D'ya Like DAGs? A Survey on Structure Learning and Causal Discovery

MATTHEW J. VOWELS, NECATI CIHAN CAMGOZ, and RICHARD BOWDEN,
CVSSP, University of Surrey, U.K.

Causal reasoning is a crucial part of science and human intelligence. In order to discover causal relationships from data, we need structure discovery methods. We provide a review of background theory and a survey of methods for structure discovery. We primarily focus on modern, continuous optimization methods, and provide reference to further resources such as benchmark datasets and software packages. Finally, we discuss the assumptive leap required to take us from structure to causality.

CCS Concepts: • **Mathematics of computing** → **Causal networks**; • **Computing methodologies** → **Machine learning**; **Causal reasoning and diagnostics;**

Additional Key Words and Phrases: Causality, causal discovery, directed acyclic graphs, DAGs, structure learning, survey

ACM Reference format:
Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. 2022. D'ya Like DAGs? A Survey on Structure Learning and Causal Discovery. *ACM Comput. Surv.* 55, 4, Article 82 (November 2022), 36 pages.
https://doi.org/10.1145/3527154

1 INTRODUCTION

Causal understanding has been described as 'part of the bedrock of intelligence' [145], and is one of the fundamental goals of science [10, 69, 184, 248–250]. It is important for a broad range of applications, including policy making [136], medical imaging [30], advertisement [22], the development of medical treatments [191], the evaluation of evidence within legal frameworks [184, 223], social science [81, 94, 254], biology [242], and many others. It is also a burgeoning topic in machine learning and artificial intelligence [6, 16, 65, 75, 144, 214, 255, 265], where it has been argued that a consideration for causality is crucial for reasoning about the world. In order to discover causal relations, and thereby gain causal understanding, one may perform interventions and manipulations as part of a randomized experiment. These experiments allow researchers or agents to identify causal relationships, but also to estimate the magnitude of these relationships.

Unfortunately, in many cases, it may not be possible to undertake such experiments due to prohibitive cost, ethical concerns, or impracticality. For example, to understand the impact of smoking, it would be necessary to force different individuals to smoke or not-smoke. Researchers are therefore often left with non-experimental, observational data. In the absence of intervention and

Authors' address: M. J. Vowels (corresponding author), N. C. Camgoz, and R. Bowden, Centre for Vision, Speech and Signal Processing 388 Stag Hill, University of Surrey, Guildford, GU2 7XH, United Kingdom; emails: {m.j.vowels, n.camgoz, r.bowden}@surrey.ac.uk.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
0360-0300/2022/11-ART82 $15.00
https://doi.org/10.1145/3527154

manipulation, observational data leave researchers facing a number of challenges: Firstly, observational datasets may not contain all relevant variables - there may exist unobserved/hidden/latent factors (this is sometimes referred to as the third variable problem). Secondly, observational data may exhibit selection bias - for example, younger patients may in general prefer to opt for surgery, whereas older patients may prefer medication. Thirdly, the causal relationships underlying these data may not be known *a priori* - for example, are genetic factors independent causes of a particular outcome, or do they mediate or moderate an outcome? These three challenges affect the discovery and estimation of causal relationships.

To address these challenges, researchers in the fields of statistics and machine learning have developed numerous methods for uncovering causal relations (causal discovery) and estimating the magnitude of these effects (causal inference) from observational data, or from a mixture of observational and experimental data. Under various (often strong) assumptions, these methods are able to take advantage of the relative abundance of observational data in order to infer causal structure and causal effects. Indeed, observational data may, in spite of the three challenges above, provide improved statistical power and generalizability compared with experimental data [44].

In this paper we review relevant background theory and provide a survey of methods which perform structure discovery (sometimes called causal induction [80]) with observational data or with a mixture of observational and experimental data. We split structure discovery algorithms into two principal groups. We only briefly discuss combinatoric/search based algorithms (Section 4), and instead focus on continuous optimization based algorithms (Section 5). In both cases, we focus on the static, non-dynamic causal discovery setting, although we will briefly discuss the dynamic, time series setting where relevant. A number of reviews, surveys and guides are already available (see e.g., [69, 92, 230]), however, these reviews cover combinatoric approaches to causal discovery, hence why we primarily focus on the recent flurry of developments in continuous optimization approaches. Furthermore, the existing reviews are relatively short, and we attempt to provide a more scoping introduction to the necessary background material. Finally, we provide references to further useful resources including datasets and openly available software packages.

The structure of this survey is as follows: Following an overview of relevant background information in Section 2, we provide an overview of approaches to structure discovery in Section 3, including a list of common evaluation metrics. In Section 4 we briefly outline a range of combinatoric approaches before focusing on continuous optimization approaches in Section 5. We begin Section 6 by referencing several additional resources including reviews, guides, datasets, and software packages. We also provide a summary and discussion of the methods covered in this section, and note various opportunities for future work and future direction. Many of the methods we review in this survey seek to discover and interpret the learned structure *causally*. Whilst this is a laudable aim, we are reminded of important commentaries (e.g., [41, 57, 105]) which argue for appropriate skepticism and care when making the leap from observation to causality via causal discovery methods. We therefore conclude Section 6, and this survey as a whole, by providing a discussion on these issues.

2 BACKGROUND - DEFINITIONS AND ASSUMPTIONS

In this section we provide working definitions of key concepts in structure discovery. We include a presentation of a common framework used in structure discovery (namely, that of structured graphical representations) as well as a number of common assumptions.

2.1 Causality and SCMs

In spite of some notable reluctance to treat graphs learned from observational data as causal [41, 57, 105], we acknowledge that it is a common and worthwhile aim, and begin by presenting

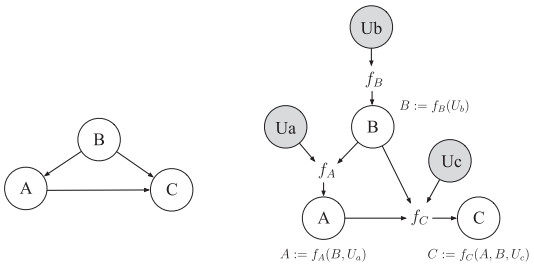

Fig. 1. Transitioning from a typical DAG representation (left) to a structural equation model (right). Grey vertices are unobserved/latent random variables.

a working definition of causality and its popular systematization in **Structural Causal Models (SCMs)**. Causality eludes straightforward definition [228], and is often characterized intuitively with examples involving fires and houses [184], firing squads [98], and bottles and rocks [99]. One definition of what is known as *counterfactual causality* is given by Lewis [143] as follows:[1]

> "We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it. Had it been absent, its effects – some of them, at least, and usually all – would have been absent as well".

Lewis' definition is counterfactual in the sense that he effectively describes 'what would have happened if the cause had been A*, given that the effect was B when the cause was A'. Seemingly, this definition is compatible with the 'Pearlian' school of causal reasoning. Specifically, in the context of what are known as SCMs:

> "Given two disjoint sets of variables X and Y, the causal effect of X on Y, denoted as... $P(y|do(x))$, is a function from X to the space of probability distributions on Y. For each realization of x of X, $P(y|do(x))$ gives the probability of $Y = y$ induced by deleting from the model $[x_i = f_i(pa_i, u_i), i = 1, \ldots, n,]$ all equations corresponding to variables in X and substituting $X = x$ in the remaining equations."

This definition [184, p. 70] requires further examination. Firstly, the model $x_i = f_i(pa_i, u_i), i = 1, \ldots, n$, is a **Structural equation/Causal Model (SEM/SCM)** which indicates assignment of the value x_i in the space of X to a function of its structural parents pa_i and exogenous noise u_i. We elaborate on what parents are (as well as children, descendants, etc.) below. Secondly, the *do* notation [184] indicates *intervention*, where the value of x is set to a specific quantity. The structure (including attributes such as *parents*) can be represented graphically using various types of graphical models (e.g., Directed Acyclic Graphs). Figure 1 shows the relationship between a DAG and a general Structural equation Model. Sometimes this SEM is also called a **Functional Causal Model (FCM)**, where the functions are assumed to represent the causal mechanisms [74]. The use of the assignment operator ':=' makes explicit the asymmetric nature of these equations. In other words, they are not to be rearranged to solve for their inputs. To transform these relationships from mathematical relationships to causal relations, the Causal Markov Condition is imposed, which assumes that the exogenous variables U are mutually independent, and that the arrows represent causal dependencies which therefore entail a (Markovian) conditional independency structure [190, p. 105-6].

The ultimate benefit of the graphical and structural model frameworks is that they, at least in principle and under some strong assumptions, enable us to use observational data to answer scientific questions such as 'how?', 'why?', and 'what if?' [186].

[1] See discussion in Menzies & Beebee [160].

2.2 Graphical Models

For background on graphical models, see work by Koller and Friedman [133]. We follow a similar formalism to Peters et al. [190] and Strobl [231]. A graph $\mathcal{G}(\mathbf{X}, \mathcal{E})$ represents a joint distribution $P_{\mathbf{X}}$ as a factorization of d variables $\mathbf{X} = \{X_1, \ldots, X_d\}$ using d corresponding *nodes/vertices* $v \in \mathbf{V}$ and connecting edges $(i, j) \in \mathcal{E}$, where (i, j) indicates an edge between v_i and v_j. If two vertices i and j are connected by an edge we call them *adjacent*, and, can also denote this in terms of the corresponding variables $\mathbf{X}$ as $X_i \to X_j$ or $X_i \leftarrow X_j$ (directed), $X_i - X_j$ (undirected), $X_i \leftrightarrow X_j$ (bidirected), $X_i \multimap X_j$ or $X_i \circ\!\!- X_j$ (partially undirected), $X_i \circ\!\!\to X_j$ or $X_i \leftarrow\!\!\circ X_j$ (partially directed), or $X_i \circ\!\!-\!\!\circ X_j$ (nondirected). A graph comprising entirely undirected edges forms a *skeleton*. It is also possible to have self-loops, although these occur relatively infrequently in the structure discovery literature. These different edge types allow us to define a range of graph types and relationships.

An *undirected path* exists if there are edges connecting two vertices regardless of the edge types between them. In contrast, a *directed path* constitutes directed edges with consistent arrowhead directions. We can define a *parent* pa_j as a vertex v_i with *child* v_j connected by a directed edge $X_i \to X_j$ such that $(i, j) \in \mathcal{E}$ but $(j, i) \notin \mathcal{E}$. Further upstream parents are *ancestors* of downstream *descendants* if there exists a directed path constituting $i_k \to j_{k+1}$ for all k in a sequence of vertices. An *immorality* or *v-structure* describes when two non-adjacent vertices are parents of a common child. A *collider* is a vertex where incoming directed arrows converge.

It is possible for *directed cycles* to occur when following a directed path results in the visitation of a vertex more than once (e.g., $X_i \to X_j \to X_k \to X_i$). Many phenomena in nature exhibit cyclic properties and feedback, and ignoring this possibility has the potential to induce bias [68, 210, 231]. However, in such cases it is important to delineate between static and dynamic (time series) settings. Assuming that the future cannot cause the past, a cyclic (dynamic) graph can encode that a previous node influence itself at a future timepoint. We briefly discuss time in Section 3.4 below, but note that causal discovery in time-based settings is not the principal focus of this survey. If all edges are directed, and there are no cycles, we have the well-known class of ***Directed Acyclic Graphs* (DAGs)**. On the other hand, if all edges are directed but there is no restriction preventing cycles, we have a ***Directed Graph* (DG)**.

2.3 The Markov Assumption, d-Separation, and d-Faithfulness

The graphs are usually assumed to fulfil the Markov property, such that the implied joint distribution factorizes according to the following recursive decomposition, characteristic of Bayesian networks [184]:

$$P(\mathbf{X}) = \prod_{i}^{d} P(X_i | pa_i) \tag{1}$$

This decomposition relates to the notion of d-separation. Two vertices X_i and X_k are d-separated by the set of vertices $\mathbf{S}$ if $X_j \in \mathbf{S}$ in any of the following structural scenarios [190]:

$$\begin{aligned} X_i \to X_j \to X_k \\ X_i \leftarrow X_j \leftarrow X_k \\ X_i \leftarrow X_j \to X_k \end{aligned} \tag{2}$$

They are also d-separated if neither X_j nor any of the descendants of X_j are in set $\mathbf{S}$ in the following structural scenario (collider):

$$X_i \to X_j \leftarrow X_k \tag{3}$$

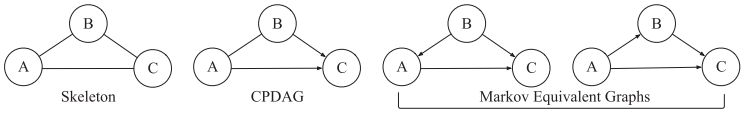

Fig. 2. Showing a skeleton, a CPDAG, and the Markov Equivalence set of graphs. Variable C is a collider, and so the direction of incoming arrows can be identified from conditional independencies.

If the DAG's d-separation properties hold (an assumpion of faithfulness - see below), they imply Markovian conditional independencies in the joint distribution, which can be denoted as $X_i \perp\!\!\!\perp_{P_X} X_k | X_j$. In terms of the DAG, disjoint (i.e., non-overlapping) sets of variables **A** and **B** are d-separated by disjoint set of variables **S** in graph $\mathcal{G}$ if $\mathbf{A} \perp\!\!\!\perp_{d-sep} \mathbf{B}|\mathbf{S}$ [190], and are, conversely d-connected if this conditional independence in the graph does not hold.

The assumption of d-faithfulness is that any conditional independencies in the joint distribution $P_{\mathbf{X}}$ are implied by the graph, according to its d-separation properties. More formally [190, p. 107], for joint distribution $P_{\mathbf{X}}$ and DAG $\mathcal{G}$, the assumption of d-faithfulness holds if $\mathbf{A} \perp\!\!\!\perp_{P_X} \mathbf{B}|\mathbf{C} \implies \mathbf{A} \perp\!\!\!\perp_{d-sep} \mathbf{B}|\mathbf{C}$. One example of a violation of d-faithfulness occurs when the influence of two paths cancel each other out, resulting in a DAG with different implied conditional independencies to those present in the joint distribution.

2.4 Markov Equivalence Class (MEC) and Completed Partially Directed Acyclic Graphs (CPDAGs)

The conditional independence constraints implied by a graph's d-separation properties are not always enough to uniquely identify it. Whether a graph can be uniquely identified is known as the problem of *identifiability*, and a significant body of work has been devoted to identifying scenarios for which the true graph is identifiable (e.g., linear functional form with non-Gaussian errors [101], or nonlinear functional forms with additive noise [100]).[2] As such, there are situations in which multiple graphs satisfy the same conditional independencies. For example, conditional independence implied by $X_i \perp\!\!\!\perp X_k | X_j$ is present in the graph $X_i \rightarrow X_j \rightarrow X_k$ as well as the graphs $X_i \leftarrow X_j \leftarrow X_k$ and $X_i \leftarrow X_j \rightarrow X_k$, in spite of the fact that these graphs have drastically different causal implications. The class of graphs which represent the same set of conditional independencies together constitute the ***Markov Equivalence Class*** **(MEC)**. Graphs belong to the same equivalence class when they have the same skeleton and the same immoralities [252].

Completed Partially Directed Acyclic Graphs **(CPDAGs)** can be used to represent an MEC. In CPDAGs, an edge is only directed if all graphs in the MEC contain the edge in that direction. Otherwise, if there is uncertainty about the direction, it is left 'non-directed' using $\circ\!\!-\!\!\circ$. One might wonder whether there are any MECs without undirected edges, and indeed there are. A collider or v-structure forms an MEC with only one valid DAG: $X_i \rightarrow X_j \leftarrow X_k$. This is because conditioning on X_j makes X_i and X_k d-connected. An example of a skeleton graph, a CPDAG, and corresponding MEC graphs are shown in Figure 2.

2.5 Assumption: Sufficiency

One of the challenges with using observational data is the assumption that all relevant data have been collected/observed. This is less problematic in the case of **Randomized Controlled Trials (RCTs)** because the randomization itself helps mitigate the effect of confounding which would

[2]One may define an SEM defined on a DAG as identifiable if there are no other SEMs that induce the same joint distribution with a different DAG [176].

otherwise imbalance the treatment and control groups.[3] In observational settings, unobserved confounding can significantly bias effect estimates (even reversing their direction). Whilst it is possible to try to infer hidden confounders from observational data using latent variable models (see e.g., [150, 255, 257, 264], a large number of causal discovery methods assume *sufficiency*, which is the assumption that there are no unobserved confounders. The assumption of sufficiency is strong and may often be inappropriate or overly restrictive. If the assumption does not hold, the set of observed variables is (causally) *insufficient* [19] and a DAG comprising only the observed variables cannot be used (and the DAG is said to not be closed under marginalization) [102].

2.6 Acyclic Directed Mixed Graphs (ADMGs) and Maximal Ancestral Graphs (MAGs) and *m*-separation

In the presence of unobserved confounding, an ***Acyclic Directed Mixed Graph* (ADMG)** may be used. ADMGs represent hidden confounding as bidirected edges. For example, the confounding relationship given by $X_i \leftarrow H \rightarrow X_j \rightarrow X_k$ can, in the absence of H, be represented in an ADMG as $X_i \leftrightarrow X_j \rightarrow X_k$.

***Maximal Ancestral Graphs* (MAGs)** can also be used to represent hidden confounding, and have the further capacity of representing selection bias (*i.e.,* as might occur when a certain subpopulation is sampled). MAGs satisfy the following three properties [2, 203, 204]: (1) there are no directed cycles (acyclicity); (2) if an edge $X_i \leftrightarrow X_j$ exists (which implies X_i is the *spouse* of X_j), then there are no directed paths between X_i and X_j; (3) if an edge $X_i - X_j$ exists (which implies X_i is the *neighbour* X_j), then X_i and X_j have no spouses or parents. This edge is used to represent selection bias (*i.e.,* where a subpopulation has been sampled according to some condition).

The definitions of ancestor and descendent translate naturally from DAGs (see above) to MAGs, as does the definition for *d*-separation, which becomes *m-separation*. In the latter case, the conditions for *d*-separation in Equations (2) and (3) hold, substituting any confounding variable relationships (e.g., $X_i \leftarrow H \rightarrow X_j$) with a bidirected arrow (e.g., $X_i \leftrightarrow X_j$). The graph is then *maximal* if for any pair of non-adjacent nodes X_i and X_j, there exists a set of nodes $\mathbf{S}$ such that $X_i, X_j \notin \mathbf{S}$ whereby X_i and X_j are *m*-separated by $\mathbf{S}$. In the presence of selection bias, a $X_i - X_j$ edge can be used. Readers are directed to [2, 203, 204] for a more detailed and formal exposition.

The assumption of *m*-faithfulness also translates naturally from *d*-faithfulness for DAGs (see above) to MAGs, according to the conditional independencies implied by *m*-separation.

2.7 Partial Ancestral Graphs (PAGs)

Similarly to how the MEC of a set of DAGs was represented using a CPDAG, the MEC for a set of MAGs can be representing using a ***Partial Ancestral Graph* (PAG)**. PAGs make use of edges $X_i \circ\!\!-\!\!\circ X_j$, $X_i \circ\!\!\rightarrow X_j$, and $X_i -\!\!\circ X_j$. Edges with arrowheads indicate that arrowheads are present in *all* MAGs in the associated MEC. A tail (i.e., an edge without either a circle mark or an arrowhead) indicates that the tail is present in all MAGs in the associated MEC. Circle marks (as with CPDAGs) indicate uncertainty in the edge mark, such that the MEC contains MAGs in which the edge mark is either a tail or an arrowhead. [92, 190]. An example of a DAG and its equivalent MAG and PAG are shown in Figure 3.

2.8 Other Definitions and Assumptions

Other types of graph used to represent causal structure include **Partially Oriented Induced Path Graphs (POIPGs)** [190, 228], **Single World Intervention Graphs (SWIGs)** [24, 201, 202],

[3] In reality, limited sample sizes (which are often encountered with expensive RCTs) can still render this issue problematic [44].

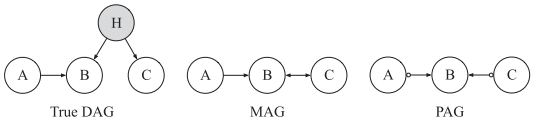

Fig. 3. Showing the relationship between the true DAG and its representation using a MAG and a PAG. Shaded vertex is a hidden/unobserved confounding variable. Adapted from [190, p. 179].

σ-connection graphs [56], undirected graphs [11], interaction and component graphs for dynamic systems [40], **Maximal Almost Ancestral Graphs (MAAGs)** [231], psi-ECs [110], Patterns [274], and arid, bow-free, and ancestral ADMGs [19]. There are also other types of assumptions relating to the functional form of the structural relationships (e.g., linear or non-linear) as well as the parametric form of the marginals and the errors (e.g., Gaussian or non-Gaussian). In the interests of brevity, we have not discussed these additional graph-types and assumptions here, but encourage interested readers to consult the listed references.

3 STRUCTURE DISCOVERY METHODS

We consider four approaches to structure discovery: constraint-based, score-based, those exploiting structural asymmetries, and those exploiting various forms of intervention.[4] We begin by introducing these four approaches. Each structure discovery method may be sub-categorized into those which seek to identify a graphical structure via combinatoric/search-based algorithms, or those which seek to identify a graphical structure via continuous optimization. Previous reviews exist for the former (e.g., [69, 92, 230]), so we primarily focus on the latter. Finally, methods may be categorized as *local*, whereby edges are tested one at a time, or *global*, whereby an entire graph candidate is tested. It can be assumed that all methods are concerned with learning from independent and identically (i.i.d) distributed data, except in the time-series case where it is each time series which is sampled i.i.d.

3.1 Constraint-Based and Score-Based Approaches

Most constraint-based approaches test for conditional independencies in the empirical joint distribution in order to construct a graph that reflects these conditional independencies.[5] According to the discussion above, there are often multiple graphs that fulfil a given set of conditional independencies, and so it is common for constraint-based approaches to output a graph representing some MEC (e.g., a PAG). Conditional independence testing represents a significant subfield in its own right, and presents many challenges. Indeed, conditional independence tests can require large sample sizes to be reliable, and Shah and Peters [217] discuss further challenges relating to the control of both Type I and Type II error rates. Examples of flexible conditional independence testing include GAN-based [15, 220], gradient boosting and neural network classifier based [216], and Kernel based [62, 278] methods.

Even though the focus of this review is continuously optimized methods, for pedagogical purposes it is worth briefly describing how one might use (e.g.,) conditional independence constraints to learn a graph. One of the most well-known algorithms is the PC algorithm [228]. The algorithm starts with a complete and undirected graph, and begins by removing edges in order to identify the skeleton. To do this, conditional independence tests are used to evaluate $X_i \perp\!\!\!\perp X_j | \mathbf{S}$, where the conditioning set is potentially empty (thereby reducing to a test for bivariate statistical

[4]There are also hybrid approaches which incorporate some combination of these classes, but we do not treat these separately.

[5]Other constraints exist, such as Verma constraints [252, 274].

independence). From here, the algorithm begins orienting/directing the edges by leveraging the fact that if a path $X_i - X_k - X_j$ exists in the skeleton such that $X_i \perp\!\!\!\perp X_j | X_k$ but $X_i \not\perp\!\!\!\perp X_j | X_k$, then we know that this path must actually be a v-structure, and the edges can be oriented as $X_i \rightarrow X_k \leftarrow X_j$. Additionally, now that v-structures have been identified and their corresponding edges have been directed, the algorithm can orient additional edges in partially directed paths which would otherwise form additional v-structures.

In contrast, score-based approaches test the validity of a candidate graph $\mathcal{G}$ according to some scoring function S. The goal is therefore stated as [190]:

$$\hat{\mathcal{G}} = \text{argmax}_{\mathcal{G}} \text{ over} \mathbf{X} S(\mathcal{D}, \mathcal{G}) \tag{4}$$

where $\mathcal{D}$ represents the empirical data for variables $\mathbf{X}$. Common scoring functions include the **Bayesian Information Criterion (BIC)** [66], the Minimum Description Length (as an approximation of Kolmogorov Complexity) [82, 114, 121], the **Bayesian Gaussian equivalent (BGe)** score [66], the **Bayesian Dirichlet equivalence (BDe)** score [91], the **Bayesian Dirichlet equivalence uniform (BDeu)** score [91], and others [103, 108, 109].

3.2 Exploiting Structural Asymmetries

There is no way to rule out scenarios whereby a joint distribution admits SCMs indicating either of the structural directions $X_i \rightarrow X_j$ or $X_i \leftarrow X_j$, thereby making the induction of causal directionality from observation alone, impossible. However, if some additional assumptions are made about the functional and/or parametric forms of the underlying true data-generating structure, then one can exploit asymmetries in order to identify the direction of a structural relationship. These asymmetries manifest in various ways, including non-independent errors, measures of complexity, and dependencies between marginals and cumulative distribution functions. Methods which exploit such asymmetries are typically *local* methods, as they are only able to test edges one at a time (pairwise/bivariate causal directionality), or to test triples (with the third variable being an unobserved confounder) [101]. They may, of course, be extended to construct full-graphs by iteratively testing pairwise relationships (see e.g., the Information-Geometric Causal Inference algorithm [112]). We now briefly provide some examples of structural asymmetries, and direct interested readers to Mooij et al. [166] for a detailed review.

3.2.1 Additive Noise. Given the linear structural equations $X = U_X$ and $Y = X + U_Y$ such that $U_Y \perp\!\!\!\perp X$, we expect the residuals from a regression on the data from this generative model to reflect the $U_Y \perp\!\!\!\perp X$ property. Interestingly, if at most X or U_Y is non-Gaussian, then the causal direction (*i.e.,* $X \rightarrow Y$) is identifiable [69, 107, 117, 190]. This is illustrated in Figure 4. The true structural relationship $X = U_X$ and $Y = X + U_Y$ is used to generate data, where U_X and U_Y are non-Gaussian (they are uniformly distributed). In plot **A**, Y is regressed onto X (aligning with the true structural directionality), and it can be seen from plot **B** that the residuals following this regression are uncorrelated with X. Conversely, and as shown in plot **C**, when X is regressed onto Y(conflicting with the true structural directionality), it can be seen in plot **D** that this results in dependence between the residuals and Y.

The example given in Figure 4 depicts the non-Gaussian, linear case. Unfortunately, the assumption that (a) the data generating process is linear and (b) that the noise are sufficiently non-Gaussian to facilitate reliable identifiability may be overly restrictive in practice. There exist the non-linear additive noise [100] as well as the post-non-linear [277] models which seek increased generality. The non-linear additive noise model assumes the data are generated according to the structural equations $X = U_X$ and $Y = f(X) + U_Y$, where f is sufficiently non-linear, whilst the post-non-linear model assumes $X = U_X$ and $Y = f_A(f_B(X) + U_Y)$. There is no assumption that either U_Y

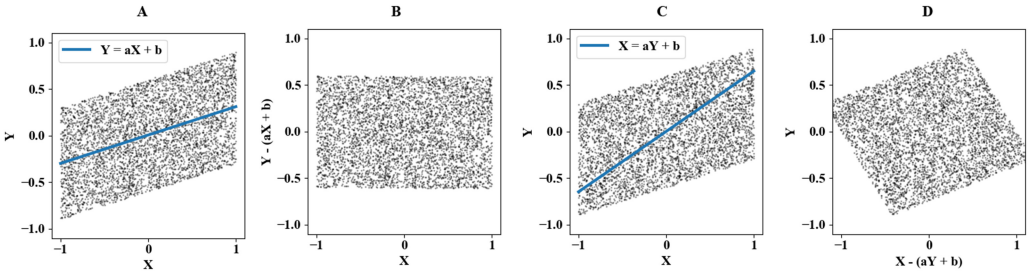

Fig. 4. The true structural relationship is $Y = X + U_Y$ and $X = U_X$ where U_X and U_Y are uniform noise sources. (A) shows the regression line when regressing Y onto X, and (B) shows the corresponding residuals plotted against X. (C) shows the regression line when regressing X onto Y and (D) shows the corresponding residuals plotted against Y. Together, these demonstrate that under the assumption of linear functional form and non-Gaussian noise, the true structural direction is identifiable as the one for which X is independent of the residuals, as indicated in (B). Example adapted from [190].

or X are Gaussian. Similarly to the linear non-Gaussian case above, both models exploit structural asymmetries that are reflected in the (in)dependence of regression residuals [69, 100, 190, 277].

3.2.2 Information Geometric Properties. From a causal perspective, the information geometric approach to identifying structural directionality takes inspiration from the concept of *independent mechanisms*. Assuming that the true structural direction is $X \rightarrow Y$, the concept of independent mechanisms holds that $P(X)$ contains no information about $P(Y|X)$, and vice versa. A common illustrative example [190] involves measurements of temperature Y at weather stations of different altitudes X. Regardless of the distribution of weather station altitudes $P(X)$, the mechanisms linking altitude to temperature (e.g., the law determining the relationship between the temperature and pressure of a gas) exist independently, and changing the temperature around a weather station does not increase its altitude.

Numerically, this scenario may be easily demonstrated by considering the inverse transform sampling method for transforming a uniform distribution $P(X)$ into a target distribution $P(Y)$ using the inverse cumulative distribution function. The uniform distribution is clearly independent of the function being used to transform it, but this independence does not hold for the transformed distribution. More generally, if $Y = f(X)$, the independence of mechanisms implies with high likelihood that P_X will be independent of the mechanism f. The corollary is that there exists dependence between P_Y and f^{-1} [112]. Assuming a structural direction $X \rightarrow Y$ via function f, the inverse function f^{-1} satisfies $\text{cov}[\log f^{-1}, p_Y] \geq 0$ [112, 115, 190].

It is worth noting various limitations to this approach, particularly with respect to its application to causal discovery in real-world systems. Firstly, it assumes that the mechanism f is deterministic. Secondly, it assumes that f is sufficiently non-linear that it may be used to identify dependence. Thirdly, real-world systems may (in addition to having non-deterministic mechanisms) demonstrate adaptation between cause and effect, such that P_X is no longer independent of f.

3.3 Interventions and Adjustment Sets

If interventional data are available, we are able to reduce the number of graphs in our MEC. An intervention can be denoted using Pearl's *do* operator [184] such that, "for each realization of x of X, $P(y|do(x))$ gives the probability of $Y = y$ induced by deleting from the model $[x_i = f_i(pa_i, u_i), i = 1..., n,]$ all equations corresponding to variables in X and substituting $X = x$ in the remaining equations." Such interventions can be hard/perfect/structural/atomic/deterministic,

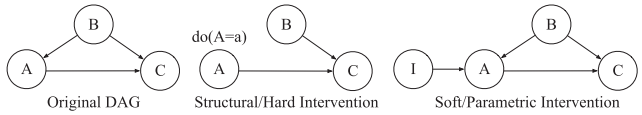

Fig. 5. Showing the differences between a hard/structural intervention (middle) and a soft/parametric intervention (right) on the original DAG (left). It can been that the parametric intervention preserves structural relationships. Adapted from [49, p. 986].

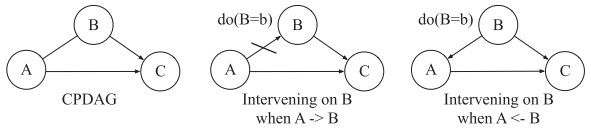

Fig. 6. Starting with the CPDAG on the left, where the structural direction between vertices A and B is unknown, intervening on B allows us to orient this edge. In the middle, the edge from A to B is removed following intervention on B (this is illustrated with the slash). On the right, an intervention on B does not remove the edge from B to A, and the effect of the intervention flows to A.

or soft/imperfect/parametric, depending on whether a variable is set to a specific value, or whether the variable and its relationship to its neighbours is modified in some way (e.g., by changing the noise distribution u). Graphically, a hard intervention can be represented by removing all incoming arrows (from parents) to a vertex, and setting that vertex to the value x [190, p. 88-91]. For a structural equation model $X = U_X$, $Y = f(X) + U_Y$ and $Z = g(X) + h(Y) + U_Z$, an intervention $Y = 4$ would entail $X = U_X$ (unmodified), $Y = 4$ (modified), and $Z = g(X) + h(4) + U_Z$ (modified). Thus it can be seen that only Y and its descendants have been affected by the intervention, leaving X unchanged. In contrast to hard interventions, a parametric intervention preserves the structure of the intervention itself, introducing an additional vertex and affecting the conditional distribution of the intervened variable. Parametric interventions also preserve any correlations deriving from unobserved/hidden confounders [49]. This difference is illustrated in Figure 5.

In order to demonstrate how interventions can be used to narrow the equivalence set (and in some cases make the true graph identifiable), consider the graphs in Figure 6. Starting with the CPDAG on the left, where the edge from A to C is undirected because the direction of the edge cannot be ascertained from conditional independencies alone. Intervening (hard) on B allows us to orient this edge by comparing the resulting distribution under intervention. If the edge is oriented $A \rightarrow B$ then the intervention has the effect of 'removing' this edge. Conversely, if the edge is oriented $B \rightarrow A$ then the intervention does nothing to remove this arrow, and the downstream variable A should change accordingly.

For a detailed review of different types of interventions and their implications, readers are directed to Eberhardt & Scheines [49]. Suffice to say there are many ways to leverage different types of intervention, including multiple interventions on different vertices, or single interventions applied to multiple nodes. Finally, there is work investigating the use of data representing unknown or uncertain interventions, whereby it is not known which variables have been intervened on [47, 124, 165, 208]. The use of intervention also yields what is known as an **Interventional Equivalence Class (IEC)**, representing the set of graphs compatible with a given intervention(s).

3.4 Causality Over Time

Whilst this review is primarily concerned with the static, non-dynamic setting, here we briefly describe some considerations for causal discovery with time series. Consider a graph $X \rightarrow Y$ for

the case where X and Y vary over time. In this scenario, a single right-arrow is not sufficient to detail whether X causes Y on an intra-timepoint[6] basis (i.e., contemporaneously), or on an inter-timepoint (i.e., lagged) basis. Indeed, at different points in time, and over different lags, the direction of causality may switch. In these cases it is common to *unroll* the graph over time, such that each instance of variables X^t and Y^t and their structural relationships over time are modelled explicitly.

There are then two types of causality considered in the context of time series. The most common (which is generally considered to be the industry standard, particularly in economics) is *Granger* causality [77]. If a variable X 'Granger-causes' Y, then it means that $Y^t \not\perp\!\!\!\perp X^{<t} | Y^{<t}$ [190, pp. 207], where $< t$ indicates timepoints previous to t. If this is the case, then the predictability of Y^t will decrease when $X^{<t}$ is removed from the model (because X contains unique information for predicting Y), when accounting for previous values of Y.

The first thing to note about Granger causality is that it tends to fail in the presence of contemporaneous effects [190, pp. 207], owing to difficulties with identifiability. The second, and perhaps more important aspect of Granger causality, is that it is only applicable if *separability* holds. Separability refers to the independence of the variables in the absence of causal interactions. Unfortunately, this is rarely the case in dynamic systems, where the current state of a variable may be heavily determined by the past of another (e.g., consider a predator-prey model, where both population levels are always functions of each other).

This failure of Granger causality was noted by Granger himself, and has motivated the development and application of *dynamic-causality*; in particular, methods deriving from Sugihara et al.'s Convergent Cross Mapping methods [232, 269]. The methods operate using *time delayed embeddings* or *shadow manifolds*. These shadow manifolds are constructed by concatenating time-lagged versions of the original time series. This process has been shown according to Takens' theorem [236], to be sufficient in recovering the dynamics of the full system even if only one, or a limited number, of observational variables are used. Dynamic or CCM-causality has shown great promise in applications to ecosystems and genetics [232, 269] where the phenomena may exhibit chaotic trajectories. The general idea behind these methods is to exploit asymmetries that exist between the compactness of neighbourhoods of points in the shadow manifolds. If $X \to Y$ in a CCM-causal sense, then points which are tightly clustered in the 'effect shadow manifold' of Y should also be tightly clustered in the 'cause shadow manifold' of X. This characteristic does not hold in the reverse direction (nor if there is no causal interaction in either direction), and this asymmetry enables us to identify causal interactions and directionality.

3.5 Evaluation Metrics

There are a number of common metrics used for evaluating the performance of causal discovery algorithms. The metrics given below are those used to evaluate the success of edge discovery. Other score metrics can be used to measure model fit (such as the log likelihood or the Bayesian Information Criterion). For a more detailed discussion on structural discovery metrics, readers are directed to work by de Jongh [43].

True Positive Rate (TPR) [92, p. 383]: Assuming an edge a_{ij} can be thresholded by $t \in (0, 1)$, TPR is defined as $\text{TPR}_t = |\{(i, j) : a_{ij} \geq t\} \cap \text{S}|/|\text{S}|$ where S is the set of ground truth edges (i.e., $\{(i, j) : a^*_{ij} = 1\}$).

False Positive Rate (FPR) [92, p. 383]: Assuming the probability of an edge a_{ij} can be thresholded by $t \in (0, 1)$, FPR is defined as $\text{FPR}_t = |\{(i, j) : a_{ij} \geq t\} \cap \hat{\text{S}}|/|\hat{\text{S}}|$ where $\hat{\text{S}}$ is the set of ground truth missing edges (i.e., $\{(i, j) : a^*_{ij} = 0\}$).

[6]It is generally accepted that an effect has to follow the cause in time, thereby precluding contemporaneous effects. However, in cases where the sampling rate is too low to capture this delay, it is reasonable to model the effects as instantaneous.

Area Over Curve (AOC) [92, p. 383]: Simply (1 − **Area Under Curve (AUC)** for the AUC of (FPR_t, TPR_t) where the threshold t is varied between 0 and 1. Either the AUC or AOC can be used as a structure discovery performance metric.

Structural Hamming Distance (SHD): Is the number of required changes to the graph for it to match the ground truth. It is the sum of missing edges, extra edges, and incorrect edges [43].

Structural Interventional Distance (SID) [189]: Is a count of the number of vertex pairs (i, j) for which the intervention $p(x_j|do(X_i = x))$ would be incorrect if the estimated graph (as opposed to the ground-truth graph) were used for what is known as the associated *adjustment set* (see Pearl [185] or Peters et al. [190]). It is therefore well suited for causal inference tasks [139, p. 7].

4 COMBINATORIC/SEARCH BASED ALGORITHMS

The number of possible DAGs increases super-exponentially with the number of variables [205]. As noted by Peters et al. [190], the number of possible DAGs for 10 variables is $> 4 \times 10^{18}$. As such, the search problem is NP-hard [33], and this will later motivate the use of continuous optimization based algorithms for graph learning.

Table 1 presents a non-exhaustive list of methods which do not use continuous optimization. In other words, they include primarily combinatoric/search-based algorithms for structure discovery. The table presents the type of approach used: constraint-based, score-based, asymmetry-based, hybrid, and sampling-based (which measure belief in a proposed graph structure by sampling from a posterior). In addition, the table provides the associated assumptions: Sufficiency (i.e., whether it assumes there are no hidden variables), Faithfulness (some methods achieve a less severe/relaxed form of faithfulness), and Acyclicity (some methods can learn feedback loops and cycles). Finally, the table indicates whether the method leverages interventions, and indicates the method's output (CPDAG, PAG, etc.).

5 CONTINUOUS OPTIMIZATION BASED ALGORITHMS

The primary focus of this survey is to review continuous optimization based methods for structure discovery. Continuous optimization methods are pervasive in the field of deep learning, whereby highly parameterized networks are optimized using variations on gradient descent [72]. The motivation for the neural network is that they do not impose restrictions on the functional form *a priori* and therefore "let the data speak" [248]. Increased computational power (particularly with the advent of GPUs) make the task of learning from large, high-dimensional datasets feasible. Recently, there have been an increasing number of methods which seek to learn structure from data, whilst leveraging the advantages of continuous optimization. This has resulted in the confluence of black-box deep learning approaches, and structure discovery. These continuous optimization approaches recast the combinatoric graph-search problem into a continuous optimization problem (specifically, an Equality Constrained Program) [279]. In Equation (5), the left hand side represents the traditional approach, which seeks the adjacency matrix $\mathbf{A}$ that minimizes some score function $S(\mathbf{A})$, subject to the implied d-vertex graph $\mathcal{G}(A)$ being in the set of valid DAGs. The right hand side represents a characterization of the continuous optimization problem which, again, seeks the adjacency matrix $\mathbf{A}$ that minimizes some score function $S(\mathbf{A})$, but this time subject to the constraint $h(\mathbf{A}) = 0$. Here, h is the function used to enforce acyclicity in the inferred graph.

$$\begin{array}{ll} \min_{\mathbf{A}\in\mathbb{R}^{d\times d}} S(\mathbf{A}) & \min_{\mathbf{A}\in\mathbb{R}^{d\times d}} S(\mathbf{A}) \\ \text{subject to } \mathcal{G}(\mathbf{A}) \in \text{DAGs} & \text{subject to } h(\mathbf{A}) = 0 \end{array} \tag{5}$$

The increased popularity of structure discovery in deep learning is not without sound motivation, with arguments that disentangled, structured, and symbolic representations are key to the next generation of AI, as well as robust cross-domain performance, transfer learning, and

Table 1. 'Combinatoric' based Algorithms for Causal Discovery (i.e., Search-based, SAT-solver)

Method	Year	Type	Suff.	Faith.	Acycl.	Interv.	Output
PC [228]	1993	constraint	yes	yes	yes	no	CPDAG
FCI [229, 275]	1995	constraint	no	yes	yes	no	POIPG
CCD [200]	1996	constraint	yes	yes	both	no	PAG
TPDA [31]	2002	constraint	yes	yes	yes	no	CPDAG
CPC [193]	2006	constraint	yes	relaxed	yes	no	CPDAG
KCL [233]	2007	constraint	yes	yes	yes	no	CPDAG
ION [241]	2008	constraint	no	yes	yes	no	PAG
IDA [153]	2009	constraint	yes	yes	yes	yes	DAG
cSAT+ [244]	2010	constraint	no	yes	yes	no	PCG
KCI-test [278]	2012	constraint	yes	yes	yes	no	CPDAG
RFCI [37]	2012	constraint	no	yes	yes	no	PAG
CHC [64]	2012	constraint	yes	yes	yes	no	PDAG
SAT [106]	2013	constraint	no	yes	no	yes	DG
Parallel-PC [141]	2014	constraint	yes	yes	yes	no	CPDAG
RPC [87]	2013	constraint	yes	yes	yes	no	CPDAG
PC-stable [36]	2014	constraint	both	yes	both	no	CPDAG
COmbINE [243]	2015	constraint	no	yes	yes	yes	summary SMCMs
backshift [208]	2015	–	no	no	no	yes	DG
IGSP [266]	2018	constraint	yes	relaxed	yes	yes	I-MEC
σ-CG [56]	2018	constraint	no	yes	no	yes	σ-connection graphs
CCI [231]	2018	constraint	no	yes	no	no	MAAG
FCI-soft [132]	2019	constraint	no	relaxed	yes	yes	I-MEC
IBSSI [32]	2020	constraint	no	yes	yes	yes	DAG
CD-NOD [104]	2020	constraint	no	yes	both	yes	–
psi-FCI [110]	2020	constraint	no	relaxed	yes	yes	Psi-EC
LCDI [274]	2020	constraint	no	yes	yes	yes	Pattern
EG [52]	2009	score	yes	yes	yes	no	BT-DAG
TWILP [183]	2014	score	yes	yes	yes	no	BT-DAG
CAM [26]	2014	score	yes	yes	yes	no	CPDAG
K2 [38]	1992	score	no	yes	yes	no	CPDAG
LB-MDL [140]	1994	score	yes	yes	yes	no	DAG
HGC [91]	1995	score	yes	yes	yes	no	CPDAG
GES [34]	2002	score	yes	yes	yes	no	CPDAG
OS [238]	2005	score	yes	yes	yes	no	DAG
HGL [90]	2005	score	yes	yes	yes	yes	CPDAG
Meinshausen [159]	2006	score	yes	–	no	no	UG
Graphical Lasso [59]	2008	score	yes	–	no	no	UG
BC [8]	2008	score	yes	–	no	no	UG
TC [187]	2008	score	yes	yes	yes	no	CPDAG
HG [89]	2008	score	yes	yes	yes	yes	DAG
Adaptive Lasso [222]	2010	score	yes	yes	yes	no	DAG
GIES [88]	2012	score	yes	yes	yes	yes	PDAG
CD [61]	2013	score	yes	yes	yes	yes	DAG
GBN learner [246]	2013	score	yes	no	yes	no	CPDAG
GES-mod [3]	2013	score	yes	yes	yes	no	CPDAG
Pen-PC [85]	2015	score	yes	yes	yes	no	CPDAG
Scalable GBN [4]	2015	score	yes	no	yes	no	DAG
K-A* [212]	2016	score	yes	yes	yes	no	DAG
NS-DIST [86]	2016	score	yes	no	yes	yes	DAG
MIP-GD [182]	2017	score	yes	yes	yes	no	CPDAG
CD2 [83]	2018	score	yes	yes	yes	yes	DAG
SP [195]	2018	score	yes	relaxed	yes	no	CPDAG
VAR [271]	2018	score	yes	yes	both	no	DG
GSF [103]	2018	score	yes	yes	yes	no	CPDAG
bQCD [235]	2020	score	yes	yes	yes	no	Bi
GCL [251]	2020	score	no	–	no	no	GCLM
GGIM [55]	2020	score	yes	no	no	no	GGIM
GYKZ [68]	2020	score	yes	yes	both	no	DG
SLARAC etc. [260]	2020	score	Granger	–	–	no	Bi
Order-MCMC [60]	2003	sampling	yes	yes	yes	no	DAG
OG [53]	2008	sampling	yes	yes	yes	yes	DAG
EE-DAG [281]	2011	sampling	yes	yes	yes	yes	DAG
ZIPBN [35]	2020	sampling	yes	no	yes	no	DAG
LiNGAM [221]	2006	asymmetries	yes	no	yes	no	DAG
LV LiNGAM [101]	2008	asymmetries	no	yes	yes	no	DAG
non-linear ANM [100]	2008	asymmetries	yes	yes	yes	no	DAG
PNL [277]	2009	asymmetries	yes	no	yes	no	DAG
CAN [113]	2009	asymmetries	no	yes	yes	no	Bi/tri
CCM [232]	2012	asymmetries	–	–	no	no	Bi
IGCI [112]	2012	asymmetries	yes	yes	yes	no	Bi
KCDC [161]	2018	asymmetries	yes	yes	yes	no	Bi
MMHC [245]	2006	hybrid	yes	yes	yes	no	DAG
ARGES [172]	2018	hybrid	yes	yes	yes	no	CPDAG

Provides indication of assumptions of sufficiency ('Suff.'), faithfulness ('Faith.'), acyclicity ('Acycl.'), as well as whether the method leverages forms of intervention ('Interv.'). 'Bi' indicates bivariate cause-effect pairs (possibly multivariate).

interpretability [16, 17, 65, 78]. Researchers have noted three primary approaches to learning representations of the world: (1) distributed, (2) structured and symbolic, and (3) a hybrid of (1) and (2). Most basic neural networks perform distributed learning and there is no clear separation of high-level semantics. Finally, we include Tables 3 and 4 which provide a comparison of a selection of methods' relative performance on the Sachs [210] and Tuebingen cause-effect pairs [164, 166] datasets, respectively. In these tables we also include some combinatoric algorithms (such as PC) for completeness.

Conversely, DAGs are highly structured and facilitate causal reasoning. However, such reasoning is only possible if one already has access to variables which represent high-level semantic concepts, which is not the case when learning from raw video data, for example. Hence, the motivation for hybrids which can be used to 'learn' or infer high-level representations as well as the structured relations between them. Examples of hybrid approaches include methods such as Recurrent Independent Mechanisms [76], graph networks [11, 12, 213], and a large body of work on scene understanding [116, 155, 171, 173, 270]. The debate as to how much structural inductive bias / constraint is required for an algorithm to reason effectively is ongoing [11, 45]. Indeed, finding a DAG to represent complex phenomena (such as natural language) is non-trivial and potentially impossible.

In this section we review (non-exhaustively) the recent evolution of continuous optimization based algorithms to structure learning, and Table 2 presents a list of the methods which are discussed. We provide summaries of performance in Tables 3 and 4 for Sachs proteins [210] and Tuebingen cause-effect pairs [164, 166], which are two common benchmark datasets. For the exposition, we categorize the methods into those which leverage acyclicity penalties, those which use neural networks, those which leverage interventions and concepts from reinforcement learning, other miscellaneous approaches, and those which are intended for time-series problems. Note that there is overlap between these categories.

5.1 Methods with an Acyclicity Penalty

The recent method DAGs with **NO TEARS (Non-combinatoric Optimization via Trace Exponential Augmented lagRangian Structure learning)** [279] is generally considered as the first to recast the combinatoric graph search problem as a continuous optimization problem (see Equation (5)), and numerous methods have adapted their principal contribution, which is the acyclicity penalty. The function/penalty h for enforcing acyclicity is derived to be:

$$h(\mathbf{A}) = tr(e^{\mathbf{A}\odot\mathbf{A}}) - d = 0 \tag{6}$$

where d is the number of vertices in the graph, 'tr' is the trace operator, and $\odot$ is the Hadamard product. In practice, $h(A)$ may be small but non-zero, and edges may require some thresholding. One of the disadvantages of this acyclicity constraint is that the matrix exponential requires $O(d^3)$ computations, and subsequent methods seek to improve on this. The structural model learnt is linear such that $X_j = a_j^T\mathbf{X} + U_j$, where a_j is the weight in the adjacency matrix corresponding with the edges into X_j, (the noise variables are not assumed to be Gaussian). NO TEARS uses a least-squares loss with an $l1$ penalty to encourage sparsity, and their objective is optimized using the Augmented Lagrangian method [174] with L-BFGS [28]. As well as synthetic data, NO TEARS is also evaluated on the proteins and phospholipid dataset by Sachs et al. [210]. Despite the fact that the formulated optimization problem does not guarantee an optimal solution, their results demonstrate close-to-optimal results on the chosen datasets.

5.1.1 Neural Network Adaptations of the Acyclicity Penalty. Since the introduction of the penalty in NO TEARS, numerous works have followed which adapt the penalty for use as a regularizer

Table 2. Continuous Optimization based Algorithms to Causal Discovery

Method	Year	Data	Form	Acycl.	Interv.	Output
CMS [152]	2014	low, dynamic/time series	NN	–	no	direction
NO TEARS [279]	2018	low	linear	yes	no	DAG
CGNN [74]	2018	low	NN	yes	no	DAG
SAM [121]	2019	low/medium	NN	yes	no	DAG
DAG-GNN [272]	2019	low	NN	yes	no	DAG
GAE [178]	2019	low	NN	yes	no	DAG
NO BEARS [142]	2019	low/medium/high	3-poly	yes	no	DAG
DEAR [219]	2020	image	NN	yes	no	–
CAN [167]	2020	low/medium/image	NN	yes	no	DAG
NO FEARS [259]	2020	low	linear	yes	no	DAG
GOLEM [177]	2020	low	linear	yes	no	DAG
ABIC [19]	2020	low	linear	yes	no	ADMG/PAG
DYNOTEARS [179]	2020	low	linear	yes	no	SVAR
SDI [124]	2020	low	NN	yes	yes	DAG
AEQ [63]	2020	Bi	NN	–	no	direction
RL-BIC [284]	2020	low	NN	yes	no	DAG
CRN [125]	2020	low	NN	yes	yes	DAG
ACD [151]	2020	low, time series	NN	Granger	no	time-series DAG
CASTLE (reg.) [138]	2020	low/medium	NN	yes	no	DAG
GranDAG [139]	2020	low	NN	yes	no	DAG
MaskedNN [176]	2020	low	NN	yes	no	DAG
CausalVAE [267]	2020	image	NN	yes	yes	DAG
CAREFL [126]	2020	low	NN	yes	no	DAG / direction
Varando [251]	2020	low	linear	yes	no	DAG
NO TEARS+ [280]	2020	low	non-linear	yes	no	DAG
ICL [258]	2020	low	NN	yes	no	DAG
LEAST [283]	2020	low/medium/high	linear	yes	no	DAG
CausalMosaic [263]	2020	Bi	NN	–	no	direction
NSM [253]	2021	video, dynamic/time series	NN	–	no	direction

'Data' indicates the dimensionality or type of the data the method has been demonstrated to handle. 'Form' indicates assumptions about the functional form (e.g., 'NN' for neural network, '3-poly' for 3rd order polynomial). 'Bi' indicates bivariate cause-effect pairs (possibly multivariate), 'low' indicates <100 vertices, 'medium' indicates >100, and 'high' indicates either dimensionality >10,000 or data which are not already projected into a causal/semantic space (e.g., image data). 'Acycl.' indicates whether the method enforces acyclicity, and 'Interv.' indicates the use of interventions during learning.

when training neural networks. DAG-GNN [272] extends NO TEARS by incorporating neural network functions f and black-box variational inference such that the score function is the **Evidence Lower BOund (ELBO)** [20, 129, 194, 199]. The method assumes faithfulness, and infers a latent posterior $\mathbf{Z}$:

$$\mathbf{Z} = f_4((\mathbf{I} - \mathbf{A}^T) f_3(\mathbf{X}) \tag{7}$$

where A is a weighted adjacency matrix, and $\mathbf{X}$ may comprise vector-valued variables. DAG-GNN recovers the observations with a decoder:

$$\mathbf{X} = f_2((\mathbf{I} - \mathbf{A}^T)^{-1} f_1(\mathbf{Z})) \tag{8}$$

Together, Equations (7) and (8) constitute a variational autoencoder [129, 199]. Noting that if f_2 is invertible, then:

$$f_2^{-1}(\mathbf{X}) = \mathbf{A}^T f_2^{-1}(\mathbf{X}) + f_1(\mathbf{Z}) \tag{9}$$

which is a generalization of the linear SEM model $\mathbf{X} = \mathbf{A}^T\mathbf{X} + \mathbf{Z}$. Acyclicity is enforced using a constraint derived from the one employed in NO TEARS [279] as:

$$tr[(\mathbf{I} + \alpha \mathbf{A} \odot \mathbf{A})^d] - d = 0 \tag{10}$$

Table 3. Comparison of **Structural Hamming Distance (SHD)** Performance on the Sachs Proteins Dataset [210], Sorted from Best (Top) to Worst (Bottom)

Method	Year	SHD
SDI [124]	2020	6
RL-BIC [284]	2020	11
MaskedNN [176]	2020	12
LEAST [283]	2020	12
CAM [26]	2014	12 [176]
GranDAG [139]	2020	13
GOLEM [177]	2020	14
DAG-GNN [272]	2019	16 [139]
NO TEARS + [280]	2020	16
SAM [121]	2019	17
PC [228]	1993	17 [139]
NO TEARS [279]	2018	19 [176]
GES [34]	2002	26 [139]

The reference next to the SHD figure indicates from where the experimental results are taken (if different from the reference for the original method).

Table 4. Comparison of Accuracy on the Tuebingen Cause-Effect Pairs Dataset [164, 166], Sorted from Best to Worst

Method	CausalMosaic [263]	AEQ [63]	CAREFL [126]	RECI [21]	IGCI [112]	ANM [100]
SHD %	83	80	73	69 [126]	61 [263]	52 [263]

The reference next to the accuracy figure indicates from where the experimental results are taken (if different from the reference for the original method).

where α acts as a hyperparameter on this constraint. This formulation of the acyclicity constraint is justified on the basis that it is preferred over a calculation that involves the matrix exponential (as appears in Equation (6)). Similarly to NO TEARS, they also use the augmented Lagrangian approach to optimization and evaluate on low-dimensional data such as the proteins and phospholipid dataset by Sachs et al. [210].

As it happens, autoencoder-type methods such as DAG-GNN are quite common in the continuous optimization based causal discovery literature. For instance, the authors of CASTLE [138] propose causal discovery as an auxiliary task which helps to regularize a supervised predictive model. The motivation is that, by identifying key causal factors, the model avoids overfitting to potential confounders which hurt model robustness and generalizability. Specifically, a neural network model attempts to identify the DAG that explains the structural relationships between the observed variables, and this task is built into an autoencoder [135] framework. Their structural model is non-parametric, following the form $X_i = f_i(pa_i, U_i)$ and using an adaptation of the NO TEARS acyclicity constraint which, they explain, also forces the autoencoder to reconstruct only the input variables which have neighbours.

Another autoencoder based method is GAE [178], which further extends the NO TEARS and DAG-GNN formulations for structure learning to facilitate non-linear structural relationships and vector-valued variables. They model structure in the same way as DAG-GNN, and draw a

connection to graph convolutional neural networks [130]:

$$f(X_j, \mathbf{A}) = f_2(\mathbf{A}^T f_1(X_j)) \tag{11}$$

where f_1 and f_2 are **multilayer perceptrons (MLPs)**. Similarly to NO TEARS, and DAG-GNN, they also use the augmented Lagrangian method with Adam [128] for constrained optimization. Their acyclicity constraint is identical to the one used in NO TEARS (Equation (6)). They demonstrate that GAE performs significantly better than NO TEARS and DAG-GNN, particularly as the number of vertices in the graph increases, and also highlight that training time is much shorter.

In a similar vain, the creators of CausalVAE [267] (yet another autoencoder-based method) argue that whilst many disentangled representation learning methods assume independence between latent factors [95, 137, 148], most latent factors behind real-world phenomena exhibit causal dependencies. They propose the use of a Variational AutoEncoder [129, 199]. The latent space of a VAE is usually parameterized by a set of exogeneous factors (often modelled as a multivariate, isotropic Gaussian). CausalVAE integrates a *Causal Layer* which transforms these exogenous latent factors into endogenous factors which reflect the causal semantics of the data. They assume a linear SEM following the form $\mathbf{Z} = \mathbf{A}^T\mathbf{Z} + \mathbf{U}$ where $\mathbf{Z}$ are the inferred latent factors following the application of the adjacency matrix $\mathbf{A}$. They integrate supervision in the form of semantic labels $\mathbf{Y}$ to condition the posterior $p(\mathbf{Z}|\mathbf{Y})$, which forces identifiability. These factors (which now reflect semantic quantities according to the provided supervision) are then passed to a masking layer, similar to the one used in MaskedNN (see below). They then apply $Z_j = g_j(\mathbf{A}_j \odot \mathbf{Z}) + U_j$ where g are nonlinear and invertible functions. $\mathbf{A}_j \odot \mathbf{Z}$ yields a vector only containing parental information, because the adjacency matrix effectively masks non-parents. The authors explain how this masking layer facilitates interventional queries. In order to learn the causal structure, they incorporate the structural inductive prior into the supervised loss function:

$$l_y = \mathbb{E}_q ||\mathbf{Y} - \sigma(\mathbf{A}^T\mathbf{Y})||_2^2 \tag{12}$$

where q is the approx posterior distribution. They incorporate the NO TEARS acyclicity constraint:

$$h(\mathbf{A}) = tr((\mathbf{I} + \mathbf{A} \odot \mathbf{A})^d) - d = 0 \tag{13}$$

The method is evaluated on the CelebA [147] dataset, as well as a synthetic data of a pendulum casting a shadow from a light. The second dataset is used to demonstrate the interventions - they intervene (for example) on the position of the light in order to demonstrate the independence of the position of the pendulum as well as the dependence with the shadow.

Temporarily moving away from autoencoder-based methods, GranDAG [139] follow the nonlinear additive noise structural model of the form $X_j = f_j(pa_j) + U_j$, where each function f_j is parameterized as a fully-connected neural network. In order to maintain an independence of mechanisms which corresponds with the independence implied by an adjacency matrix, they formulate *neural network paths* and a *connectivity matrix*, resembling previous work by Germain et al. [67]. The connectivity matrix $\mathbf{C}_j$ is essentially the matrix product of all neural network weights in a single neural network (i.e., parameterizing one f_j). This product results in $\mathbf{C}_j \in \mathbb{R}^{m \times d}$ where m is the number of parameters needed to specify a chosen distribution for X_j (e.g., a Gaussian has two parameters), and d is the number of variables. If $\mathbf{C}_{j,ki} = 0$ then the input i is independent of output k for variable X_j. Note that f_j takes as input X_{-j} (where the variable of interest j is masked to zero). The connectivity matrix is then used to define their weighted adjacency matrix, such that the adjacency matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ depends on all neural network weights from all neural networks. They define the weighted adjacency matrix and substitute it into the NO TEARS acyclicity constraint (Equation (6)). For learning they employ the augmented Lagrangian formulation, using a

log-likelihood score function, and threshold the resulting edges for $\mathbf{A}_{ij}$ close to zero. They demonstrate that their algorithm exceeds the performance of combinatoric approaches such as PC [228], as well as other continuous optimization based algorithms e.g., NO TEARS and DAG-GNN.

The researchers behind MaskedNN [176] also attempt to improve on NO TEARS [279] using neural networks. They assume an additive noise SEM of the form $X_j = f_j(pa_j)+U_j$, and explain how their method can be directly extended from handling scalar variables to vector valued variables. They provide a discussion on identifiability (something which a number of methods in both the combinatoric and continuous optimization literature tend to omit). They provide an overview of the gradual evolution from NO TEARS (which assumes linear SEMs), via DAG-GNN [272], GAE [178] and GraNDAG [139] (which handle non-linear SEMs), but highlight that these methods do not provide an in depth discussion about identifiability. They also highlight that the use of non-linear transformations on the adjacency matrices in DAG-GNN and GAE may affect their causal interpretability. MaskedNN uses a binary adjacency matrix $\mathbf{A}$ (rather than weighted), which is integrated into their SEM as: $X_j = h_j(\mathbf{A}_j \odot \mathbf{X}) + U_j$ and refer to this as an **Augmented SEM (ASEM)**. Their discussion on identifiability states that their method can learn a Super-graph of the true graph, and further utilize thresholding and Causal Additive Model [26] based pruning to remove spurious edges under mild conditions. They leverage the Gumbel-softmax trick [111, 154] to incorporate discrete learning (in view of the binary adjacency matrix) into an augmented Lagrangian 1st order continuous optimization based algorithm with an Adam optimizer [128]).

Finally, SAM [119, 121] is another neural network approach that is intended to address the limitations of CGNN (see below). Specifically, the limitations are CGNN's quadratic complexity (due to the calculation of the MMD), and the scalability issues that arise due to CGNNs use of a greedy-search. SAM addresses these two limitations with the use of adversarial training [73], and by making the mechanism which optimizes the DAG part of end-to-end training. Their score function is a log-likelihood loss with two model complexity regularizers: one that penalizes the model, on a per-vertex basis, by an amount proportional to the number of vertex parents; and one which acts as neural network parameter/weight decay. It uses an acyclicity regularizer similar to the NO TEARS penalty [279] to encourage DAG-ness:

$$\sum_{k=1}^{d} \frac{tr(A_j)}{k!} = 0 \tag{14}$$

Here, A is what they call a structural gate, which performs the same function as an adjacency matrix. The neural network parameterization of the structural equation model is $X_j = L_{j,H+1} \circ \sigma \circ \ldots L_{j,1}([\mathbf{a}_j \odot \mathbf{X}, U_j]$. In words, the stack of H neural network layers L_H and non-linearities σ for each variable j is used as the function over the Hadamard product between the data $\mathbf{X}$ and a binary vector form of the adjacency matrix A, s.t. $a_{i,j} = 1$ iff there is an edge $X_i \rightarrow X_j$. They provide a detailed theoretical analysis of their method, showing how the global training objective constitutes a combination of a structural component (which seeks the CPDAG) and a functional component (which exploits asymmetries). They assume both faithfulness and sufficiency, and evaluate on a range of low to medium dimensionality datasets (the highest number of dimensions is approximately 6,000 (DREAM5 [156]).

5.1.2 Improving the Acyclicity Penalty. As described above, the NO TEARS acyclicity penalty/constraint can be expensive to compute, and a number of methods have focussed on addressing this particular issue. The NO BEARS method [142] reformulates the acyclicity constraint by using the spectral radius of a matrix. Normally the spectral radius also requires $O(d^3)$ operations, but they present an approximation that takes only $O(d^2)$. The spectral radius is the maximum magnitude of eigenvalues, and the authors show how it forms an upper bound on the original NO

TEARS acyclicity constraint. Rather than using neural networks to increase the flexibility of the structural functions, they use a polynomial (order 3) regression. NO BEARS is demonstrated to scale well even on data with as many as 12,800 vertices.

Similarly, the authors of LEAST [283] propose a new acyclicity constraint intending to improve upon the $O(d^3)$ cost of NO TEARS [279]. To do this, they first consider:

$$h(\mathrm{S}) = tr(e^{\mathrm{S}}) - d = 0 \tag{15}$$

to be the NO TEARS constraint, where $\mathbf{S} = \mathbf{A} \odot \mathbf{A}$. This was subsequently altered by [272] to:

$$g(\mathrm{S}) = tr((\mathbf{I} + \mathrm{S})^d) - d = tr\left(\sum_{k=1}^{d} \binom{d}{k} \mathrm{S}^k\right) = 0 \tag{16}$$

on the basis that $e^{\mathrm{S}} = \sum_{i=0}^{\inf} \frac{\mathrm{S}^k}{k!}$, where k is the length of a cycle. The authors of LEAST argue that both of these have drawbacks relating to $O(d^3)$ complexity, as well as storage of e^S. They note that NO BEARS [142] framed the problem in terms of a spectral radius (the absolute value of the largest Eigenvalue of S). However, this also requires $O(d^3)$ computation, so they derive an upper bound $\bar{\delta}$ on this spectral radius as:

$$\begin{aligned} \bar{\delta}^{(k)} &= \sum_{i=1}^{d} b^{(k)}[i] \quad \text{where} \\ b^{(k)} &= (r(\mathrm{S}^{(k)}))^{\alpha} \odot (c(\mathrm{S}^{(k)}))^{1-\alpha} \quad \text{and} \\ \mathrm{S}^{(k+1)} &= (D^{(k)})^{-1} \mathrm{S}^{(k)} D^{(k)} \quad \text{and} \\ D^{(k)} &= \mathrm{Diag}(b^{(k)}) \end{aligned} \tag{17}$$

Combining a computable form for this upper bound with the least squares objective and $l1$ regularization, they show that this new objective is nearer to $O(d)$, and trains between 5 and 15 times faster than NO TEARS. Note that edge thresholding is still required. They demonstrate the benefits of this speedup by evaluating on both small graphs, as well as graphs with as many as 160,000 vertices.

A closely related approach is NO FEARS [259], which provides a detailed analysis of the acyclicity constraint of NO TEARS (Equation (6)) and show that, following the augmented Lagrangian optimization, it is not guaranteed to converge to a feasible solution of the intend constraint (i.e., when $h(\mathbf{A}) = 0$). Instead of a constraint that depends on $\mathbf{A} \odot \mathbf{A}$, they propose one that depends only on the absolute value $|\mathbf{A}|$, on the basis that there is a connection with the $l1$ penalty and sparsity. Following some modifications to make the absolute value function differentiable, the authors modify existing algorithms with knowledge derived through theoretic analysis, and show their proposal to improve all baselines (including combinatoric approaches).

Following in a similar vain to other works such as NO BEARS, DAG-GNN, and NO FEARS, the authors of GOLEM [177] note that NO TEARS uses a least-squares score function, and improve on this by proposing a score function that directly maximizes the data likelihood. The authors show that in the linear Gaussian case and under mild assumptions (such as faithfulness), a likelihood-based objective with 'soft' sparsity regularization is sufficient to asymptotically identify a quasi-equivalent (see original paper for definition) DAG and that a hard acyclicity constraint is not required. Further, in the linear non-Gaussian scenario, they explain how an acyclicity constraint is not needed in the asymptotic regime, although it may be necessary with finite samples. Finally, they explain how it is sufficient to have a 'soft' acyclicity penalty, instead of a hard constraint, which greatly reduces the complexity of the optimization problem. They propose their own objective, including a likelihood based score with an $l1$ regularizer and soft acyclicity constraint, which they optimize

using Adam [128]. Some post-processing is undertaken to threshold edges in order to guarantee acyclicity. The primary distinctions from NO TEARS are, therefore, (a) the likelihood based score function, and (b) the use of a soft (rather than hard) acyclicity penalty.

Finally, a number of the same authors from NO TEARS have since revisited their original work. We refer to this later work as NO TEARS+ [280], which seeks to extend NO TEARS acyclicity constraint to handle nonparametric, general models of the form $g_j(f_j(X))$ (which subsumes additive noise models, linear models, and generalized linear models). They integrate a multi-layer perceptron into their derived framework (as well as a number of other variations). This model does not utilize an adjacency matrix, and thus they frame acyclicity in terms of partial derivatives (an idea they attribute to Rosasco et al. [206]) such that $[\mathbf{W}(f)]_{kj} := ||\partial_k f_j||_2$. This states that the dependency structure between variable k and the function f_j (which is described by the DAG represented in matrix $\mathbf{W}$) is the $l2$ norm of the partial derivative of f_j with respect to X_k.

5.2 Methods Without an Acyclicity Penalty

Even though the NO TEARS acyclicity penalty has found popular assimilation into numerous methods, there exist alternative ways to learn a causal graph, without necessarily imposing such a constraint. We categorize these methods into those which use neural networks as well as interventions or reinforcement learning approaches.

5.2.1 Neural Network Methods without an Acyclicity Penalty. There are a number of methods which leverage the flexible function approximation capabilities of neural networks, without applying an acyclicity regularizer. Not all of these methods learn an entire, multivariate DAG at once. For instance, Causal Mosaic [263] is a neural network based non-linear Independent Component Analysis which is intended for discriminating the causal direction between bivariate pairs. The method is motivated by the fact that for each cause-effect pair there may be a mixture of similar underlying mechanisms. As such, they form a 'mosaic' ensemble of non-linear models for predicting the direction of unseen cause-effect pairs. They demonstrate state of the art performance on the Tuebingen cause-effect pairs dataset [164, 166]. Similarly, the authors of the AEQ method [63] develop a score function based on an autoencoder's reconstruction error for discovering the directionality of vector valued cause-effect pairs, and do not utilise an acyclicity constraint. Their key result is that the SEM $Y = g(f(X), U)$ only holds in one direction if X and Y are vectors and g and f are neural network functions. They extend this result to univariate X by creating multivariate versions of the variable based on a sorted concatenation of slices of the original. The complexity of this multivariate surrogate is then measured using an autoencoder reconstruction error (they use an $l2$ loss). For a cause-effect pair, the variable with the higher loss is likely to be the cause. In the case where the original variables are multivariate, they propose an adversarial conditional independence method that discriminates between joint distributions and the product of the marginals.

Other neural network methods seek to estimate the full DAG. One such method, CGNN [74], combines neural networks with hill-climbing or Tabu search. The neural networks are used to learn the functions mapping variables (e.g., see the SEM breakdown in Figure 1), where the variables themselves are selected according to the output of a greedy-search algorithm. The networks are trained using the Adam [128] optimizer with a **Maximum Mean Discrepancy (MMD)** [79] score function. During training, the edges are directed in order to minimize this discrepancy, and following training, the graph is adjusted to remove cycles. CGNN incorporates a hill-climbing search algorithm to optimize the structure of the DAG, and then the network optimization resumes. This training cycle is repeated to convergence, and each edge has an associated score representing its contribution to the global fit. They use a thresholding function to regularize the number of edges

in the graph. Finally, their method includes a means to identify possible hidden confounding, by leveraging the fact that confounding can be modelled as correlations/associations between the (otherwise) exogenous latent random variables.

Another neural network method which does not use the NO TEARS acyclicity constraint is DEAR [219], which combines a **Variational AutoEncoder (VAE)** [129, 199] with an adversarial loss [73] in order to infer a latent space with "causal" structure. Strictly, this is not a causal discovery method, because they assume the 'super-graph' is given, and they learn the associated weights and parameters. The latent space is given as supervision in the form of labels for the generative factors. The latent structure is defined as:

$$z = f((\mathbf{I} - \mathbf{A}^T)^{-1} h(\epsilon)) \tag{18}$$

Here, A is a weighted adjacency matrix, f and h are neural networks, and ϵ is noise sampled from a prior distribution. DEAR is notable for its use of high-dimensional data with semantic labels. It maps from image data to the structured latent space, where the labels provide a form of supervision.

In CAREFL [126], the authors combine causal discovery with the deep learning framework known as normalizing flows [131, 198]. Normalizing flows provide a means to construct generative models which have the capacity to model complex densities using invertible transformations of a basic and tractable density. They enable the exact computation of the log-likelihood (which constitute their learning objective) via the use of the change of variables formula and inverse log Jacobian determinant. Specifically, they use autoregressive flows, which are a form of normalizing flow for which the transformations are affine and have simple, lower-triangular Jacobians [127, 158]. The authors consider an SEM in terms of a *causal ordering*, whereby, according to the SEM/DAG, there exists a permutation of the vertices that corresponds to the order of specified dependencies. For example, a parent vertex precedes a child vertex in the causal ordering. The generic additive noise SEM $X_j = f_j(pa_j) + U_j$ can be written in terms of a causal ordering π as $X_j = f_j(\mathbf{X}_{<\pi(j)}) + U_j$ (which is assumed for CAREFL), where $X_{<\pi(j)}$ represents variables that precede X_j in the causal order (including its parents). This latter form is shown to bear resemblance to the autoregressive flow model with a few constraints. The CAREFL method is shown to be flexible enough to answer both counterfactual and interventional queries. As well as outputting a DAG, the method can also be used to judge causal direction by using the log-likelihood to score different directions.

Finally, the authors of ICL [258] focus on the problem of structure discovery under the missing-data setting, and provide definitions and examples of three types of missingness: **Missing At Random (MAR)**, **Missing Completely At Random (MCAR)**, and **Missing Not At Random (MNAR)**. They propose the use of **Generative Adversarial Networks (GANs)** [73] and **Variational AutoEncoders (VAEs)** [129, 199]. ICL takes incomplete data and simultaneously imputes the missing data using the GAN, in order to match the generated distribution to the empirical distribution. The task of the discriminator in the GAN is to differentiate between observed versus generated data. The skeleton graph is estimated using a method following DAG-GNN [272]. Following this, the edges in the skeleton are oriented following a method proposed by Cai et al. [29] which is based on the additive noise model for causal direction identification.

5.2.2 Interventional and Reinforcement Learning Methods. There exist a number of methods which 'act' upon the graph being learned. By evaluating modifications of the graph, these methods can be considered to fall under the umbrella of interventional or reinforcement learning methods. SDI [124], for instance, is a neural network method which attempts to discover structure using data which have been subject to unknown interventions. SDI assumes faithfulness and sufficiency, and is restricted to discrete, categorical variables with no missingness; it assumes the available

interventions are sparse and only effect a single (possibly unknown) variable; the interventions may be soft; and there are no compounding interventions (i.e., only one or fewer interventions occur in the data). The method is trained in three stages which repeat until convergence. The first stage is concerned with updating the functional parameters (those which map between vertices). The procedure involves randomly drawing data samples and graph configurations, and optimizing the functional parameters using the log-likelihood as a score function. In the second stage, the structural parameters are updated (those which model the edges between vertices), and interventional (unknown) data are sampled. The variable subject to intervention is predicted using a simple heuristic; namely, that the variable exhibiting the greatest reduction in log-likelihood is predicted on the basis that it is a poor fit to the observational distribution. Given a new set of interventional data and sampled graphs, these graphs can be scored whilst masking the intervened variable. In the third stage, and following [18], the REINFORCE algorithm [261] is used to update the discrete structural parameters. The method is evaluated on low-dimensional data ($d < 100$), and is shown to exceed state-of-the-art on a number of benchmark datasets.

Similarly, Ke et al. [125] propose a meta-learning neural network method which also incorporates interventions and leverages continuous representations of graphs. Training is split into episodes where, for each episode, a graph is proposed and used to generate data for the duration of the episode. The episode is further split into k time points, and for each time point a random intervention is undertaken on the graph and data is generated. The model is then asked to predict the outcome of the intervention, and thereby ends up 'learning' the causal relationships between the variables in the graph. They also propose a **Causal Relational Network (CRN)**, which accumulates information about the interventions and graphs over time (similar to an LSTM [96]). They use a graph decoder (the gradients from which are not backpropagated to the rest of the network) in order to validate the graph's continuous representation against the ground truth graph. It is shown that CRNs learn new causal models quickly and efficiently. Interestingly, there is no discussion about (a)cyclicity, but nonetheless their intention is to learn DAGs.

CAN [167] is a **Generative Adversarial Network (GAN)** [73] that facilitates interventional sampling from a structural graph (which the authors refer to as a causal graph) at inference time. It comprises a Label Generation Network, which learns a graph from the dataset labels, and a Conditional Image Generation Network, which generates the images conditioned on the interventional distribution specified by the user at inference time. Their generator is a function of an adjacency matrix applied to the noise vectors as $\mathbf{X} = G((\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{Z})$ where X is a sample from the joint distribution, $\mathbf{Z}$ is random noise, A is a weighted adjacency matrix, $\mathbf{I}$ is the identity matrix, and G is the non-linear generator function. In order to impose acyclicity, they leverage an equality constraint [272], such that acyclicity occurs if Equation (10) is satisfied. As well as evaluating CAN on CelebA [147] image data (including the generation of interventional samples), they also evaluate it on the more traditional CHILD [227] and Alarm [14] datasets showing competitive performance.

Finally, the authors of RL-BIC [284] explicitly take a reinforcement learning approach to causal discovery. They generate directed graphs using an encoder-decoder neural network model, which forms the 'actor'. The output of the encoder-decoder is the proposed graph, which is scored using the BIC in order to generate a reward signal. A critic is used to update the proposed graphs and therefore also to drive the optimization of the neural network parameters. They assume an additive noise model $X_i = f_i(pa_i) + U_i$ as well as faithfulness and causal sufficiency. Their output graph is represented using a binary adjacency matrix. They mask out (i, i) edges to prevent self-loops, and incorporate an adapted form of the NO TEARS acyclicity penalty (see Equation (6)). In order to guarantee acyclicity (in the event that $h(\mathbf{A})$ is small but non-zero), they augment it with a hard indicator function penalty that acts on whether the graph is a valid DAG or not. All generated graphs are stored during training, and the one with the best score is chosen, and this graph is

finally pruned to reduce false discovery. The method is trained using poly-gradient method and REINFORCE [261] with Adam [128], and evaluated on relatively small graphs (≤30 nodes).

5.3 Miscellaneous Continuous Optimization Approaches without an Acyclicity Penalty

We briefly discuss two methods which use neither acyclicity penalties, nor neural networks. Firstly, Varando [251] proposes a proximal gradient [180] optimization objective that yields a linear SEM and corresponding DAG. The method derives the novel objective by framing the learning problem in terms of sparse matrix factorization, and the resulting method NODAG is shown to be both effective and efficient. Secondly, ABIC [19] extends the continuous optimization paradigm to discover various types (ancestral, arid, bow-free) of ADMGs which account for unmeasured confounding. In the linear SEM case, unmeasured confounders manifest as correlated errors, which are represented in a *second* adjacency matrix. They present three differentiable constraints which can be used to discover a particular type of ADMG. They use the BIC criterion as the primary objective/score function. The parameters are optimized using a Residual Iterative Conditional Fitting algorithm [46].

5.4 Time Series Approaches

Despite time series causal discovery not being the principal focus of this survey, for completeness we briefly describe a number of continuous optimization methods intended for this domain. DYNOTEARS [179] seeks to discover structure in time series data, which is a topic we have not covered in detail in this survey. By using second order optimization, DYNOTEARS seeks to learn a **Structural Vector AutoRegressive (SVAR)** model, which is also a form of dynamic Bayesian network. This is argued to be important on the basis that temporal dynamics are an essential part of real-world systems, which cannot be captured using a static graph model. They assume that variables potentially affect each other both contemporaneously, and in a time-lagged manner. DYNOTEARS is therefore not strictly Granger causal, because it accounts for contemporaneous effects [190, p. 203-208]. They model two adjacency matrices, $\mathbf{W}$ and $\mathbf{A}$, for the intra-slice and inter-slice graph edges, respectively. Because the edges represented in $\mathbf{A}$ only go forward in time, only $\mathbf{W}$ needs an acyclicity constraint. They use the same constraint as NO TEARS (see Equation (6) and the section above), and incorporate it into an augmented Lagrangian problem which is optimized using L-BFGS-B [282]. Following optimization, and similarly to other methods using acyclicity constraints, they threshold edges with weights close to 0. DYNOTEARS is evaluated on S&P 500 returns data with 97 vertices, and on DREAM4 with 100 vertices [157].

In contrast, ACD [151] is a Granger-causality non-linear time-series method which leverages black-box variational inference [20, 129, 194, 199] to infer a latent posterior graph. Granger causality assumes there are no contemporaneous effects [190, p. 203-208]. The method is demonstrated to perform well under hidden confounding (and so does not assume causal sufficiency). ACD learns from samples with different causal relationships but shared dynamics. This is motivated using an example from neuroscience. They use the encoder to infer the causal graph from a particular sample, and a decoder which models the dynamics and takes past samples and the inferred graph in order to predict the future. Specifically, for sample $\mathbf{X}_s$ with graph encoder f and decoder dynamics model g, the future is predicted as $\mathbf{X}_s^{t+1} = g(\mathbf{X}_s^{\leq t}, f(\mathbf{X}_s))$.

Finally, there exist two continuously optimized approaches which seek to discovery causes in the dynamic and chaotic time series regime. Firstly, the CMS algorithm [152] is intended to identify causal directionality between time varying variables in dynamic systems. The method is inspired by convergent cross mapping methods [232, 269] which operate using *time delayed embeddings* or *shadow manifolds*. In order to ascertain whether two variables X and Y from a time varying

Table 5. Python and R Packages for General Causal Inference and Discovery

Method	Keywords & Software
causaleffect [240]	general causality, R
daggity [237]	general causality, R
dosearch [239]	causal effect identification, R
Causal Discovery Toolbox [120]	causal discovery, Python
pcalg [122]	causal discovery, R
bnlearn [215]	causal discovery, R
rEDM [181]	dynamic modeling and convergent cross mapping, R
DoWhy [218]	general causality, Python
CausalImpact [25]	intervention, time series, R
causal-cmd [262]	general causality, Python (py-causal) & JAVA + CLI
CausalNex [192]	general causality, Python

CLI = command line interface.

dynamical system are causally related, CMS uses a radial basis function neural network to map between the shadow embeddings for X and Y. The asymmetry in the error when mapping from X to Y, compared with the error when mapping from Y to X, is used as a proxy to infer causal directionality. Note that this method is only demonstrated to work when the input variables are univariate, but may be used multiple times to ascertain the causal structure of more than two observational variables. Secondly, NSM [253] provides a step-wise approach to discovering the causal structure in video data representing objects varying in location. The method works by first deriving a neural network embedding/representation of the visual state (such as object coordinates/locations), and then evaluating the cross-map strength between the time series of objects in this visual state. NSM is demonstrated to correctly identify the correct graph in synthetic video data with three objects.

6 SUMMARY AND DISCUSSION

We have attempted to present the relevant background, definitions, assumptions, approaches to causal discovery, and common evaluation metrics, as well as providing a brief review of combinatoric methods, and a detailed review of continuous optimization based methods. In terms of additional resources, a range of software packages exist for undertaking causal inference and structure discovery and we have provided a list in Table 5 for convenience. Also, in Table 6 we provide a list of datasets used for causal discovery. Finally, we encourage readers to explore various additional references and commentaries. These include: a discussion of the relevance of causality to machine learning [71, 175, 214]; commentaries on the nature of causality [143, 160]; alternative reviews on causal inference and causal discovery [69, 84, 92, 230, 268]; reviews with a focus on time-series causal inference and discovery [51]; frameworks for dynamical SCMs with ODEs [163, 188]; guides on the foundations for causal discovery [48]; some example applications [123, 146, 149, 211, 284]; and textbooks on causal inference and causal discovery [184, 190, 228].

6.1 Opportunities and Future Directions

One of the main advantages to combinatoric approaches to structure discovery relates to the provision of guarantees for identifying the true graph, or at least the true equivalence class. This advantage comes at a significant cost, however, because such approaches are limited to low-dimensional problems (or low-cardinality graphs) due to the super-exponential search space. One might expect, then, that even though the continuous optimization approaches are confronted with a non-trivial, non-convex solution space, they might at least scale to larger problems. Unfortunately, and as

Table 6. A List of Datasets That Have Been Used for Testing Structure Discovery Methods

Dataset	Vertices	Notes
Multi-body Interaction [145]	–	up to 5 moving balls with physical interactions/relations
Fabric deformation [145]	–	applying forces to different fabrics
Cause-effect pairs [235]	2	bivariate distributions
Cause-effect pairs [162]	2	bivariate distributions
Cause-effect pairs (Tuebingen) [164, 166]	2	bivariate distributions
SynTReN [247]	user specified	synthetic gene expression data
Sachs [210]	11	proteins and phospholipids in human cells
Scale-Free Graphs [9]	user specified	preferential attachment graph generation law
Erdos-Rényi Graphs (e.g. [139])	user specified	adds edges with probability $p = \frac{2e}{d^2-d}$
Linear, GP Add, GP Mix, Sigmoid Add and Sigmoid Mix	–	mixed graph data
CausalWorld [1]	–	comprehensive robotics dataset
MPI3D [70]	–	visual disentanglement dataset
Pendulum-light-shadow [267]	–	image data
Phase coupled oscillator [5]	–	physical relations
NetSim [225]	user specified	fMRI data simulation
Temperature [151]		
BnLearn [215]	–	Repository
DREAM series [156, 157]	up to 6000	simulated and in-vivo gene regulation networks
Causality 4 Climate [209]	–	climate change time series competition data
Archaeology [103]	8	archaeology data
S&P500	500	time series/stock returns
CauseMe [168]	–	Repository/benchmarking platform

can be seen from Table 2, most continuous optimization approaches have only been evaluated on low-dimensional problems. This seems to be due to the fact that the most common acyclicity constraint, namely the one in Equation (6) from NO TEARS [279], contains a term that requires $O(d^3)$ computations. This has motivated the development of higher-efficiency acyclicity constraints for continuous optimization approaches to structure discovery, such as the one in LEAST [283]. One further way to alleviate the issues when confronted with high-dimensional problems is to encode the data into a lower-dimensional representation. This was undertaken in CausalVAE [267], who applied the NO TEARS constraint to a graph operating in low-dimensional representation space. Whilst this approach works well for non-semantic data (such as pixel data from images), it might not be useful in situations whereby the data are both high-dimensional *and* semantic (as with gene regulation data in the DREAM5 dataset [156]). In the latter case, encoding semantic data into a new subspace may or may not be meaningful, and will likely depend on the domain of application.

In terms of what we consider to present the most opportunity for future work, we note that there are relatively few continuous optimization approaches which seek to learn structured, semantic representations from non-semantic, high-dimensional data such as video or image data (exceptions include CausalVAE [267] and DEAR [219], and related works on scene understanding include [13, 27, 50, 54]). Interestingly, the field of reinforcement learning, which involves the interaction of learning agents with each other and their environment, has been relatively slow on the uptake of causal perspectives [7, 42]. Ashton [7] even notes that one of the seminal texts on reinforcement learning [234] makes no explicit reference to causality throughout the entire text. As such, the application of causal discovery to reinforcement learning presents significant opportunity.[7] Finally, whilst there were numerous combinatoric methods which are designed to handle unobserved confounding and/or cyclicity (e.g., CCD [200], backshift [208], CCI [231]), there are relatively few such continuous optimization approaches. Given the complexity of time-varying real-world phenomena and the potential for cycles, we note the opportunity to develop continuous optimization methods which can operate in a broader class of scenarios.

[7]Some exceptions include [39, 42, 93, 123, 170, 197, 207, 226, 273, 276].

6.2 The Causal Leap

It was mentioned in Section 1, that a causal perspective is crucial to the empirical sciences as well as for improving machine learning methods. More fundamentally, as humans we are interested in how to reason about and interact in a world full of causal interactions. In general, the pursuit of causality is essential to understanding the world and our universe. However, it is fraught with difficulty, and below we finish with a discussion on some of the criticisms and warnings relating to this otherwise laudable pursuit.

We now take the time to discuss how structure discovery methods take us from a structural association (albeit, an association which may exhibit directional asymmetry) to that of a causal association. What is there to suggest that learning or identifying such a graphical or structural model is equivalent to learning or identifying causes and generative structure in reality? In order to interpret graphical models causally, the the **Causal Markov Condition (CMC)** [228] is often assumed. However, in our view (and see also [57]) the CMC simply represents an uninformative re-branding of the regular Markov condition (which describes the conditional independence properties of the graph), with the additional interpretation of the arrows as directed causal dependencies. As Dawid [41, p. 83] argues, "there is no reason to believe [the causal implications of the CMC] hold in complete generality". It should be clear that the conditional independence properties of DAGs play a foundational role in causal discovery. However, as Dawid [2008] states in his work *Beware of the DAG!*: ". . . for conditional independence the arrows are nothing but incidental construction features supporting the d-separation semantics."

The use of structural equations gets us somewhat closer to where we want to be when seeking to represent causality, than do graphical models alone. This is because the structural equation formalism can be more specific and informative than its simpler (yet intuitive) graphical counterpart [190, p. 106]. Nonetheless, as with graphical models, the interpretation of structural equations as structural causal models cannot be made without strong and often untestable assumptions. Applying these strong assumptions to structural or graphical models incites some harsh criticism. Indeed, Korb & Wallace [1997] caricature research into causal discovery as "a glorious perversion" akin to the "search for the philosopher's stone" [134, p. 551].

Such criticisms are important to assimilate, and they remind us to be careful when using statistical/causal models to draw inference about the nature of reality. In particular, even if a graphical model bears resemblance to our own conception of a phenomenon, it may not be an appropriate or fair way to represent complex social constructs (e.g., gender or race), representing what Freidman described as a biased attempt to "quantify the qualitative, make discrete the continuous, or formalize the nonformal" [58]. For instance, it is not clear what it means to be able to manipulate/intervene on someone's race, independently of their other attributes, or indeed at all. In general, we need a thorough understanding of what a variable is *supposed* to represent, and whether it actually represents it at all (both a problem of ontology and epistemology) before we perform meaningful inference. However, a sufficiently clear understanding may be difficult if not impossible to attain.

The prevalence of reports of systemic bias arising from automated decision processes is increasing, and an awareness for sources of bias is critical in undertaking fair and equitable machine learning [97, 169, 224, 256]. Just because causal discovery methods define themselves as 'causal', does not mean there are not significant difficulties associated with taking the leap from data to reality. Indeed, blindly interpreting structured models as robustly representing causal quantities can be immensely problematic. We appreciate Dawid's [41] reference to Bourdieu who warns of "sliding from the model of reality to the reality of the model" [23]. Furthermore, score-based approaches in particular have recently been highlighted to be highly sensitive to

data scaling [118, 196], making it difficult to rely on such methods for robust structure learning, regardless of whether the structure is interpreted causally or not.

In spite of these warnings, causal discovery methods may still be used productively, particularly for exploratory purposes (e.g., in providing candidate causal links for further investigation and validation) [41]. Furthermore, the combination of observational and interventional/experimental data may provide us with opportunities to uniquely *identify* models which, at least under various assumptions, correspond with some true external cause-effect relationships. More broadly, shifting from naive associational and purely predictive machine learning models to models informed by causal structure, may bring concomitant improvements in robustness and generalizability. If researchers maintain a cautious approach when making the leap from modelling to causality, structure discovery methods can be used in support of the endeavour to further human understanding.

REFERENCES

[1] O. Ahmed, F. Träuble, A. Goyal, A. Neitz, M. Wütrich, Y. Bengio, B. Schölkopf, and S. Bauer. 2020. CausalWorld: A robotic manipulation benchmark for causal structure and transfer learning. arXiv:2010.04296v1 (2020).

[2] R. A. Ali, T. S. Richardson, and P. Spirtes. 2009. Markov equivalence for ancestral graphs. *The Annals of Statistics* 37, 5B (2009), 2808–2837.

[3] J. I. Alonso-Barba, L. de la Ossa, J. A. Gámez, and J. M. Puerta. 2013. Scaling up the greedy equivalence algorithm by constraining the search space of equivalence classes. *Internat. J. Approx. Reason.* 54 (2013).

[4] B. Aragam and Q. Zhou. 2015. Concave penalized estimation of sparse Gaussian Bayesian networks. *Journal of Machine Learning Research* 16 (2015).

[5] H. Araki (Ed.). [n.d.]. *International Symposium on Mathematical Problems in Theoretical Physics.* Springer-Verlag, New York, Chapter Lecture Notes in Physics. 39.

[6] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. 2020. Invariant risk minimization. arXiv:1907.02893v3 (2020).

[7] H. Ashton. 2020. Causal Campbell-Goodhart's law and reinforcement learning. arXiv:2011.01010v1 (2020).

[8] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. 2008. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research* 9 (2008).

[9] A.-L. Barabási and R. Albert. 1999. Emergence of scaling in random networks. *Science* (1999).

[10] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. 2020. On Pearl's Hierarchy and the foundations of causal inference. *ACM Special Reports* (2020).

[11] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santori, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu. 2018. Relational inductive biases, deep learning, and graph networks. arXiv:1806.01261v2 (2018).

[12] D. Beaini, S. Passaro, V. Létourneau, W. L. Hamilton, G. Corso, and P. Liò. 2020. Directional graph networks. arXiv:2010.02863v2 (2020).

[13] M. B. Bear, C. Fan, D. Mrowca, Y. Li, S. Alter, A. Nayebi, J. Schwartz, L. Fei-Fei, J. Wu, J. B. Tenenbaum, and D. L. K. Yamis. 2020. Learning physical graph representations from visual scenes. arXiv:2006.12373v2 (2020).

[14] I. A. Beinlich, H. J. Suermondt, R. M. Chavez, and G. F. Cooper. 1989. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. *Proceedings of the 2nd European Conference on Artificial Intelligence in Medicine* (1989).

[15] A. Bellot and M. van der Schaar. 2019. Conditional independence testing using generative adversarial networks. arXiv:1907.04068v1 (2019).

[16] Y. Bengio. 2019. The consciousness prior. arXiv:1709.08568v2 (2019).

[17] Y. Bengio, A. Courville, and P. Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2013).

[18] Y. Bengio, T. Deleu, N. Rahaman, N. R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pal. 2019. A meta-transfer objective for learning to disentangle causal mechanisms. arXiv:1901.10912v2 (2019).

[19] R. Bhattacharya, T. Nagarajan, D. Malinsky, and I. Shpitset. 2020. Differentiable causal discovery under unmeasured confounding. arXiv:2010.06978v1 (2020).

[20] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. 2018. Variational inference: A review for statisticians. arXiv:1601.00670v9 (2018).

[21] P. Bloebaum, D. Janzing, T. Washio, S. Shimizu, and B. Schölkopf. 2018. Cause-effect inference by comparing regression errors. *Int. Conf. on Artificial Intelligence and Statistics* (2018), 900–909.

[22] L. Bottou, J. Peters, J. Quinonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research* 14 (2013).
[23] P. Bourdieu. 1977. *Outline of a Theory of Practice*. Cambridge University Press, Cambridge.
[24] A. Breskin, S. R. Cole, and M. G. Hudgens. 2018. A practical example demonstrating the utility of single-world intervention graphs. *Epidemiology* 29, 3 (2018).
[25] K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S. L. Scott. 2015. Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics* 9 (2015), 247–274.
[26] P. Bühlmann, J. Peters, and J. Ernest. 2014. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics* 42, 6 (2014), 2526–2556.
[27] C. P. Burgess, L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick, and A. Lerchner. 2019. MONet: Unsupervised scene decomposition and representation. arXiv:1901.11390v1 (2019).
[28] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* (1995).
[29] R. Cai, J. Qiao, K. Zhang, Z. Zhang, and Z. Hao. 2019. Causal discovery with cascade nonlinear additive noise model. *Proceedings of the 28th International Joint Conference on Artificial Intelligence* (2019).
[30] D. C. Castro, I. Walker, and B. Glocker. 2019. Causality matters in medical imaging. arXiv:1912.08142v1 (2019).
[31] J. Cheng, R. Greiner, J. Kelly, D. Bell, and W. Liu. 2002. Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence* 132, 1–2 (2002).
[32] D. Chicharro, M. Besserve, and S. Panzeri. 2020. Causal learning with sufficient statistics: An information bottleneck approach. arXiv:2010.05375v1 (2020).
[33] D. M. Chickering. 1996. *Learning from Data. Lecture Notes in Statistics, Vol 112*. Springer, New York, Chapter Learning Bayesian networks is NP-complete.
[34] D. M. Chickering. 2002. Optimal structure identification with greedy search. *Journal of Machine Learning Research* 3 (Nov. 2002).
[35] J. Choi, R. Chapkin, and Y. Ni. 2020. Bayesian causal structure learning with zero-inflated Poisson Bayesian networks. *34th Conference on Neural Information Processing Systems* (2020).
[36] D. Colombo and M. H. Maathuis. 2014. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research* 15 (2014).
[37] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. 2012. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics* 40, 1 (2012).
[38] G. F. Cooper and E. Herskovits. 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 9 (1992).
[39] O. Corcoll and R. Vicente. 2020. Disentangling causal effects for hierarchical reinforcement learning. arXiv:2010.01351v1 (2020).
[40] B. Cummins, T. Gedeon, and L. Spendlove. 2015. On the efficacy of state space reconstruction methods in determining causality. *SIAM Journal on Applied Dynamical Systems* 14, 1 (2015), 335–381.
[41] A. P. Dawid. 2008. Beware of the DAG! *NeurIPS Workshop on Causality* (2008).
[42] P. de Haan, D. Jayaraman, and S. Levine. 2019. Causal confusion in imitation learning. *33rd Conference on Neural Information Processing Systems* (2019).
[43] M. de Jongh and M. J. Druzdzel. 2009. A comparison of structural distance measures for causal Bayesian network models. *Recent Advances in Intelligent Information Systems* (2009).
[44] A. Deaton and N. Cartwright. 2018. Understanding and misunderstanding randomized controlled trials. *Social Science and Medicine* 210 (2018), 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005
[45] D. Ding, F. Hill, A. Santoro, and M. Botvinick. 2020. Object-based attention for spatio-temporal reasoning: Outperforming neuro-symbolic models with flexible distributed architectures. arXiv:2012.08508v1 (2020).
[46] M. Drton, M. Eichler, and T. S. Richardson. 2009. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research* 10, 10 (2009).
[47] D. Eaton and K. Murphy. 2007. Exact Bayesian structure learning from uncertain interventions. *AISTATS* (2007).
[48] F. Eberhardt. 2017. Introductions to the foundations of causal discovery. *Int. J. Data Sci. Anal.* 3 (2017).
[49] F. Eberhardt and R. Scheine. 2006. Interventions and causal inference. *Philosophy of Science* 74, 5 (2006), 981–995.
[50] S. Ehrhardt, O. Groth, A. Monszpart, M. Engelcke, I. Posner, N. Mitra, and A. Vedaldi. 2020. RELATE: Physically plausible multi-object scene synthesis using structured latent spaces. arXiv:2007.01272v1 (2020).
[51] M. Eichler. 2013. Causal inference with multiple time series: Principles and problems. *Philosophical Transactions of the Royal Society* 371, 1997 (2013).
[52] G. Elidan and S. Gould. 2009. Learning bounded treewidth Bayesian networks. *Advances in Neural Information Processing Systems* 21 (2009).

[53] B. Ellis and W. H. Wong. 2008. Learning causal Bayesian network structures from experimental data. *J. Amer. Statist. Assoc.* 103 (2008).
[54] M. Engelcke, A. R. Kosiorek, O. P. Jones, and I. Posner. 2019. Genesis: Generative scene inference and sampling with object-centric latent representations. arXiv:1907.13052v2 (2019).
[55] K. Fitch. 2020. Learning directed graphical models from Gaussian data. arXiv:1906.08050v3 (2020).
[56] P. Forré and J. M. Mooij. 2018. Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. arXiv:1807.03024v1 (2018).
[57] D. Freedman and P. Humphreys. 1999. Are there algorithms that discover causal structure? *Synthese* 121 (1999).
[58] B. Friedman. 1996. Value-sensitive design. *Interactions* 3, 6 (1996).
[59] J. Friedman, T. Hastie, and R. Tibshirani. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9 (2008).
[60] N. Friedman and D. Koller. 2003. Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning* 50 (2003).
[61] F. Fu and Q. Zhou. 2013. Learning sparse causal Gaussian networks with experimental intervention: Regularization and coordinate descent. *J. Amer. Statist. Assoc.* 108, 501 (2013).
[62] K. Fukumizu and A. Gretton. 2008. Kernel measures of conditional dependence. *Electronic Proceedings of Neural Information Processing Systems* (2008).
[63] T. Galanti, O. Nabati, and L. Wolf. 2020. A critical view of the structural causal model. arXiv:2002.10007v1 (2020).
[64] J. Gámez, J. L. Mateo, and J. Puerta. 2012. One iteration CHC algorithm for learning Bayesian networks: An effective and efficient algorithm for high dimensional problems. *Progress in Artificial Intelligence* 1 (2012).
[65] A. D. Garcez and L. C. Lamb. 2020. Neurosymbolc AI: The 3rd wave. arXiv:2012.05876v2 (2020).
[66] D. Geiger and D. Heckerman. 1994. Learning Gaussian networks. *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence* (1994).
[67] M. Germain, K. Gregor, I. Murray, and H. Larochelle. 2015. MADE: Masked autoencoder for distribution estimation. *Proceedings of the 32nd International Conference on Machine Learning* (2015).
[68] A. E. Ghassami, A. Yang, N. Kiyavash, and K. Zhang. 2020. Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs. arXiv:1910.12993v3 (2020).
[69] C. Glymour, K. Zhang, and P. Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in Genetics* 10 (2019).
[70] M. W. Gondal, M. Wütrich, D. Miladinovic, F. Locatello, M. Breidt, V. Volchkov, J. Akpo, O. Bachem, B. Schölkopf, and S. Bauer. 2019. On the transfer of inductive bias from simulation to the real world: A new disentanglement dataset. arXiv:1906.03292 (2019).
[71] M. Gong. 2020. Bridging causality and learning: How do they benefit from each other? *Proceedings of the 29th International Joint Conference on Artificial Intelligence* (2020).
[72] I. Goodfellow, Y. Bengio, and A. Courville. 2016. *Deep Learning*. MIT Press, Cambridge, Massachusetts.
[73] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative adversarial nets. arXiv:1406.2661 (2014).
[74] O. Goudet, D. Kalainathan, P. Caillou, I. Guyon, D. Lopez-Paz, and M. Sebag. 2018. Learning functional causal models with generative neural networks. arXiv:1709.05321v3 (2018).
[75] A. Goyal and Y. Bengio. 2020. Inductive biases for deep learning of higher-level cognition. arXiv:2011.15091v2 (2020).
[76] A. Goyal, A. Lamb, J. Hoffmann, S. Sodhani, S. Levine, Y. Bengio, and B. Schölkopf. 2019. Recurrent independent mechanisms. arXiv:10893v2 (2019).
[77] C. W. J. Granger. 1980. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control* 2, 1 (1980), 329–352.
[78] K. Greff, S. van Steenkiste, and J. Schmidhuber. 2020. On the binding problem in artificial neural networks. arXiv:2012.05208v1 (2020).
[79] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. 2007. A kernel method for the two-sample problem. *Neural Information Processing Systems* (2007).
[80] T. L. Griffiths and J. B. Tenenbaum. 2009. Theory-based causal induction. *Psychological Review* 116, 4 (2009), 661–716. https://doi.org/10.1037/a0017201
[81] M. P. Grosz, J. M. Rohrer, and F. Thoemmes. 2020. The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science* (2020), 1–13. https://doi.org/10.1177/1745691620921521
[82] P. D. Grünwald and P. M. Vitányi. 2008. *Handbook of the Philosophy of Information*. North Holland, Chapter Algorithms information theory.
[83] J. Gu, F. Fu, and Q. Zhou. 2018. Penalized estimation of directed acyclic graphs from discrete data. arXiv:1403.2310v4 (2018).

[84] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu. 2020. A survey of learning causality with data: Problems and methods. *ACM Comput. Surv.* 1, 1 (2020).
[85] M. J. Ha, W. Sun, and J. Xie. 2015. PenPC: A two-step approach to estimate the skeletons of high-dimensional directed acyclic graphs. *Biometrics* (2015).
[86] S. W. Han, G. Chen, M.-S. Cheon, and H. Zhong. 2016. Estimation of directed acyclic graphs through two-stage adaptive lasso for gene network inference. *J. Amer. Statist. Assoc.* 111, 515 (2016).
[87] N. Harris and M. Drton. 2013. PC algorithm for nonparanormal graphical models. *Journal of Machine Learning Research* 14 (2013).
[88] A. Hauser and P. Bühlmann. 2012. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research* 13 (2012).
[89] Y.-B. He and Z. Geng. 2008. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research* 9 (2008).
[90] Y.-B. He, Z. Geng, and X. Liang. 2005. Learning causal structures based on Markov equivalence class. *International Conference on Algorithmic Learning Theory* (2005).
[91] D. Heckerman, D. Geiger, and D. M. Chickering. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20 (1995).
[92] C. Heinze-Deml, M. H. Maathuis, and N. Meinshausen. 2018. Causal structure learning. *Annual Review of Statistics and Its Application* 5 (2018).
[93] T. Herlan. 2020. Causal variables from reinforcement learning using generalized Bellman equations. arXiv:2010.15745v1 (2020).
[94] M. Hernan. 2018. The c-word: Scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health* 108, 5 (2018), 625–626. https://doi.org/10.2105/AJPH.2018.304337
[95] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. 2017. Beta-VAE: Learning basic visual concepts with a constrained variational framework. *ICLR* (2017).
[96] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1–32.
[97] K. Holstein, J. W. Vaughan, H. Daume III, M. Dudik, and H. Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? arXiv:1812.05239v2 (2019).
[98] M. Hopkins and J. Pearl. 2003. Clarifying the usage of structural models for commonsense causal reasoning. *Proceedings of AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning* (2003).
[99] M. Hopkins and J. Pearl. 2005. Causality and counterfactuals in the situation calculus. *Proceedings of the 7th International Symposium on Logical Formalization of Commonsense Reasoning* (2005).
[100] P. O. Hoyer, D. Janzing, J. M. Mooij, and J. Peters. 2008. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems* (2008).
[101] P. O. Hoyer, S. Shimizu, A. J. Kerminen, and M. Palviainen. 2008. Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *Int. J. Approx. Reasoning* 49, 2 (2008).
[102] Z. Hu and R. Evans. 2020. Faster algorithms for Markov equivalence. *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence* (2020).
[103] B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour. 2018. Generalized score functions for causal discovery. *KDD* (2018).
[104] B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero , C. Glymour, and B. Schölkopf. 2020. Causal discovery from heterogeneous/nonstationary data with independent changes. arXiv:1901.01672v5 (2020).
[105] P. Humphreys and D. Freedman. 1996. The grand leap. *The British Journal for the Philosophy of Science* 47 (1996).
[106] A. Hyttinen, P. O. Hoyer, F. Eberhardt, and M. Jarvisalo. 2013. Discoverying cyclic causal models with latent variables: A general SAT-based procedure. *Proceedings of the 29th Conf. on Uncertainty in Artificial Intelligence* (2013).
[107] A. Hyvärinen, J. Karhunen, and E. Oja. 2001. *Independent Component Analysis*. John Wiley and Sons, Inc.
[108] A. Hyvärinen and S.n.M. Smith. 2013. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research* 14 (2013), 111–152.
[109] S. Imoto, T. Goto, and S. Miyano. 2002. Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. *Pacific Symposium on Biocomputing* 175–186 (2002).
[110] A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim. 2020. Causal discovery from soft interventions with unknown targets: Characterization and learning. *34th Conference on Neural Information Processing Systems* (2020).
[111] E. Jang, S. Gu, and B. Poole. 2017. Categorical reparameterization with Gumbel-Softmax. arXiv:1611.01144v5 (2017).
[112] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniusis, B. Steudel, and B. Schölkopf. 2012. Information-geometric approach to inferring causal directions. *Artificial Intelligence* 182–183 (2012).
[113] D. Janzing, J. Peters, J. Mooij, and B. Scholkopf. 2009. Identifying confounders using additive noise models. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence* (2009).

[114] D. Janzing and B. Schölkopf. 2010. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory* 56, 10 (2010).
[115] D. Janzing, B. Steudel, N. Shajarisales, and B. Schölkopf. 2015. *Measures of Complexity*. Springer, Germany, Chapter justifiying information geometric causal inference.
[116] J. Jiang and S. Ahn. 2020. Generative neurosymbolic machines. arXiv:2010.12152v1 (2020).
[117] A. M. Kagan, Y. V. Linnik, and C. R. Rao. 1973. *Characterization Problems in Mathematical Statistics*. Wiley, New York, NY.
[118] M. Kaiser and M. Sipos. 2021. Unsuitability of NOTEARS for causal graph discovery. arXiv:2104.05441 (2021).
[119] D. Kalainathan. 2019. *Generative Neural Networks to Infer Causal Mechanisms: Algorithms and Applications*. Ph.D. Dissertation. Universite Paris Sud.
[120] D. Kalainathan and O. Goudet. 2019. Causal discovery toolbox: Uncover causal relationships in Python. arXiv:1903.02278v1 (2019).
[121] D. Kalainathan, O. Goudet, I. Guyon, D. Lopez-Paz, and M. Sebag. 2020. Structural agnostic modeling: Adversarial learning of causal graphs. arXiv:1803.04929v3 (2020).
[122] M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann. 2012. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software* 47, 11 (2012).
[123] N. R. Ke, A. Didolkar, S. Mittal, A. Goyal, G. Lajoie, S. Bauer, D. Rezende, Y. Bengio, M. Mozer, and C. Pal. 2021. Systematic evaluation of causal discovery in visual model based reinforcement learning. *arXiv preprint arXiv:2107.00848* (2021).
[124] N. R. Ke, O. Bilaniuk, A. Goyal, S. Bauer, H. Larochelle, B. Schölkopf, M. C. Mozer, C. Pal, and Y. Bengio. 2020. Learning neural causal models from unknown interventions. arXiv:1910.01075v2 (2020).
[125] N. R. Ke, J. X. Xang, J. Mitrovic, M. Szummer, and D. J. Rezende. 2020. Amortized learning of neural causal representations. *ICLR Causal Learning for Decision Making Workshop* (2020).
[126] I. Khemakem, R. P. Monti, R. Leech, and A. Hyvärinen. 2020. Causal autoregressive flows. arXiv:2011.02268v1 (2020).
[127] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. 2016. Improved variational inference with inverse autoregressive flow. *29th Conference on Neural Information Processing Systems* (2016).
[128] D. P. Kingma and J. L. Ba. 2017. Adam: A method for stochastic optimization. arXiv:1412.6980v9 (2017).
[129] D. P. Kingma and M. Welling. 2014. Auto-encoding variational Bayes. arXiv:1312.6114v10 (2014).
[130] T. N. Kipf and M. Welling. 2017. Semi-supervised classification with graph convolutional networks. *ICLR* (2017).
[131] I. Kobysev, S. J. D. Prince, and M. A. Brubaker. 2020. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).
[132] M. Kocaoglu, A. Jaber, K. Shanmugam, and E. Bareinboim. 2019. Characterization and learning of causal graphs with latent vaiables from soft interventions. *33rd Conference on Neural Information Processing Systems* (2019).
[133] D. Koller and N. Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, Massachusetts.
[134] K. B. Korb and C. S. Wallace. 1997. In search of the philosopher's stone: Remarks on Humphreys and Freedman's critique of causal discovery. *British Journal for the Philosophy of Science* (1997), 543–553.
[135] M. A. Kramer. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal* 37, 2 (1991).
[136] N. Kreif and K. DiazOrdaz. 2019. Machine learning in policy evaluation: New tools for causal inference. arXiv:1903.00402v1 (2019).
[137] A. Kumar, P. Sattigeri, and A. Balakrishnan. 2018. Variational inference of disentangled latent concepts from unlabeled observations. arXiv:1711.00848v3 (2018).
[138] T. Kyono, Y. Zhang, and M. van der Schaar. 2020. CASTLE: Regularization via auxiliary causal graph discovery. *34th Conference on Neural Information Processing Systems* (2020).
[139] S. Lachapelle, P. Brouillard, T. Deleu, and S. Lacoste-Julien. 2020. Gradient-based neural DAG learning. arXiv:1906.02226v2 (2020).
[140] W. Lam and F. Bacchus. 1994. Learning Bayesian belief networks: An approach based on the MDL principle. *Computational Intelligence* 10, 3 (1994).
[141] T. D. Le, T. Hoang, L. Li, J. Liu, H. Liu, and S. Hu. 2014. A fast PC algorithm for high dimensional causal discovery with multi-core PCs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 13, 9 (2014).
[142] H.-C. Lee, M. Danieletto, R. Miotto, S. T. Cherng, and J. T. Dudley. 2020. Scaling structural learning with NO-BEARS to infer causal transcriptome networks. *Pacific Symposium on Biocomputing* 25 (2020).
[143] D. Lewis. 1973. Causation. *The Journal of Philosophy* 70, 17 (1973), 556–567.
[144] J. Li, L. Liu, T. D. Le, and J. Liu. 2020. Accurate data-driven prediction does not mean high reproducibility. *Nature Machine Intelligence* (2020).

[145] Y. Li, A. Torralba, A. Anandkumar, D. Fox, and A. Garg. 2020. Causal discovery in physical systems from videos. arXiv:2007.00631v2 (2020).
[146] L. Lin, M. Sperrin, D. A. Jenkins, G. P. Martin, and N. Peek. 2020. A systematic review of causal methods enabling predictions under hypothetical interventions. arXiv:2011.09815v1 (2020).
[147] Z. Liu, P. Luo, X. Wang, and X. Tang. 2015. Deep learning face attributes in the wild. *Proceedings of ICCV* (2015).
[148] F. Locatello, S. Bauer, M. Lucic, G. Ratsch, S. Gelly, B. Scholkopf, and O. Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. arXiv:1811.12359v3 (2019).
[149] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Schölkopf, and L. Bottou. 2016. Discovery causal signals in images. *CVPR* (2016).
[150] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling. 2017. Causal effect inference with deep latent-variable models. *31st Conference on Neural Information Processing Systems* (2017).
[151] S. Lowe, D. Madras, R. Zemel, and M. Welling. 2020. Amortized causal discovery: Learning to infer causal graphs from time-series data. arXiv:2006.10833v1 (2020).
[152] H. Ma, K. Aihara, and L. Chen. 2014. Detecting causality from nonlinear dynamics with short-term time series. *Scientific Reports* 4, 7464 (2014).
[153] M. H. Maathuis, M. Kalisch, and P. Bühlmann. 2009. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics* 37 (2009).
[154] C. J. Maddison, A. Mnih, and Y. W. Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. arXiv:1611.00712v3 (2017).
[155] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *ICLR* (2019).
[156] D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, M. Kellis, J. J. Collins, and G. Stolovitzky. 2012. Wisdom of crowds for robust gene network inference. *Nature Methods* 9, 8 (2012), 796–804.
[157] D. Marbach, T. Schaffter, C. Mattiussi, and D. Floreano. 2009. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology* 16, 2 (2009), 229–239.
[158] J. Marino, L. Chen, J. He, and S. Mandt. 2019. Improving sequential latent variable models with autoregressive flows. *2nd Symposium on Advances in Approximate Bayesian Inference* (2019).
[159] N. Meinshausen and P. Bühlmann. 2006. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34 (2006).
[160] P. Menzies and H. Beebee. 2020. *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Chapter Counterfactual Theories of Causation.
[161] J. Mitrovic, D. Sejdinovic, and Y. W. Teh. 2018. Causal inference via kernel deviance measures. arXiv:1804.04622v1 (2018).
[162] J. M. Mooij and D. Janzing. 2010. Distinguishing cause from effect. *JMLR Workshop and Conference Proceedings* 6 (2010), 147–156.
[163] J. M. Mooij, D. Janzing, and B. Schölkopf. 2013. From ordinary differential equations to structural causal models: The deterministic case. *UAI* (2013).
[164] J. M. Mooij, D. Janzing, J. Zscheischler, and B. Schölkopf. 2014. CauseEffectPairs repository. https://webdav.tuebingen.mpg.de/cause-effect/.
[165] J. M. Mooij, S. Magliacane, and T. Claassen. 2016. Joint causal inference from multiple contexts. arXiv:1611.10351 (2016).
[166] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Scholkopf. 2016. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research* 17, 32 (2016), 1–102.
[167] R. Moraffah, B. Moraffah, M. Karami, A. Raglin, and H. Liu. 2020. Causal adversarial network for learning conditional and interventional distributions. arXiv:2008.11376v2 (2020).
[168] J. Munoz-Marí, G. Mateo, J. Runge, and G. Camps-Valls. 2020. CauseMe: An online system for benchmarking causal discovery methods. *In Preparation* (2020).
[169] R. Nabi, D. Malinsky, and I. Shpitser. 2019. Optimal training of fair predictive models. arXiv:1910.04109v1 (2019).
[170] S. Nair, Y. Zhu, S. Savarese, and L. Fei-Fei. 2019. Causal induction from visual observations for goal directed tasks. arXiv:1910.01751v1 (2019).
[171] L. Nanbo, C. Eastwood, and R. B. Fisher. 2020. Learning object-centric representations of multi-object scenes from multiple views. *34th Conference on Neural Information Processing Systems* (2020).
[172] P. Nandy, A. Hauser, and M. H. Maathuis. 2018. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics* 46, 6A (2018).
[173] C. Nash, S. M. A. Eslami, C. Burgess, I. Higgins, D. Zoran, T. Weber, and P. Battaglia. 2017. The multi-entity variational autoencoder. *31st Conference on Neural Information Processing Systems* (2017).

[174] A. Nemirovsky. 1999. *Optimization II: Numerical Methods for Nonlinear Continuous Optimization*. Israel Institute of Technology.
[175] E. C. Neto. 2020. Towards causality-aware predictions in static machine learning tasks: The linear structural causal model case. arXiv:2001.03998v1 (2020).
[176] I. Ng, Z. Fang, S. Zhu, Z. Chen, and J. Wang. 2020. Masked gradient-based causal structure learning. arXiv:1910.08527v2 (2020).
[177] I. Ng, A. E. Ghassami, and K. Zhang. 2020. On the role of sparsity and DAG constraints for learning linear DAGs. *34th Conference on Neural Information Processing Systems* (2020).
[178] I. Ng, S. Zhu, Z. Chen, and Z. Fang. 2019. A graph autoencoder approach to causal structure learning. *NeurIPS Workshop* (2019).
[179] R. Pamfil, N. Sriwattanaworachai, S. Desai, P. Pilgerstorfer, P. Beaumont, K. Georgatzis, and B. Aragam. 2020. DYNOTEARS: Structure learning from time-series data. *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics* (2020).
[180] N. Parikh and S. Boyd. 2014. Proximal algorithms. *Found. Trends Optim.* 1, 3 (2014), 127–239.
[181] J. Park, C. Smith, G. Sugihara, E. Deyle, E. Saberski, and H. Ye. 2021. rEDM: Empirical dynamic modeling. (2021). https://CRAN.R-project.org/package=rEDM.
[182] Y. W. Park and D. Klabjan. 2017. Bayesian network learning via topological order. arXiv:1701.05654v2 (2017).
[183] P. Parviainen, H. S. Farahani, and J. Lagergren. 2014. Learning bounded tree-width Bayesian networks using integer linear programming. *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics* (2014).
[184] J. Pearl. 2009. *Causality*. Cambridge University Press, Cambridge.
[185] J. Pearl. 2012. On a class of bias-amplifying variables that endanger effect estimates. arXiv:1203.3503 (2012).
[186] J. Pearl and D. Mackenzie. 2018. *The Book of Why*. Penguin Books.
[187] J. P. Pellet and A. Elisseeff. 2008. Using Markov blankets for causal structure learning. *Journal of Machine Learning Research* 9 (2008).
[188] J. Peters, S. Bauer, and N. Pfister. 2020. Causal models for dynamical systems. arXiv:2001.06208v1 (2020).
[189] J. Peters and P. Bühlmann. 2015. Structural intervention distance (SID) for evaluating causal graphs. *Neural Computation* (2015).
[190] J. Peters, D. Janzing, and B. Schölkopf. 2017. *Elements of Causal Inference*. MIT Press, Cambridge, Massachusetts.
[191] M. Petersen, L. Balzer, D. Kwarsiima, N. Sang, G. Chamie, J. Ayieko, J. Kabami, A. Owaraganise, T. Liegler, F. Mwangwa, and K. Kadede. 2017. Association of implementation of a universal testing and treatment intervention with HIV diagnosis, receipt of antiretroviral therapy, and viral suppression in East Africa. *Journal of American Medical Association* 317, 21 (2017), 2196–2206. https://doi.org/10.1001/jama.2017.5705
[192] QuantumBlack Labs. 2020. CausalNex. .
[193] J. Ramsey, J. Zhang, and P. Spirtes. 2006. Adjacency-faithfulness and conservative causal inference. *Proceedings of the 22nd International Conference on Uncertainty in Artificial Intelligence* (2006).
[194] R. Ranganath, S. Gerrish, and D. M. Blei. 2013. Black box variational inference. arXiv:1401.0118v1 (2013).
[195] G. Raskutti and C. Uhler. 2018. Learning directed acyclic graph models based on sparsest permutations. *ISI Journal for the Rapid Dissemination of Statistics Research* (2018).
[196] A. G. Reisach, C. Seiler, and S. Weichwald. 2021. Beware of the simulated DAG! varsortability in additive noise models. *arXiv preprint arXiv:2102.13647* (2021).
[197] D. J. Rezende, I. Danihelka, G. Papamakarios, N. R. Ke, R. Jiang, T. Weber, K. Gregor, H. Merzic, F. Viola, J. Wang, J. Mitrovic, F. Besse, I. Antonoglou, and L. Buesing. 2020. Causally correct partial models for reinforcement learning. *ICLR* (2020).
[198] D. J. Rezende and S. Mohamed. 2016. Variational inference with normalizing flows. arXiv:1505.05770v6 (2016).
[199] D. J. Rezende, S. Mohamed, and D. Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. arXiv:1401.4082 (2014).
[200] T. Richardson. 1996. A discovery algorithm for directed cyclic graphs. *Proceedings of the 12th International Conference on Uncertainty in Artificial Intelligence* (1996).
[201] T. S. Richardson and J. M. Robins. 2013. Single world intervention graphs: A primer. *Working Paper Number 128, Center for Statistics and the Social Sciences, University of Washington* (2013).
[202] T. S. Richardson and J. M. Robins. 2013. Single world intervention graphs (SWIGS): A unification of the counterfactual and graphical approaches to causality. *Working Paper Number 128, Center for Statistics and the Social Sciences, University of Washington* (2013).
[203] T. Richardson and P. Spirtes. 2002. Ancestral graph Markov models. *The Annals of Statistics* 30, 4 (2002), 962–1030.
[204] T. S. Richardson and P. Spirtes. 2003. *Highly Structured Stochastic Systems*. Oxford University Press, Oxford, Chapter Causal inference via ancestral graph models.

[205] R. W. Robinson. 1973. *New Directions in the Theory of Graphs*. Academic Press, New York, Chapter Counting labeled acyclic digraphs.
[206] L. Rosasco, S. Villa, S. Mosci, M. Santoro, and A. Verri. 2013. Nonparametric sparsity and regularization. *Journal of Machine Learning Research* (2013).
[207] S. Ross, G. J. Gordon, and J. A. Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics* (2011).
[208] D. Rothenhausler, C. Heinze, J. Peters, and N. Meinshausen. 2015. backShift: Learning causal cyclic graphs from unknown shift interventions. *Advances in Neural Information Processing Systems* (2015).
[209] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, et al. 2019. Inferring causation from time series in earth system sciences. *Nature Communications* 10, 2553 (2019).
[210] K. Sacks, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308, 5721 (2005), 523–529.
[211] N. Sani, D. Malinsky, and I. Shpitser. 2020. Explaining the behavior of black-box prediction algorithms with causal learning. arXiv:2006.02482v1 (2020).
[212] M. Scanagatta, G. Corani, C. P. de Campos, and M. Zaffalon. 2016. Learning treewidth-bounded Bayesian networks with thousands of variables. *30th Conference on Neural Information Processing Systems* (2016).
[213] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. 2009. The graph neural network model. *IEEE Transactions on Neural Networks* 20, 1 (2009), 61–80.
[214] B. Scholkopf. 2019. Causality for machine learning. arXiv:1911.10500v1 (2019).
[215] M. Scutari. 2017. Bayesian network constraint-based structure learning algorithms: Parallel and optimized implementations in the bnlearn R package. *Journal of Statistical Software* 77, 2 (2017).
[216] R. Sen, A. T. Suresh, K. Shanmugam, A. G. Dimakis, and S. Shakkottai. 2017. Model-powered conditional independence test. *31st Conference on Neural Information Processing Systems* (2017).
[217] R. D. Shah and J. Peters. 2020. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics* 48, 3 (2020).
[218] A. Sharma and E. Kiciman. 2020. DoWhy: An end-to-end library for causal inference. arXiv:2011.04216 (2020).
[219] X. Shen, F. Liu, H. Dong, Q. Lina, Z. Chen, and T. Zhang. 2020. Disentangled generative causal representation learning. arXiv:2010.02637v1 (2020).
[220] C. Shi, T. Xu, and W. Bergsma. 2020. Double generative adversarial networks for conditional independence testing. arXiv:2006.02615v1 (2020).
[221] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7 (2006).
[222] A. Shojaie and G. Michailidis. 2010. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika* 97 (2010).
[223] B. Siegerink, W. den Hollander, M. Zeegers, and R. Middelburg. 2016. Causal inference in law: An epidemiological perspective. *European Journal of Risk Regulation* 7, 1 (2016), 175–186. https://doi.org/10.1017/S1867299X0000547X
[224] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. 2019. How can we fool LIME and SHAP? Adversarial attacks on post hoc explanation methods. arXiv:1911.02508v1 (2019).
[225] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich. 2011. Network modelling methods for fMRI. *NeuroImage* 54 (2011), 875–891.
[226] S. A. Sontakke, A. Mehrjou, L. Itti, and B. Schölkopf. 2020. Causal curiosity: RL agents discovering self-supervised experiments for causal representation learning. arXiv:2010.03110v1 (2020).
[227] D. J. Spiegelhalter and R. G. Cowell. 1992. *Bayesian Statistics* (4th ed.). Clarendon Press, Oxford, Chapter learning in probabilistic expert systems.
[228] P. Spirtes, C. Glymour, and R. Scheines. 2000. *Causation, Prediction, and Search* (2nd ed.). MIT Press, Cambridge, Massachusetts.
[229] P. Spirtes, C. Meek, and T. Richardson. 1995. Causal inference in the presence of latent variables and selection bias. *Proc. 11th Conf. on Uncertainty in AI* 499–506 (1995).
[230] P. Spirtes and K. Zhang. 2016. Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics* 3, 3 (2016). https://doi.org/10.1186/s40535-016-0018-x
[231] E. V. Strobl. 2018. A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias. arXiv:1805.02087v1 (2018).
[232] G. Sugihara, R. May, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch. 2012. Detecting causality in complex ecosystems. *Science* 338 (2012).
[233] X. Sun, D. Janzing, B. Schölkopf, and K. Fukumizu. 2007. A kernel-based causal learning algorithm. *ICML* (2007).
[234] R. S. Sutton and A.G. Barto. 2018. *Reinforcement Learning: An Introduction*. Bradford Books / MIT Press, Cambridge, Massachusetts.

[235] N. Tagasovska, V. Chavez-Demoulin, and T. Vatter. 2020. Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery. *Proceedings of the 37th International Conference on Machine Learning* (2020).
[236] F. Takens. 1981. *Dynamical Systems and Turbulence, Lecture notes in Mathematics 898*. Springer, Berlin, Chapter detecting strange attractors in turbulence.
[237] J. Textor, B. van der Zander, M. K. Gilthorpe, M. Liskiewicz, and G. T. H. Ellison. 2016. Robust causal inference using directed acyclic graphs: The R package 'dagitty'. *International Journal of Epidemiology* (2016).
[238] M. Teyssier and D. Koller. 2005. Ordering-based search: A simple and effective algorithm for learning Bayesian networks. *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence* (2005).
[239] S. Tikka, A. Hyttinen, and J. Karvanen. 2020. dosearch: causal effect identification from multiple incomplete data sources. https://cran.r-project.org/package=dosearch.
[240] S. Tikka and J. Karvanen. 2017. Identifying causal effects with the R package causaleffect. *Journal of Statistical Software* 76, 12 (2017), 1–30.
[241] R. E. Tillman, D. Danks, and C. Glymour. 2008. Integrating locally learned causal structures with overlapping variables. *Advances in Neural Information Processing Systems* (2008).
[242] S. Triantafillou, V. Lagani, C. Heinze-Deml, A. Schmidt, J. Tegner, and I. Tsamardinos. 2017. Predicting causal relationships from biological data: Applying automated causal discovery on mass cytometry data of human immune cells. *Nature Scientific Reports* 7 (2017).
[243] S. Triantafillou and I. Tsamardinos. 2015. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research* 16 (2015).
[244] S. Triantafillou, I. Tsamardinos, and Tollis. I. 2010. Learning causal structure from overlapping variable sets. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* (2010).
[245] I. Tsamardinos, L. E. Brown, and C. F. Aliferis. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning* 65 (2006).
[246] S. van de Geer and P. Bühlmann. 2013. l0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics* 41, 2 (2013).
[247] T. Van den Bulcke, K. Van Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. De Moor, and Marchal K. 2006. SynTReN: A generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics* 7, 43 (2006).
[248] M. J. van der Laan and S. Rose. 2011. *Targeted Learning - Causal Inference for Observational and Experimental Data.* Springer International, New York.
[249] M. J. van der Laan and S. Rose. 2018. *Targeted Learning in Data Science.* Springer International, Switzerland.
[250] M. J. van der Laan and R. J. C. M. Starmans. 2014. Entering the era of data science: Targeted learning and the integration of statistics and computational data analysis. *Advances in Statistics* (2014).
[251] G. Varando. 2020. Learning DAGs without imposing acyclicity. arXiv:2006.03005v1 (2020).
[252] T. Verma and J. Pearl. 1991. Equivalence and synthesis of causal models. *Computer Science Department, UCLA* (1991).
[253] M. J. Vowels, N. C. Camgoz, and R. Bowden. 2021. Shadow-mapping for unsupervised neural causal discovery. *IEEE Conference on Computer Vision and Pattern Recognition Causality in Vision Workshop* (2021).
[254] M. J. Vowels. 2021. Misspecification and unreliable interpretations in psychology and social science. *Psychological Methods* (2021). https://doi.org/10.1037/met0000429
[255] M. J. Vowels, N. C. Camgoz, and R. Bowden. 2020. Targeted VAE: Structured inference and targeted learning for causal parameter estimation. *Under Review* (2020).
[256] M. J. Vowels, N. C. Camgoz, and R. Bowden. 2020. NestedVAE: Isolating common factors via weak supervision. *Conference on Computer Vision and Pattern Recognition* (2020).
[257] Y. Wang and D. M. Blei. 2019. The blessings of multiple causes. arXiv:1805.06826v3 (2019).
[258] Y. Wang, V. Menkovski, H. Wang, X. Du, and M. Pechenizkiy. 2020. Causal discovery from incomplete data: A deep learning approach. *Association for the Advancement of Artificial Intelligence* (2020).
[259] D. Wei, T. Gao, and Y. Yu. 2020. DAGs with No Fears: A closer look at continuous optimization for learning Bayesian networks. *34th Conference on Neural Information Processing Systems* (2020).
[260] S. Weichwald, M. E. Jakobsen, P. B. Mogensen, L. Petersen, N. Thams, and G. Varando. 2020. Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values. arXiv:2002.09573v1 (2020).
[261] R. J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* 8 (1992), 229–256.
[262] C. K. Wongchokprasitti, H. Hochheiser, J. Espino, E. Maguire, B. Andrews, M. Davis, and C. Inskip. 2019. pycausal. *bd2kccd/py-causal* (2019). https://doi.org/10.5281/zenodo.3592985
[263] P. A. Wu and K. Fukumizu. 2020. Causal mosaic: Cause-effect inference via nonlinear ICA and ensemble method. *AISTATS* 108 (2020).

[264] F. Xie, R. Cai, B. Huang, C. Glymour, Z. Hao, and K. Zhang. 2020. Generalized independent noise condition for estimating latent variable causal graphs. *NeurIPS* (2020).
[265] G. Xu, T. D. Duong, Q. Li, S. Liu, and X. Wang. 2020. Causality learning: A new perspective for interpretable machine learning. arXiv:2006.16789 (2020).
[266] K. D. Yang, A. Katcoff, and C. Uhler. 2018. Characterizing and learning equivalence classes of causal DAGs under interventions. *ICML* (2018).
[267] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang. 2020. CausalVAE: Disentangled representation learning via neural structural causal models. arXiv:2004.08697v4 (2020).
[268] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang. 2020. A survey on causal inference. arXiv:2002.02770 (2020).
[269] H. Ye, E. Deyle, L. J. Gilarranz, and G. Sugihara. 2015. Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific Reports* 5, 14750 (2015).
[270] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum. 2020. CLEVRER: Collision events for video representation and reasoning. arXiv:1910.01442v2 (2020).
[271] W. C. Young, K. Y. Yeung, and A. E. Raftery. 2018. Identifying dynamical time series model parameters from equilibrium samples, with application to gene regulatory networks. *Statistical Modelling* 19, 4 (2018).
[272] Y. Yu, J. Chen, T. Gao, and M. Yu. 2019. DAG-GNN: DAG structure learning with graph neural networks. *Proceedings of the 36th International Conference on Machine Learning* (2019).
[273] A. Zhang, Z. C. Lipton, L. Pineda, K. Azizzadenesheli, A. Anandkumar, L. Itti, J. Pineau, and T. Furlanello. 2019. Learning causal state representations of partially observable environments. arXiv:1906.10437v1 (2019).
[274] C. Zhang, B. Chen, and J. Pearl. 2020. A simultaneous discover-identify approach to causal inference in linear models. *Proceedings of the 34th International Conference on Artificial Intelligence* (2020).
[275] J. Zhang. 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence* 172, 16–17 (2008), 1873–1896. https://doi.org/10.1016/j.artint.2008.08.001
[276] J. Zhang, D. Kumor, and E. Bareinboim. 2020. Causal imitation learning with unobserved confounders. *34th Conference on Neural Information Processing Systems* (2020).
[277] K. Zhang and A. Hyvärinen. 2009. On the identifiability of the post-nonlinear causal model. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (2009), 647–655.
[278] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. 2012. Kernel-based conditional independence test and application in causal discovery. arXiv:1202.3775 (2012).
[279] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing. 2018. DAGs with NO TEARS: Continuous optimization for structure learning. arXiv:1803.01422v2 (2018).
[280] X. Zheng, C. Dan, B. Aragam, P. Ravikumar, and E. P. Xing. 2020. Learning sparse nonparametric DAGs. arXiv:1909.13189v2 (2020).
[281] Q. Zhou. 2011. Multi-domain sampling with applications to structural inference of Bayesian networks. *J. Amer. Statist. Assoc.* 106 (2011).
[282] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. 1997. Algorithm 778: L-BFGS-B Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Soft.* 23, 4 (1997), 550–560.
[283] R. Zhu, A. Pfadler, Z. Wu, Y. Han, X. Yang, F. Ye, Z. Qian, J. Zhou, and B. Cui. 2020. Efficient and scalable structure learning for Bayesian networks: Algorithms and applications. arXiv:2012.03540v1 (2020).
[284] S. Zhu, I. Ng, and Z. Chen. 2020. Causal discovery with reinforcement learning. arXiv:1906.04477v4 (2020).

Received 3 March 2021; revised 19 November 2021; accepted 14 March 2022