Distributions over DAGs for Causal Discovery: Limitations of Expressiveness

Simon Rittel[1,2,3] (✉) and Sebastian Tschiatschek[4]

[1] Ludwig-Maximilians-Universität München, Germany simon.rittel@lmu.de
[2] Munich Center for Machine Learning (MCML), Munich, Germany
[3] UniVie Doctoral School Computer Science, Vienna, Austria
[4] Faculty of Computer Science, University of Vienna, Austria
sebastian.tschiatschek@univie.ac.at

Abstract. Bayesian approaches for causal discovery can—in principle—quantify uncertainty in the prediction of the underlying causal structure, typically modeled by a directed acyclic graph (DAG). Various semi-implicit models for parametrized distributions over DAGs have been proposed, but their limitations have not been studied thoroughly. In this work, we focus on the expressiveness of parametrized distributions over DAGs in the context of causal discovery. We show several limitations of candidate models in a theoretical analysis and validate them empirically in supervised settings. To overcome these limitations, we propose using mixture models of the considered distributions over DAGs.

This workshop paper has been updated to an extended version [25].

1 Introduction

Causal discovery, also known as *causal structure learning* (CSL), is the task of uncovering cause and effect relations among modeled variables based on observed data [7, 29]. The inferred structures govern the translation of a causal estimand into a statistical one that can be measured from data and, hence, provide a basis for causal inference [15, 19]. Errors in causal discovery can imply different statistical estimands and bias the analysis of causal queries. This dependence highlights the importance of quantifying the uncertainty of a predicted causal structure. Bayesian CSL goes beyond identifying only a single, potentially incorrect, causal structure and rather models the uncertainty over the true causal structure by a distribution. It incorporates prior knowledge in the form of a prior distribution and in principle enables the computation of a posterior distribution over possible causal structure when new evidence is presented. Typical assumptions for causal discovery are causal sufficiency, the absence of selection bias, and acyclicity of the causal graph, allowing to represent the causal structure by a *directed acyclic graph* (DAG). When modeling the data-generating mechanism for such acyclic causal model, two sources of error can be distinguished. The *approximation error* results from the finite amount of evidence for the prediction of the possibly nonlinear relations among the observed variables and decreases

with increasing data size. Conversely, the model choices for the functional relationships and the distribution over the causal graph introduce a *model error* that cannot be overcome with more data. Novel Bayesian CSL algorithms often differ not only in a single model choice, but multiple ones and seldom identify in ablation studies for all of them which key detail is responsible for the reported increase in performance in contrast to competing algorithms. While a branch of research focuses exclusively on the functional relationships and parametric forms that allow identification of cause and effect pairs [8, 9, 12, 27, 34], distributions over causal graphs, e.g., DAGs, have received only little attention yet.

Contributions In our work, we investigate and compare the expressiveness of distributions over DAGs used in recent Bayesian CSL algorithms [1, 3–5, 13, 23, 32]. We highlight their limitations to assign equal probabilities to graphs of the same Markov equivalence class (Example 1 or capture dependences between edges in the graphs (Example 2) and compare their ability to match synthetic graph distributions. In addition, we provide experimental and theoretical evidence that probabilistic models [5, 23, 32] and mixtures of them are more expressive simpler particle distributions over a substantially higher number of graphs. We believe our work will be helpful to researchers and practitioners alike by demonstrating shortcomings of recently proposed distributions over DAGs that can limit the applicability of Bayesian CSL and proposing mixture models as an effective countermeasure. In addition, we provide a recipe to evaluate the probability mass function of a given graph, a directed path or a subgraph, that is induced by a generative model by marginalization over an auxiliary random structure using importance samples.

Structure of the paper We begin in section 2 with the presentation of the considered distributions over DAGs and discuss their theoretical limitations. In section 3, we outline an efficient evaluation of generative models based on importance sampling and derive analytically the minimal statistical divergences for particle distributions with K graphs. Finally, we apply both techniques to supervised learning of different target distributions in section 4 and conclude our paper in section 5. Due to page limitation, we explain in Appendix section A the used notation and provide necessary background on causal discovery.

2 Distributions over DAGs

The number of DAGs for a given number of variables D is finite, but super-exponential in D [18, 30]. Consequently, using categorical distributions that distribute the probability mass freely among all graphs quickly becomes infeasible, even for small D. However, for many applications, it is not necessary to specify arbitrary probabilities to all DAGs, i.e. a smaller degree of freedom may suffice. For causal discovery, the posterior distribution is expected to concentrate on graphs that are *similar* to the ground truth graph according to which the observed data was generated. This motivates the design of probabilistic models

Table 1: Overview of different candidates for distributions over DAGs alongside the number of learnable parameters which depends of the number of variables D, particles K, hidden neurons H_N as well as embedding and key size H_E and H_K. Their default values are reported in section E

Graph model	Figure	Equation	# Learnable parameters
DAG with independent edges	1a	(2)	$D(D-1)$
Graph particles	1b	(3)	$K\left(D(D-1)+1\right)$
DPM-DAG [23]	1c	(5)	$(D-1)+D(D-1)$
ARCO-DAG [32]	1c	(6)	$H_N(D^2+1)+(H_N+1)D+D(D-1)$
GFlowNet-DAG [5]	1e	(7)	$2DH_E+7(H_LD^2+16(H_EH_K+H_K^2))+4H_E^2$

for graphs that are flexible enough to model any possible DAG $\boldsymbol{G}$ and assign probability mass to some candidate graphs but require substantially fewer parameters than a general categorical distribution.
In the following, we discuss different models for distributions over DAGs proposed in recent works and investigate their expressiveness. To enhance comprehensibility, we introduce them by their generative model and provide the corresponding probability mass functions in section B. As a summary, we visualize their generative models in Figure 1 and provide a list of the number of learnable parameters for each model in Table 1. To avoid limitations by specific functional relationships or subjective prior knowledge, we consider independence relations as a general desideratum. In particular, we begin the investigation of the expressiveness of the candidate graph distributions with the following Markov equivalence class that serves as a running example.

Example 1. For a parametric linear model with exogenous Gaussian noise, the true causal graph is identifiable only up to its Markov equivalence class [20]. Consider the simple case with three variables and the chain graph $\mathrm{X}_1 - \mathrm{X}_2 - \mathrm{X}_3$ as the Markov equivalence class (MEC). Its edges cannot be further oriented even in the asymptotic limit of infinite data. To represent the corresponding uncertainty over the true DAG, all graphs of this MEC, the common cause, $\boldsymbol{G}^{(1)} : \mathrm{X}_1 \leftarrow \mathrm{X}_2 \rightarrow \mathrm{X}_3$, and the two causal chains, $\boldsymbol{G}^{(2)} : \mathrm{X}_1 \rightarrow \mathrm{X}_2 \rightarrow \mathrm{X}_3$ and $\boldsymbol{G}^{(3)} : \mathrm{X}_1 \leftarrow \mathrm{X}_2 \leftarrow \mathrm{X}_3$, should be assigned the probability $\frac{1}{3}$.

2.1 Independent edges

The arguably simplest probabilistic model for a distribution over a directed, not necessarily acyclic graph $\boldsymbol{A}$ consists of a product over independent Bernoulli probabilities, each of them modeling a possible directed edge $\mathrm{X}_i \rightarrow \mathrm{X}_j$:

$$\mathbf{A} \sim q_{\boldsymbol{\phi}}(\boldsymbol{A}) \quad \text{with} \quad \mathrm{A}_{ij} \sim q_{\phi_{ij}}(A_{ij}) \,, \tag{1}$$

where $\boldsymbol{\phi}$ are the parameters of the Bernoulli distributions over the edges. While self-loops can be directly ruled out by setting $\forall i \in [D] : \mathrm{A}_{ii} = 0$, the random graphs $\mathbf{A}$ drawn from this distribution can still have cycles. For continuous optimization, acyclicity of a point predictor can be enforced using a nonnegative,

differentiable constraint function $h : \mathcal{R}^{D\times D} \to \mathcal{R}_+$ that evaluates to zero for any acyclic graph and otherwise to some positive value that quantifies the deviation, e.g., number of closed cycles [2, 17, 33, 35]. For probabilistic models, h can be introduced within an exponential prefactor to the unconstrained probability $q_{\boldsymbol{\phi}}(\boldsymbol{A})$, i.e.,

$$q_{\boldsymbol{\phi},\lambda}(\boldsymbol{G}) \propto \exp(-\lambda h(\boldsymbol{G}))\, q_{\boldsymbol{\phi}}(\boldsymbol{G})\ . \tag{2}$$

For a sufficiently high prefactor λ, the resulting distribution $\tilde{q}_{\boldsymbol{\phi},\lambda}(\boldsymbol{G})$ assigns only negligible probability mass to any cyclic graph. For an independent factorization over the edges of the graph, this comes at the cost of expressivity, since the resulting model is locked to some ordering of the nodes [24]. Note that having a set of interdependent parameters $\boldsymbol{\phi}_{ij}$ does not prevent this, since the realizations of the edges have to be coupled, e.g., two non-zero probabilities for $P_{\boldsymbol{\phi}_{ij}}(G_{ij})$ and $P_{\boldsymbol{\phi}_{ji}}(G_{ji})$ result in a positive probability for the cycle $\mathrm{X}_i \to \mathrm{X}_j \to \mathrm{X}_i$ of length two. In Example 1, each chain implies a total order and the common cause a partial order, all of them being incompatible with each other. Hence, such distribution with independent edge probabilities can only concentrate its probability mass on one of the three graphs resulting in a skewed uncertainty measure.

2.2 Particle representations & mixture model

Lorch et al. [13] circumvents this limitation by modeling the graph posterior $p(\boldsymbol{G}|\mathcal{D})$ by a set of K particles, each representing a single DAG $\boldsymbol{G}^{(k)}$. The term particle distribution originates from the idea of approximating a continuous density function by discrete probability masses. In the context of approximating a discrete distribution, it rather refers to a simplified model that constrains the support to fewer possible outcomes, e.g., K DAGs. The corresponding generative model can be expressed as

$$\mathrm{k} \sim q_{\boldsymbol{w}}(k)\,, \quad \mathbf{G} = \boldsymbol{G}^{(\mathrm{k})}\ , \tag{3}$$

where $\boldsymbol{w}$ are the unnormalized weights defining the probability of each particle. As an alternative to the constraint function $h(\boldsymbol{G})$, a total order induced by the permutation $\boldsymbol{\pi}$ over $\mathbf{X}$ naturally constrains a graph to be acyclic, i.e., the adjacency matrix of the permuted graph is upper-triangular. Instead of modeling K deterministic graphs with some positive probability mass, K permutations allow to model a random upper-triangular matrix $\mathbf{U}$. Consequently, we no longer refer to it as a particle distribution, but a probabilistic mixture model. Denoting $\boldsymbol{\Pi}$ as the permutation matrix corresponding to $\boldsymbol{\pi}$, the permuted random matrix $\mathbf{U}^{(\boldsymbol{\Pi}^{(k)})}$ generates a DAG $\mathbf{G}$:

$$\mathrm{k} \sim q_{\boldsymbol{w}}(k)\,, \quad \mathbf{U} \sim q_{\boldsymbol{\phi}}(\boldsymbol{U})\,, \quad \mathbf{G} = \mathbf{U}^{(\boldsymbol{\Pi}^{(k)})} \coloneqq {\boldsymbol{\Pi}^{(k)}}^T \mathbf{U} \boldsymbol{\Pi}^{(k)}\ . \tag{4}$$

The expressivity of the mixture model can be further increased by also modeling different distributions for the upper-triangular matrix, i.e., $\mathbf{U}^{(k)}$. Although

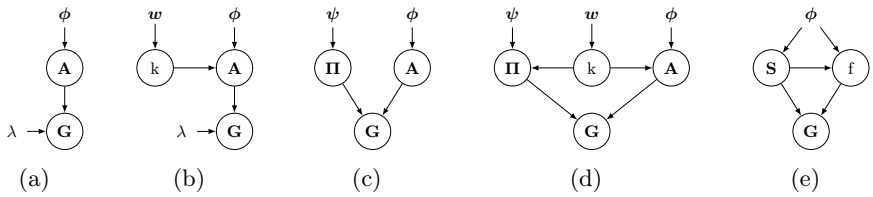


Fig. 1: Generative models of the candidate distributions over DAGs in section 2. *(a)* Adjacency matrix $\mathbf{A}$ with independent edges, *(b)* K graph distribution extending the former by a categorical distribution over k, *(c)* order-based model with permutation matrix $\mathbf{\Pi}$ to mask $\mathbf{A}$, *(d)* K order-based model, *(e)* sequence-based model adding edges autoregressively until the DAG is final (f = 1).

smaller than the number of possible DAGs, with $D!$ the number of possible permutations is still super-exponential. Hence, only a very small fraction of all K permutations can be represented by the previous model highlighting the limitation of particle distributions. In the running Example 1 there are 25 possible DAGs and the stated MEC consisting only of the 3 graphs, however, they are each admissible under a different total order. This illustrates that here at least $K = 3$ particles are required. Note that while the number of different MECs is still super-exponential, most of them contain only very few graphs and almost 27 % of all classes consist of only a single graph [6]. The upper bound on the size of a MEC is given by the completely connected chain graph that summarizes $D!$ graphs each of them following a different permutation. Evidently, splitting the probability mass equally among a huge number of graphs may be analytically correct, albeit it provides a poor trade-off between expressivity and computational efficiency represented by the number of learnable parameters that scales with K. In case the distribution concentrates its probability mass on graphs that are admissible for a low number of permutations, e.g., sparse graphs with multiple components, such a model may suffice.

2.3 Probabilistic models over orders

Several works apply the idea of an order-based search [31] in a probabilistic generative model [1, 3, 4, 23]. Their shared underlying idea is to learn a total order of the variables that imposes an acyclicity constraint on the (sampled) adjacency matrix $\mathbf{A}$. Following the description of DPM-DAG [23], we outline its generative model and provide the corresponding graphical model in Figure 1c. The generative process starts by drawing a total order that defines a permutation $\mathbf{\Pi}$ of the variables. A permuted upper-triangular matrix of ones $\boldsymbol{M}$ then defines a random acyclicity matrix $\mathbf{M}^{(\mathbf{\Pi})}$ that is used to mask a sample of an unconstrained adjacency matrix $\mathbf{A}$ as modeled in Equation (2), i.e.,

$$\mathbf{\Pi} \sim q_\psi(\boldsymbol{\Pi}), \quad \mathbf{A} \sim q_\phi(\boldsymbol{A}), \quad \mathbf{G} = \mathbf{M}^{(\mathbf{\Pi})} \coloneqq \left(\mathbf{\Pi}^T \boldsymbol{M} \mathbf{\Pi}\right). \tag{5}$$

Example 1 shows the limitation of the *Plackett-Luce* (PL) distribution [21] for causal discovery [23, 32] that is used for sampling the permutation . Both chains imply a total order of the three variables, but reversed ones. In the PL model, their permutations can only receive the same probability mass in the case of uniform weights for all three variables, yet the probability for the two causal chains is then upper-bounded by $\frac{1}{6}$ and cannot take the value of $\frac{1}{3}$. Some remaining probability mass is then concentrated on other graphs that do not belong to the MEC. This motivated the ARCO-DAG model [32], an autoregressive model over causal orders that computes the weights for the categorical sampling without replacement conditionally on the previous drawn sequence, i.e.,

$$q_{\boldsymbol{\psi}}(\pi) = \prod_{d=1}^{D} q_{\boldsymbol{\psi}^{(d)}}\big(\pi(d)\big) \qquad \text{with} \quad \boldsymbol{\psi}^{(d)} = f_{\boldsymbol{\psi}}\left(\{\pi(i)\}_{i=1}^{d}\right) \ . \tag{6}$$

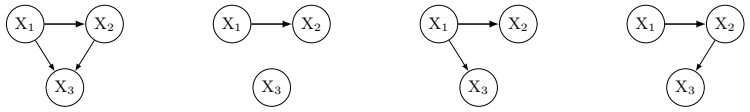
The function $f_{\boldsymbol{\psi}} : \mathcal{R}^{D\times D} \mapsto [D]$ takes a permutation matrix where some rows are still zeros as input and predicts the weights for sampling the next variable in the order. For their experiments, the authors of ARCO applied a multilayer perceptron with a single hidden layer with $H > D$ neurons. To avoid marginalizing over the set of possible parents under a sampled permutation as in Toth et al. [32], the same distribution over an unconstrained graph may be used yielding Equation (5) but with an autoregressive distribution over the permutation. In case of Example 1 where the probability mass should be split equally on all graphs of the MEC $\mathrm{X}_1 - \mathrm{X}_2 - \mathrm{X}_3$, this is consistent with setting $P_{\mathbf{A}_{13}}(1) = P_{\mathbf{A}_{31}}(1) = 0$ and $P_{\mathbf{A}_{ij}}(1) = 1$ otherwise.

2.4 Fully autoregressive model

In all covered candidate models for distributions over DAGs, the unmasked edges are still sampled from independent Bernoulli distributions as stated in Equation (2). Besides acyclicity, the edges of the causal graph drawn from these distributions are not coupled limiting the models' expressivity.

Example 2. Consider the four DAGs depicted in Figure 2 with $\boldsymbol{G}^{(1)}$ being the true causal graph. Assume the functional relationship between X_1 and X_2 is identifiable and, hence, the cumulative probability of all graphs that contain this edge should be very high in the posterior. If the structural equation for variable X_3, $f_{\mathrm{X}_3}(\mathrm{X}_1, \mathrm{X}_2, \epsilon_3)$, depends only on both variables X_1 and X_2, but not a single one alone, e.g., $f_3 = [\mathrm{X}_1 > a][\mathrm{X}_2 > b] + \epsilon_3$, then the edges are consequently coupled. This implies that the posterior concentrate its probability mass on $\boldsymbol{G}^{(1)}$ and $\boldsymbol{G}^{(2)}$ that misses X_1 and X_3 as causes of X_2. The graphs $\boldsymbol{G}^{(3)}$ and $\boldsymbol{G}^{(4)}$ with only X_1 or X_2 as causes of X_3 should be assigned small probability in this setting.

Deleu et al. [5] construct a DAG by adding edges sequentially, we refer to their graphical model as the GFlowNet-DAG model. A sampled sequence of distinct edges $\boldsymbol{S} = (S_1, \ldots, S_E) \in \mathcal{S}$ uniquely defines a DAG $\boldsymbol{G}$ over a mapping



(a) $\boldsymbol{G}^{(1)}$ (b) $\boldsymbol{G}^{(2)}$ (c) $\boldsymbol{G}^{(3)}$ (d) $\boldsymbol{G}^{(4)}$

Fig. 2: Graphs of Example 2 that share the highlighted edge between X_1 and X_2.

$g : \mathcal{S} \mapsto \mathcal{G}$. Note that $E!$ different sequences can yield the same graph G where $E := |\boldsymbol{S}|$ equals its number of edges $|\boldsymbol{G}|$. The underlying sampling is based on a transformer architecture $t_{\boldsymbol{\phi}} : \mathcal{G} \mapsto \mathcal{R}^{D \times D}$ that autoregressively computes the parameters φ defining the probabilities for potential new edges, $\varphi := t_{\boldsymbol{\phi}}(S_{:i-1})$. At each sampling step i, acyclicity is enforced by masking of the edges that would create a cycle, i.e., the transitive closure of the adjacency matrix of the graph that decodes ancestral relations. After the addition of a new edge, the ancestral mask is updated and the iterative sampling continues until a (binary) stop signal $\mathrm{f}_i = 1$ is sampled indicating that the graph sample is final, or the DAG is fully connected:

$$\mathbf{S}_i \sim q_{\boldsymbol{\phi}}(\boldsymbol{S}_i | \boldsymbol{S}_{:i-1}) \,, \quad \mathrm{f}_i \sim q_{\boldsymbol{\phi}}(f_i | \mathbf{S}_{:i-1}) \tag{7}$$

The probability of stopping $q_\phi(1|\boldsymbol{S}_{:i-1})$ theoretically guarantees that each DAG can be assigned an arbitrary probability mass provided that the function $t_{\boldsymbol{\phi}}$ has the capacity to model the parameter for the transition probabilities between different states exactly.

3 Evaluation

In contrast to Bayesian causal discovery algorithms where an observed data set is the basis for unsupervised learning of the causal graph, in this work we focus on a comparison of different models for distributions over DAGs w.r.t. their expressiveness for known target distributions. The supervised setting is motivated by shielding of other sources of error, including any bias due to the choice of functional relationships, prior distributions over the data, or the size of the observed training data. To demonstrate the limitations of the candidate models for distribution over DAGs obtained in section 2, we fit each candidate model $q_{\boldsymbol{G}}$ to a specified target distribution $p_{\boldsymbol{G}}$. The target distribution is either derived by the MEC class from Example 1, by the coupling of edges from Example 2 or a synthetically generated distribution that arises by concentrating the probability mass around a target graph using the structural Hamming distance (SHD). For training of the parameters $\boldsymbol{\phi}$ with gradient descent, we take the forward KL divergence between the target distribution $p_{\mathbf{G}}$ and the candidate distribution $q_{\mathbf{G}}$ as loss function and approximate it using samples from the target distribution. Due to the limited support of particle distributions, we evaluate the fitted candidate distribution with the reverse KL divergence $D_{\mathrm{KL}}(q_{\boldsymbol{G}} \| p_{\boldsymbol{G}})$ and the total variation

distance $D_{\mathrm{TV}}(q_{\boldsymbol{G}} \| p_{\boldsymbol{G}})$. In contrast to its backward formulation, the forward KL divergence estimates the log-likelihood ratio using samples from the target distribution. During supervised training with gradient descent, this ensures finding the region of the support where the probability mass is concentrated. To illustrate this important detail consider the GFlowNet model where the generation of a DAG starts with the empty graph. If the probability mass is concentrated around graphs with a moderate number of edges, almost no signal for learning is provided for graphs with only one or two of its edges.

3.1 Particle distributions

Lemma 1 (Minimal statistical distances for the particle distribution). *The reverse Kullback-Leibler (KL) divergence as well as the total variation, Hellinger and Bhattacharyya distance between a discrete target distribution $p_{\mathbf{G}}$ and a particle representation as candidate model $q_{\mathbf{G}}$ are minimized if the particles of the candidate $q_{\mathbf{G}}$ represent the items with the highest probability mass in the target distribution $p_{\mathbf{G}}$ and have their normalized probability mass.*

The minimal reverse KL divergence and total variation distance are given by:

$$\min_{q \in \mathcal{Q}} D_{\mathrm{KL}}(q_{\mathbf{G}} \| p_{\mathbf{G}}) = -\log \sum_{k=1}^{K} p(\boldsymbol{G}^{(k)}) \,, \quad \min_{q \in \mathcal{Q}} D_{\mathrm{TV}}(q_{\mathbf{G}} \| p_{\mathbf{G}}) = 1 - \sum_{k=1}^{K} p(\boldsymbol{G}^{(k)})$$
$$\text{with} \quad \forall k \in [1, K] : \boldsymbol{G}^{(k)} = \operatorname*{arg\,max}_{\boldsymbol{G} \in \mathcal{G} \setminus \{\boldsymbol{G}^{(i)}\}_{i<k}} p(\boldsymbol{G}) \,. \tag{8}$$

We present the derivation of all minima and the proof of Lemma 1 in section C.

3.2 Semi-implicit generative models

A latent generative model for $\mathbf{G}$ defines a sampling procedure using some auxiliary random objects $\mathbf{Y} \sim q_{\boldsymbol{\phi}}$. A mechanism g maps the latent samples $\mathbf{Y}$ to a DAG, i.e., $\mathbf{G} = g(\mathbf{Y})$, and induces the joint distribution $q(\mathbf{G}, \mathbf{Y})$. For DAGs $\mathbf{Y}$ can be a combination of a permutation π and an unconstrained adjacency matrix $\mathbf{A}$ or alternatively a sequence of edges $\mathbf{S}$. Such implicit distribution cannot be directly evaluated in a graph $\boldsymbol{G}^*$ sampled from the target distribution since g is in general not injective. The joint distribution has to be marginalized either by explicit summation over $\boldsymbol{Y}$ or approximated by samples of $\mathbf{Y}$, i.e.,

$$q_{\boldsymbol{\phi}}(\boldsymbol{G}^*) = \sum_{\boldsymbol{Y}} \big[\boldsymbol{G}^* = g(\boldsymbol{Y})\big] q_{\boldsymbol{\phi}}(\boldsymbol{G}, \boldsymbol{Y}) \,. \tag{9}$$

We leverage importance sampling to efficiently estimate the marginal probability of $\boldsymbol{G}^*$ using a low number of samples of $\mathbf{Y}^*$ that only generates samples of the target graph $\mathbf{G}^* = g(\boldsymbol{Y}^*)$:

$$q_{\boldsymbol{\phi}}(\boldsymbol{G}^*) = \sum_{\boldsymbol{Y}^* \colon f(\boldsymbol{Y}^*) = \boldsymbol{G}^*} q_{\boldsymbol{\phi}}(\boldsymbol{G}^*, \boldsymbol{Y}^*) = \mathbb{E}_{(\boldsymbol{G}^*, \mathbf{Y}^*) \sim q^*_{\mathbf{G}}} \left[\frac{q_{\boldsymbol{\phi}}(\boldsymbol{G}^*, \mathbf{Y}^*)}{q^*_{\boldsymbol{\phi}}(\boldsymbol{G}^*, \boldsymbol{Y}^*)} \right] \,. \tag{10}$$

The proposal distribution $q^*_{\boldsymbol{\phi}}$ is derived from the distribution $q_{\boldsymbol{\phi}}$ by constraining the discrete samples of $\mathbf{Y}$ to $\boldsymbol{Y}^*$. Note that due to the parameter sharing of $\boldsymbol{\phi}$, the proposal distribution is not constant during training and is updated jointly with the candidate distribution at every iteration. We provide further details on the implementation in section D.

4 Experiments

In our experiments, we compare the following models from section 2 and Table 1:

- distributions over a fixed number of DAGs (K graph particle distribution),
- *reinforced probabilistically masked DAG* (RPM-DAG), a discretized version of DPM-DAG [23] that is trained with the score-function gradient estimator,
- ARCO-DAG, the autoregressive model over the causal order [32],
- Mixture models of RPM- and ARCO-DAG, as depicted in Figure 1d,
- GFlowNet-DAG[5], an autoregressive model adding edges sequentially.

We minimize the forward KL divergence with the AdamW Optimizer [14] over 1000 optimization steps. Further details and the used hyperparameters for the supervised training of each model are provided in section E.

4.1 Markov equivalent graphs

The first experiment consists of the MEC represented by the chain graph $A-B-C$ motivated in Example 1. The target distribution assigns all three graphs of the equivalence class the same probability of a third. In Table 2a we report the mean of the reverse KL divergence and the total variation distance together with their standard deviations error over 20 independent runs for the RPM-DAG, ARCO-DAG, and GFlowNet-DAG model. For the particle distributions we compute the optimal empirical values according to Equation (8). The results align with the theoretical analysis that RPM-DAG fails to distribute the probability equally among all graphs of the MEC. The observed preference for the common cause $\boldsymbol{G}^{(1)}$ is a model bias arising from the fact that $\boldsymbol{G}^{(1)}$ induces only a partial order that is compatible with two total ones. For $K = 3$, a particle distribution can theoretically match the target distribution, whereas a reduced number of modeled graphs is limited to an overconfidence estimation that is outperformed by ARCO-DAG and GFlowNet-DAG. Our experimental result for a mixture of 3 RPM-DAG models further indicates that these optimal theoretical values are not reached during model training. With increasing mixture models, the statistical divergence only decreases slowly, while their standard deviations increases initially, before they decrease again. The standard errors underline the validity of the reported mean values, since they are by a factor of $\sqrt{20}$ smaller than the reported standard deviations. The high variance is due to some outlier, illustrating that there is a regime where training is less stable. By contrast, the results for ARCO-DAG demonstrate that it can overcome the limitations of RPM-DAG by a simple autoregressive model for the permutation weights. In comparison to the

Table 2: Statistical divergences and graph probabilities for different candidate models $q_{\mathbf{G}}$ and underlying target distribution $p_{\mathbf{G}}$. Empirical means and standard deviations are reported for 20 runs, except for the particle representations (analytical values using Equation (8)).

(a) Example 1: MEC class $\mathrm{X}_1 - \mathrm{X}_2 - \mathrm{X}_3$.

$q_{\mathbf{G}}$	$D_{\mathrm{KL}}(q_{\mathbf{G}} \Vert p_{\mathbf{G}})$	$D_{\mathrm{TV}}(q_{\mathbf{G}} \Vert p_{\mathbf{G}})$	$q(\boldsymbol{G}_1)$	$q(\boldsymbol{G}_2)$	$q(\boldsymbol{G}_3)$
$p_{\mathbf{G}}$	0	0	$0.\bar{3}$	$0.\bar{3}$	$0.\bar{3}$
1 graph particle	1.0986	$0.\bar{6}$	1	0	0
2 graph particles	0.4055	$0.\bar{3}$	0.5	0.5	0
3 graph particles	0	0	$0.\bar{3}$	$0.\bar{3}$	$0.\bar{3}$
RPM-DAG	0.32673 ± 0.00004	0.34377 ± 0.00003	0.53318 ± 0.00001	0.16145 ± 0.00002	0.16145 ± 0.00002
5-RPM-DAG	0.05 ± 0.12	0.05 ± 0.11	0.36 ± 0.07	0.31 ± 0.06	0.07 ± 0.06
10-RPM-DAG	0.0013 ± 0.0009	0.003 ± 0.008	0.335 ± 0.007	0.333 ± 0.001	0.331 ± 0.009
ARCO-DAG	0.00370 ± 0.00005	0.0038 ± 0.0003	0.3322 ± 0.0007	0.3320 ± 0.0007	0.3321 ± 0.0006
5-ARCO-DAG	0.0021 ± 0.0006	0.003 ± 0.004	0.333 ± 0.004	0.332 ± 0.005	0.333 ± 0.002
10-ARCO-DAG	0.0015 ± 0.0006	0.022 ± 0.0021	0.333 ± 0.002	0.3327 ± 0.0017	0.3324 ± 0.0012
GFlowNet-DAG	0.004 ± 0.015	0.02 ± 0.04	0.328 ± 0.026	0.34 ± 0.04	0.328 ± 0.023

(b) Example 2: Coupled edges $\mathrm{X}_1 \to \mathrm{X}_3 \leftarrow \mathrm{X}_2$ and identifiable causal effect $\mathrm{X}_1 \to \mathrm{X}_2$.

$q_{\mathbf{G}}$	$D_{\mathrm{KL}}(q_{\mathbf{G}} \Vert p_{\mathbf{G}})$	$D_{\mathrm{TV}}(q_{\mathbf{G}} \Vert p_{\mathbf{G}})$	$q(\boldsymbol{G}_1)$	$q(\boldsymbol{G}_2)$	$q(\boldsymbol{G}_3) + q(\boldsymbol{G}_4)$
$p_{\mathbf{G}}$	0	0	0.3	0.6	0.1
1 graph particle	0.5108	0.4	0	1	0
2 graph particles	0.1054	0.1	$0.\bar{3}$	$0.\bar{6}$	0
3 graph particles	0.0513	0.05	0.3158	0.6316	0.0526
4 graph particles	0	0	0.3	0.6	0.1
RPM-DAG	0.333 ± 0.009	0.36 ± 0.01	0.126 ± 0.012	0.414 ± 0.023	0.44 ± 0.02
2-RPM-DAG	0.0024 ± 0.0008	0.0046 ± 0.0047	0.2986 ± 0.0006	0.5970 ± 0.0046	0.1004 ± 0.0011
ARCO-DAG	0.328846 ± 0.000002	0.355704 ± 0.000012	0.122402 ± 0.000015	0.42189 ± 0.00003	0.45447 ± 0.00002
2-ARCO-DAG	0.00126 ± 0.00005	0.0021 ± 0.0017	0.2999 ± 0.0020	0.5993 ± 0.0020	0.1004 ± 0.0002
GFlowNet-DAG	0.0001 ± 0.0003	0.004 ± 0.004	0.30 ± 0.05	0.599 ± 0.004	0.099 ± 0.004

GFlowNet-DAG model, it yields a considerable lower total variation distance, is more stable and much faster in training as well as evaluation, due to the reduced number of model parameters (see Table 1). For mixtures of 10 models, the situation slightly reverses with 10-RPM-DAG yielding a smaller statistical distance than 10-ARCO-DAG. Note that the structure consisting of only three variables also appears in graphs with more variables and comes w.l.o.g.. In the case of a graph with C unconnected graph components, each consisting of a graph of the simple MEC class, its MEC contains 3^C graphs. Under assumed ideal conditions and a supervised setting, a particle method picks K graphs of them with equal probability. Then Equation (8) implies high statistical divergences that are not competitive with parametrized distributions over orders, e.g., RPM-DAG, that scale well for unconnected components.

4.2 Coupled edges

Dependent edges in the posterior as in Example 2 motivate the second experiment in which we consider the corresponding graphs of Figure 2 with the following probabilities. The target distribution concentrates 60% of the probability mass on the graph $\boldsymbol{G}^{(2)}$ that misses X_1 and X_2 as parents of X_3, further 30% are assigned to the true causal graph $\boldsymbol{G}^{(1)}$. To represent a coupling of the two causes, the graphs $\boldsymbol{G}^{(3)}$ and $\boldsymbol{G}^{(4)}$ that miss either of these parents get only a

Table 3: Target distribution around the MAP graph $\boldsymbol{G}_0$. Probabilities degrees with decreasing SHD.

SHD	$p(\mathcal{G}_{\text{SHD}})$	$\lvert\mathcal{G}_{\text{SHD}}\rvert$	$p(\boldsymbol{G} \in \mathcal{G}_{\text{SHD}})$
0	0.15000	1	0.15000
1	0.42500	8	0.05313
2	0.22500	28	0.00804
3	0.10000	61	0.00164
4	0.05000	94	0.00053
5	0.01581	111	0.00014
6	0.01439	101	0.00014
7	0.01068	75	0.00014
8	0.00627	44	0.00014
9	0.00242	17	0.00014
10	0.00043	3	0.00014

Table 4: Statistical divergences of the candidate models for the posterior distribution in Table 3.

$q_{\mathbf{G}}$	$D_{\text{KL}}(q_{\mathbf{G}} \Vert p_{\mathbf{G}}) \downarrow$	$D_{\text{TV}}(q_{\mathbf{G}} \Vert p_{\mathbf{G}}) \downarrow$
10 graph particles	0.5395	0.4170
50 graph particles	0.1969	0.1787
100 graph particles	0.1042	0.0989
250 graph particles	0.0426	0.0417
RPM-DAG	0.133 ± 0.008	0.166 ± 0.007
2-RPM-DAG	0.107 ± 0.009	0.149 ± 0.009
10-RPM-DAG	0.041 ± 0.005	0.09 ± 0.01
ARCO-DAG	0.087 ± 0.009	0.146 ± 0.007
2-ARCO-DAG	0.079 ± 0.010	0.137 ± 0.008
10-ARCO-DAG	0.058 ± 0.007	0.114 ± 0.007
GFlowNet-DAG	0.25 ± 0.17	0.25 ± 0.08

probability of 0.05 each. In Table 2b the corresponding metrics for 20 runs are listed, following the same reporting as in subsection 4.1. Both RPM-DAG and ARCO-DAG, assign the graphs $\boldsymbol{G}^{(3)}$ and $\boldsymbol{G}^{(4)}$ too high probability as they fail to account for the coupling of the edges $\text{X}_1 \to \text{X}_3$ and $\text{X}_2 \to \text{X}_3$. Due to its autoregressive model over edges, GFlowNet-DAG can approximate the target posterior with very high accuracy in terms of individual graph probabilities as well as the two evaluated statistical divergences. By design, a particle distribution is not constrained by dependent edges. Since 90% of the probability is concentrated on two graphs, a reasonable approximation can be obtained by using a mixture of only two graphs in this particular example. However, our results show that both probabilistic mixture model, with two components of RPM- or ARCO-DAG models each, outperform the simple mixture of graphs substantially.

4.3 Concentration of posterior mass

For an identifiable FCM and a high number of samples, the posterior should be concentrated on graphs that show a high likelihood. In the idealized setting of perfect regression of a child on its parents, the likelihood can be expected to peak for the true underlying graph $\boldsymbol{G}_0$ provided that parameter uncertainty or regularization prevents superfluous edges that are not in $\boldsymbol{G}_0$. This implies that high-scoring graphs are 'similar' to $\boldsymbol{G}_0$ implied by the FCM. In the absence of an analytic posterior that motivates such similarity, we generate a synthetic target distribution around the assumed *maximimum-a-posteriori* (MAP) graph $\boldsymbol{G}_0$ depicted in Figure 5 that has positive support for all 543 possible DAGs with 4 nodes. We start by assigning probability masses to the groups of graphs that have a *structural Hamming distance* (SHD) up to 4, i.e., 4 different entries in the adjacency matrix, and split the remaining probability mass equally among all graphs with greater SHD. For the five groups with SHD up to 4, we distribute their cumulated probability masses equally among all graphs within the respective group. The target distribution is summarized in Table 3 which lists the probability masses for the groups of graphs with the same SHD along side individual ones. We report in Table 4 the mean values and standard deviations for the statistical divergences over 20 independent runs for all candidate

models, except the particle representations for which the optimal values are computed analytically using Equation (8). The results highlight the expressivity of the RPM-DAG and its extension ARCO-DAG that both outperform a particle distribution consisting of the 50 graphs with the highest probability in the target distribution that account for a cumulative probability of 86.3%. A single ARCO-DAG yields lower statistical divergences than a mixture of 2 RPM-DAG models in this very setting. When comparing mixture models with $K \geq 3$, it can be observed that mixtures of RPM-DAG yields better results than ARCO-DAG. Although we trained in an idealized very low-dimensional setting with samples drawn from the target distribution, the performance of the GFlowNet-DAG model does not match its performance from subsection 4.1 and 4.2. We conjecture that the gradients from a changing variety of different graphs do not yield a sufficiently stable training signal that is necessary to tune the transformer network.

5 Conclusion

Bayesian causal discovery promises uncertainty quantification in the prediction of the causal graph. We reviewed several candidate models for distribution over DAGs and investigated the limitations of their expressivity. To minimize confounding factors such as the influence of functional relationships or the size of the training set, we considered an idealized supervised setting that can also be used to specify prior distributions. We showed in our theoretical analysis and experimental results that all considered candidate models, besides the autoregressive GFlowNet-DAG model, are theoretically not sufficiently expressive to match simple target distributions in which edges are coupled. While GFlowNet-DAG satisfies the theoretical requirements, we cannot verify its expressiveness in a low-dimensional experiment where graph samples are drawn from the target distribution with support over all possible DAGs. Since causal structure learning is typically an unsupervised problem, this poses a major limitation to this model.

Coupled edges due to interaction effects pose a general challenge in causal structure learning and affect most algorithms—not only Bayesian approaches. Our results further suggest that a mixture of a small number of simple probabilistic models as RPM-DAG models may approximate distributions with coupled edges sufficiently well in practical applications and outperform particle distributions with a moderate number of modeled graphs. Its extension, ARCO-DAG is more expressive w.r.t. probabilistic causal orders and shows competitive performance to GFlowNet-DAG, while having a substantially lower number of parameters and is more stable in training. When comparing mixture models of RPM- and ARCO-DAG, the simpler model outperforms the latter more expressive model for 10 mixture components consistently in all experiments. Therefore, we conjecture that a mixture of RPM-DAG models is particularly suited to scale Bayesian causal discovery to higher numbers of variables.

Bibliography

[1] Annadani, Y., Pawlowski, N., Jennings, J., Bauer, S., Zhang, C., Gong, W.: BayesDAG: Gradient-based posterior inference for causal discovery. In: Advances in Neural Information Processing System, vol. 36 (2023)

[2] Bello, K., Aragam, B., Ravikumar, P.: DAGMA: learning DAGs via M-matrices and a log-determinant acyclicity characterization. In: Advances in Neural Information Processing Systems, vol. 35 (2022)

[3] Charpentier, B., Kibler, S., Günnemann, S.: Differentiable DAG sampling. In: International Conference on Learning Representations, 10, Openreview (2022)

[4] Cundy, C., Grover, A., Ermon, S.: BCD nets: Scalable variational approaches for Bayesian causal discovery. In: Advances in Neural Information Processing Systems, vol. 34, pp. 7095–7110 (2021)

[5] Deleu, T., Góis, A., Emezue, C.C., Rankawat, M., Lacoste-Julien, S., Bauer, S., Bengio, Y.: Bayesian structure learning with generative flow networks. In: Conference on Uncertainty in Artificial Intelligence, Proceedings of Machine Learning Research, vol. 180, pp. 518–528, PMLR (2022)

[6] Gillispie, S.B., Perlman, M.D.: Enumerating markov equivalence classes of acyclic digraph models. In: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, pp. 171–177, Morgan Kaufmann (2001)

[7] Glymour, C., Zhang, K., Spirtes, P.: Review of causal discovery methods based on graphical models. Frontiers in Genetics **10**, 524 (2019)

[8] Hoyer, P.O., Janzing, D., Mooij, J.M., Peters, J., Schölkopf, B.: Nonlinear causal discovery with additive noise models. In: Advances in Neural Information Processing Systems, vol. 21, pp. 689–696 (2008)

[9] Immer, A., Schultheiss, C., Vogt, J.E., Schölkopf, B., Bühlmann, P., Marx, A.: On the identifiability and estimation of causal location-scale noise models. In: International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 202, pp. 14316–14332, PMLR (2023)

[10] Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: International Conference on Learning Representations, 5, Openreview (2017)

[11] Koller, D., Friedman, N.: Probabilistic Graphical Models - Principles and Techniques. MIT Press (2009)

[12] Loh, P., Bühlmann, P.: High-dimensional learning of linear causal networks via inverse covariance estimation. Journal of Machine Learning Research **15**(1), 3065–3105 (2014)

[13] Lorch, L., Rothfuss, J., Schölkopf, B., Krause, A.: DiBS: Differentiable Bayesian structure learning. In: Advances in Neural Information Processing Systems, vol. 34, pp. 24111–24123 (2021)

[14] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations, 7, OpenReview (2019)

[15] Lundberg, I., Johnson, R., Stewart, B.M.: What is your estimand? Defining the target quantity connects statistical evidence to theory. American Sociological Review **86**(3), 532–565 (2021)
[16] Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: A continuous relaxation of discrete random variables. In: International Conference on Learning Representations, 5, Openreview (2017)
[17] Nazaret, A., Hong, J., Azizi, E., Blei, D.: Stable differentiable causal discovery. In: International Conference on Machine Learning, OpenReview (2024)
[18] OEIS Foundation Inc.: The On-Line Encyclopedia of Integer Sequences (2023), published electronically at `http://oeis.org/A003024`
[19] Pearl, J.: Causality. Cambridge University Press, 2nd edn. (2009)
[20] Peters, J., Janzing, D., Schölkopf, B.: Elements of causal inference: Foundations and learning algorithms. The MIT Press (2017)
[21] Plackett, R.L.: The analysis of permutations. Journal of the Royal Statistical Society Series C: Applied Statistics **24**(2), 193–202 (1975)
[22] Prillo, S., Eisenschlos, J.: Softsort: A continuous relaxation for the argsort operator. In: International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 119, pp. 7793–7802, PMLR (2020)
[23] Rittel, S., Tschiatschek, S.: Specifying prior beliefs over DAGs in deep Bayesian causal structure learning. In: European Conference on Artificial Intelligence, Frontiers in Artificial Intelligence and Applications, vol. 372, pp. 1962–1969, IOS Press (2023)
[24] Rittel, S., Tschiatschek, S.: On differentiable Bayesian causal structure learning. In: UAI Workshop on Causal Inference (2024)
[25] Rittel, S., Tschiatschek, S.: Expressiveness of parametrized distributions over DAGs for causal discovery. Transactions on Machine Learning Research (2025)
[26] Shah, R.D., Peters, J.: The hardness of conditional independence testing and the generalised covariance measure. The Annals of Statistics **48**(3), 1514–1538 (2020)
[27] Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A.J.: A linear non-Gaussian acyclic model for causal discovery. In: Journal of Machine Learning Research, vol. 7, pp. 2003–2030, JMLR (2006)
[28] Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction and Search. Adaptive Computation and Machine learning, MIT Press, 2nd edn. (2000)
[29] Squires, C., Uhler, C.: Causal structure learning: A combinatorial perspective. Foundations of Computational Mathematics pp. 1–35 (2022)
[30] Stanley, R.P.: Acyclic orientations of graphs. Discrete Mathematics **5**(2), 171–178 (1973)
[31] Teyssier, M., Koller, D.: Ordering-based search: A simple and effective algorithm for learning Bayesian networks. In: Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence, pp. 548–549, AUAI Press (2005)
[32] Toth, C., Knoll, C., Pernkopf, F., Peharz, R.: Effective Bayesian causal inference via structural marginalisation and autoregressive orders. In: ICML Workshop on Structured Probabilistic Inference & Generative Modeling (2024)

[33] Yu, Y., Chen, J., Gao, T., Yu, M.: DAG-GNN: DAG structure learning with graph neural networks. In: International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 97, pp. 7154–7163, PMLR (2019)

[34] Zhang, K., Hyvärinen, A.: On the identifiability of the post-nonlinear causal model. In: Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, pp. 647–655, AUAI Press (2009)

[35] Zheng, X., Aragam, B., Ravikumar, P., Xing, E.P.: DAGs with NO TEARS: continuous optimization for structure learning. In: Advances in Neural Information Processing Systems, vol. 31, pp. 9492–9503 (2018)

A Preliminaries

A.1 Probability and random variables

We denote scalar random variables as y, random vectors as $\mathbf{y}$, and random matrices as $\mathbf{Y}$. Single elements of a random vector or matrix are scalars and written as y or Y respectively. The probability of a discrete random variable y taking the value y is expressed by the probability mass function $p_{\mathrm{y}}(y) \coloneqq \mathbb{P}(\mathrm{y} = y)$. When clear from the context, we omit the random variable in the subscript of a probability mass function in favor of the parameters $\boldsymbol{\theta}$ of the distribution to increase readability, e.g., $p_{\boldsymbol{\theta}}(\mathrm{y})$. We mainly apply this concise notation in the main text to introduce the generative models. With a slight abuse of terminology, we refer to a distribution P_{y} by its induced probability mass function $p_{\mathrm{y}}(y) = P_{\mathrm{y}}(\{y\})$. To distinguish approximations of the true target distribution, we denote model distributions by $q_{\mathrm{y}}(y) = Q_{\mathrm{y}}(\mathrm{y})$ or simply $q_{\boldsymbol{\theta}}(y)$.

A.2 Functional causal models

A *functional causal model* (FCM) is defined as the triple $\mathcal{M}_{\mathbf{X}} \coloneqq \{\mathbf{X}, (\boldsymbol{\epsilon}, P_{\boldsymbol{\epsilon}}), \boldsymbol{f}\}$ consisting of a set of endogenous random variables $\mathbf{X}$, a set of exogenous noise variables $\boldsymbol{\epsilon}$ with joint probability distribution $P_{\boldsymbol{\epsilon}}$ and a set of deterministic functions $\boldsymbol{f}$, all three indexed by $[D] \coloneqq \{1, ..., D\}$. Each endogenous variable X_d of the model is generated by a function of a subset of the endogenous variables $\mathbf{X}$ and its exogenous noise ε_d, i.e.,

$$\mathrm{X}_d \coloneqq f_d(\mathbf{X}, \varepsilon_d) = f_d(\mathrm{Pa}_{\boldsymbol{G}}(\mathrm{X}_d), \varepsilon_d) \qquad \forall d \in [D] \; . \tag{11}$$

The distribution of X_d is implicitly defined as the pushforward measure of $P_{\boldsymbol{\epsilon}}$ through the *causal mechanism* f_d. The structure induced by the direct functional dependencies is often restricted to be acyclic such that it can be represented by a *directed acyclic graph* DAG or—equivalently—its adjacency matrix $\boldsymbol{G} \in \mathcal{G} \subset \{0, 1\}^{D \times D}$ with a one-to-one correspondence between random variables and nodes. Let $\sim d$ denote the index set $[D] \setminus \{d\}$, then the d-th column of $\boldsymbol{G}$ encodes the parents $\mathrm{Pa}_{\boldsymbol{G}}(\mathrm{X}_d) \subseteq \mathbf{X}_{\sim d}$ of a node/random variable X_d, i.e., the subset of $\mathbf{X}_{\sim d}$ that has a direct influence on X_d via f_d. Ancestors of a random variable, $\mathrm{An}_{\boldsymbol{G}}(\mathrm{X}_d)$, have a directed path to X_d in the causal graph $\boldsymbol{G}$ and indirectly influence X_d. Their causal effect is mediated by at least one parent of X_d (possibly themselves), it can be blocked by conditioning on all its parents $\mathrm{Pa}_{\boldsymbol{G}}(\mathrm{X}_d)$. Children $\mathrm{Ch}_{\boldsymbol{G}}(\mathrm{X}_d)$ and descendants $\mathrm{De}_{\boldsymbol{G}}(\mathrm{X}_d)$ are affected by changes of the corresponding node X_d and are defined complementary to parents and ancestors.

A.3 Causal structure learning

The objective of causal discovery is to learn the underlying causal graph $\boldsymbol{G}$ from observed random variables $\mathbf{X}$ that encodes the causal effects implied by the FCM $\mathcal{M}_{\mathbf{X}}$. Throughout this work, we assume causal sufficiency, i.e., all endogenous variables $\mathbf{X}$ are observable and the exogenous noise variables $\boldsymbol{\epsilon}$ are mutually independent. This implies that all dependencies and independencies between the observed values of the random variables $\mathbf{X}$ result from their causal effects over the functions $\boldsymbol{f}$ and not from some latent common causes, i.e., an unobserved shared ancestors of them. In addition, we assume that all samples in the data set $\mathcal{D} \coloneqq \{\boldsymbol{X}^{(n)}\}_{n=1}^{N}$ are generated i.i.d. from the FCM $\mathcal{M}_{\mathbf{X}}$ without any selection bias for the generated samples, e.g., there is no conditioning on unobserved confounders.

Without any assumptions on the functions $\boldsymbol{f}$, the *Markov equivalence class* (MEC) of a causal graph $\boldsymbol{G}$ can be consistently identified from $\mathcal{D}$ using *conditional independence* (CI) tests [28]. This equivalence class contains all DAGs that entail the same observed independence relations and can be compactly represented by a *completely partially directed acyclic graph* (CPDAG), an acyclic mixed graph that has the same adjacencies, $\mathrm{X}_i - \mathrm{X}_j$, and unshielded colliders, $\mathrm{X}_i \to \mathrm{X}_k \leftarrow \mathrm{X}_j$, as the underlying true graph $\boldsymbol{G}$, but with some of its edges remaining undirected. While CI tests can be easily parallelized and their required overall number can be sequentially restricted by the individual test results as in the PC algorithm [28], the combination of the uncertainty attached to each test is non-trivial. Moreover, CI testing for continuous random variables lacks statistical power against alternatives and suffers from the curse of dimensionality [26].

A.4 Bayesian causal discovery

Alternatively, score-based algorithms for causal discovery optimize a scalar quantity that is typically derived from the likelihood of the observed data $p_{\boldsymbol{\Theta}}(\mathcal{D}|\boldsymbol{G})$. Since the maximum likelihood $p_{\boldsymbol{\Theta}^*}(\mathcal{D}|\boldsymbol{G})$ can only improve with increasing number of allowed parents as covariates, regularization of the number of edges is required to prevent estimating a fully connected DAG. Bayesian causal structure learning is also based on the likelihood, but models the full data generating process for the so called evidence $p(\mathcal{D})$ that includes the uncertainty over model parameters $\boldsymbol{\Theta}$ and a prior distribution over the causal graph $p(\boldsymbol{G})$:

$$p(\mathcal{D}) = \int \int \prod_{n=1}^{N} p(\boldsymbol{X}^{(n)}|\boldsymbol{\Theta}) p(\boldsymbol{\Theta}|\boldsymbol{G}) p(\boldsymbol{G}) \, \mathrm{d}\boldsymbol{\Theta} \, \mathrm{d}\boldsymbol{G} = \int p(\mathcal{D}|\boldsymbol{G}) p(\boldsymbol{G}) \, \mathrm{d}\boldsymbol{G} \quad (12)$$

The marginal likelihood $p(\mathcal{D}|\boldsymbol{G})$ arises by averaging over the (conditional) prior distribution of model parameters $p(\boldsymbol{\Theta}|\boldsymbol{G})$. In contrast to the maximum likelihood estimate $p_{\theta^*}(\mathcal{D}|\boldsymbol{G})$, it avoids overfitting to the noise of the data $\mathcal{D}$ [11]. The posterior distribution over the causal $p(\boldsymbol{G}|\mathcal{D})$ that quantifies the uncertainty over the true causal graph is obtained by Bayes' formula:

$$p(\boldsymbol{G}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{G}) p(\boldsymbol{G})}{p(\mathcal{D})} \, . \quad (13)$$

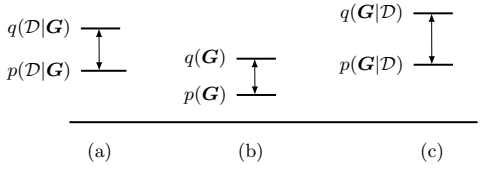


Fig. 3: Modeling error in Bayesian causal discovery due to approximations q of *(a)* the marginal likelihood, *(b)* the model for the prior and *(c)* the posterior over the DAG $\boldsymbol{G}$.

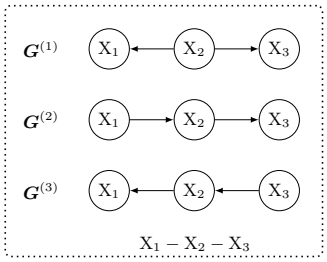


Fig. 4: Graphs of the MEC from Example 1 along its CPDAG.

Computing the evidence involves marginalization over all DAGs and parameters of the likelihood and is intractable. Therefore, approximate methods for Bayesian inference are needed. In the following we focus on variational inference where the joint distribution $p(\boldsymbol{G}, \mathcal{D})$ is approximated by a generative model $q(\mathcal{D}|\boldsymbol{G})p(\boldsymbol{G})$, and a variational family of distributions $\mathcal{Q} := \{q_{\boldsymbol{\phi}}(\boldsymbol{G}|\mathcal{D}) \,|\, \boldsymbol{\phi} \in \mathbb{R}\}$ is specified for the posterior $q(\boldsymbol{G}|\mathcal{D})$. Both steps involve parametric distributions over DAGs.

A.5 Bayesian model error

Given a fixed data set $\mathcal{D}$, deviations of a modeled posterior $q(\boldsymbol{G}|\mathcal{D})$ to the true posterior $p(\boldsymbol{G}|\mathcal{D})$ can be backtracked to three different model errors as depicted in Figure 3. The modeled marginal likelihood $q(\mathcal{D}|\boldsymbol{G})$ may induce a bias, with the particular case of the maximum likelihood estimate leading to an overconfident prediction. Throughout this work, we assume a flexible model for the marginal likelihood and focus on the role of the distribution over DAGs instead. The model for the prior distribution $q(\boldsymbol{G})$ directly constrains the belief that domain experts can express over the causal structure, but the subjective bias vanishes in the asymptotic limit as long as it assigns each DAG some positive probability mass. Lastly, the true posterior may not be in the variational family Q, i.e. cannot be expressed by $q(\boldsymbol{G}|\mathcal{D})$ which poses a direct constrain for the outcome of the Bayesian analysis.

B Probability mass functions over DAGs

For conciseness, we present in section 2 only the generative models for DAGs that implicitly induce a probability distribution. In the following, we provide their analytic probability mass functions.

B.1 Independent Edges

$$q_{\mathbf{A}}(\boldsymbol{A}) = \prod_{i \neq j}^{D} q_{\mathrm{A}_{ij}}(A_{ij}) \tag{14}$$

$$\begin{aligned} q_{\mathbf{G}}(\boldsymbol{G}) =& \frac{1}{Z} \exp\big(-\lambda h(\boldsymbol{G})\big) q_{\mathbf{A}}(\boldsymbol{G}) \\ \text{with } \; Z \coloneqq& \sum_{\boldsymbol{G} \in \mathcal{G}} \exp\big(-\lambda h(\boldsymbol{G})\big) q_{\mathbf{A}}(\boldsymbol{G}) \; . \end{aligned} \tag{15}$$

B.2 Particle distribution

$$q_{\mathbf{G}}(\mathbf{G}) = \sum_{k=1}^{K} \Big[\boldsymbol{G}^{(k)} = \boldsymbol{G}\Big] \, q_{\mathrm{k}}(k) \; , \tag{16}$$

B.3 Probabilistically masked DAG

$$\begin{aligned} q_{\mathbf{G}}(\boldsymbol{G}) =& \sum_{\boldsymbol{\Pi}, \boldsymbol{A}} \Big[\boldsymbol{M}^{(\boldsymbol{\Pi})} \circ \boldsymbol{A} = \boldsymbol{G}\Big] q_{\boldsymbol{\Pi}}(\boldsymbol{\Pi}) \, q_{\mathbf{A}}(\boldsymbol{A}) \\ \text{with } \; \boldsymbol{M}^{(\boldsymbol{\Pi})} \coloneqq& \boldsymbol{\Pi}^T \boldsymbol{M} \boldsymbol{\Pi} \; . \end{aligned} \tag{17}$$

Sampling without replacement from a categorical distribution with some fixed weights $\{w_d\}_{d=1}^{D}$ is equivalent to drawing a permutation over $[D]$ and known as the Plackett-Luce distribution [21]. At each sampling stage d the categorical weights for the remaining variables that were not yet sampled are normalized :

$$q_{\pi}(\boldsymbol{\pi}) = \prod_{d=1}^{D-1} q_{\pi(d)}\big(\pi(d)\big) = \prod_{d=1}^{D-1} \frac{\psi_{\boldsymbol{\pi}(d)}}{\sum_{i=1}^{D} \psi_i - \sum_{j<d} \psi_{\boldsymbol{\pi}(j)}} \; . \tag{18}$$

A differentiable, continuous relaxation of the discrete sampling of permutation matrices $\boldsymbol{\Pi}$ can be obtained by pairing the Gumbel-Softmax trick [10, 16] with Softsort [22].

B.4 Mixture of DAGs with probabilistic entries

$$q_{\mathbf{G}}(\boldsymbol{G}) = \sum_{k=1}^{K} \sum_{\boldsymbol{U}} \left[\boldsymbol{U}^{(\boldsymbol{\Pi}^{(k)})} = \boldsymbol{G}\right] q_{\mathbf{U}}(\boldsymbol{U})\, q_{\mathrm{k}}(k)$$
$$\text{with } \boldsymbol{U}^{(\boldsymbol{\Pi}^{(k)})} \coloneqq {\boldsymbol{\Pi}^{(k)}}^{T} \boldsymbol{U} \boldsymbol{\Pi}^{(k)} \,. \tag{19}$$

B.5 Autoregressive model over all potential edges

Denoting the mapping between the sequence of edges $\boldsymbol{S} \in \mathcal{S}$ to a graph $\boldsymbol{G}$ by $g : \mathcal{S} \mapsto \{0,1\}^{D \times D}$, the probability of sampling a graph $\boldsymbol{G}$ equals:

$$q_{\mathbf{G}}(\boldsymbol{G}) = \sum_{\boldsymbol{S} \in \mathcal{S}} q_{\mathbf{G},\mathbf{S}}(\boldsymbol{G}, \boldsymbol{S}) = \sum_{\boldsymbol{S} \in \mathcal{S}} \left[\boldsymbol{G} = g(\boldsymbol{S})\right] q_{\mathbf{S}}(\boldsymbol{S}) \,. \tag{20}$$

$$q_{\mathbf{S}}(\boldsymbol{S}) = \left(\prod_{i=1}^{E} q_{\mathrm{S}_i | \mathrm{S}_{:i-1}}(S_i | S_{:i-1}) \right) \left(\prod_{j=1}^{E-1} q_{\mathrm{f}_j | \mathbf{S}_{:j-1}}(0 | S_{:j-1}) \right) q_{\mathrm{f}_E | \mathrm{S}_{:E-1}}(1 | \mathrm{S}_{E-1}) \,. \tag{21}$$

The sequential process of a GFlowNet (GFN) itself defines a DAG where all non-leafe nodes represent the states of an unfinished graph $\boldsymbol{G}^{(i)} = g(\boldsymbol{S}_{:i})$. Starting with the empty graph state as the root node and probability 1, it splits the probability of a state representing a preliminary graph among its children who are either other preliminary graphs with an additional edge or finalized graphs.

C Particle distributions

In the following we present a proof of Lemma 1 and derive the minimal values for the four statistical distances between a discrete target distribution p and a reduced distribution q with only K particles.

C.1 Minimal total variation distance

The total variation distance D_{TV} between the candidate and target distribution, $q_{\mathbf{G}}$ and $p_{\mathbf{G}}$, can be rewritten as:

$$\begin{aligned} D_{\mathrm{TV}}(q_{\mathbf{G}} \| p_{\mathbf{G}}) &= \frac{1}{2} \sum_{\boldsymbol{G} \in \mathcal{G}} |q(\boldsymbol{G}) - p(\boldsymbol{G})| \\ &= \frac{1}{2} \Bigg(\underbrace{\sum_{\boldsymbol{G} \in \mathcal{G} \setminus \{\boldsymbol{G}^{(k)}\}_{k=1}^{K}} p(\boldsymbol{G})}_{= 1 - \sum_{k=1}^{K} p(\boldsymbol{G}^{(k)})} + \sum_{k=1}^{K} \left| q(\boldsymbol{G}^{(k)}) - p(\boldsymbol{G}^{(k)}) \right| \Bigg) \,. \end{aligned} \tag{22}$$

In the RHS of Equation (22) the sum over all graphs is split into two terms. The graphs that are not in the support of the particle distribution $q_{\mathbf{G}}$ lead to a distance that is independent of the weights for the particles. The second term measures the absolute deviations for the probabilities of the K graphs that are assigned non-zero probabilities in $q_{\mathbf{G}}$.
For equality of the probability masses $p(\boldsymbol{G}^{(k)})$ and $q(\mathbf{G}^{(k)})\,\forall k \in [K]$, it would be minimized. Due to the normalization constraint, the accumulated probability mass of the not modeled graphs has to be distributed among the K graphs in $q_{\mathbf{G}}$, leading to $q(\boldsymbol{G}^{(k)}) \geq p(\boldsymbol{G}^{(k)})\,\forall k \in [K]$:

$$\begin{aligned}\min_{q in \mathcal{Q}} D_{\mathrm{TV}}(q_{\mathbf{G}} \| p_{\mathbf{G}}) &= \frac{1}{2}\left(1 - \sum_{k=1}^{K} p(\boldsymbol{G}^{(k)}) + \left(\sum_{k=1}^{K} q(\boldsymbol{G}^{(k)}) - p(\boldsymbol{G}^{(k)})\right)\right) \\ &= 1 - \sum_{k=1}^{K} p(\boldsymbol{G}^{(k)})\,. \end{aligned} \tag{23}$$

Consequently, both sums in the previous equation amount precisely to the loss of the accumulated probability mass for the graphs not modeled by $q_{\mathbf{G}}$ and the total variation distance is minimized when $q_{\mathbf{G}}$ assigns non-zero probability to the K graphs with the highest probability in $p_{\mathbf{G}}$:

$$\forall k \in [K]: \quad \boldsymbol{G}^{(k)} = \underset{\boldsymbol{G} \in \mathcal{G} \setminus \{\boldsymbol{G}^{(i)}\}_{i<k}}{\arg\max} \; p_{\mathbf{G}}(\boldsymbol{G})\,. \tag{24}$$

In general their corresponding probabilities $q(\boldsymbol{G}(k))$ are not unique. The approximated probability distribution $q_{\mathbf{G}}$ with the minimal D_{TV} to $p_{\mathbf{G}}$ does not constrain how the additional probability mass is distributed.

In principle the probability of a single graph could account for the missing accumulated probability mass yielding only a single biased value, but distorting the relative probabilities. This limits the utility of such an approximated distribution heavily and motivates the following constraint:

$$\forall i \neq j \in [K]: \quad \frac{q(\boldsymbol{G}^{(i)})}{q(\boldsymbol{G}^{(j)})} = \frac{p(\boldsymbol{G}^{(i)})}{p(\boldsymbol{G}^{(j)})}\,. \tag{25}$$

Preserving the relative probabilities of $p_{\mathbf{G}}$ in $q_{\mathbf{G}}$ uniquely defines $q_{\mathbf{G}}$ by normalized probabilities

$$q(\boldsymbol{G}^{(k)}) = \frac{p(\boldsymbol{G}^{(k)})}{\sum_{j=1}^{K} p(\boldsymbol{G}^{(j)})}\,, \tag{26}$$

and yields a approximation that is for all K graphs overconfident w.r.t to the approximated distributed $p_{\mathbf{G}}$, i.e., $q(\boldsymbol{G}^{(k)}) > p(\boldsymbol{G}^{(k)})\,\forall k \in [K] \subset |\mathcal{G}|$. □

C.2 Minimal Kullback-Leibler divergence

In contrast to the total variation distance, the KL divergence is not symmetric. Due to the smaller support of the candidate distribution $q_{\mathbf{G}}$ compared to the

target distribution over all DAGs $p_{\mathbf{G}}$ only the reverse KL divergence, $D_{\mathrm{KL}}(q|p)$, is properly defined. Imposing normalized probabilities as in Equation (26)to preserve the relative probabilities of the graphs in the target distribution $p_{\mathbf{G}}$ simplifies the KL divergence to:

$$\begin{aligned} D_{\mathrm{KL}}(q_{\mathbf{G}} \| p_{\mathbf{G}}) &= \sum_{\boldsymbol{G} \in \mathcal{G}} q(\boldsymbol{G}) \log \frac{q(\boldsymbol{G})}{p(\boldsymbol{G})} = \sum_{k=1}^{K} q(\boldsymbol{G}^{(k)}) \log \frac{q(\boldsymbol{G}^{(k)})}{p(\boldsymbol{G}^{(k)})} \\ &= -\sum_{k=1}^{K} q(\boldsymbol{G}^{(k)}) \log \sum_{j=1}^{K} p(\boldsymbol{G}^{(j)}) = -\log \sum_{j=1}^{K} p(\boldsymbol{G}^{(j)}) \,. \end{aligned} \tag{27}$$

Hence, the constrained KL divergence is minimized, when $q_{\mathbf{G}}$ assigns non-zero probabilities to the K graphs with the highest probabilities in $p_{\mathbf{G}}$ as in Equation (3). □

C.3 Minimal Hellinger & Bhattacharyya distance

The derivation of the minimal Hellinger distance D_{H} follows analog to the one in subsection C.1 with the same constraint of preserved relative probabilities from Equation (26):

$$\begin{aligned} {D_{\mathrm{H}}}^2(q_{\mathbf{G}} \| p_{\mathbf{G}}) &= \frac{1}{2} \sum_{\boldsymbol{G}} \left(\sqrt{q(\boldsymbol{G})} - \sqrt{p(\boldsymbol{G})} \right)^2 \\ &= \frac{1}{2} \left(\sum_{\boldsymbol{G} \in \mathcal{G} \setminus \{\boldsymbol{G}^{(k)}\}_{k=1}^{K}} p(\boldsymbol{G}) \; + \; \sum_{k=1}^{K} \left(\sqrt{q(\boldsymbol{G}^{(k)})} - \sqrt{p(\boldsymbol{G}^{(k)})} \right)^2 \right) \\ &= \frac{1}{2} \left(1 + \sum_{k=1}^{K} q(\boldsymbol{G}^{(k)}) - 2 \sqrt{q(\boldsymbol{G}^{(k)}) p(\boldsymbol{G}^{(k)})} \right) \\ &= \frac{1}{2} \left(2 - \frac{2}{\sqrt{\sum_{j=1}^{K} p(\boldsymbol{G}^{(j)})}} \sum_{k=1}^{K} p(\boldsymbol{G}^{(k)}) \right) \\ &= 1 - \sqrt{\sum_{k=1}^{K} p(\boldsymbol{G}^{(k)})} \,. \end{aligned} \tag{28}$$

The Bhattacharyya distance D_{B} can be directly derived from the Hellinger distance D_{H} by:

$$\begin{aligned} D_{\mathrm{B}}(q_{\mathbf{G}} \| p_{\mathbf{G}}) &= -\ln \left(\sum_{\boldsymbol{G} \in \mathcal{G}} \sqrt{q(\boldsymbol{G}) p(\boldsymbol{G})} \right) = -\ln \left(1 - {D_{\mathrm{H}}}^2(q_{\mathbf{G}} | p_{\mathbf{G}}) \right) \\ &= -\frac{1}{2} \ln \sum_{k=1}^{K} p(\boldsymbol{G}^{(k)}) \,. \end{aligned} \tag{29}$$

Both are minimized by selecting the graphs with the hightest probability in $p_{\mathbf{G}}$ as in Equation (24). □

D Importance sampling for parametrized distributions with an auxiliary discrete structure

Implicit generative models for discrete structures such as DAGs samples substructures from categorical distributions. The weights of discrete objects, e.g., edges, that do not appear in some target structure can be masked to 0. Since the relative probabilities of admissible substructures are preserved then, it is equal in probability to sampling form the unconstrained distribution and rejecting any sample that is not admissible under the target structure. The resulting proposal distribution $q_\phi^\star$ restricts the sample space to this target structure and is consequently optimal to the candidate distribution q_ϕ at any optimization step due to the shared parameters. In comparison direct sampling has higher variance, in particular for structures with very low probability in the target distribution.
While we present the proposed importance sampling to evaluate the probability of a full graph, its application to directed paths or subgraphs is straight-forward, following the same rational.

Order-based model For order-based models as described in subsection 2.3 and depicted in Figure 1c, e.g., RPM-DAG and ARCO-DAG, the weights of the Bernoulli distribution for any edges that do not appear in the target graph can be set to zero, i.e., $\forall i, j$ with $\boldsymbol{G}_{ij}^\star = 0 : \phi_{ij} = 0$. In addition, the sampling of a total order in Equation (18) can be constrained to the (potentially partial) order induced by the target graph by setting the weights of all variables to 0 as long as their parents have not been sampled as a predecessor in the order.

Sequence-based model Sampling of the next edge in a sequence as described in subsection 2.4 and depicted in Figure 1e can be directly constrained to the edges that appear in the target graph. The ancestral masking that ensures acyclicity remains unaffected.

E Experimental Parameters

In the experiments in subsection 4.1 and 4.2, we evaluate the forward KL divergence in all three, respectfully four, graphs with positive probability mass. The probability of a graph under the candidate model is approximated by Equation (10) using *importance samples* (IS). To speed up the calculations during training, the number of IS is chosen to be smaller than for the final evaluation of the trained model. The learning rate was chosen by a grid search over $\{1, 5\times10^{-1}, 1\times10^{-1}, 5\times10^{-2}, 1\times10^{-2}, 5\times10^{-3}, 1\times10^{-3}, 5\times10^{-4}, 1\times10^{-4}\}$. Training was performed over 1000 optimization steps to account for the instability of GFlowNet-DAG. We document the used hyperparameters in Table 5.
Due to the low number of DAGs, our first experiment (Section 5.1) allows to evaluate all 25 graphs within a single forward pass during training (hence, 1 optimization step equals 1 epoch). To account for the higher number of possible sequences, we average the probability of sampling a target graph from the model

distribution using 10 importance samples for GFlowNet-DAG in contrast to a single one. For evaluation of each graph probability we drew 100 importance samples for all methods. In the second experiment (Section 5.2) we had to limit the number of samples from the target distribution for a single optimization step. Due to the high number of trainable parameters we decided to go with 25 instead of 100 samples from the target distribution and 5 instead of 10 for the GFlowNet-DAG model. By contrast, we doubled the number of importance samples to 20 during evaluation of each of the 543 DAGs.

GFlowNet-Dag The transformer architecture for the GFlowNet-Dag model follows the official implementation provided on `https://github.com/tristandeleu/jax-dag-gflownet`, as default parameters it uses an embedding size of $H_E = 128$, a key size of $H_K = 64$ and $H_L = 7$ layers of Transformer blocks.

ARCO-DAG The autoregressive neural network of ARCO-DAG consists of a simple two layer perceptron with $H_N = 30$ hidden neurons and ReLU-activations. It follows the official implementation provided on `https://github.com/chritoth/bci-arco-gp/`.

Table 5: Hyperparameters.

(a) For the experiments in subsection 4.1 and 4.2.

Model $q_{\mathbf{G}}$	Learning rate	# forward KL samples	# IS for training	# IS for evaluation
RPM-DAG	5×10^{-1}	"all"	1	100
ARCO-DAG	5×10^{-1}	"all"	1	100
GFlowNet-DAG	1×10^{-3}	"all"	10	100

(b) For the experiments in subsection 4.3.

Model $q_{\mathbf{G}}$	Learning rate	# forward KL samples	# IS for training	# IS for evaluation
RPM-DAG	5×10^{-2}	100	10	10
ARCO-DAG	5×10^{-2}	100	10	10
GFlowNet-DAG	1×10^{-3}	25	5	20

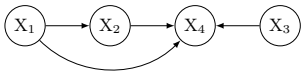


Fig. 5: True graph $\boldsymbol{G}_0$ as the MAP graph of the posterior in Table 3.