causal inference* tasks, where the causal graph is assumed to be known and the task is to estimate interventional distributions. An interesting future direction is to also explore the effect of interventional data on the complexity of causal structure learning.

***Experimental design for causal structure learning*** In this review article, we have focused on causal structure learning in a *passive* setting, where we are given a dataset, or possibly several datasets from different interventions or contexts. However, in many scientific settings, such as biology, where interventions such as genetic or chemical perturbations can readily be performed, an important component of causal discovery is the choice of what data to gather [63]. This leads us to consider *experimental design* approaches for causal structure learning, where an experimenter may pick interventions (and their values) in an effort to identify the underlying causal structure. Several approaches have been proposed for a variety of settings. In the *non-adaptive* setting, the experimenter picks all interventions at once. In [43] it is shown that, in the absence of any preexisting observational data, $p - 1$ interventions are sufficient and in the worst-case necessary for identifying the underlying causal structure over $p$ variables. Other work in the non-adaptive setting considers the presence of background knowledge (e.g., from observational data) [79], differences in costs between interventions [92, 105], and a *fixed-budget* setting [56].

Alternatively, the *adaptive* setting allows the experimenter to observe the outcome of each intervention before picking the next intervention. He and Geng [73] and Hauser and Bühlmann [71] propose greedy approaches for the adaptive setting, picking new interventions based on some measure of either expected or worst-case information

gain. While these approaches are designed for the *noiseless* setting, in which an infinite amount of data is gathered from each intervention, more recent works [98, 164] explore greedy approaches in the *noisy* setting. [66] shows that strategies which maximize expected information gain can be exponentially suboptimal in the number of interventions that they use, and propose the *Central Node* algorithm for settings where the essential graph is a tree. They show that this algorithm is a 2-approximation to the optimal adaptive strategy. Follow-up work [152] adapts this algorithm to a more general class of essential graphs, provides a characterization of the number of single-node interventions needed by an oracle to identify a causal graph, and shows that their algorithm uses within a logarithmic factor of this number of interventions.

In between the non-adaptive and adaptive settings, [4] considers the active *batched* setting, in which the experimenter observes the outcome of a batch of interventions before picking the next batch of interventions. Recent work [161] establishes novel submodularity properties for greedy objectives in this settings, allowing for efficient optimization over the choice of interventions in each batch. Taken together, these recent advances suggest several future directions, including (1) characterizing the number of multi-target interventions needed by an oracle in the adaptive case [125], (2) approximation guarantees for experimental design, compared to either oracles or optimal strategies, and (3) experimental design in settings with latent confounding [2, 94], selection bias, and cycles.

***Targeted causal structure learning*** Thus far, we have focused on the problem of causal structure learning as an end in itself; i.e., in both the passive and active settings discussed, the desired output was a causal graph (or equivalence class). However, ultimately, a major motivation for causal structure learning is to use the causal model in downstream tasks. A task of considerable importance is *policy evaluation*, i.e., predicting the effect of an action. The overall goal of task can be phrased as estimating a specific *functional* of an interventional distribution defined by a structural causal model $M$. Then two principal subtasks are (1) determining whether this functional is *identifiable* by transforming it into a functional of the available distributions and (2) estimating the resulting functional from samples. When the only available distribution is the observational distribution defined by $M$, possibly with some variables unobserved, the first subtask is covered by the *ID algorithm* [146] and its variants [147].

More generally, data might be available from some set of interventional distributions defined by $M$, or from observational and interventional distributions associated to some related structural causal model $M'_1, \ldots, M'_K$. The relation between these structural causal models is encoded using a *selection diagram*, and the task of using the selection diagram to identify the functional is covered by a rich literature on *transportability* [9, 10, 37, 104]. Once the target functional is transformed into a functional of the available distributions, it becomes essential to estimate the functional in a sample-efficient way. This has been extensively studied in the literature on *semiparametric efficiency* [15, 135], double machine learning [85], and targeted machine learning [139], also covered in a recent review [90]. Thus far, causal structure learning and policy evaluation have been studied as separate tasks: The output of causal structure learning is a causal graph, while the input to policy evaluation is a causal graph or selection diagram. Therefore, the current approach to using policy evaluation tasks when the graph is

unknown would be to first perform causal structure learning, then to use the methods discussed for policy evaluation. It is likely that this approach is not optimally sample-efficient—the two steps should be "aware" of each other, i.e., causal structure learning should be performed in a way that is targeted toward the downstream task.

The problem of targeted causal structure learning remains mostly unexplored, with a few notable exceptions. In the adaptive experimental design setting, [4] considers targeted learning of any property of the underlying graph, and [188] considers targeted learning of a "matching" intervention, which affects the system in some desired way. In the batched data setting, [12, 172, 189] consider targeted learning of the *difference* between two DAG models, instead of the DAG models themselves. All of these works demonstrate computational and statistical benefits to targeted learning over untargeted structure learning, indicating that this is an important and promising direction.

***Causal structure in reinforcement learning*** Policy evaluation is also an important task in reinforcement learning, where the policy is a *sequence* of actions that can depend on the state of the environment. The overlap between reinforcement learning and causality has been recently explored in the simple setting of *multi-armed bandits*, where an agent's actions do not affect the state of the environment. By assuming that actions correspond to interventions in a known causal graph, the effects of different actions become related, allowing for better regret bounds [102, 116]. If the causal graph is not assumed to be known, there is an additional exploration–exploitation trade-off that needs to be taken into account, which has been considered in recent work [18, 97, 108]. Since certain parts of the causal graph might not be relevant to predicting the effect of an action on some reward, the reinforcement learning setting is another case in which *targeted* structure learning may be more efficient.

**Acknowledgements** Chandler Squires was partially supported by an NSF Graduate Fellowship. Caroline Uhler was partially supported by NSF (DMS-1651995), ONR (N00014-17-1-2147 and N00014-22-1-2116), the MIT-IBM Watson AI Lab, MIT J-Clinic for Machine Learning and Health, the Eric and Wendy Schmidt Center at the Broad Institute, and a Simons Investigator Award.

**Funding** Open Access funding provided by the MIT Libraries



## References

1. Acharya, J., Bhattacharyya, A., Daskalakis, C., Kandasamy, S.: Learning and testing causal models with interventions. Advances in Neural Information Processing Systems **31** (2018)
2. Addanki, R., Kasiviswanathan, S., McGregor, A., Musco, C.: Efficient intervention design for causal discovery with latents. In: International Conference on Machine Learning, pp. 63–73. PMLR (2020)
3. Agrawal, R., Squires, C., Prasad, N., Uhler, C.: The DeCAMFounder: Non-linear causal discovery in the presence of hidden variables. arXiv preprint arXiv:2102.07921 (2021)

4. Agrawal, R., Squires, C., Yang, K., Shanmugam, K., Uhler, C.: ABCD-strategy: Budgeted experimental design for targeted causal structure discovery. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 3400–3409. PMLR (2019)
5. Agrawal, R., Uhler, C., Broderick, T.: Minimal I-MAP MCMC for scalable structure discovery in causal DAG models. In: International Conference on Machine Learning, pp. 89–98. PMLR (2018)
6. AhmadiTeshnizi, A., Salehkaleybar, S., Kiyavash, N.: Lazyiter: a fast algorithm for counting Markov equivalent DAGs and designing experiments. In: International Conference on Machine Learning, pp. 125–133. PMLR (2020)
7. Alonso-Barba, J.I., Gámez, J.A., Puerta, J.M., et al.: Scaling up the greedy equivalence search algorithm by constraining the search space of equivalence classes. International journal of approximate reasoning **54**(4), 429–451 (2013)
8. Andrews, B., Ramsey, J., Cooper, G.F.: Learning high-dimensional directed acyclic graphs with mixed data-types. In: The 2019 ACM SIGKDD Workshop on Causal Discovery, pp. 4–21. PMLR (2019)
9. Bareinboim, E., Pearl, J.: Transportability of causal effects: Completeness results. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 26, pp. 698–704 (2012)
10. Bareinboim, E., Pearl, J.: Transportability from multiple environments with limited experiments: Completeness results. Advances in neural information processing systems **27** (2014)
11. Bartlett, M., Cussens, J.: Integer linear programming for the Bayesian network structure learning problem. Artificial Intelligence **244**, 258–271 (2017)
12. Belyaeva, A., Squires, C., Uhler, C.: DCI: Learning causal differences between gene regulatory networks. Bioinformatics **btab167** (2021)
13. Bernstein, D., Saeed, B., Squires, C., Uhler, C.: Ordering-based causal structure learning in the presence of latent variables. In: International Conference on Artificial Intelligence and Statistics, pp. 4098–4108. PMLR (2020)
14. Bernstein, M., Tetali, P.: On sampling graphical Markov models. arXiv preprint arXiv:1705.09717 (2017)
15. Bhattacharya, R., Nabi, R., Shpitser, I.: Semiparametric inference for causal effects in graphical models with hidden variables. arXiv preprint arXiv:2003.12659 (2020)
16. Bhattacharya, R., Nagarajan, T., Malinsky, D., Shpitser, I.: Differentiable causal discovery under unmeasured confounding. In: International Conference on Artificial Intelligence and Statistics, pp. 2314–2322. PMLR (2021)
17. Bhattacharyya, A., Gayen, S., Kandasamy, S., Raval, V., Vinodchandran, N.: Efficient inference of interventional distributions. arXiv preprint arXiv:2107.11712 (2021)
18. Bilodeau, B., Wang, L., Roy, D.M.: Adaptively exploiting d-separators with causal bandits. arXiv preprint arXiv:2202.05100 (2022)
19. Bongers, S., Blom, T., Mooij, J.: Causal modeling of dynamical systems. arXiv preprint arXiv:1803.08784 (2018)
20. Bongers, S., Forré, P., Peters, J., Mooij, J.M.: Foundations of structural causal models with cycles and latent variables. The Annals of Statistics **49**(5), 2885–2915 (2021)
21. Brenner, E., Sontag, D.: SparsityBoost: a new scoring function for learning Bayesian network structure. In: Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, pp. 112–121 (2013)
22. Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., Drouin, A.: Differentiable causal discovery from interventional data. Advances in Neural Information Processing Systems **33**, 21865–21877 (2020)
23. Bühlmann, P., Peters, J., Ernest, J., et al.: CAM: Causal additive models, high-dimensional order search and penalized regression. Annals of statistics **42**(6), 2526–2556 (2014)
24. Cai, R., Qiao, J., Zhang, K., Zhang, Z., Hao, Z.: Causal discovery from discrete data using hidden compact representation. Advances in neural information processing systems **2018**, 2666 (2018)
25. Cai, R., Xie, F., Glymour, C., Hao, Z., Zhang, K.: Triad constraints for learning causal structure of latent variables. Advances in neural information processing systems **32** (2019)
26. Campbell, J.: An interventionist approach to causation in psychology. Causal learning: Psychology, philosophy and computation pp. 58–66 (2007)
27. Castelletti, F., Peluso, S.: Network structure learning under uncertain interventions. Journal of the American Statistical Association (just-accepted), 1–28 (2022)
28. Chen, R., Dash, S., Gao, T.: Integer programming for causal structure learning in the presence of latent variables. In: International Conference on Machine Learning, pp. 1550–1560. PMLR (2021)

29. Chickering, D.M.: A transformational characterization of equivalent Bayesian network structures. In: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, pp. 87–98 (1995)
30. Chickering, D.M.: Learning Bayesian networks is NP-complete. In: Learning from data, pp. 121–130. Springer (1996)
31. Chickering, D.M.: Optimal structure identification with greedy search. Journal of machine learning research **3**(Nov), 507–554 (2002)
32. Chickering, M.: Statistically efficient greedy equivalence search. In: Conference on Uncertainty in Artificial Intelligence, pp. 241–249. PMLR (2020)
33. Chobtham, K., Constantinou, A.C.: Bayesian network structure learning with causal effects in the presence of latent variables. In: International Conference on Probabilistic Graphical Models, pp. 101–112. PMLR (2020)
34. Claassen, T., Mooij, J.M., Heskes, T.: Learning sparse causal models is not NP-hard. In: Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, pp. 172–181 (2013)
35. Colombo, D., Maathuis, M.H., Kalisch, M., Richardson, T.S.: Learning high-dimensional directed acyclic graphs with latent and selection variables. The Annals of Statistics pp. 294–321 (2012)
36. Cooper, G.F., Yoo, C.: Causal discovery from a mixture of experimental and observational data. In: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, pp. 116–125 (1999)
37. Correa, J., Bareinboim, E.: General transportability of soft interventions: Completeness results. Advances in Neural Information Processing Systems **33**, 10902–10912 (2020)
38. Cussens, J.: Bayesian network learning with cutting planes. In: Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011), pp. 153–160. AUAI Press (2011)
39. Cussens, J.: GOBNILP: Learning Bayesian network structure with integer programming. In: M. Jaeger, T.D. Nielsen (eds.) Proceedings of the 10th International Conference on Probabilistic Graphical Models, *Proceedings of Machine Learning Research*, vol. 138, pp. 605–608. PMLR (2020). http://proceedings.mlr.press/v138/cussens20a.html
40. Daly, R., Shen, Q., Aitken, S.: Learning Bayesian networks: approaches and issues. The knowledge engineering review **26**(2), 99–157 (2011)
41. Drton, M., Maathuis, M.H.: Structure learning in graphical modeling. Annual Review of Statistics and Its Application **4**, 365–393 (2017)
42. Eaton, D., Murphy, K.: Exact Bayesian structure learning from uncertain interventions. In: Artificial intelligence and statistics, pp. 107–114. PMLR (2007)
43. Eberhardt, F., Glymour, C., Scheines, R.: On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In: Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence, pp. 178–184 (2005)
44. Eberhardt, F., Scheines, R.: Interventions and causal inference. Philosophy of science **74**(5), 981–995 (2007)
45. Ellis, B., Wong, W.H.: Learning causal Bayesian network structures from experimental data. Journal of the American Statistical Association **103**(482), 778–789 (2008)
46. Evans, R.J.: Graphs for margins of Bayesian networks. Scandinavian Journal of Statistics **43**(3), 625–648 (2016)
47. Evans, R.J.: Margins of discrete Bayesian networks. The Annals of Statistics **46**(6A), 2623–2656 (2018)
48. Forré, P., Mooij, J.M.: Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. arXiv preprint arXiv:1807.03024 (2018)
49. Friedman, N., Koller, D.: Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. Machine learning **50**(1), 95–125 (2003)
50. Friedman, N., Nachman, I.: Gaussian process networks. In: Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence, pp. 211–219 (2000)
51. Frot, B., Nandy, P., Maathuis, M.H.: Robust causal structure learning with some hidden variables. arXiv preprint arXiv:1708.01151 (2017)
52. Ganian, R., Hamm, T., Talvitie, T.: An efficient algorithm for counting Markov equivalent DAGs. Artificial Intelligence **304**, 103648 (2022)
53. Gao, M., Aragam, B.: Efficient Bayesian network structure learning via local Markov boundary search. Advances in Neural Information Processing Systems **34** (2021)
54. Gao, M., Tai, W.M., Aragam, B.: Optimal estimation of Gaussian DAG models. arXiv preprint arXiv:2201.10548 (2022)

55. Geiger, D., Heckerman, D., et al.: Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. The Annals of Statistics **30**(5), 1412–1440 (2002)
56. Ghassami, A., Salehkaleybar, S., Kiyavash, N., Bareinboim, E.: Budgeted experiment design for causal structure learning. In: International Conference on Machine Learning, pp. 1724–1733. PMLR (2018)
57. Ghassami, A., Salehkaleybar, S., Kiyavash, N., Zhang, K.: Counting and sampling from Markov equivalent DAGs using clique trees. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 3664–3671 (2019)
58. Ghassami, A., Yang, A., Kiyavash, N., Zhang, K.: Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs. In: International Conference on Machine Learning, pp. 3494–3504. PMLR (2020)
59. Ghoshal, A., Honorio, J.: Information-theoretic limits of Bayesian network structure learning. In: Artificial Intelligence and Statistics, pp. 767–775. PMLR (2017)
60. Ghoshal, A., Honorio, J.: Learning identifiable Gaussian Bayesian networks in polynomial time and sample complexity. Advances in Neural Information Processing Systems **30** (2017)
61. Gillispie, S.B., Lemieux, C.: Enumerating Markov Equivalence Classes of Acyclic Digraph Models. In: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, pp. 171–177 (2001)
62. Giudici, P., Castelo, R.: Improving Markov chain Monte Carlo model search for data mining. Machine learning **50**(1), 127–158 (2003)
63. Glocker, B., Musolesi, M., Richens, J., Uhler, C.: Causality in digital medicine. Nature Communications **12**(1) (2021)
64. Glymour, C., Zhang, K., Spirtes, P.: Review of causal discovery methods based on graphical models. Frontiers in genetics **10**, 524 (2019)
65. Gordon, S.L., Mazaheri, B., Rabani, Y., Schulman, L.J.: Identifying mixtures of Bayesian network distributions. arXiv preprint arXiv:2112.11602 (2021)
66. Greenewald, K., Katz, D., Shanmugam, K., Magliacane, S., Kocaoglu, M., Boix Adsera, E., Bresler, G.: Sample efficient active learning of causal trees. Advances in Neural Information Processing Systems **32** (2019)
67. Grzegorczyk, M., Husmeier, D.: Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. Machine Learning **71**(2-3), 265 (2008)
68. Halpern, Y., Horng, S., Sontag, D.: Anchored discrete factor analysis. arXiv preprint arXiv:1511.03299 (2015)
69. Harris, N., Drton, M.: PC algorithm for nonparanormal graphical models. Journal of Machine Learning Research **14**(11) (2013)
70. Hauser, A., Bühlmann, P.: Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. The Journal of Machine Learning Research **13**(1), 2409–2464 (2012)
71. Hauser, A., Bühlmann, P.: Two optimal strategies for active learning of causal models from interventional data. International Journal of Approximate Reasoning **55**(4), 926–939 (2014)
72. He, Y., Jia, J., Yu, B.: Counting and exploring sizes of Markov equivalence classes of directed acyclic graphs. The Journal of Machine Learning Research **16**(1), 2589–2609 (2015)
73. He, Y.B., Geng, Z.: Active learning of causal networks with intervention experiments and optimal designs. Journal of Machine Learning Research **9**(Nov), 2523–2547 (2008)
74. Heinze-Deml, C., Maathuis, M.H., Meinshausen, N.: Causal structure learning. Annual Review of Statistics and Its Application **5**, 371–391 (2018)
75. Heinze-Deml, C., Peters, J., Meinshausen, N.: Invariant causal prediction for nonlinear models. Journal of Causal Inference **6**(2) (2018)
76. Hoyer, P., Janzing, D., Mooij, J.M., Peters, J., Schölkopf, B.: Nonlinear causal discovery with additive noise models. Advances in neural information processing systems **21**, 689–696 (2008)
77. Hu, Z., Evans, R.: Faster algorithms for Markov equivalence. In: Conference on Uncertainty in Artificial Intelligence, pp. 739–748. PMLR (2020)
78. Hyttinen, A., Eberhardt, F., Hoyer, P.O.: Learning linear cyclic causal models with latent variables. The Journal of Machine Learning Research **13**(1), 3387–3439 (2012)
79. Hyttinen, A., Eberhardt, F., Hoyer, P.O.: Experiment selection for causal discovery. Journal of Machine Learning Research **14**, 3041–3071 (2013)
80. Hyttinen, A., Eberhardt, F., Järvisalo, M.: Constraint-based causal discovery: Conflict resolution with answer set programming. In: UAI, pp. 340–349 (2014)

81. Hyttinen, A., Hoyer, P.O., Eberhardt, F., Jarvisalo, M.: Discovering cyclic causal models with latent variables: A general SAT-based procedure. arXiv preprint arXiv:1309.6836 (2013)
82. Jaakkola, T., Sontag, D., Globerson, A., Meila, M.: Learning Bayesian network structure using LP relaxations. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 358–365. JMLR Workshop and Conference Proceedings (2010)
83. Jaber, A., Kocaoglu, M., Shanmugam, K., Bareinboim, E.: Causal discovery from soft interventions with unknown targets: Characterization and learning. Advances in neural information processing systems **33**, 9551–9561 (2020)
84. Jernite, Y., Halpern, Y., Sontag, D.: Discovering hidden variables in noisy-or networks using quartet tests. Advances in Neural Information Processing Systems **26** (2013)
85. Jung, Y., Tian, J., Bareinboim, E.: Estimating identifiable causal effects on Markov equivalence class through double machine learning. In: International Conference on Machine Learning, pp. 5168–5179. PMLR (2021)
86. Kalisch, M., Bühlman, P.: Estimating high-dimensional directed acyclic graphs with the PC-algorithm. Journal of Machine Learning Research **8**(3) (2007)
87. Kano, Y., Shimizu, S., et al.: Causal inference using nonnormality. In: Proceedings of the international symposium on science of modeling, the 30th anniversary of the information criterion, pp. 261–270 (2003)
88. Katz, D., Shanmugam, K., Squires, C., Uhler, C.: Size of interventional Markov equivalence classes in random DAG models. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 3234–3243. PMLR (2019)
89. Ke, N.R., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Schölkopf, B., Mozer, M.C., Pal, C., Bengio, Y.: Learning neural causal models from unknown interventions. arXiv preprint arXiv:1910.01075 (2019)
90. Kennedy, E.H.: Semiparametric doubly robust targeted double machine learning: a review. arXiv preprint arXiv:2203.06469 (2022)
91. Kivva, B., Rajendran, G., Ravikumar, P., Aragam, B.: Learning latent causal graphs via mixture oracles. Advances in Neural Information Processing Systems **34** (2021)
92. Kocaoglu, M., Dimakis, A., Vishwanath, S.: Cost-optimal learning of causal graphs. In: International Conference on Machine Learning, pp. 1875–1884. PMLR (2017)
93. Kocaoglu, M., Dimakis, A.G., Vishwanath, S., Hassibi, B.: Entropic causal inference. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
94. Kocaoglu, M., Shanmugam, K., Bareinboim, E.: Experimental design for learning causal graphs with latent variables. Advances in Neural Information Processing Systems **30** (2017)
95. Koivisto, M., Sood, K.: Exact Bayesian structure discovery in Bayesian networks. The Journal of Machine Learning Research **5**, 549–573 (2004)
96. Koster, J.T.: Marginalizing and conditioning in graphical models. Bernoulli pp. 817–840 (2002)
97. de Kroon, A.A., Belgrave, D., Mooij, J.M.: Causal discovery for causal bandits utilizing separating sets. arXiv preprint arXiv:2009.07916 (2020)
98. von Kügelgen, J., Rubenstein, P.K., Schölkopf, B., Weller, A.: Optimal experimental design via Bayesian optimization: active causal structure learning for Gaussian process networks. arXiv preprint arXiv:1910.03962 (2019)
99. Kuipers, J., Moffa, G.: Partition MCMC for inference on acyclic digraphs. Journal of the American Statistical Association **112**(517), 282–299 (2017)
100. Kummerfeld, E., Ramsey, J.: Causal clustering for 1-factor measurement models. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1655–1664 (2016)
101. Lachapelle, S., Brouillard, P., Deleu, T., Lacoste-Julien, S.: Gradient-Based Neural DAG Learning. In: International Conference on Learning Representations (2019)
102. Lattimore, F., Lattimore, T., Reid, M.D.: Causal bandits: Learning good interventions via causal inference. Advances in Neural Information Processing Systems **29** (2016)
103. Lauritzen, S., Sadeghi, K.: Unifying Markov properties for graphical models. The Annals of Statistics **46**(5), 2251–2278 (2018)
104. Lee, S., Correa, J.D., Bareinboim, E.: Generalized transportability: Synthesis of experiments from heterogeneous domains. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence (2020)

105. Lindgren, E., Kocaoglu, M., Dimakis, A.G., Vishwanath, S.: Experimental design for cost-aware learning of causal graphs. Advances in Neural Information Processing Systems **31** (2018)
106. Linusson, S., Restadh, P., Solus, L.: Greedy causal discovery is geometric. arXiv preprint arXiv:2103.03771 (2021)
107. Lorch, L., Rothfuss, J., Schölkopf, B., Krause, A.: DiBS: Differentiable Bayesian Structure Learning. Advances in Neural Information Processing Systems **34** (2021)
108. Lu, Y., Meisami, A., Tewari, A.: Causal bandits with unknown graph structure. Advances in Neural Information Processing Systems **34** (2021)
109. Maathuis, M., Drton, M., Lauritzen, S., Wainwright, M.: Handbook of graphical models. CRC Press (2018)
110. Madigan, D., York, J., Allard, D.: Bayesian graphical models for discrete data. International Statistical Review/Revue Internationale de Statistique pp. 215–232 (1995)
111. Meek, C.: Causal inference and causal explanation with background knowledge. In: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, pp. 403–410 (1995)
112. Meek, C.: Graphical models: Selecting causal and statistical models. Ph.D. thesis, PhD thesis, Carnegie Mellon University (1997)
113. Meinshausen, N., Hauser, A., Mooij, J.M., Peters, J., Versteeg, P., Bühlmann, P.: Methods for causal inference from gene perturbation experiments and validation. Proceedings of the National Academy of Sciences **113**(27), 7361–7368 (2016)
114. Mooij, J., Claassen, T., et al.: Constraint-based causal discovery with partial ancestral graphs in the presence of cycles. Proceedings of Machine Learning Research **124** (2020)
115. Mooij, J.M., Magliacane, S., Claassen, T.: Joint causal inference from multiple contexts (2020)
116. Nair, V., Patil, V., Sinha, G.: Budgeted and non-budgeted causal bandits. In: International Conference on Artificial Intelligence and Statistics, pp. 2017–2025. PMLR (2021)
117. Nandy, P., Hauser, A., Maathuis, M.H., et al.: High-dimensional consistency in score-based and hybrid structure learning. Annals of Statistics **46**(6A), 3151–3183 (2018)
118. Niinimäki, T., Parviainen, P., Koivisto, M.: Structure discovery in Bayesian networks by sampling partial orders. The Journal of Machine Learning Research **17**(1), 2002–2048 (2016)
119. Ogarrio, J.M., Spirtes, P., Ramsey, J.: A hybrid causal search algorithm for latent variable models. In: Conference on probabilistic graphical models, pp. 368–379. PMLR (2016)
120. Park, G., Park, H.: Identifiability of generalized hypergeometric distribution (ghd) directed acyclic graphical models. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 158–166. PMLR (2019)
121. Parviainen, P., Koivisto, M.: Bayesian structure discovery in Bayesian networks with less space. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 589–596. JMLR Workshop and Conference Proceedings (2010)
122. Pearl, J.: Causality. Cambridge university press (2009)
123. Peters, J., Bühlmann, P.: Identifiability of Gaussian structural equation models with equal error variances. Biometrika **101**(1), 219–228 (2014)
124. Peters, J., Janzing, D., Schölkopf, B.: Elements of causal inference: foundations and learning algorithms. The MIT Press (2017)
125. Porwal, V., Srivastava, P., Sinha, G.: Almost Optimal Universal Lower Bound for Learning Causal DAGs with Atomic Interventions. arXiv preprint arXiv:2111.05070 (2021)
126. Radhakrishnan, A., Solus, L., Uhler, C.: Counting Markov equivalence classes by number of immoralities. In: 33rd Conference on Uncertainty in Artificial Intelligence. AUAI Press Corvallis (2017)
127. Radhakrishnan, A., Solus, L., Uhler, C.: Counting Markov equivalence classes for DAG models on trees. Discrete Applied Mathematics **244**, 170–185 (2018)
128. Rajendran, G., Kivva, B., Gao, M., Aragam, B.: Structure learning in polynomial time: Greedy algorithms, Bregman information, and exponential families. Advances in Neural Information Processing Systems **34** (2021)
129. Ramsey, J., Spirtes, P., Zhang, J.: Adjacency-faithfulness and conservative causal inference. In: Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, pp. 401–408 (2006)
130. Rantanen, K., Hyttinen, A., Järvisalo, M.: Discovering causal graphs with cycles and latent confounders: an exact branch-and-bound approach. International Journal of Approximate Reasoning **117**, 29–49 (2020)

131. Raskutti, G., Uhler, C.: Learning directed acyclic graph models based on sparsest permutations. Stat **7**(1), e183 (2018)
132. Richardson, T., Spirtes, P., et al.: Ancestral graph Markov models. The Annals of Statistics **30**(4), 962–1030 (2002)
133. Richardson, T.S., Evans, R.J., Robins, J.M., Shpitser, I.: Nested Markov properties for acyclic directed mixed graphs. arXiv preprint arXiv:1701.06686 (2017)
134. Rothenhäusler, D., Heinze, C., Peters, J., Meinshausen, N.: Backshift: Learning causal cyclic graphs from unknown shift interventions. Advances in Neural Information Processing Systems **28** (2015)
135. Rotnitzky, A., Smucler, E.: Efficient adjustment sets for population average treatment effect estimation in non-parametric causal graphical models. arXiv preprint arXiv:1912.00306 (2019)
136. Saeed, B., Belyaeva, A., Wang, Y., Uhler, C.: Anchored causal inference in the presence of measurement error. In: Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence (2020)
137. Saeed, B., Panigrahi, S., Uhler, C.: Causal structure discovery from distributions arising from mixtures of DAGs. In: International Conference on Machine Learning, pp. 8336–8345. PMLR (2020)
138. Schmidt, M., Niculescu-Mizil, A., Murphy, K., et al.: Learning graphical model structure using l1-regularization paths. In: AAAI, vol. 7, pp. 1278–1283 (2007)
139. Schuler, M.S., Rose, S.: Targeted maximum likelihood estimation for causal inference in observational studies. American journal of epidemiology **185**(1), 65–73 (2017)
140. Schulte, O., Frigo, G., Greiner, R., Khosravi, H.: The IMAP hybrid method for learning Gaussian Bayes nets. In: Canadian Conference on Artificial Intelligence, pp. 123–134. Springer (2010)
141. Shah, R.D., Frot, B., Thanei, G.A., Meinshausen, N.: Right singular vector projection graphs: fast high dimensional covariance matrix estimation under latent confounding. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **82**(2), 361–389 (2020)
142. Shah, R.D., Peters, J.: The hardness of conditional independence testing and the generalised covariance measure. The Annals of Statistics **48**(3), 1514–1538 (2020)
143. Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A., Jordan, M.: A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research **7**(10) (2006)
144. Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P.O., Bollen, K.: DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. The Journal of Machine Learning Research **12**, 1225–1248 (2011)
145. Shpitser, I., Evans, R.J., Richardson, T.S., Robins, J.M.: Introduction to nested Markov models. Behaviormetrika **41**(1), 3–39 (2014)
146. Shpitser, I., Pearl, J.: Identification of joint interventional distributions in recursive semi-Markovian causal models. In: proceedings of the 21st national conference on Artificial intelligence-Volume 2, pp. 1219–1226 (2006)
147. Shpitser, I., Wood-Doughty, Z., Tchetgen, E.J.T.: The proximal ID algorithm. arXiv preprint arXiv preprint arXiv:2108.06818 (2021)
148. Solus, L., Wang, Y., Uhler, C.: Consistency Guarantees for Greedy Permutation-Based Causal Inference Algorithms. Biometrika (2021). https://doi.org/10.1093/biomet/asaa104. Asaa104
149. Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D.: Causation, prediction, and search. MIT press (2000)
150. Spirtes, P., Meek, C., Richardson, T.: Causal inference in the presence of latent variables and selection bias. In: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, pp. 499–506 (1995)
151. Spirtes, P., Richardson, T.: A polynomial time algorithm for determining DAG equivalence in the presence of latent variables and selection bias. In: Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics, pp. 489–500 (1996)
152. Squires, C., Magliacane, S., Greenewald, K., Katz, D., Kocaoglu, M., Shanmugam, K.: Active structure learning of causal DAGs via directed clique trees. Advances in Neural Information Processing Systems **33**, 21500–21511 (2020)
153. Squires, C., Wang, Y., Uhler, C.: Permutation-based causal structure learning with unknown intervention targets. In: Conference on Uncertainty in Artificial Intelligence, pp. 1039–1048. PMLR (2020)
154. Squires, C., Yun, A., Nichani, E., Agrawal, R., Uhler, C.: Causal structure discovery between clusters of nodes induced by latent factors. In: First Conference on Causal Learning and Reasoning (2022)
155. Strobl, E.V.: A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias. International Journal of Data Science and Analytics **8**(1), 33–56 (2019)

156. Strobl, E.V.: Improved causal discovery from longitudinal data using a mixture of DAGs. In: The 2019 ACM SIGKDD Workshop on Causal Discovery, pp. 100–133. PMLR (2019)
157. Strobl, E.V.: The global Markov property for a mixture of DAGs. arXiv preprint arXiv:1909.05418 (2019)
158. Strobl, E.V., Zhang, K., Visweswaran, S.: Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. Journal of Causal Inference **7**(1) (2019)
159. Studeny, M.: Probabilistic conditional independence structures. Springer Science & Business Media (2006)
160. Studenỳ, M., Hemmecke, R., Lindner, S.: Characteristic imset: a simple algebraic representative of a Bayesian network structure. In: Proceedings of the 5th European workshop on probabilistic graphical models, pp. 257–264. HIIT Publications (2010)
161. Sussex, S., Uhler, C., Krause, A.: Near-optimal multi-perturbation experimental design for causal structure learning. Advances in Neural Information Processing Systems **34** (2021)
162. Talvitie, T., Koivisto, M.: Counting and sampling Markov equivalent directed acyclic graphs. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 7984–7991 (2019)
163. Tian, J., Pearl, J.: Causal discovery from changes. In: Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence, pp. 512–521 (2001)
164. Tigas, P., Annadani, Y., Jesson, A., Schölkopf, B., Gal, Y., Bauer, S.: Interventions, where and how? experimental design for causal models at scale. arXiv preprint arXiv:2203.02016 (2022)
165. Triantafillou, S., Tsamardinos, I.: Constraint-based causal discovery from multiple interventions over overlapping variable sets. The Journal of Machine Learning Research **16**(1), 2147–2205 (2015)
166. Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing Bayesian network structure learning algorithm. Machine learning **65**(1), 31–78 (2006)
167. Tsirlis, K., Lagani, V., Triantafillou, S., Tsamardinos, I.: On scoring maximal ancestral graphs with the max–min hill climbing algorithm. International Journal of Approximate Reasoning **102**, 74–85 (2018)
168. Uhler, C., Raskutti, G., Bühlmann, P., Yu, B.: Geometry of the faithfulness assumption in causal inference. The Annals of Statistics pp. 436–463 (2013)
169. Verma, T., Pearl, J.: Causal networks: Semantics and expressiveness. In: Machine intelligence and pattern recognition, vol. 9, pp. 69–76. Elsevier (1990)
170. Verma, T., Pearl, J.: Equivalence and synthesis of causal models. In: Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, pp. 255–270 (1990)
171. Wang, Y., Solus, L., Yang, K., Uhler, C.: Permutation-based causal inference algorithms with interventions. In: 31st Annual Conference on Neural Information Processing Systems, NIPS 2017, Long Beach, United States, 4 December 2017 through 9 December 2017, vol. 2017, pp. 5823–5832. Neural information processing systems foundation (2017)
172. Wang, Y., Squires, C., Belyaeva, A., Uhler, C.: Direct estimation of differences in causal graphs. Advances in neural information processing systems **31** (2018)
173. Wang, Y.S., Drton, M.: High-dimensional causal discovery under non-Gaussianity. Biometrika **107**(1), 41–59 (2020)
174. Wermuth, N.: Probability distributions with summary graph structure. Bernoulli **17**(3), 845–879 (2011)
175. Wienöbst, M., Bannach, M., Liskiewicz, M.: Polynomial-time algorithms for counting and sampling Markov equivalent DAGs. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 12198–12206 (2021)
176. Xie, F., Cai, R., Huang, B., Glymour, C., Hao, Z., Zhang, K.: Generalized independent noise condition for estimating latent variable causal graphs. Advances in Neural Information Processing Systems **33**, 14891–14902 (2020)
177. Yang, K., Katcoff, A., Uhler, C.: Characterizing and learning equivalence classes of causal DAGs under interventions. In: International Conference on Machine Learning, pp. 5541–5550. PMLR (2018)
178. Yu, Y., Chen, J., Gao, T., Yu, M.: DAG-GNN: DAG structure learning with graph neural networks. In: International Conference on Machine Learning, pp. 7154–7163. PMLR (2019)
179. Yuan, C., Malone, B.: Learning optimal Bayesian networks: A shortest path perspective. Journal of Artificial Intelligence Research **48**, 23–65 (2013)
180. Zhalama, Zhang, J., Eberhardt, F., Mayer, W.: SAT-Based Causal Discovery under Weaker Assumptions. In: UAI (2017)

181. Zhang, J., Spirtes, P.: Strong faithfulness and uniform consistency in causal inference. In: Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence, pp. 632–639 (2002)
182. Zhang, J., Spirtes, P.: A transformational characterization of Markov equivalence for directed acyclic graphs with latent variables. In: Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence, pp. 667–674 (2005)
183. Zhang, J., Spirtes, P.: The three faces of faithfulness. Synthese **193**(4), 1011–1027 (2016)
184. Zhang, K., Gong, M., Ramsey, J., Batmanghelich, K., Spirtes, P., Glymour, C.: Causal discovery in the presence of measurement error: Identifiability conditions. arXiv preprint arXiv:1706.03768 (2017)
185. Zhang, K., Hyvärinen, A.: On the identifiability of the post-nonlinear causal model. In: 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009), pp. 647–655. AUAI Press (2009)
186. Zhang, K., Peters, J., Janzing, D., Schölkopf, B.: Kernel-based conditional independence test and application in causal discovery. In: Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, pp. 804–813 (2011)
187. Zhang, M., Jiang, S., Cui, Z., Garnett, R., Chen, Y.: D-VAE: A variational autoencoder for directed acyclic graphs. arXiv preprint 1904.11088 (2019)
188. Zhang, V., Squires, C., Uhler, C.: Matching a desired causal state via shift interventions. Advances in Neural Information Processing Systems **34** (2021)
189. Zhao, B., Wang, Y.S., Kolar, M.: Direct estimation of differential functional graphical models. Advances in Neural Information Processing Systems **32** (2019)
190. Zheng, X., Aragam, B., Ravikumar, P.K., Xing, E.P.: DAGs with NO TEARS: Continuous Optimization for Structure Learning. Advances in Neural Information Processing Systems **31** (2018)