Amortized Inference for Causal Structure Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Learning causal structure poses a combinatorial search problem that typically involves evaluating structures with a score or independence test. The resulting search is costly, and designing suitable scores or tests that capture prior knowledge is difficult. In this work, we propose to *amortize causal structure learning*. Rather than searching over structures directly, we train a variational inference model to predict the causal structure from observational or interventional data. This allows us to bypass both the search over graphs and the hand-engineering of suitable score functions. Instead, our inference model acquires domain-specific inductive biases for causal discovery solely from data generated by a simulator. The architecture of our inference model emulates permutation invariances that are crucial for statistical efficiency in structure learning, which facilitates generalization to significantly larger problem instances than seen during training. On synthetic data and semisynthetic gene expression data, our models exhibit robust generalization capabilities when subject to substantial distribution shifts and significantly outperform existing algorithms, especially in the challenging genomics domain.

1 Introduction

Learning the causal structure among the variables we observe is a fundamental task in various scientific disciplines (Spirtes et al., 2000; Pearl, 2009). However, inferring this causal structure from observations of the variables is a difficult inverse problem. The solution space of potential causal structures, usually modeled as directed graphs, grows superexponentially with the number of variables. To infer a causal structure, standard methods have to search over potential graphs, usually maximizing either a graph scoring function or testing for conditional independences (Heinze-Deml et al., 2018).

A key limitation of search approaches to causal discovery is the difficulty of specifying realistic inductive biases. Score-based methods use strong assumptions about the data-generating process, such as linearity (Shimizu et al., 2006), specific noise models (Hoyer et al., 2008; Peters and Bühlmann, 2014), and the absence of measurement error (e.g., Scheines and Ramsey 2016; Zhang et al. 2017), which are difficult to verify (Dawid, 2010; Reisach et al., 2021). Conversely, constraint-based methods do not have enough domain-specific inductive bias. Even with oracle conditional independence tests, they are limited to identifying equivalence classes that may be exponentially large (He et al., 2015b). Moreover, the search over directed graphs itself introduces unwanted bias and artifacts (Colombo et al., 2014). The intractable search space ultimately imposes hard constraints on the causal structure, e.g., the node degree (Spirtes et al., 2000; Koivisto and Sood, 2004), which limits the suitability of search in real-world domains.

In the present work, we propose to *amortize* causal structure learning. In other words, our goal is to optimize an inference model to solve the inverse problem of predicting a causal structure from a provided dataset. This proposal is motivated by the dichotomy that, given a causal model, it is easy to generate realistic data; the inverse direction, that is, inferring the causal structure from data, is hard. Many high-impact domains of causal discovery can be realistically forward-simulated, for example, the chemical ODEs and technical noise of gene regulatory networks (Schaffter et al., 2011; Dibaeinia and Sinha, 2020), fMRI brain responses (Buxton, 2009; Bassett and Sporns, 2017), and chemical

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

kinetics (Anderson and Kurtz, 2011; Wilkinson, 2018). We can leverage such simulators to train an inference model in causal discovery for specific data-generating processes. In doing so, amortized inference allows us to specify domain-specific inductive biases that are not easily represented by a graph scoring function and bypass the problems of structure search.

The central challenge of amortizing causal discovery consists in designing a learning rule that generalizes out-of-distribution. This generalization is essential because any simulated data distribution is unlikely to cover all real-world distribution shifts. To quantify the epistemic uncertainty in this inference process, we propose to train an amortized variational inference model for causal discovery (AVICI) by approximating the posterior over causal graphs given data from a specified data-generating distribution, such as a simulator. The model architecture of AVICI is permutation invariant and equivariant with respect to the observation and variable dimensions of the provided dataset, respectively, and generalizes to significantly larger problem instances than seen during training.

On synthetic data and semisynthetic expression data of gene regulatory networks, our approach significantly outperforms existing algorithms for causal discovery, often by a large margin. Moreover, we demonstrate that our inference models induce calibrated uncertainties and robust behavior when subject to substantial distribution shifts of graphs, mechanisms, noise, and problem sizes, which makes our trained models both reliable and versatile for future downstream use. In particular, AVICI was the only method to infer plausible causal structure from noisy gene expression data, advancing the frontiers of structure discovery in fields such as molecular biology.

2 Background and Related Work

2.1 Causal Structure

In this work, we define a causal structure in the same way as Mooij et al. (2016). Models such as causal Bayesian networks and structural causal models (Pearl, 2009; Peters et al., 2017) are less well-suited to describing systems with feedback loops, which we consider practically relevant. Specifically, we define the causal structure G as the directed graph whose nodes correspond to d variables $\mathbf{x} = (x_1, \dots, x_d)$ and whose edges represent all *direct causal* effects among the variables. A variable x_i has a direct causal effect on x_j if intervening on x_i affects the outcome of x_j independent of the other variables $\mathbf{x}_{\setminus ij} := \mathbf{x} \setminus \{x_i, x_j\}$, i.e., there exists $a \neq a'$ such that

$$p(x_j \,|\, \mathrm{do}(x_i = a, \mathbf{x}_{\setminus ij} = \mathbf{c})) \neq p(x_j \,|\, \mathrm{do}(x_i = a', \mathbf{x}_{\setminus ij} = \mathbf{c})) \tag{1}$$

for some $\mathbf{c}$. An *intervention* $\mathrm{do}(\cdot)$ denotes any active manipulation of the generative process of the observations $\mathbf{x}$. An example of interventions are gene knockouts in genomics, in which transcription rates of genes are externally set to zero. G is an abstraction of the physical processes generating the data that describes the system's behavior under arbitrary interventions (Peters et al., 2017; Rubenstein et al., 2017; Schölkopf, 2019). We emphasize that G is not necessarily acyclic; however, we later show how to apply our approach when G is constrained to be acyclic. In this work, we assume causal sufficiency, i.e., that $\mathbf{x}$ contains all common causal parents of the variables x_i (Peters et al., 2017).

2.2 Related Work

Classical methods for causal structure learning search over causal graphs and evaluate them using a likelihood or conditional independence tests (Chickering, 2003; Kalisch and Bühlman, 2007; Hauser and Bühlmann, 2012; Zheng et al., 2018; Heinze-Deml et al., 2018). Other methods combine constraint- and score-based ideas (Tsamardinos et al., 2006) or use the noise properties of an SCM that is postulated to underlie the data-generating process (Shimizu et al., 2006; Hoyer et al., 2008).

Deep learning has been applied to causal inference, e.g., for the estimation of treatment effects (Shalit et al., 2017; Louizos et al., 2017; Yoon et al., 2018) and in instrumental variable analysis (Hartford et al., 2017; Bennett et al., 2019). In causal structure learning, neural networks have been used primarily to model nonlinear causal mechanisms among the variables (Goudet et al., 2018; Yu et al., 2019; Lachapelle et al., 2020; Brouillard et al., 2020; Lorch et al., 2021). While amortized inference has gained increasing popularity in recent years (Amos, 2022), prior work applying it to causal discovery only studied narrowly defined subproblems such as the bivariate case (Lopez-Paz et al., 2015) and fixed causal mechanisms for each task (Löwe et al., 2022) or only used observational correlation coefficients for prediction (Li et al., 2020). In parallel work, Ke et al. (2022) independently perform causal discovery as a supervised learning task. There are significant

differences in their approach, in particular, our inference models generalize to significantly larger problem sizes, even when trained on realistic genomics data.

3 AVICI: Amortized Variational Inference for Causal Discovery

3.1 The Variational Objective

To amortize causal structure learning, we define a data-generating distribution $p(D)$ that models the domain over which structure learning is performed. From a generative perspective, the observations $D = \{\mathbf{x}^1, \ldots, \mathbf{x}^n\} \sim p(D)$ are produced by sampling from a distribution over causal structures $p(G)$ and then obtaining realizations of a data-generating mechanism $p(D \mid G)$. It is important to highlight that $p(D \mid G)$ does not need to represent an ancestral sampling procedure over a directed acyclic graph. In particular, real-world systems are often more naturally modeled at different granularities or as dynamical systems (Mooij et al., 2013; Hoel et al., 2013; Schölkopf, 2019). Rather, $p(D \mid G)$ characterizes a data-generating process whose causal graph G formalizes all direct causal effects (Eq. 1) of the system's mechanisms, that is, how intervening on certain variables influences the realizations of others (Mooij et al., 2016; Peters et al., 2017; Rubenstein et al., 2017).

Having observed a set of observations D in our domain, our goal is to approximate the true posterior over causal structures $p(G \mid D)$ with a variational distribution $q(G; \theta)$. To amortize this inference task for the data-generating distribution $p(D)$, we optimize an inference model f_ϕ to predict the variational parameters θ that minimize the expected *forward KL divergence* from the intractable posterior $p(G \mid D)$ to $q(G; \theta)$ for $D \sim p(D)$:

$$\min_{\phi} \; \mathbb{E}_{p(D)} \; D_{KL}\big(p(G \mid D) \big\| q\big(G; f_\phi(D)\big)\big) \tag{2}$$

Since it is not tractable to compute the true posterior in (2), we make use of ideas by Barber and Agakov (2004) and rewrite the expected forward KL to obtain an equivalent, tractable objective:

$$\begin{aligned} \mathbb{E}_{p(D)} \; D_{KL}\big(p(G \mid D) \big\| q(G; f_\phi(D))\big) &= \mathbb{E}_{p(D)} \; \mathbb{E}_{p(G \mid D)}[\log p(G \mid D) - \log q(G; f_\phi(D))] \\ &= -\mathbb{E}_{p(G)} \; \mathbb{E}_{p(D \mid G)}[\log q(G; f_\phi(D))] + \text{const.} \end{aligned} \tag{3}$$

The constant does not depend on ϕ, so we can maximize $\mathcal{L}(\phi) := \mathbb{E}_{p(G)} \; \mathbb{E}_{p(D \mid G)}[\log q(G; f_\phi(D))]$, which allows us to amortize variational inference for causal discovery (AVICI). While the data-generating distribution $p(D) = \mathbb{E}_{p(G)}[p(D \mid G)]$ can be arbitrarily complex, the objective $\mathcal{L}$ is tractable whenever we have access to the ground-truth causal graph G underlying the generative process of D, i.e., to samples from the joint distribution $p(G, D)$. In practice, $p(G)$ and $p(D \mid G)$ can thus be specified by a simulator for the domain of interest. Since the quality of simulators in many domains may improve substantially over time, the sim-to-real bias of the inference model f_ϕ can reduce along the way.

From an information-theoretic view, the objective (2) maximizes a variational lower bound on the mutual information $I[G; D]$ between the causal structure G and the observations D (Barber and Agakov, 2004). Starting from the definition of mutual information, we obtain

$$\begin{aligned} I[G; D] = H[G] - H[G \mid D] &= H[G] + \mathbb{E}_{p(G,D)}[\log p(G \mid D)] \\ &\geq H[G] + \mathbb{E}_{p(G,D)}[\log q(G; f_\phi(D))] = H[G] + \mathcal{L}(\phi) \end{aligned} \tag{4}$$

where the entropy $H[G]$ is constant. The bound is tight if $\mathbb{E}_{p(D)} D_{KL}(p(G \mid D) \| q(G; f_\phi(D))) = 0$.

3.2 Likelihood-Free Inference using the Forward KL

The AVICI objective in (3) intentionally targets the forward KL $D_{KL}(p \,\|\, q(\,\cdot\,; \theta))$, which requires optimizing $\mathbb{E}_{p(G,D)}[\log q(G; \theta)]$. This choice implies that we both model the density $q(G; \theta)$ explicitly and have access to samples from the true data-generating distribution. As long as *samples* from $p(G, D)$ are available to estimate $\mathcal{L}(\phi)$, minimizing the forward KL in (2) enables us to infer causal structures in arbitrarily complex domains—that is, even domains where it is difficult to specify an explicit likelihood $p(D \mid G)$. Furthermore, the forward KL typically yields more reliable uncertainty estimates since it does not suffer from the variance underestimation problems common to the reverse KL (Bishop and Nasrabadi, 2006).

In contrast, variational inference usually optimizes the reverse KL $D_{KL}(q(\,\cdot\,; \theta) \,\|\, p)$ with respect to θ, which involves the reconstruction term $\mathbb{E}_{q(G;\theta)}[\log p(D \mid G)]$ (Blei et al., 2017). This objective

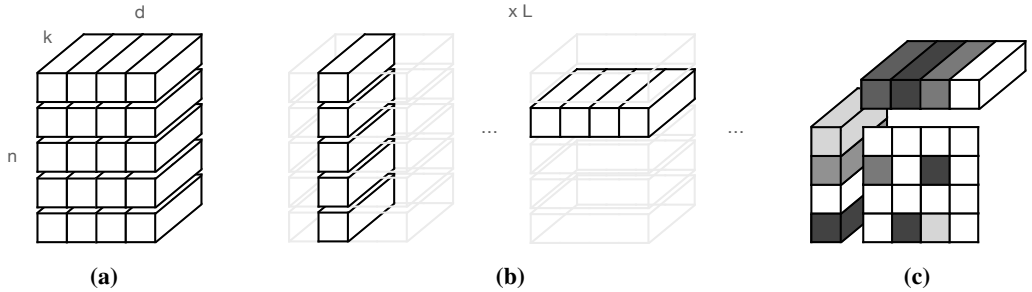


Figure 1: Model architecture. **(a)** Our model maps the input to a three dimensional tensor of shape $n \times d \times k$ and remains permutation in- and equivariant over axes n and d, respectively. **(b)** Each of the L blocks first self-attends over axis n and then over d, sharing parameters across the other axis. **(c)** The inner product of two variables' representations models the probability of a direct causal effect.

requires a tractable marginal likelihood $p(D \mid G)$. Unless inferring the mechanism parameters jointly (Brouillard et al., 2020; Lorch et al., 2021), this requirement limits inference to conjugate models with linear Gaussian or categorical mechanisms that assume zero measurement error (Geiger and Heckerman, 1994; Heckerman et al., 1995), which are not justified in practice (Friston et al., 2000; Schaffter et al., 2011; Runge et al., 2019; Dibaeinia and Sinha, 2020). Furthermore, unless the noise scale is learned jointly, likelihoods can be sensitive to the measurement scale of $\mathbf{x}$ (Reisach et al., 2021).

In the following section, we describe a choice for the variational distribution $q(G; \theta)$ and the inference model f_ϕ that predicts θ given D. After that, we detail our training procedure for optimizing the model parameters ϕ as well as for learning the inductive bias of acyclicity.

4 Inference Model

4.1 Variational Family

While any inference model that defines a density is feasible for maximizing the objective in (3), we opt to use a factorized variational family in this work.

$$q(G; \theta) = \prod_{i,j} q(g_{i,j}; \theta_{i,j}) \;\text{ with }\; g_{i,j} \sim \text{Bern}(\theta_{i,j}) \tag{5}$$

The inference model f_ϕ maps a dataset D corresponding to n samples $\{\boldsymbol{o}^1, \dots, \boldsymbol{o}^n\}$ to a d-by-d matrix θ parameterizing the variational approximation of the causal graph posterior. In addition to the joint observation $\mathbf{x}^i = (x_1^i, \dots, x_d^i)$, each sample $\boldsymbol{o}^i = (o_1^i, \dots, o_d^i)$ may contain interventional information for each variable. When interventions or gene knockouts are performed, we set $o_j^i = (x_j^i, u_j^i)$ and $u_j^i \in \{0, 1\}$ indicating whether variable j was intervened upon in sample i. Other settings could be encoded analogously, e.g., when the intervention targets are unknown.

4.2 Model Architecture

To maximize statistical efficiency, f_ϕ should be constrained to satisfy the invariances inherent to the task of learning a causal structure from data. Firstly, f_ϕ should be *permutation invariant* across the sample dimension (axis n). Shuffling the samples should have no influence on the prediction, i.e., for any permutation π, $f_\phi(\pi(\{\boldsymbol{o}\})) = f_\phi(\{\boldsymbol{o}\})$. Moreover, f_ϕ should be *permutation equivariant* across the variable dimension (axis d). Reordering the variables should permute the predicted causal edge probabilities, i.e., $f_\phi(\{\boldsymbol{o}_{\pi(1:d)}\})_{i,j} = f_\phi(\{\boldsymbol{o}_{1:d}\})_{\pi(i),\pi(j)}$. Lastly, f_ϕ should apply to any $d, n \geq 1$.

In the following, we show how to parameterize f_ϕ as a neural network that encodes these properties. After first mapping each o_j^i to a real-valued vector using a position-wise linear layer, f_ϕ operates over a continuous, three-dimensional tensor of n rows for the observations, d columns for the variables, and feature size k. Figure 1 illustrates the key components of the architecture.

Attending over axes n and d The core of f_ϕ is composed of $L = 8$ identical blocks. Each block consists of four residual layers, where the first and third apply multi-head self-attention and the second and fourth position-wise feed-forward networks, similar to the vanilla Transformer encoder (Vaswani et al., 2017). To enable information flow across all $n \times d$ tokens of the representation, the model alternates in attending over the observation and the variable dimension (Kossen et al., 2021). Specifically, the first self-attention block attends over axis n, treating axis d as a batch dimension; the second attends over axis d, treating axis n as a batch dimension. This message-passing schema keeps memory in check and results in a moderate model size because parameters are shared across the batched axes. Moreover, since multi-head attention is permutation equivariant and modules are shared across non-attended axes, the representation is permutation equivariant over axes n and d at all times (Lee et al., 2019b).

Causal graph model After building up a representation tensor from the input using the attention blocks, we max-pool over the observation axis n to obtain a representation $(\mathbf{z}^1, \ldots, \mathbf{z}^d)$ consisting of one vector $\mathbf{z}^i \in \mathbb{R}^k$ for each causal variable. Following Lorch et al. (2021), we use two position-wise linear layers to map each $\mathbf{z}^i$ to two embeddings $\mathbf{u}^i, \mathbf{v}^i \in \mathbb{R}^k$. After normalizing the embeddings $\mathbf{u}^i, \mathbf{v}^i$ to an ℓ_2-norm of one (Radford et al., 2021), we model the probability of each edge in the causal graph with an inner product:

$$\theta_{i,j} = \sigma\Big(\tau\, \mathbf{u}^i \cdot \mathbf{v}^j + b\Big) \tag{6}$$

where σ is the logistic function, b a learned bias, and τ a positive scale that is learned in $\log$ space. Since max-pooling is invariant to permutations and since the graph model in (6) permutes with respect to axis d, the inference model f_ϕ satisfies the required permutation invariance over axis n and permutation equivariance over axis d.

4.3 Acyclicity

There is no reason to generally exclude feedback loops in G. Cyclic causal effects often occur, e.g., when modeling stationary distributions of dynamical systems. However, certain domains may be more accurately modeled by acyclic causal structures (Peters et al., 2017; Rubenstein et al., 2017). While the variational family in (5) does not allow enforcing it directly, we can optimize for acyclicity through ϕ. Whenever the acyclicity prior is justified, we amend the optimization problem in (2) with the constraint that q only models acyclic graphs in expectation:

$$\mathcal{F}(\phi) := \mathbb{E}_{p(D)}\left[h(f_\phi(D))\right] = 0 \tag{7}$$

The function h is zero if and only if the predicted edge probabilities induce an acyclic graph. We use the insight by Lee et al. (2019a), who show that acyclicity is equivalent to the spectral radius ρ, i.e., the largest absolute eigenvalue, of the predicted matrix being zero. We use power iteration to approximate and differentiate through the largest eigenvalue of $f_\phi(D)$ (Golub and Van der Vorst, 2000; Lee et al., 2019a):

$$h(W) := \rho(W) \approx \frac{\mathbf{a}^\top W \mathbf{b}}{\mathbf{a}^\top \mathbf{b}} \quad \text{where for } t \text{ steps:} \quad \begin{array}{l} \mathbf{a} \leftarrow \mathbf{a}^\top W \,/\, \|\mathbf{a}^\top W\|_2 \\ \mathbf{b} \leftarrow W\mathbf{b} \,/\, \|W\mathbf{b}\|_2 \end{array} \tag{8}$$

and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ are initialized randomly. Since a few steps t are sufficient in practice, (8) scales with $O(d^2)$ and is significantly more efficient than alternative $O(d^3)$ constraints based on matrix powers (Zheng et al., 2018; Yu et al., 2019). We do not backpropagate gradients with respect to ϕ through $\mathbf{a}, \mathbf{b}$.

4.4 Optimization

Combining the objective in (3) with our inference model (5), we can directly use stochastic optimization to train the parameters ϕ of our neural network model. The expectations over $p(G, D)$ inside $\mathcal{L}$ and $\mathcal{F}$ are approximated using samples from the data-generating process of the domain, e.g., a simulator. In practice, the available data can be further augmented using bootstrapping. When imposing an acyclicity prior on G, we solve the constrained program $\max_\phi \mathcal{L}(\phi)$ s.t. $\mathcal{F}(\phi) = 0$ through its dual formulation (Nandwani et al., 2019):

Algorithm 1 Training the AVICI inference model f_ϕ

Parameters: ϕ variational, λ dual, η step size
while not converged **do**
  for l **steps do**
    $\Delta\phi \propto \nabla_\phi\big(\mathcal{L}(\phi) - \lambda\mathcal{F}(\phi)\big)$
  $\lambda \leftarrow \lambda + \eta\mathcal{F}(\phi)$

$$\min_\lambda \max_\phi \; \mathcal{L}(\phi) - \lambda\mathcal{F}(\phi) \tag{9}$$

Algorithm 1 summarizes the general optimization procedure for q_ϕ, which converges to a local optimum under regularity conditions on the learning rates (Jin et al., 2020). Without an acyclicity constraint, training reduces to the primal updates of ϕ with $\lambda = 0$.

5 Experimental Setup

Evaluating causal discovery algorithms is difficult since there are few interesting real-world datasets that come with ground-truth causal structure. Often, the believed ground truths may be incomplete or change as expert knowledge improves (Schaffter et al., 2011; Mooij et al., 2020). Following prior work, we deal with this difficulty by evaluating our approach using simulated data with known causal structure and by controlling for various aspects of the task. In Appendix E, we additionally report results on a real-world proteomics dataset (Sachs et al., 2005).

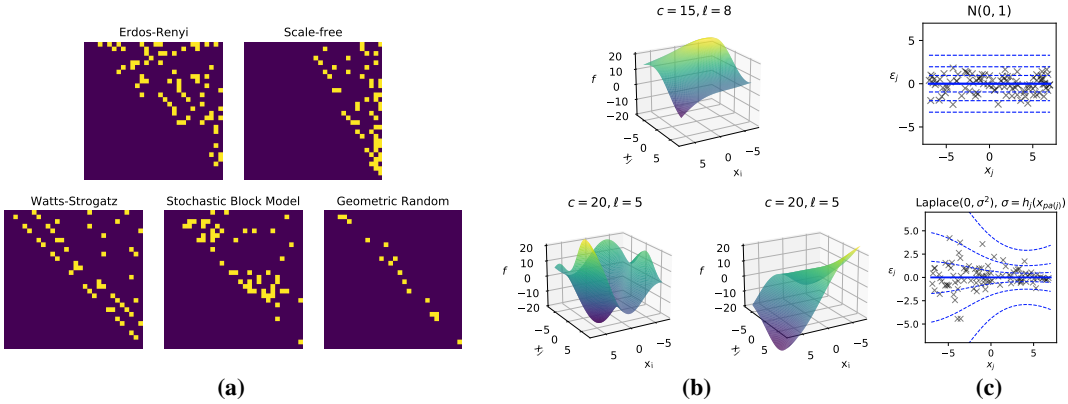


Figure 2: Moving out-of-distribution in the RFF domain. Randomly sampled data-generating components of the nonlinear SCM domain during training $p(D)$ (top) and o.o.d. evaluation $\tilde{p}(D)$ (bottom). For visualization, the adjacency matrices **(a)** are topologically sorted, the causal mechanisms **(b)** have two parents, where c and ℓ are output and length scales of the underlying GP, and the noise **(c)** is shown as a function of one parent, where dashed lines indicate 0.66, 0.95, and 0.999 coverage.

5.1 Domains and Simulated Components

Our goal is to evaluate our inference models' ability to learn inductive biases in previously intractable domains and to generalize robustly beyond the data distribution $p(D)$ seen during training. We study three domains: two classes of structural causal models (SCMs) as well as semisynthetic single-cell expression data of gene regulatory networks (GRNs). To study the generalization of AVICI beyond the training distribution $p(D)$, we carefully construct a test distribution $\tilde{p}(D)$ that incurs substantial shift from $p(D)$ in terms of the underlying causal structures, mechanisms, and noise components.

Data-generating processes $p(D \mid G)$ The data-generating process of SCMs is defined by recursively sampling $x_j \leftarrow f_j(\mathbf{x}_{\text{pa}(j)}, \epsilon_j)$ based on a set of mechanisms f_j and exogenous noise distributions $p_j(\epsilon_j)$ for each variable. We consider the two domains of linear functions (LINEAR) and of nonlinear functions of random Fourier features (RFF) that correspond to functions drawn from a Gaussian process with squared exponential kernel (Rahimi and Recht, 2007). In the out-of-distribution (o.o.d.) setting $\tilde{p}(D)$, we sample the linear function and kernel parameters from the tails of $p(D)$ and unseen value ranges. Moreover, while simulating additive, homoscedastic Gaussian noise ϵ_j in the training distribution $p(D)$, we employ heteroscedastic Cauchy and Laplacian noise o.o.d., where $f_j(\mathbf{x}_{\text{Pa}(j)}, \epsilon_j) = f_j(\mathbf{x}_{\text{pa}(j)}) + h_j(\mathbf{x}_{\text{pa}(j)})\epsilon_j$ for some random, input-dependent function h_j. Interventions are performed by clipping a variable x_j to a random value during data generation.

In addition to SCMs, we consider the challenging domain of GRNs (GRN) using the simulator of Dibaeinia and Sinha (2020). Contrary to SCMs, single-cell gene expression samples correspond to draws from the steady state of a stochastic dynamical system that varies between cell types (Huynh-Thu and Sanguinetti, 2019). Like LINEAR and RFF, the simulator parameters considered o.o.d. are drawn from significantly wider ranges than for $p(D)$. The simulator additionally models several technical noise effects, resulting in the observation of highly zero-inflated count matrices D. When moving o.o.d., we use the noise levels of different single-cell RNA sequencing technologies, as calibrated by Dibaeinia and Sinha (2020) on real datasets. In GRN, interventions correspond to gene knockouts and are performed by forcing the transcription rate of a variable x_j to zero.

Whenever a method infers causal structure from interventional data in our experiments, half of the dataset consists of observational data and half of single-variable interventions. In LINEAR and RFF, these single-variable interventions are performed on a subset of target variables containing 50% of the nodes. Since GRN is more challenging, the interventional data here contains single-node interventions on all variables. All datasets are standardized (Reisach et al., 2021).

Causal structures $p(G)$ Following prior work, we use random graph models and known biological networks to generate ground-truth causal structures. In all three domains, the training data distribution $p(D)$ is induced by simple Erdős-Rényi and scale-free graphs of various degrees (Erdős and Rényi, 1959; Barabási and Albert, 1999). In the o.o.d. setting, the data distributions $\tilde{p}(D)$ of the LINEAR and RFF domains are simulated using causal structures from unseen graph models: the Watts-Strogatz model, capturing small-world phenomena (Watts and Strogatz, 1998); the stochastic block model, generalizing Erdős-Rényi to community structures (Holland et al., 1983); and geometric random graphs, modeling connectivity based on spatial distance (Gilbert, 1961). In the GRN domain, we use subgraphs of the known *S. cerevisiae* and *E. coli* GRNs and their effect signs whenever known.

To extract these subgraphs, we use the procedure by Marbach et al. (2009) to maintain structural patterns like motifs and modularity (Ravasz et al., 2002; Shen-Orr et al., 2002).

To illustrate the distribution shift from $p(D)$ to $\tilde{p}(D)$, Figure 2 shows a set of graph, mechanism, and noise distribution samples in the RFF domain. The parameter configurations and functions defining $p(D)$ and $\tilde{p}(D)$ in the three domains are given in Appendix A. We also provide details on the simulator by Dibaeinia and Sinha (2020) and subgraph extraction (Marbach et al., 2009).

5.2 Evaluation Metrics

To asses how well a predicted causal structure reflects the ground-truth graph, we report the structural Hamming distance (SHD) and the structural intervention distance (SID) (Peters and Bühlmann, 2015). While the SHD simply reflects the graph edit distance, the SID quantifies the closeness of two graphs in terms of their interventional adjustment sets. For these metrics, as well as for edge precision, recall, and F1 score, we collapse the posterior probabilities predicted by AVICI to hard predictions using a decision threshold of 0.5. In addition, we follow existing work and evaluate the uncertainty estimates by computing the areas under the precision-recall curve (AUPRC) and the receiver operating characteristic (AUROC) (Friedman and Koller, 2003). More details on the metrics are given in Appendix B.

5.3 Inference Model Configuration

We train three inference models overall, one for each domain, and perform all experiments on these three trained models. Performance could be further improved by training models for more narrow settings, e.g., specializing solely on inference from observational data. However, to demonstrate the versatility of AVICI, we use the same models for the evaluation with only observational data as well as with a combination of observational and interventional data. The datasets sampled from $p(D)$ during training have $d = 2$ to 50 variables and $n = 200$ samples, where with probability 0.5, these observations contain 50 interventional samples. All experiments throughout this paper are conducted on datasets that AVICI has never seen during training, regardless of whether we evaluate the predictive performance in-distribution or o.o.d.

The inference models in the three domains share identical hyperparameters for the architecture and optimization, except for the dropout rate. We add the acyclicity constraint for the SCM domains LINEAR and RFF. Details on the optimization and architecture are given in Appendix C.

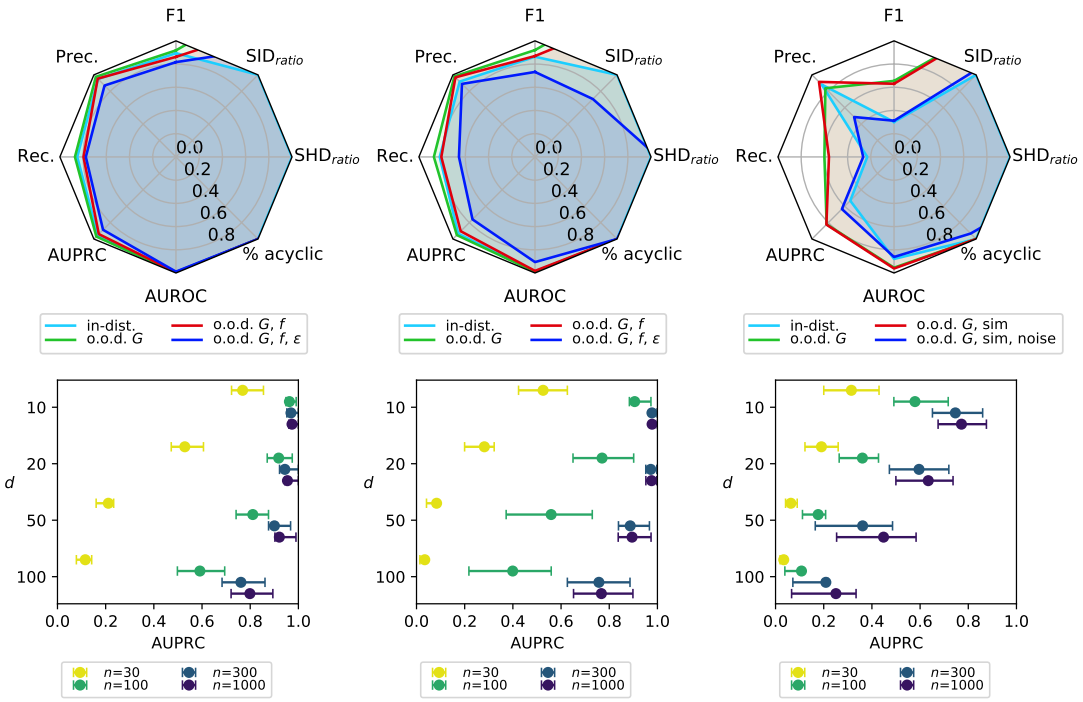
6 Experimental Results

6.1 Out-Of-Distribution Generalization

In the first part of our experiments, we study the generalization capabilities of our inference models across the spectrum of test distributions described in Section 5.1. This analysis consists of two parts. First, we test the model's sensitivity to queries with o.o.d. data. Second, we investigate how well it generalizes to different problem sizes.

Sensitivity to distribution shift We perform causal discovery from $n = 1000$ observations in systems of $d = 30$ variables. Starting from the training distribution $p(D)$, we incrementally introduce the described o.o.d. shifts in the causal structures, causal mechanisms, and finally noise, where fully o.o.d. corresponds to $\tilde{p}(D)$. The top row of Figure 3 visualizes the results of an empirical sensitivity analysis. The radar plots disentangle how combinations of the three o.o.d. aspects, i.e., graphs, mechanisms, and noise, affect the empirical performance of the AVICI models differently on the three domains LINEAR, RFF, and GRN. In addition to the metrics in Section 5.2, we also report the percentage of predicted graphs that are acyclic. To make the structural distances comparable, we use their ratios $\text{SID}_{\text{ratio}} := \text{SID}_{\text{in-dist.}}/\text{SID}$ and $\text{SHD}_{\text{ratio}}$, respectively. Given this, all metrics are in a similar value range and higher values indicate better performance.

As LINEAR is the easiest domain, AVICI performs very well in all metrics and hardly suffers from performance degradation due to distribution shift. In contrast, GRN is the most challenging, with performance degrading more significantly far o.o.d. We observe that AVICI can perform better under certain distribution shifts than in-distribution, e.g., in GRN. This is because AVICI is empirically better at predicting edges adjacent to large-degree nodes, a predominant feature of the GRN graph distribution extracted from *E. coli* and *S. cerevisiae* that is not present in Erdős-Rényi graphs of the training distribution. The acyclicity constraint is perfectly satisfied for LINEAR and RFF.



(a) LINEAR **(b)** RFF **(c)** GRN

Figure 3: Generalization properties of the inference model f_ϕ. Top row plots show performance metrics of AVICI under increasing distributional shift given $n = 1000$ observations for $d = 30$ variables. $\text{SID}_{\text{ratio}}$ is defined as $\text{SID}_{\text{in-dist.}}/\text{SID}$ and analogously for SHD. Bottom row shows the in-distribution AUPRC for various d as we vary the number of observations provided to AVICI. All values are the mean over fifteen random task instances and AVICI predicts from mixed observational and interventional data. For all metrics, higher is better. Error bars indicate the interquartile range.

We observe that AUPRC and AUROC do not suffer as much from distributional shift as the metrics based on thresholded point estimates. This indicates that the marginal posterior edge probabilities convey useful information about how certain AVICI is about individual edge predictions and how likely it is that the thresholded prediction is wrong. Overall, AVICI is surprisingly robust, even under substantial distribution shifts of all aspects of the data distribution at the same time. In Appendix E.1

Generalization to different problem sizes In addition to the sensitivity to distribution shift, we are interested in the ability of our inference models to generalize to unseen problem sizes in terms of the number of variables d and the number of data points n. The bottom row of Figure 3 illustrates the AUPRC for the edge predictions of AVICI when varying d and n on unseen in-distribution data. The predictions improve with the number of data points while exhibiting diminishing marginal improvement when seeing additional data. Moreover, the performance decreases smoothly as the number of variables increases and the task becomes harder. Most importantly, this robust behavior can be observed well beyond the settings ($n = 200$ and $d \leq 50$) used during training.

6.2 Benchmarking

Next, we compare AVICI with existing causal discovery algorithms that use purely observational data or a combination of observational and interventional data. The benchmark is performed on the fully o.o.d. domain distributions $\tilde{p}(D)$, i.e., under distribution shifts on causal graphs, mechanisms, and noise distributions w.r.t. the training distribution of AVICI. For each class of methods, we consider at least one constraint-based and one score-based approach as well as one recent algorithm using neural networks to model nonlinear mechanisms.

In the observational setting, we compare AVICI with the PC algorithm (Spirtes et al., 2000), greedy equivalence search (GES) (Chickering, 2003), LiNGAM (Shimizu et al., 2006), and DAG-GNN (Yu et al., 2019). When using interventional data, we compare with IGSP (Wang et al., 2017), greedy interventional equivalence search (GIES) (Hauser and Bühlmann, 2012), and DCDI (Brouillard et al., 2020). For each domain, we independently calibrate important hyperparameters of each baseline method in a grid search on held-out task instances. When computing the evaluation metrics, we favor

Table 1: Benchmark results ($d = 30$ variables). Mean SID (↓) and F1 score (↑) with standard error of all methods on 30 random task instances. Methods in the top section use only observational data, in the bottom section both observational and interventional data. We highlight the best result of each section and those within its 95% confidence interval according to an unequal variances t-test.

	LINEAR		RFF		GRN	
Algorithm	SID	F1	SID	F1	SID	F1
GES	**215.6** (35.0)	0.548 (0.03)	**346.3** (44.4)	0.285 (0.03)	**573.6** (29.2)	**0.058** (0.01)
LiNGAM	413.4 (48.4)	0.369 (0.04)	410.3 (47.6)	0.238 (0.02)	**617.5** (31.7)	**0.044** (0.01)
PC	400.5 (53.7)	0.338 (0.03)	**370.1** (51.2)	0.421 (0.03)	**594.0** (30.0)	**0.061** (0.01)
DAG-GNN	474.5 (50.8)	0.154 (0.01)	425.3 (50.2)	0.221 (0.03)	**588.7** (36.6)	**0.078** (0.02)
AVICI (ours)‖	**145.6** (21.5)	**0.672** (0.04)	**255.1** (48.2)	**0.618** (0.06)	**641.7** (34.7)	0.000 (0.00)
GIES	**120.8** (26.2)	0.736 (0.03)	**304.8** (44.0)	0.338 (0.04)	545.5 (26.9)	0.092 (0.01)
IGSP	244.0 (34.4)	0.559 (0.02)	374.1 (45.0)	0.407 (0.04)	597.4 (31.7)	0.057 (0.01)
DCDI	383.5 (45.1)	0.327 (0.03)	**282.8** (46.3)	0.409 (0.04)	590.9 (30.6)	0.075 (0.02)
AVICI (ours)‖	**110.9** (19.3)	**0.819** (0.02)	**192.7** (44.8)	**0.707** (0.06)	**416.9** (47.1)	**0.338** (0.06)

‖ Point estimate using decision threshold 0.5

methods that only predict (interventional) Markov equivalence classes by orienting undirected edges correctly when present in the ground truth. Details on the baselines are given in Appendix D.

Table 1 shows the SID and F1 scores of all methods given $n = 1000$ observations for $d = 30$ variables. The AVICI model trained on LINEAR outperforms all baselines, both given only observational and in combination with interventional data, despite operating under significant distribution shift. Only GIES achieves comparable accuracy. The same holds for RFF, where constraint-based and nonlinear approaches outperform GES and GIES, whose scores assume linearity, but ultimately do not reach the accuracy set by AVICI. In the GRN domain, the inductive bias encoded by classical methods reaches its limits. However, provided interventional data, AVICI can use its learned inductive bias to infer plausible causal structures from the noisy gene expressions, even under distribution shift, which is a promising step towards reliable structure discovery in fields like molecular biology. Even without gene knockout data, the edge probabilities predicted by AVICI achieve nontrival AUROC and AUPRC while classical methods predict close to randomly (Table 7 in Appendix E; see also Dibaeinia and Sinha 2020; Chen and Mar 2018).

Results for $d = 100$ variables are given in Appendix E.2. In Appendix E.3 and E.4, we also report benchmarking results on a real-world proteomics dataset (Sachs et al., 2005) and give additional metrics that evaluate the uncertainty quantification of AVICI and the baselines.

7 Discussion

We proposed AVICI, a method for inferring causal structure by performing amortized variational inference over an arbitrary data-generating distribution. Our approach leverages the insight that inductive biases crucial for statistical efficiency in structure learning might be more easily encoded in a domain simulator than in an inference technique. This is reflected in our experiments, where AVICI solves structure learning problems in complex domains intractable for existing approaches (Dibaeinia and Sinha, 2020). Our method can likely be extended to other typically difficult domains, including structure learning settings where we cannot assume causal sufficiency (Bhattacharya et al., 2021).

Using AVICI comes with several trade-offs. First, while optimizing the dual program empirically induces acyclicity, this constraint is not satisfied with certainty. Moreover, similar to most amortized inference techniques (Amos, 2022), AVICI gives no theoretical guarantees of performance. Some classical methods can do so in the infinite sample limit given specific assumptions on the data-generating process (Peters et al., 2017). However, future work might achieve guarantees for AVICI that are similar to learning theory results for the bivariate causal discovery case (Lopez-Paz et al., 2015).

Finally, while causal discovery with AVICI amounts to a single forward pass of f_ϕ at inference time, the inference models can be intensive to train. To mitigate this training cost and make our results accessible to other researchers, we release the trained models studied in this work. Our experiments demonstrate that our inference models are highly robust to simulator misspecification, suggesting that the trained models could be useful even in causal structure learning tasks outside the domains studied in this paper. In this context, fine-tuning a pretrained AVICI model on labeled real-world datasets is a promising avenue of future work.

References

Amos, B. (2022). Tutorial on amortized optimization for learning to optimize over continuous domains. *arXiv preprint arXiv:2202.00665*.

Anderson, D. F. and Kurtz, T. G. (2011). Continuous time markov chain models for chemical reaction networks. In *Design and analysis of biomolecular circuits*, pages 3–42. Springer.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barber, D. and Agakov, F. (2004). The IM algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16(320):201.

Bassett, D. S. and Sporns, O. (2017). Network neuroscience. *Nature neuroscience*, 20(3):353–364.

Bennett, A., Kallus, N., and Schnabel, T. (2019). Deep generalized method of moments for instrumental variable analysis. *Advances in neural information processing systems*, 32.

Bhattacharya, R., Nagarajan, T., Malinsky, D., and Shpitser, I. (2021). Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322. PMLR.

Bishop, C. M. and Nasrabadi, N. M. (2006). Approximate inference. In *Pattern recognition and machine learning*, volume 4, chapter 10. Springer.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. (2018). JAX: composable transformations of Python+NumPy programs. http://github.com/google/jax.

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. (2020). Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877.

Buxton, R. B. (2009). *Introduction to functional magnetic resonance imaging: principles and techniques*. Cambridge university press.

Chen, S. and Mar, J. C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC bioinformatics*, 19(1):1–21.

Chickering, D. M. (2003). Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554.

Colombo, D., Maathuis, M. H., et al. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782.

Dawid, A. P. (2010). Beware of the DAG! In *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, pages 59–86.

DeGroot, M. H. and Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22.

Dibaeinia, P. and Sinha, S. (2020). Sergio: a single-cell expression simulator guided by gene regulatory networks. *Cell Systems*, 11(3):252–271.

Erdős, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae*, 6:290–297.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine learning*, 31(1):1–38.

Friedman, N., Goldszmidt, M., and Wyner, A. (1999). Data analysis with bayesian networks: A bootstrap approach. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 196–205, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Friedman, N. and Koller, D. (2003). Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. *Machine Learning*, 50(1):95–125.

Friston, K. J., Mechelli, A., Turner, R., and Price, C. J. (2000). Nonlinear responses in fmri: the balloon model, volterra kernels, and other hemodynamics. *NeuroImage*, 12(4):466–477.

Geiger, D. and Heckerman, D. (1994). Learning gaussian networks. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*, UAI'94, page 235–243, San Francisco, CA, USA.

Gilbert, E. N. (1961). Random plane networks. *Journal of the society for industrial and applied mathematics*, 9(4):533–543.

Golub, G. H. and Van der Vorst, H. A. (2000). Eigenvalue computation in the 20th century. *Journal of Computational and Applied Mathematics*, 123(1-2):35–65.

Goudet, O., Kalainathan, D., Caillou, P., Guyon, I., Lopez-Paz, D., and Sebag, M. (2018). Causal generative neural networks. *arXiv preprint arXiv:1711.08936*.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR.

Hauser, A. and Bühlmann, P. (2012). Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1):2409–2464.

He, K., Zhang, X., Ren, S., and Sun, J. (2015a). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

He, Y., Jia, J., and Yu, B. (2015b). Counting and exploring sizes of markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 16(1):2589–2609.

Heckerman, D., Geiger, D., and Chickering, D. M. (1995). Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, 20(3):197–243.

Heinze-Deml, C., Maathuis, M. H., and Meinshausen, N. (2018). Causal structure learning. *Annual Review of Statistics and Its Application*, 5:371–391.

Hennigan, T., Cai, T., Norman, T., and Babuschkin, I. (2020). Haiku: Sonnet for JAX. http://github.com/deepmind/dm-haiku.

Hoel, E. P., Albantakis, L., and Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49):19790–19795.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137.

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

Huynh-Thu, V. A. and Sanguinetti, G. (2019). Gene regulatory network inference: an introductory survey. In *Gene Regulatory Networks*, pages 1–23. Springer.

Jin, C., Netrapalli, P., and Jordan, M. (2020). What is local optimality in nonconvex-nonconcave minimax optimization? In *International conference on machine learning*, pages 4880–4889. PMLR.

Kalainathan, D., Goudet, O., and Dutta, R. (2020). Causal discovery toolbox: Uncovering causal relationships in python. *J. Mach. Learn. Res.*, 21:37–1.

Kalisch, M. and Bühlman, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3).

Ke, N. R., Chiappa, S., Wang, J., Bornschein, J., Weber, T., Goyal, A., Botvinic, M., Mozer, M., and Rezende, D. J. (2022). Learning to induce causal structure. *arXiv:2204.04875*.

Koivisto, M. and Sood, K. (2004). Exact Bayesian structure discovery in Bayesian networks. *The Journal of Machine Learning Research*, 5:549–573.

Kossen, J., Band, N., Lyle, C., Gomez, A. N., Rainforth, T., and Gal, Y. (2021). Self-attention between datapoints: Going beyond individual input-output pairs in deep learning. *Advances in Neural Information Processing Systems*, 34:28742–28756.

Lachapelle, S., Brouillard, P., Deleu, T., and Lacoste-Julien, S. (2020). Gradient-based neural dag learning. In *International Conference on Learning Representations*.

Lee, H.-C., Danieletto, M., Miotto, R., Cherng, S. T., and Dudley, J. T. (2019a). Scaling structural learning with no-bears to infer causal transcriptome networks. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020*, pages 391–402. World Scientific.

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., and Teh, Y. W. (2019b). Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR.

Li, H., Xiao, Q., and Tian, J. (2020). Supervised whole dag causal discovery. *arXiv preprint arXiv:2006.04697*.

Lopez-Paz, D., Muandet, K., Schölkopf, B., and Tolstikhin, I. (2015). Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*, pages 1452–1461. PMLR.

Lorch, L., Rothfuss, J., Schölkopf, B., and Krause, A. (2021). DiBS: Differentiable Bayesian structure learning. *Advances in Neural Information Processing Systems*, 34.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30.

Löwe, S., Madras, D., Zemel, R., and Welling, M. (2022). Amortized causal discovery: Learning to infer causal graphs from time-series data. *arXiv preprint arXiv:2006.10833*, 140:1–24.

Marbach, D., Schaffter, T., Mattiussi, C., and Floreano, D. (2009). Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of computational biology*, 16(2):229–239.

Mooij, J. M., Janzing, D., and Schölkopf, B. (2013). From ordinary differential equations to structural causal models: The deterministic case. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 440–448, Arlington, Virginia, USA. AUAI Press.

Mooij, J. M., Magliacane, S., and Claassen, T. (2020). Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. (2016). Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204.

Nandwani, Y., Pathak, A., and Singla, P. (2019). A primal dual formulation for deep learning with constraints. *Advances in Neural Information Processing Systems*, 32.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. PMLR.

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

Peters, J. and Bühlmann, P. (2014). Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228.

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

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

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555.

Reisach, A. G., Seiler, C., and Weichwald, S. (2021). Beware of the simulated DAG! varsortability in additive noise models. *Advances in Neural Information Processing Systems*.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Rubenstein, P. K., Weichwald, S., Bongers, S., Mooij, J. M., Janzing, D., Grosse-Wentrup, M., and Schölkopf, B. (2017). Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, page ID 11.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., et al. (2019). Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):1–13.

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

Schaffter, T., Marbach, D., and Floreano, D. (2011). Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270.

Scheines, R. and Ramsey, J. (2016). Measurement error and causal discovery. In *CEUR workshop proceedings*, volume 1792, page 1. NIH Public Access.

Schölkopf, B. (2019). Causality for machine learning. *arXiv preprint arXiv:1911.10500*.

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, 31(1):64–68.

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

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass, 2nd ed edition.

Squires, C., Belyaeva, A., Karnik, S., Saeed, B., Jablonski, K. P., and Uhler, C. (2018). Causaldag. https://github.com/uhlerlab/causaldag.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y., Solus, L., Yang, K., and Uhler, C. (2017). Permutation-based causal inference algorithms with interventions. *Advances in Neural Information Processing Systems*, 30.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442.

Wilkinson, D. J. (2018). *Stochastic modelling for systems biology*. Chapman and Hall/CRC.

Yoon, J., Jordon, J., and Van Der Schaar, M. (2018). Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., and Hsieh, C.-J. (2019). Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*.

Yu, Y., Chen, J., Gao, T., and Yu, M. (2019). DAG-GNN: DAG structure learning with graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7154–7163. PMLR.

Zhang, K., Gong, M., Ramsey, J., Batmanghelich, K., Spirtes, P., and Glymour, C. (2017). Causal discovery in the presence of measurement error: Identifiability conditions. *arXiv preprint arXiv:1706.03768*.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See in particular the discussion section for the trade-offs with using our approach versus other causal structure learning algorithms.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We see our contribution as a purely methodological contribution to causal structure learning that is unlikely to have any direct negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provide supplemental material with a README file containing instructions for reproducing the results.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We give an overview of the key details of the experiment setup in the main body of the paper. Full technical details for reproducing the results can be found in the appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We report error bars for all experiments. For our method, results are reported for a fixed model so error bars capture only noise due to randomness across tasks, since retraining is too computationally expensive.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We use a simulator by Dibaeinia and Sinha (2020) to train and evaluate our models (GNU General Public License v3.0). For approaches that we compare our method against, we use code assets that implement those approaches. These assets are Causal Discovery Toolbox (MIT Licence) (Kalainathan et al., 2020), Brouillard et al. (2020) (MIT License), Yu et al. (2019) (Apache License 2.0), the CausalDAG package (Squires et al., 2018) (3-Clause BSD license). The E.coli and Yeast subnetworks used in experiments are from Schaffter et al. (2011) (MIT License).
   (b) Did you mention the license of the assets? [Yes] Listed above.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The code to reproduce our experiments is referenced in the supplementary material. We will also release the trained models from our experiments.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]