Conditional Causal Discovery

Cixuan Zhang[1] **Benjie Wang**[1]

[1]Department of Computer Science, University of California, Los Angeles

Abstract

Multivariate causal discovery is the process of inferring a causal graphical model given data generated from that model. Many methods have been developed both for deriving point estimates as well as quantifying uncertainty over causal graphs. However, integrating complex knowledge or "what-if" scenarios, such as an observed causal effect value, into the discovery process remains a significant challenge. In this paper, we follow the Bayesian approach to causal discovery, and propose to cast such problems as conditional inference. The key computational challenge is that such knowledge may be very unlikely, making approaches based on naïve sampling infeasible: that is, there may be no sample which satisfies the condition. To overcome this, we leverage techniques from the rare event estimation literature. Our empirical results on synthetic data illustrate the efficacy of our approach where baselines fail to accurate capture the conditional distribution, and we illustrate its application to the real-world Sachs protein dataset.

1 MOTIVATION

In causal discovery, one seeks to recover the causal generative process responsible for generating some observed data. Often, however, it is not possible to reliably recover a single causal graph, for a number of reasons. First, with only a finite number of samples, statistical uncertainty can make it impossible to recover even the Markov-equivalence class of the ground-truth graph with high confidence. Second, even with unlimited observational data, identifiability limits remain: all DAGs within a Markov-equivalence class entail the same distribution, so one can recover at best that class, not a unique graph. Third, in high-dimensional problems the combinatorial explosion of the DAG search space poses a substantial computational challenge.

Bayesian causal discovery offers a principled way to tackle the first two obstacles by replacing the goal of recovering a single causal graph with representing uncertainty over the true causal graph through a posterior distribution over possible DAGs. However, representing and reasoning over the super-exponential space of DAGs is an even greater computational challenge. As such, it is of vital importance to be able to effectively incorporate any additional knowledge into the inference process both to reduce uncertainty as well as ease the computational burden. This can be viewed as *conditioning* the posterior distribution of DAGs on some event representing this knowledge.

For concreteness, in this work, we focus on the problem of *conditioning on an observed causal effect*. In many experimental settings we possess hybrid information: observational measurements for every variable and one or more interventional estimates of the causal effect between a chosen pair of nodes. This naturally poses an inverse question:

Which graphical structures—and which specific causal paths—can generate a causal effect at least as large as the one we measured?

Real-world decisions often hinge not just on some observed/measured quantitative causal effect, but the explanation behind how those effects occur. Moreover, such explanations can help domain experts to interpret any experimental observations and guide further investigation such as in choosing intervention targets for experiments.

Additionally, when experimental data about a causal effect of interest is not available, it can be useful to analyze the uncertainty implied by the model posterior. For example, regulators might be interested in *tail probabilities*: the probability that a causal effect in the population exceeds some (large) threshold that would merit further investigation and/or action (e.g. the effect of smoking on cancer).

The central challenge in applying existing Bayesian causal

Accepted for the 1st Workshop on Causal Abstractions and Representations (CAR) at UAI 2025 (CAR 2025).

discovery algorithms is that they are not able to integrate conditional information effectively (e.g. observed causal effect) in the posterior inference process. As we find in our experiments, conditioning post-hoc on the observed information (e.g. by filtering out posterior samples) can lead to inaccurate inference, either when the observed causal effect has low posterior probability or if the Bayesian causal discovery method exhibits significant bias (e.g. variational approaches).

Answering conditional causal queries effectively requires concentrating the sampling effort precisely on the parameter space's subregion that is compatible with the target effect threshold. Our proposed framework, which combines a rare-event estimation method known as Multilevel Splitting with Structure MCMC, does exactly that. Multilevel splitting tackles conditioning by progressively sampling DAGs and parameters conditioned on a series of intermediate levels, progressively guiding samples toward larger causal effects. By the final level, we have amassed sufficiently many DAG–parameter samples whose causal effect $\mathrm{CE}(i{\to}j)$ exceeds the target. The sampling process simultaneously provides estimates of $\Pr\big[\mathrm{CE}(i{\to}j) > t\big]$ for any causal effect threshold t, and offer a concrete answer set for practitioners seeking structural explanations.

Contributions. By integrating multilevel splitting to enable conditioning in Bayesian structure learning, we provide a practical toolset for:

1. integrating knowledge into the Bayesian causal discovery process;
2. rigorously quantifying the posterior probability of extreme causal effects;
3. retrieving informative DAG and path examples that explain *why* a large causal effect appears

2 RELATED WORK

Bayesian causal discovery casts causal discovery as a posterior inference problem, given a prior over causal graphs and a likelihood function. This enables one to capture uncertainty over the underlying causal graph. A very popular approach to the inference problem is to utilize Markov chain Monte Carlo (MCMC) sampling over graphs, or higher level abstractions such as orders [Friedman and Koller, 2003, Kuipers and Moffa, 2017, Viinikka et al., 2020, Giudice et al., 2023]. There have also been many other approaches which attempt to infer a variational posterior over graphs [Annadani et al., 2021, Lorch et al., 2021, Cundy et al., 2021, Wang et al., 2022, Deleu et al., 2022, Rittel and Tschiatschek, 2023, Toth et al., 2024]. The goal is typically either to produce a sample of graphs representing uncertainty over causal relations, or to use Bayesian model averaging to perform *Bayesian causal inference* [Toth et al., 2022]. However, these methods typically do not allow one to condition on arbitrary properties, such as the presence of a large causal effect. Our work develops an effective new inference strategy for these conditional posterior inference problems, where existing methods fail to produce accurate probability estimates and samples.

3 PRELIMINARIES

Causal Bayesian Networks A Bayesian network (BN) (G, Θ) is a probabilistic model $p(\boldsymbol{X})$ over d variables $\boldsymbol{X} = \{X_1, ..., X_d\}$, specified using the directed acyclic graph (DAG) G, which encodes conditional independencies in the distribution p, and Θ, which parameterizes the mechanisms (conditional probability distributions) constituting the Bayesian network. The conditional probabilities take the form $p(X_i|\mathrm{pa}_G(X_i), \Theta_i)$, giving rise to the joint data distribution:

$$p(\boldsymbol{X}|G, \Theta) = \prod_i p(X_i|\mathrm{pa}_G(X_i), \Theta_i)$$

where $\mathrm{pa}_G(X)$ denotes the parents of X in G. In this work, we will focus on the specific case of linear Gaussian models, under which the distribution is given by the structural equation $\boldsymbol{X} = \boldsymbol{X}B + \boldsymbol{\epsilon}$, where $B \in \mathbb{R}^{d\times d}$ is a matrix of real weights parameterizing the mechanisms, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{b}, \Sigma)$ where $\boldsymbol{b} \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}_{\geq 0}$ is a diagonal matrix of noise variances. In particular, for a given DAG G, we have $B_{ij} = 0$ for all i, j such that i is not a parent of j in G.

Whereas Bayesian networks typically only express probabilistic (conditional independence) information, causal Bayesian networks [Spirtes et al., 2000, Pearl, 2009] are additionally imbued with a causal interpretation, where, intuitively, the directed edges in G represent direct causation. More formally, causal BNs can predict the effect (change in joint distribution) of interventions in the system, where some mechanism is changed, for instance by setting a variable X to some value x independent of its parents. In the linear Gaussian case, the average causal effect (ACE) of variable X_i on variable X_j, written $CE(i \to j|G, B)$, is given by:

$$CE(i \to j|G, B) = (I - B)^{-1}_{ij}$$

Bayesian Causal Discovery *Causal discovery* [Koller and Friedman, 2009, Glymour et al., 2019] is the problem of inferring the DAG G of the (causal) Bayesian network responsible for generating some given data $\mathcal{D}$. Typically, strong assumptions are required for causal discovery; in this work, we make the common assumption of causal sufficiency, meaning that there are no latent (unobserved) confounders. Even given this assumption, it is often not possible to reliably infer the causal DAG, whether due to limited data, or non-identifiability within a Markov equivalence class.

Instead of learning a single DAG, Bayesian approaches to causal discovery express uncertainty over structures in a unified fashion, through defining a prior $p(G)$ and (marginal) likelihood $p(\mathcal{D}|G)$ over directed graphs G.

The graph prior is typically chosen to penalize larger parent sets. In this work, we adopt a standard independent Bernoulli edge prior $p(G) \propto p_{\text{edge}}^{|E|}(1-p_{\text{edge}})^{d(d-1)/2-|E|}$ for some $p_{\text{edge}} \in (0,1)$. In the case of linear Gaussian models, the typical choice of prior over the parameters leads to a closed-form marginal likelihood $p(\mathcal{D}|G)$ known as the *BGe* score [Geiger and Heckerman, 1994, 2002]. The posterior over parameters $p(B|\mathcal{D},G)$ then factorizing as $p(B|G,\mathcal{D}) = \prod_i p(B_i|G_i,\mathcal{D})$ where each component is a multivariate t-distribution [Viinikka et al., 2020]. The overall posterior can then be computed as:

$$p(G,B \mid \mathcal{D}) \;\propto\; p(G)\,p(\mathcal{D} \mid G)\,p(B \mid G,\mathcal{D}). \qquad (1)$$

4 CONDITIONAL CAUSAL DISCOVERY

In this paper, we tackle the related problems of estimating the probability of some property $\mathcal{E}$ and drawing samples of graphs and edge weights conditional on that property. In particular, we will consider estimating the probability of (and sampling from) *extreme causal effects*. Specifically, for fixed (i,j) and a threshold $t > 0$ of interest, define the property

$$\mathcal{A}_t = \big\{(G,B) : \mathrm{CE}(i \to j \,|\, G,B) > t\big\}.$$

The quantity we wish to estimate is the posterior probability

$$p\big(\mathcal{A}_t \mid \mathcal{D}\big) \;=\; \iint_{\mathcal{A}_t} p(G,B \mid \mathcal{D})\; dB\, dG,$$

When the threshold t is large, this probability can become extremely small, e.g. $< 10^{-8}$. As such, sampling from the posterior (e.g., running a MCMC chain targeting the posterior) may struggle to return any samples satisfying the property, leading to high-variance estimates.

We therefore adopt the multilevel splitting (MLS) strategy [Kahn and Harris, 1951, Guyader et al., 2011], which decomposes the estimation problem into a series of simpler conditional estimation problems. In particular, given a sequence of levels $-\infty = L_0 < L_1 < \ldots < L_K$ where $L_K = t$, one aims to estimate the probability by decomposing into:

$$p\big(\mathcal{A}_t \mid \mathcal{D}\big) = p\big(\mathcal{A}_{L_0} \mid \mathcal{D}\big) \prod_{i=1}^{K} p\big(\mathcal{A}_{L_i} \mid \mathcal{A}_{L_{i-1}}, \mathcal{D}\big) \qquad (2)$$

Since all (G,B) pairs are in A_{L_0}, $p\big(\mathcal{A}_{L_0} \mid \mathcal{D}\big) = 1$. MLS iteratively constructs sets of samples from the conditional distributions $p(G,B|\mathcal{A}_{L_i},\mathcal{D})$ starting from $i = 0$.

Adaptive MLS In adaptive multilevel splitting (AMLS) [Cérou and Guyader, 2007], the levels $L_1,\ldots,L_K$ are set adaptively instead of being chosen by the user. On a high-level, this estimation is performed as follows:

1. Given the current level L_k, run an inner MCMC (Sec. 4.1) for every particle (G,B) to obtain a new set of samples conditionally distributed on $\mathrm{CE} > L_k$. The inner chain length m controls the mixing of the chain.
2. Compute all n CE values for the samples, and set L_{k+1} to the q-th upper quantile (e.g. $q = 0.9$).
3. Estimate $\alpha_k := p\big(\mathcal{A}_{L_{k+1}} \mid \mathcal{A}_{L_k},\mathcal{D}\big)$ by the empirical survival fraction $\widehat{\alpha}_k = \#\{\mathrm{CE} > L_{k+1}\}/n$.
4. Resample with replacement the surviving particles so that the population size returns to n; copy their current states as starting points for the next iteration.

Algorithm 1 summarizes the overall approach, which we call *MLS-parameter* as it operates over the joint (G,B) space. The algorithm enables us to simultaneously estimate the probabilities for a set of target thresholds $t_1 < \ldots < t_T$ in one execution of the algorithm (though these estimates will be correlated).

Besides estimating the probability, the multilevel splitting procedure also provides as a byproduct a sample of graphs and edge weights conditional on the causal effect being greater than t for any threshold t. This can then be used e.g. to interpret possible explanations for a large causal effect or for Bayesian model averaging.

Particle initialisation We begin with n DAGs $\{G_r^{(0)}\}_{r=1}^n$ drawn i.i.d. from the structural prior $p(G)$. For each of the DAGs, we draw a posterior weight matrix $B_r^{(0)} \sim P(B \mid G_r^{(0)},\mathcal{D})$ which follows a multivariate t-distribution as previously mentioned.

4.1 INNER METROPOLIS–HASTINGS OVER JOINT (G,B)

In the inner MCMC update part of Algorithm 1, we run m Metropolis-Hastings (MH) steps as follows. In each step, we traverse the joint (G,B) space based on two reversible proposal distributions:

- *Structure move*: draw (G',B') by adding, deleting or reversing a single edge; when a new edge $(u \to v)$ is created, we sample an initial weight from $N(\mu,\sigma^2)$, where $\mu = 0$ and $\sigma = 1$.
- *Weight move*: pick an existing edge $(u \to v)$ and perturb B_{uv} by a zero-mean Gaussian step of variance τ^2, where $\tau = \eta \cdot \max(1, |B_{uv}|)$. Here, $\eta > 0$ is a step size hyperparameter that scales the perturbation

Algorithm 1: Adaptive Multilevel Splitting for $p(x) = \Pr\big[\mathrm{CE}_{i\to j} > x\big]$ (single run, multiple thresholds $\mathbf{x}$)

Input : dataset $\mathcal{D}$; CE threshold list $\mathbf{t} = (t^{(1)}, \ldots, t^{(T)})$; particle count n; inner MCMC length m; quantile q; max. outer iterations $K_{\max}$

Output : estimates $\{\widehat{p}(t^{(i)})\}_{i=1}^{T}$; survivor particle set

```
// Initialization
```

for $r \leftarrow 1$ **to** n **do**
  sample $G_r^{(0)} \sim \mathrm{DAGPrior}(p_{\mathrm{edge}})$
  sample $B_r^{(0)} \sim P\big(B \mid G_r^{(0)}, \mathbf{X}\big)$
end

$k \leftarrow 0$; $L_0 \leftarrow -\infty$; $S \leftarrow 0$; $i \leftarrow 0$;

```
// Core Algorithm
```

while $k < K_{\max}$ & $i \le T$ **do**
  for $r \leftarrow 1$ **to** n **do**
    run m MH steps in Sec.4.1
  end
  Compute $\theta_r \leftarrow \mathrm{CE}_{i\to j}(G_r^{(k)}, B_r^{(k)})$
  $L_{k+1} \leftarrow q$-quantile of $\{\theta_r\}_{r=1}^{n}$
  $\widehat{\alpha}_k \leftarrow \#\{\theta_r > L_{k+1}\}/n$
  while $L_{k+1} \ge t^{(i)}$ **do**
    calculate $\beta = \#\{\theta_r > t^{(i)}\}/n$
    $\widehat{p}\big(t^{(i)}\big) \leftarrow \exp(S + log(\beta))$;
    $i \leftarrow i + 1$
  end
  $S \leftarrow S + log(\widehat{\alpha}_k)$
  resample survivors to restore n particles
  $k \leftarrow k + 1$
end

Result: $\{\widehat{p}(t^{(i)})\}_{i=1}^{T}$ and level-k particle set

relative to the current weight. In all experiments, we set $\eta = 0.8$, so $\tau^2 = 0.64 \cdot \max\left(1,\ |B_{uv}|\right)^2$

The proposal is accepted with probability

$$\min\Big\{1,\ \frac{p(G', B' \mid \mathcal{D}, \mathcal{A}_{L_k})\, q\big((G', B') \to (G, B)\big)}{p(G, B \mid \mathcal{D}, \mathcal{A}_{L_k})\, q\big((G, B) \to (G', B')\big)}\Big\},$$

Here, q represents the probability of making the proposed move. The target distribution $p(G, B \mid \mathcal{D}, \mathcal{A}_{L_k}) \propto p(G, B \mid \mathcal{D}) \mathbb{1}_{\mathrm{CE}(i\to j|G,B) > L_k}$, where $p(G, B \mid \mathcal{D})$ is given by Eq. (1). This ensures that we always have a set of samples above with causal effect above the current level L_k.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate all methods on synthetic linear-Gaussian data sets generated from random Erdős–Rényi DAGs at four dimensions:

$$d \in \{4,\ 8,\ 16,\ 32\}.$$

As for the ground truth edge weights, we independently sample $\tilde{B}_{ij} \sim \mathcal{N}(2, 1)$ for every node pair (i, j), then set $B = \tilde{B} \odot G$, so that $B_{ij} = \tilde{B}_{ij}$ if $G_{ij} = 1$ and $B_{ij} = 0$ otherwise. The choice of mean edge weight of 2 is chosen to induce potentially large causal effects. We generated an observational dataset $\mathcal{D}$ of size $n = 1000$ sampled from this causal model.

Given a randomly generated ground truth graph and edge weights (G, B), we choose a pair (i, j) of nodes, calculate the ground truth causal effect $t_1 := CE(i \to j|G, B)$, and then choose a sequence of ascending thresholds (described in detail in Appendix A.3):

$$t_1 < t_2 < \cdots < t_T.$$

All methods are then tasked with estimating $p(\mathcal{A}_{t_i}|\mathcal{D}) = p(\mathrm{CE}(i \to j) > t_i \mid \mathcal{D})$ for every i.

Compared methods. We implement and compare the following methods:

1. *Exhaustive enumeration:* Enumerating all DAGs (for the $d = 4$ experiment only) and sampling 4000 weight samples per graph;
2. *OrderSPN [Wang et al., 2022]*: This method constructs an approximate represention of the posterior as a probabilistic circuit. For evaluation, we take 5000 graphs sampled from the OrderSPN model, and sample 200 edge weights from the posterior given each graph.
3. *Single MCMC chain*: This is a single, long Metropolis–Hastings chain over both graphs and edge weights, run for 500000 MCMC iterations, burn-in 10%.
4. *DiBS [Lorch et al., 2021]*: DiBS is a variational approach that returns samples jointly over the graph and edge weights. We take 10000 graph and edge weight samples (100 runs with 100 samples for each run as the runtime scales quadratically in the number of samples).
5. *MLS–parameter*: For our MLS-parameter method, we use $n = 400$ particles, $m = 4000$ MCMC iterations per level, $K_{max} = 15$.
6. *MLS–structure*: We also test a variant of MLS which operates just over the DAG space using the posterior $p(G|\mathcal{D}) \propto p(G)p(\mathcal{D}|G)$, and where the causal effect of a graph is defined by the expected CE under the posterior over edge weights $p(B|G, \mathcal{D})$. As with MLS-parameter, we use $n = 400$ particles, $m = 4000$ MCMC iterations per level, $K_{max} = 15$.

5.2 SMALL SCALE VALIDATION ($d = 4$)

We begin by validating each method in $d = 4$ setting, where we can still explicitly enumerate every directed acyclic

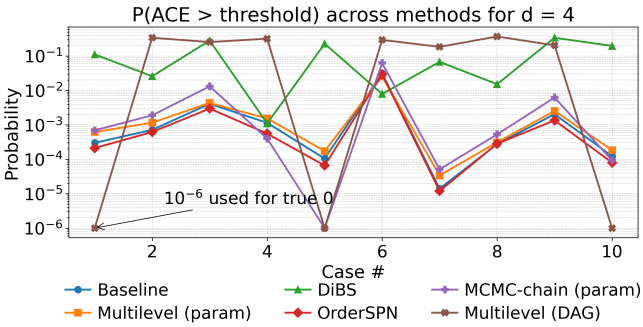

Figure 1: Estimated tail probabilities $P(\mathrm{CE} > x_t)$ for ten node pairs—one pair from each of ten randomly generated $d = 4$ graphs—with exhaustive enumeration as the ground-truth reference.

graph (DAG) ($|\mathcal{G}_{d=4}| = 543$). We can thus very accurately compute the tail probabilities $p(\mathrm{CE}(i, j) > t)$ and treat these values as a gold-standard baseline.

Figure 1 shows the estimated tail probabilities $\widehat{p}_t = P\big(\mathrm{CE} > t \mid \mathbf{X}\big)$ for ten random node pairs across ten data sets, with the exhaustive enumeration (Baseline, blue) as ground truth. We observe that:

- *DiBS* (green) consistently overestimates, returning probabilities between 10^{-2} and 10^{-1} in 8/10 cases. This systematic bias can be explained by the fact that DiBS is a variational method and thus has no convergence guarantees;
- *MLS–structure* (brown) fails to accurately capture the probability, which is unsurprising as it operates over structures (and average causal effects given structures);
- *Single MCMC chain* (purple) matches the ground-truth in 9/10 cases indicating adequate mixing; Case 5 is the lone outlier (which we analyse in Appendix A.2).
- *MLS–parameter* (orange) and *OrderSPN* (red) both track the ground-truth closely in all 10 cases, spanning probabilities over four orders of magnitude ($10^{-6} \to 10^{-2}$);

5.3 LARGE SCALE EVALUATION ($d = 8$, $d = 16$, AND $d = 32$)

We now consider scaling to higher dimensions, where exhaustive enumeration of all possible DAGs is no longer feasible and so we no longer have ground-truth probabilities. Based on the results from the low-dimensional case, we will focus primarily on OrderSPN, the long MCMC chain, and MLS-parameter. We assess rare-event performance on synthetic datasets with $d \in \{8, 16, 32\}$. For each dimension we plot the tail probability $P(\mathrm{CE}(i, j) > t)$ for different values of t, using a log-scale for the probability. A robust method should decay smoothly rather than collapse to zero. We additionally provide error bars for MLS–parameter on

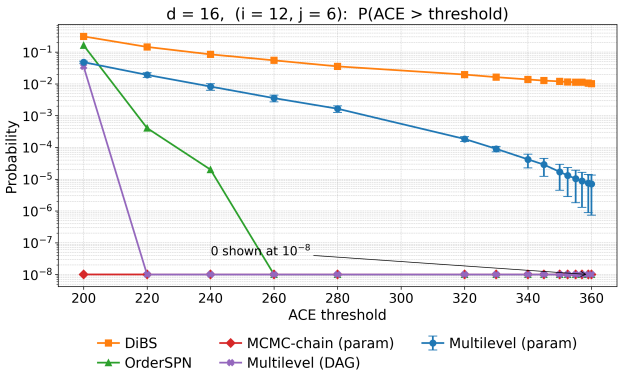

Figure 2: Estimated tail-probabilities for a representative node pair in the $d = 16$ setting. MLS-parameter shows standard error (s.e.) over 5 runs.

the $d = 8$ and $d = 16$ synthetic datasets, which appear in Figs. 8 and 2. Error bars show $\bar{p}_t \pm \mathrm{SE}_t$ from 5 independent MLS runs.From Figures 8, 2, and 9, we can make the following observation on each model.

- *MLS–parameter* (blue) shows a steady decline in probability from 10^{-1} down to values smaller than 10^{-6} for the larger causal effect thresholds.
- *DiBS* (orange) repeats the overestimation seen at d=4 and suggests its variational posterior is also not sufficiently accurate in higher dimensions;
- *OrderSPN* (green) follows MLS–parameter at the first few thresholds but then drops to zero abruptly, indicating that the OrderSPN posterior representation is not able to cover the low-probability graphs with large causal effect in higher dimensions.
- *MLS–structure* (purple) cannot capture extreme causal effects for similar reasons for the small-d case;
- *Single MCMC chain* (red) exhibits a (perhaps surprising) failure to capture any graph/weights with large causal effects. Upon further analysis, we found that the chain struggles to jump between causal effect modes indicating poor mixing even with 500K steps.

In summary, MLS-parameter is the only method that is able to accurately estimate the probability for large causal effect thresholds (as shown by the tight error bars). All other methods except DiBS also fail to return any samples conditional on large causal effects; however, as we have seen, DiBS is not accurate even in the low-d setting.

5.4 EFFECT OF TRAINING SAMPLE SIZE

In practice, the amount of data available to fit a causal model varies widely. To illustrate how the probability of causal effects varies in these settings, we fixed the $d = 16$ setting from Section 5.3 (target pair $(12, 6)$) and re–ran the sampler

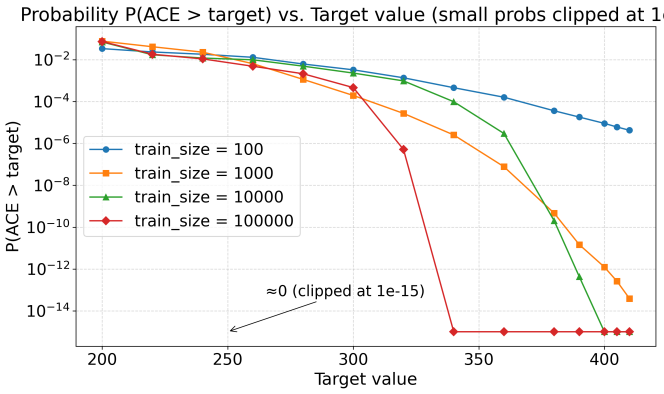

Figure 3: Effect of training-set size on $\Pr\big(\mathrm{CE}_{12,6} > t\big)$. Each curve is an independent MLS run with identical hyper-parameters but different numbers of observations.

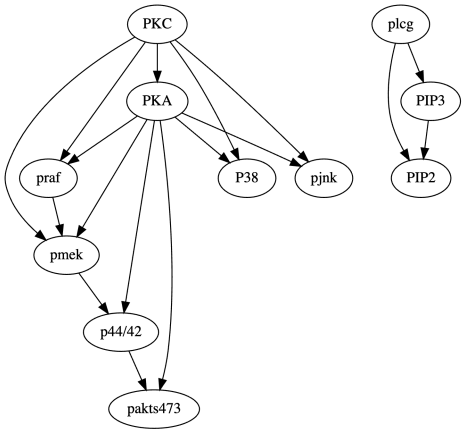

Figure 4: Ground-truth DAG for the Sachs protein dataset [Sachs et al., 2005].

with training sizes

$$n \in \{100,\ 1\,000,\ 10\,000,\ 100\,000\}.$$

Figure 3 demonstrates the resulting tail–probability curves; probabilities smaller than 10^{-15} are clipped to that value. The results shows a clear and intuitive pattern: as the training size increases, the posterior concentrates around a smaller set of graphs/edge weights that best explain data. As we have chosen thresholds much larger than the true causal effect, the probability of causal effect exceeding these thresholds becomes smaller as the training size increases.

5.5 SACHS PROTEIN DATASET

We now evaluate on the real-world Sachs dataset [Sachs et al., 2005], which provides 7 466 measurements of expression levels of 11 proteins and phospholipids. The DAG in Figure 4 is the consensus ground-truth causal graph. In this experiment, rather than trying to recover this ground-truth graph, we are interested in answering the what-if question:

Table 1: Node index ↔ protein name.

Index	Protein	Index	Protein
0	praf	6	pakts473
1	pmek	7	PKA
2	plcg	8	PKC
3	PIP2	9	P38
4	PIP3	10	pjnk
5	p44/42		

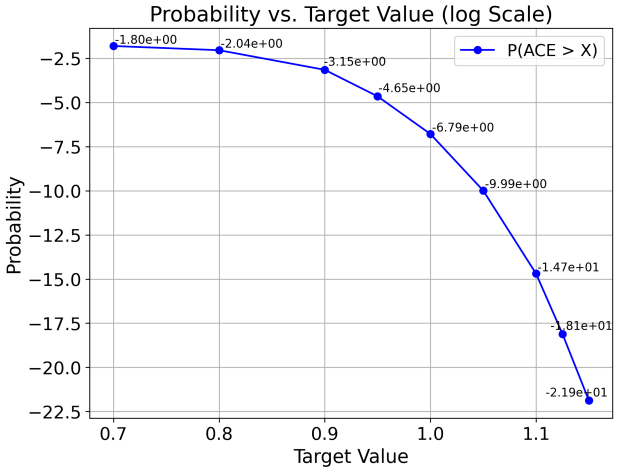

Figure 5: MLS tail probability for $\mathrm{CE}_{8,1}$ (PKC → pmek) in log scale. Empirical CE: $x_{\text{emp}} \approx 0.89$.

> *"When an unusually large causal effect is observed/hypothesized, which pathways form the most plausible explanation?"*

In particular, we illustrate this for one node pair (node number mapped to variables in Table 1) in the Sachs network:

$$(8 \to 1)\ [\text{PKC} \to \text{pmek}].$$

In the true graph, we have four paths between 8 and 1:

$$8 \to 1, \quad 8 \to 0 \to 1, \quad 8 \to 7 \to 1, \quad 8 \to 7 \to 0 \to 1$$

We now apply our MLS-parameter method, which in addition to probability estimates also provides a set of samples from the posterior conditional on the causal effect being greater than some threshold. In order to analyse the effect of each path, we compute (i) in how many of the samples the path appears; and (ii) among the samples where the path appears, the average contribution to the causal effect (given by the product of edge weights along the path). We test nine thresholds between $t = 0.70$ and $t = 1.15$. Figure 5 shows the tail probability is fairly large for $0.70 \le x \le 0.90$ but drops sharply once $x > 0.90$.

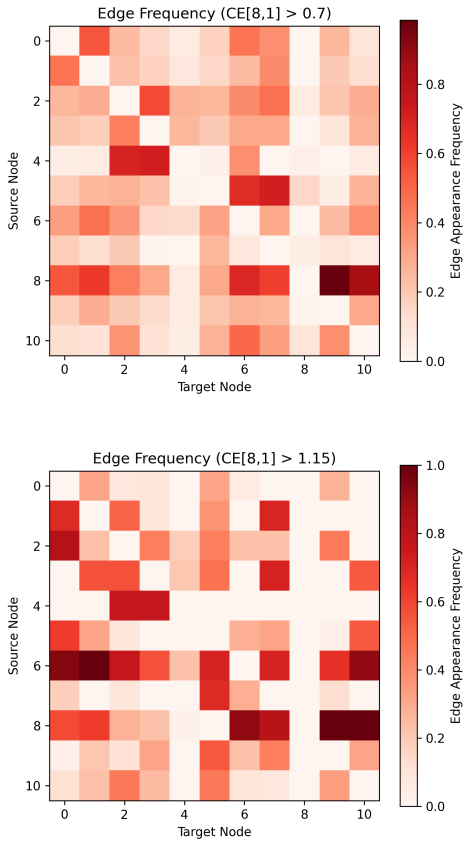

Figure 6: Edge-frequency heat-maps for $(8, 1)$ at $x = 0.70$ (top) and $x = 1.15$ (bottom).

Table 2: Most frequent paths occurring in samples at $t = 0.70$ (out of 66 graphs)

Occurrences	Path	Contribution
41	$(8, 1)$	0.1857
28	$(8, 0, 1)$	0.4580
26	$(8, 6, 1)$	0.1335
19	$(8, 9, 1)$	0.5942
12	$(8, 2, 1)$	0.07711
8	$(8, 3, 2, 1)$	0.07487

Causal Paths At the lowest threshold, $t = 0.70$, Table 2 shows a large collection of $(8 \to \ldots \to 1)$ paths. The MLS posterior comprises 66 graphs containing 414 distinct $8 \rightsquigarrow 1$ paths. Although the direct edge $(8 \to 1)$ is already the most common—present in 41 of the 66 graphs—its mean contribution (0.19) remains below the threshold. Several indirect paths, such as $(8 \to 0 \to 1)$, $(8 \to 9 \to 1)$ and $(8 \to 6 \to 1)$, occur almost as often and achieve markedly higher average effects (0.46, 0.59, and 0.13, respectively). Hence, no single edge/path is entirely responsible for the causal effect at this level; the heat map in Fig. 6 (left) reflects this significant amount of uncertainty.

Raising the threshold to $t = 1.15$ increases edge certainty,

Table 3: Top positive and negative paths contributions $t = 1.15$ (out of 101 graphs)

Occurrences	Path	Contribution
Most Frequent Paths		
92	$(8, 6, 1)$	0.1995
63	$(8, 1)$	0.4863
57	$(8, 6, 3, 1)$	0.0667
32	$(8, 6, 9, 3, 1)$	0.0021
29	$(8, 0, 1)$	0.4994
Most Positive Contribution Paths		
3	$(8, 10, 6, 0, 1)$	0.6533
21	$(8, 9, 1)$	0.5843
23	$(8, 9, 6, 0, 1)$	0.5781
29	$(8, 0, 1)$	0.4994
63	$(8, 1)$	0.4863
3	$(8, 9, 0, 1)$	0.4716
Most Negative Contribution Paths		
3	$(8, 10, 6, 5, 0, 1)$	-0.3090
23	$(8, 6, 0, 1)$	-0.2593
23	$(8, 9, 5, 0, 1)$	-0.0904
23	$(8, 2, 3, 5, 0, 1)$	-0.0384
63	$(8, 10, 6, 5, 1)$	-0.0327

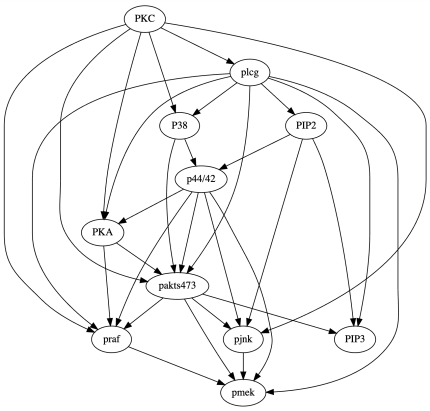

Figure 7: Sampled graph from $t = 1.15$

as shown in Fig. 6 (right). Path frequency now concentrates sharply (Table 3): path $(8 \to 6 \to 1)$ appears in 92 of the 101 graphs, while the direct path $(8 \to 1)$ is almost as frequent (63/101). Notably, only two paths — $(8 \to 0 \to 1)$ and $(8 \to 1)$ — have both high frequency and high positive contribution, and they are all among the four ground-truth paths between nodes 8 and 1, indicating that the sampled graphs, even under an extreme threshold, still follow key features of the underlying causal structure.

Taking one sample graph (Fig.7) as an example, we find 45 distinct paths from node 8 (PKC) to node 1 (pJNK), yet their contributions to the causal effect differ greatly, as shown in Table 4, which groups the paths into major positive (each adding at least 0.05) and negative contributors (each subtracting at least 0.02). Among them, path $(8 \to 9 \to 6 \to 0 \to 1)$ alone supplies about 50% of the total positive impact, while path $(8 \to 0 \to 1)$ contributes another 22%, and together they account for more than 70% of the effect. Meanwhile the largest negative path, $(8 \to 6 \to 0 \to 1)$, removes about 24% of the positive effect.

We emphasize that while these pathways are not testable (and in fact not necessarily present in the true graph), they can help narrow down the search space for a human expert.

Table 4: Dominant positive and negative $8 \to 1$ paths in the sampled graph at $t = 1.15$

Path	Edge–product expression	Contribution
	Largest positive contributions	
$(8, 9, 6, 0, 1)$	$5.06 \times 0.108 \times 0.714 \times 1.481$	0.58
$(8, 0, 1)$	0.239×1.481	0.35
$(8, 2, 6, 0, 1)$	$0.696 \times 0.197 \times 0.714 \times 1.481$	0.14
$(8, 2, 0, 1)$	$0.696 \times 0.136 \times 1.481$	0.14
	Largest negative contributions	
$(8, 6, 0, 1)$	$(-0.251) \times 0.714 \times 1.481$	-0.27
$(8, 9, 5, 0, 1)$	$5.06 \times 0.009 \times (-1.333) \times 1.481$	-0.08
$(8, 6, 1)$	$(-0.251) \times 0.148$	-0.04
$(8, 2, 3, 5, 0, 1)$	$0.696 \times 1.597 \times 0.016 \times (-1.333) \times 1.481$	-0.03

For instance, a domain scientist could focus first on these high-impact pathways—e.g., by selectively perturbing praf (node 0) or pakts473 (node 6) and observing the change in plcg activation—before turning to the minor paths whose individual contributions are an order of magnitude smaller.

6 CONCLUSIONS

Summary. We introduced a novel framework for causal discovery conditional on additional knowledge, such as a pairwise causal effect. Our method uses the adaptive multilevel-splitting (MLS) framework to infer the posterior probability over causal effects while also producing interpretable graph samples consistent with the causal effect property. We show that, compared with baselines, our MLS–parameter method remains accurate from $d = 4$ to $d = 32$; while returning interpretable samples explaning the observed causal effect.

Limitations and future work. In this study, our empirical results have been focused on (i) linear Gaussian models and (ii) conditioning on a single causal effect pair; but our framework is not restricted to these cases. Future work could thus study other properties of interest (such as a combination of observed causal effects between different node pairs). Further, to improve the MCMC mixing during each level, one could consider incorporating more sophisticated MCMC schemes over the DAG space [Kuipers and Moffa, 2017, Viinikka et al., 2020].

References

Yashas Annadani, Jonas Rothfuss, Alexandre Lacoste, Nino Scherrer, Anirudh Goyal, Yoshua Bengio, and Stefan Bauer. Variational causal networks: Approximate bayesian inference over causal structures. *arXiv preprint arXiv:2106.07635*, 2021.

Frédéric Cérou and Arnaud Guyader. Adaptive multilevel splitting for rare event analysis. *Stochastic Analysis and Applications*, 25(2):417–443, 2007.

Chris Cundy, Aditya Grover, and Stefano Ermon. Bcd nets: Scalable variational approaches for bayesian causal discovery. *Advances in Neural Information Processing Systems*, 34:7095–7110, 2021.

Tristan Deleu, António Góis, Chris Emezue, Mansi Rankawat, Simon Lacoste-Julien, Stefan Bauer, and Yoshua Bengio. Bayesian structure learning with generative flow networks. In *Uncertainty in Artificial Intelligence*, pages 518–528. PMLR, 2022.

Nir Friedman and Daphne Koller. Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. *Machine learning*, 50:95–125, 2003.

Dan Geiger and David Heckerman. Learning gaussian networks. In *Uncertainty in Artificial Intelligence*, pages 235–243. Elsevier, 1994.

Dan Geiger and David Heckerman. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics*, 30(5):1412–1440, 2002.

Enrico Giudice, Jack Kuipers, and Giusi Moffa. A bayesian take on gaussian process networks. *Advances in Neural Information Processing Systems*, 36:56602–56614, 2023.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00524.

Arnaud Guyader, Nicolas Hengartner, and Eric Matzner-Løber. Simulation and estimation of extreme quantiles and extreme probabilities. *Applied Mathematics & Optimization*, 64(2):171–196, 2011.

Herman Kahn and Theodore E Harris. Estimation of particle transmission by random sampling. *National Bureau of Standards applied mathematics series*, 12:27–30, 1951.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

Jack Kuipers and Giusi Moffa. Partition mcmc for inference on acyclic digraphs. *Journal of the American Statistical Association*, 112(517):282–299, 2017.

Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. Dibs: Differentiable bayesian structure learning. *Advances in Neural Information Processing Systems*, 34:24111–24123, 2021.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Simon Rittel and Sebastian Tschiatschek. Specifying prior beliefs over dags in deep bayesian causal structure learning. In *ECAI 2023: 26th European Conference on Artificial Intelligence, including 12th Conference on Prestigious Applications of Intelligent Systems, PAIS 2023-Proceedings*, pages 1962–1969. IOS Press, 2023.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Christian Toth, Lars Lorch, Christian Knoll, Andreas Krause, Franz Pernkopf, Robert Peharz, and Julius Von Kügelgen. Active bayesian causal inference. *Advances in Neural Information Processing Systems*, 35: 16261–16275, 2022.

Christian Toth, Christian Knoll, Franz Pernkopf, and Robert Peharz. Effective bayesian causal inference via structural marginalisation and autoregressive orders. *arXiv preprint arXiv:2402.14781*, 2024.

Jussi Viinikka, Antti Hyttinen, Johan Pensar, and Mikko Koivisto. Towards scalable bayesian learning of causal dags. *Advances in Neural Information Processing Systems*, 33:6584–6594, 2020.

Benjie Wang, Matthew R Wicker, and Marta Kwiatkowska. Tractable uncertainty for structure learning. In *International Conference on Machine Learning*, pages 23131–23150. PMLR, 2022.

A APPENDIX

A.1 ADDITIONAL PLOTS FOR $d = 8$ AND $d = 32$

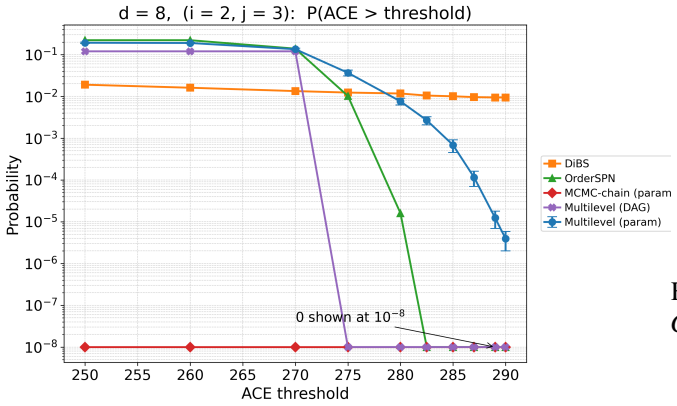

Figure 8: Estimated tail–probabilities for a representative node pair in the $d = 8$ setting. MLS-parameter shows standard error (s.e.) over 5 runs.

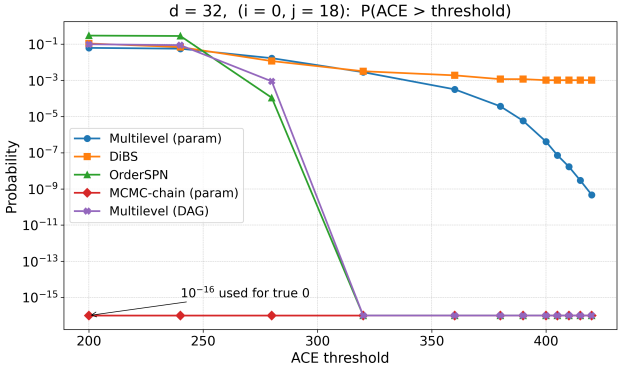

Figure 9: Estimated tail-probabilities for a representative node pair in the $d = 32$ setting.

A.2 INVESTIGATION OF MCMC CHAIN ON CASE 5 ($d = 4$)

In Section 5.2, we noted that running a single, long MCMC chain can fail to capture the distribution over causal effects accurately. From Fig 10, we see that true distribution of causal effect follows a multimodal pattern, and the causal effect of the ground truth graph lies in the right-most peak centered around 25. As Fig 11 shows, the single MCMC chain's distribution is centered around 8. This can be explained by the fact that the initial particle lies in this peak, and the MCMC chain fails to mix because of the distinct qualitative structure of the graphs corresponding to the different peaks.

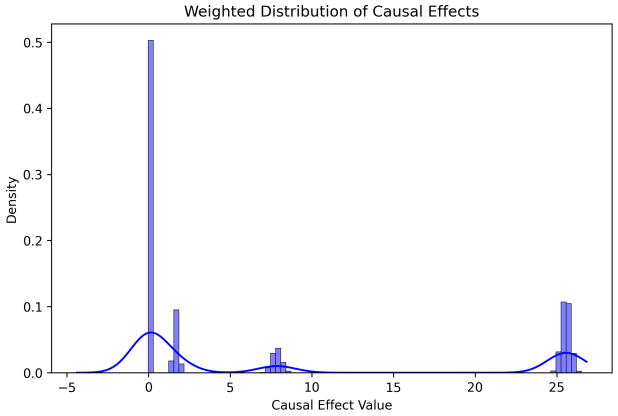

Figure 10: Ground-truth posterior distribution of $\mathrm{ACE}_{3,0}$ in *Case 5*

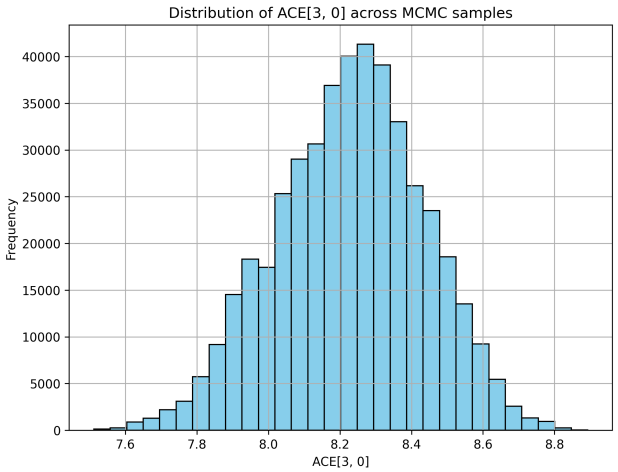

Figure 11: Posterior of $\mathrm{ACE}_{3,0}$ with $p_{\text{edge}} = 0.2$

A.3 TARGET–THRESHOLD CONSTRUCTION

We select the target causal effects thresholds for evaluation as follows, in order to test the ability of the methods to estimate probabilities of causal effects much larger than the true effect.

1. **True effect**: Let t_0 denote the *true* pair-wise CE for the node pair (i, j); this is treated as the starting point.
2. **Pilot run**: Execute a short, low-budget adaptive multilevel-splitting (MS) pilot. Identify a threshold t' for which the estimated tail probability already lies in the rare-event regime (about 10^{-8} or less).
3. **Threshold grid**: Form an ascending sequence

$$t_0 = t_1 \ < \ t_2 \ < \ \ldots \ < \ t_T = t',$$

spaced geometrically so that thresholds become denser as they approach $t_{\max}$; this focuses evaluation effort where $\Pr[\mathrm{CE} > t]$ is smallest.