Interpretable Causal Discovery via Causal-Effect Constraints

Cixuan Zhang[1] **Guy Van den Broeck**[2] **Benjie Wang**[2]

[1]Computer Science Dept., Yale University, New Haven, Connecticut, USA
[2]Computer Science Dept., University of California, Los Angeles, Los Angeles, California, USA

Abstract

Causal discovery aims to uncover the underlying causal relationships given data generated from a system. The goal, however, is not merely to predict causal edges given data, but also to be able to interpret and explain either observed or hypothesized phenomena, such as a particularly large causal effect. We consider this task of *conditional* causal discovery and cast it as a Bayesian inference problem, in which we target the posterior over causal graphs and parameters conditional on an event such as a causal-effect constraint. Unfortunately, this poses a computational challenge: existing approaches to Bayesian causal discovery struggle when the event has small posterior mass. To address this, we adapt rare-event estimation techniques to perform inference the joint graph-parameter space. Our method gradually drives a particle population toward the constrained region while maintaining samples that approximate the conditional posterior. Empirical evaluation on synthetic graphs validates the accuracy of our approach at small and large scales, and we show in a case study on the Sachs protein dataset how our method can be used to aid scientific exploration by providing pathway-level summaries.

1 INTRODUCTION

Causal discovery is often motivated not only by prediction, but by the need to obtain interpretable and actionable explanations of how a system works. In scientific domains, a practitioner may not simply ask which graph is most probable given the data; rather, they may ask which causal mechanisms could explain a particular domain-relevant phenomenon. For example, in the well-known Sachs study of protein interactions [Sachs et al., 2005], one may want to understand which directed pathways could support an unusually large effect from one protein to another, and whether such pathways suggest plausible interventions or follow-up experiments.

In this work, we formulate this type of "what-if" analysis as a *conditional* form of causal discovery. Given observational data, we seek causal structures that are both statistically plausible and consistent with a user-specified constraint, such as a large causal effect from node i to node j. This presupposes that there could be many graphs that explain the data well, and as such requires a treatment of epistemic uncertainty. We take a Bayesian approach to this problem in which we represent uncertainty over graph structure and parameters. Rather than committing to a single estimated graph, we consider the posterior distribution over directed acyclic graphs (G) and parameters (θ). Our target is the posterior conditioned on an event determined by the graph and parameters, such as the event that the signed causal effect from i to j exceeds a threshold t. The goal is to characterize the causal structures and directed pathways that remain plausible under the causal-effect constraint, while also estimating the posterior probability of the event itself.

Conditional causal discovery poses a computational challenge that is not well handled by current tools. In particular, the event of interest may be rare under the unconstrained posterior, especially when the causal effect threshold t is large, or when we impose a combination of constraints. Standard posterior samplers based on MCMC may produce few or no samples satisfying the constraint, and simple rejection-based conditioning becomes inefficient. Even when constraint-satisfying samples are obtained, poor mixing can lead to unreliable statistical summaries of e.g. different causal paths.

We therefore cast conditional causal discovery under strong (e.g. extreme-effect) constraints as a rare-event posterior inference problem. Our approach integrates adaptive mul-

[0]Our code is available at `https://github.com/ZCX031116/MLS-Framework`.

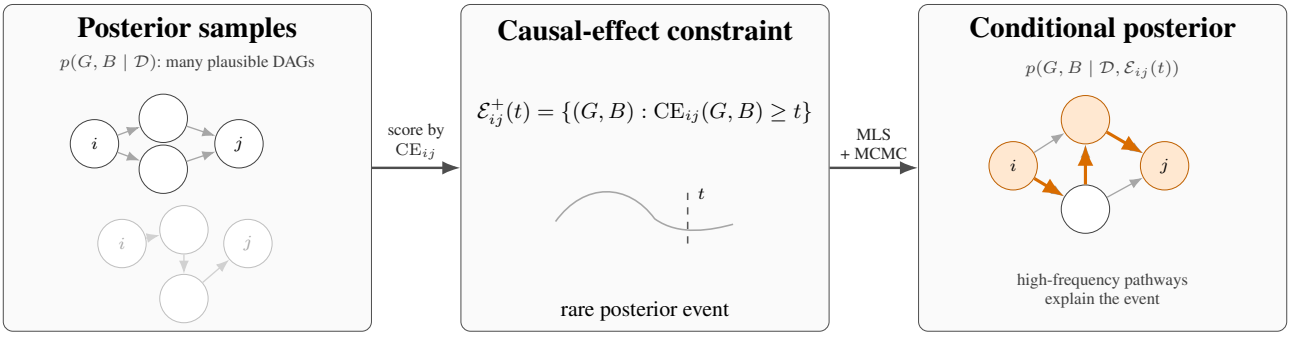


Figure 1: Overview of conditional causal discovery. Starting from the unconstrained Bayesian posterior over causal graphs and edge weights, we condition on a user-specified extreme causal-effect event and obtain a constrained posterior whose samples can be summarized at the pathway level.

tilevel splitting (AMS) [Cérou and Guyader, 2007] with an MCMC kernel over the joint graph–parameter space. Starting from posterior samples obtained from any Bayesian causal discovery method, our method progressively tightens the effect threshold while maintaining a representative particle population, producing both tail-probability estimates and conditional posterior samples. These samples can then be used to summarize which edges and directed pathways are most characteristic of the extreme-effect regime.

Our contributions are as follows:

- We formulate conditional causal discovery as posterior inference under extreme-effect constraints, targeting both constrained posterior samples and posterior tail probabilities.
- We provide a practical rare-event inference procedure that can estimate small posterior tail probabilities while producing representative graph–parameter samples from the corresponding constrained posterior.
- We validate the method on linear-Gaussian benchmarks with ($d \in 4, 8, 16, 32$) and present a Sachs dataset case study showing how extreme-effect conditioning enables pathway-level interpretation and hypothesis generation.

2 RELATED WORK

Causal discovery under background knowledge or structural constraints has been widely studied. Such constraints often encode qualitative statements about graph structure, including required or forbidden edges, paths, or ancestral relations [Meek, 1995, Borboudakis and Tsamardinos, 2012, Anand et al., 2023]. Prior work has considered both identifiability and algorithmic procedures for learning graphs subject to these constraints [Chen et al., 2016, Chen and Darwiche, 2024]. A related line of work uses interventional data as an additional source of information beyond pure observations [Hauser and Bühlmann, 2015, Brouillard et al., 2020]. Instead, we are primarily motivated by interpretability rather than encoding fixed knowledge. Our technical approach differs in that the constraint is encoded flexibly as a quantitative score function, and in particular we encode constraints corresponding to the value of a causal-effect functional rather than a local structural statement. Moreover, rather than seeking exact identification of a single graph, we target a conditional Bayesian posterior that retains uncertainty over both graphical and parametric structure.

Another line of work combines causal discovery with extreme value theory to infer causal directions when causal mechanisms are most visible in the tails of the observed distribution [Gnecco et al., 2021, Pasche et al., 2023, Bodik et al., 2024]. In these approaches, extremes are primarily a feature of the data-generating distribution and are used for identifiability or structure recovery. Our motivation is different: in our setting, the extreme event is a user-specified constraint on a causal-effect functional, and the inferential target is the posterior distribution conditioned on that constraint. Thus, our goal is not to identify causal directions from heavy-tailed observations, but to characterize which posterior graphs, parameters, and pathways explain an unusually large or small causal effect.

For posterior inference over causal graphs, a common approach is to use Markov chain Monte Carlo (MCMC) sampling over DAGs or higher-level representations such as orders [Friedman and Koller, 2003, Kuipers and Moffa, 2017, Viinikka et al., 2020, Giudice et al., 2023]. More recent work has developed efficient DAG-space MCMC samplers with locally informed and adaptive proposals, including PARNI-DAG [Liang et al., 2023], which constructs adaptive random neighborhoods guided by posterior information and can exploit a pre-tuned skeleton to improve scalability. These graph posterior samples are useful not only for representing structural uncertainty, but also for downstream causal inference; recent benchmarking work evaluates Bayesian causal discovery methods through downstream treatment-effect estimation [Emezue et al., 2023].

Alternative approaches approximate the posterior via variational inference 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]. These methods aim either to generate graph samples reflecting structural uncertainty or to support Bayesian model averaging for causal inference [Toth et al., 2022].

Our computational approach depends on the rare-event simulation literature, particularly adaptive multilevel splitting [Cérou and Guyader, 2007]. Splitting methods estimate small probabilities by introducing intermediate thresholds and repeatedly propagating a particle population toward rarer events. We adapt this idea to Bayesian causal discovery by defining levels through a causal-effect score and using MCMC moves over the joint graph–parameter space. This produces both an estimate of the posterior probability of the extreme-effect event and samples from the corresponding conditional posterior.

3 PRELIMINARIES

Causal Bayesian Networks A Bayesian network (BN) (G, θ) is a probabilistic model $p(\boldsymbol{X})$ over d variables $\boldsymbol{X} = X_1, \ldots, X_d$, specified by a directed acyclic graph (DAG) G and mechanism parameters θ. The graph encodes conditional independencies, while θ_i parameterizes the conditional distribution of X_i given its parents. The joint distribution factorizes as

$$p(\boldsymbol{X} \mid G, \theta) = \prod_{i=1}^{d} p\left(X_i \mid \mathrm{pa}_G(X_i), \theta_i\right),$$

where $\mathrm{pa}_G(X_i)$ denotes the parents of X_i in G.

In this paper, the generic mechanism parameters θ are instantiated by a linear-Gaussian structural equation model. Specifically, the observed variables satisfy $\boldsymbol{X} = \boldsymbol{X}B + \boldsymbol{\epsilon}$, where $B \in \mathbb{R}^{d\times d}$ is the weighted adjacency matrix and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{b}, \Sigma)$, with $\boldsymbol{b} \in \mathbb{R}^d$ and diagonal $\Sigma \in \mathbb{R}^{d\times d}_{\geq 0}$. For a given DAG G, we impose $B_{ij} = 0$ whenever i is not a parent of j in G.

Causal Bayesian networks [Spirtes et al., 2000, Pearl, 2009] add a causal interpretation to the directed edges in G: they describe how the joint distribution changes under interventions. In the linear-Gaussian SEM, the signed total causal effect from X_i to X_j is the derivative of the post-intervention mean of X_j with respect to an intervention on X_i. It has the closed form

$$\mathrm{CE}_{ij}(G, B) \triangleq \mathrm{CE}(i \to j \mid G, B) = \left[(I - B)^{-1}\right]_{ij} \quad (1)$$

following the standard total-effect formula for linear structural equation models [Sobel, 1990]. Since G is acyclic, B is nilpotent after a topological ordering, so $(I - B)^{-1} = I + B + B^2 + \cdots + B^{d-1}$. Thus, the (i, j) entry of $(I - B)^{-1}$ aggregates the products of edge weights along all directed paths from X_i to X_j. This path-sum interpretation is useful below because our conditional posterior summaries focus on which edges and directed pathways explain unusually large or small total effects.

Throughout the rest of the paper, we write a graph–weight state as $Z = (G, B)$, and abbreviate the signed causal effect as $\mathrm{CE}_{ij}(Z) = \mathrm{CE}_{ij}(G, B)$. For a target ordered pair (i, j) and threshold $t > 0$, we use the phrase "extreme causal effect" to mean that the total causal effect falls in a user-specified posterior tail region. In particular, we define the right- and left-tail events

$$\begin{aligned} \mathcal{E}^{+}_{ij}(t) &= \{Z : \mathrm{CE}_{ij}(Z) \geq t\}, \\ \mathcal{E}^{-}_{ij}(t) &= \{Z : \mathrm{CE}_{ij}(Z) \leq -t\}. \end{aligned} \quad (2)$$

When the sign is clear from context, we write $\mathcal{E}_{ij}(t)$ for either tail event. These events play two roles in our inference problem. First, we estimate their posterior probability, such as $\mathbb{P}(\mathcal{E}^{\pm}_{ij}(t) \mid \mathcal{D})$. Second, we use them to define the constrained posterior $p\left(G, B \mid \mathcal{D}, \mathcal{E}^{\pm}_{ij}(t)\right)$, from which we draw graph–weight samples for pathway-level summaries.

Bayesian Causal Discovery *Causal discovery* [Koller and Friedman, 2009, Glymour et al., 2019] is the problem of inferring the DAG G responsible for generating an observed dataset $\mathcal{D}$. We make the common assumption of causal sufficiency, meaning that there are no latent confounders. Even under this assumption, a single DAG may not be reliably identifiable from finite observational data due to sampling uncertainty and Markov equivalence. Bayesian causal discovery therefore represents uncertainty through a posterior distribution over graphs and parameters rather than committing to a single structure.

We place a user-specified prior $p(G)$ over DAGs and use the BGe marginal likelihood $p(\mathcal{D} \mid G)$ for linear-Gaussian models [Geiger and Heckerman, 1994, 2002]. The method itself does not require a particular graph prior; in the experiments, we use a sparse Erdős–R'enyi DAG prior, with the exact sparsity settings reported in Appendix A.1. Given G, the posterior over edge weights factorizes by node:

$$p(B \mid G, \mathcal{D}) = \prod_{j=1}^{d} p\left(B_{\mathrm{pa}_G(X_j), j} \mid G, \mathcal{D}\right),$$

where coefficients outside the parent set of X_j are fixed to zero and each nonzero incoming-coefficient block follows a multivariate t-distribution [Viinikka et al., 2020]. This node-wise factorization is used later by the MCMC mutation kernel: when a graph proposal changes a node's parent set, only the affected incoming-coefficient blocks need to be refreshed from their conditional posterior.

The joint posterior over graph–weight states is

$$p(G, B \mid \mathcal{D}) \propto p(G), p(\mathcal{D} \mid G), p(B \mid G, \mathcal{D}). \quad (3)$$

We write $\pi(Z) = \pi(G, B) = p(G, B \mid \mathcal{D})$ for this unconstrained posterior in the method section for compactness.

4 CONDITIONAL CAUSAL DISCOVERY

In this section, we formulate conditional causal discovery as posterior inference under user-specified causal-effect constraints, and then describe how we sample from the resulting conditional posterior. The main challenge is that the constraint may have low probability under the posterior. If ordinary posterior samples almost never satisfy the event, then rejection sampling gives both unstable probability estimates and too few constrained samples for pathway-level interpretation. Our solution is to convert each query into a scalar score $h(Z)$, where larger values indicate greater progress toward the desired event, and then use adaptive multilevel splitting to reach the event through a sequence of easier conditional problems.

4.1 CONDITIONAL CAUSAL DISCOVERY AS A SCORE-LEVEL PROBLEM

Let $Z = (G, B)$ denote a graph–weight state and let $\pi(Z) = p(G, B \mid \mathcal{D})$ be the unconstrained posterior from Eq. equation 3. Throughout this section, we use $\pi(\mathcal{A})$ to denote the posterior mass of an event $\mathcal{A}$. A conditional query is specified by a score function $h : \mathcal{Z} \to \mathbb{R}$ and a final level $\lambda_\star$. The score-level event, its posterior probability, and the corresponding constrained posterior are

$$\begin{aligned} \mathcal{A}_{\lambda_\star} &= \{Z : h(Z) \geq \lambda_\star\}, \\ p_\star &= \pi(\mathcal{A}_{\lambda_\star}), \\ \pi_\star(Z) &= \pi(Z \mid \mathcal{A}_{\lambda_\star}). \end{aligned} \tag{4}$$

This formulation separates the scientific query from the computational procedure: once the query has been written through h and $\lambda_\star$, the same sampler can be applied to single effects, left-tail effects, or multiple simultaneous constraints.

For a single ordered pair (i, j) and threshold $t > 0$, the right-tail event $\mathcal{E}^+_{ij}(t)$ is represented by choosing $h(Z) = \mathrm{CE}_{ij}(Z)$ and $\lambda_\star = t$. The left-tail event $\mathcal{E}^-_{ij}(t)$ is represented by choosing $h(Z) = -\mathrm{CE}_{ij}(Z)$ and $\lambda_\star = t$. Thus both tails are written as the same score-level event $\mathcal{A}_{\lambda_\star} = \{Z : h(Z) \geq \lambda_\star\}$, which avoids requiring separate algorithms for positive and negative extreme effects.

We also support conjunctions of multiple causal-effect constraints. Let $\mathcal{C} = \{(i_c, j_c, \bowtie_c, t_c, \kappa_c)\}_{c=1}^{C}$ be a collection of C constraints, where $\bowtie_c \in \{\geq, \leq\}$ gives the inequality direction, $t_c \in \mathbb{R}$ is the threshold, and $\kappa_c > 0$ is an optional scale factor. Define $\delta_c = +1$ when $\bowtie_c = \geq$ and $\delta_c = -1$ when $\bowtie_c = \leq$. The normalized margin of constraint c and the aggregate score are

$$\begin{aligned} m_c(Z) &= \frac{\delta_c\big(\mathrm{CE}_{i_c j_c}(Z) - t_c\big)}{\kappa_c}, \\ h_{\mathcal{C}}(Z) &= \min_{1 \leq c \leq C} m_c(Z). \end{aligned} \tag{5}$$

The margin $m_c(Z)$ is nonnegative exactly when the c-th constraint is satisfied. Therefore $h_{\mathcal{C}}(Z) \geq 0$ means that all constraints are satisfied, so the joint event is $\mathcal{E}_{\mathcal{C}} = \{Z : h_{\mathcal{C}}(Z) \geq 0\}$. Interval constraints can be represented by including both a lower-bound and an upper-bound inequality.

4.2 ADAPTIVE MULTILEVEL SPLITTING FOR RARE POSTERIOR EVENTS

Extreme-effect events can have very small posterior probability. A direct posterior sampler with M samples will produce only about $Mp_\star$ constrained samples on average, which may be close to zero if $p_\star$ is small. Adaptive multilevel splitting addresses this by replacing one difficult rare-event problem with a sequence of easier conditional problems. Intuitively, particles are first asked to reach a moderately high score level, then a higher one, and so on until they reach the target level $\lambda_\star$. At each stage, particles that have made sufficient progress are retained and resampled, while MCMC mutation restores diversity within the current truncated posterior.

Let $\lambda_0 < \lambda_1 < \cdots < \lambda_K = \lambda_\star$ be increasing score levels, with $\lambda_0 = -\infty$ corresponding to the unconstrained posterior. Define $\mathcal{A}_k = \{Z : h(Z) \geq \lambda_k\}$. Because these events are nested, the rare-event probability decomposes as

$$p_\star = \pi(\mathcal{A}_K) = \prod_{k=0}^{K-1} \pi(\mathcal{A}_{k+1} \mid \mathcal{A}_k). \tag{6}$$

The advantage is that each factor in this product can be much larger than $p_\star$ itself, making it estimable with a moderate number of particles.

We use the adaptive version of multilevel splitting [Kahn and Harris, 1951, Guyader et al., 2011, Cérou and Guyader, 2007], which chooses intermediate levels from the particle population rather than requiring them to be fixed in advance. At level k, the particle population $\mathcal{P}_k = \{Z_{k,n}\}_{n=1}^{N}$ is first mutated so that it approximately follows the level-truncated posterior

$$\pi_{\lambda_k}(Z) \propto \pi(Z)\mathbf{1}\{h(Z) \geq \lambda_k\}. \tag{7}$$

We then evaluate the scores $h(Z_{k,n})$ and choose the next level as an empirical quantile:

$$\begin{aligned} \widetilde{\lambda}_{k+1} &= Q_{1-\rho}\big(\{h(Z_{k,n})\}_{n=1}^{N}\big), \\ \lambda_{k+1} &= \min\{\widetilde{\lambda}_{k+1}, \lambda_\star\}. \end{aligned} \tag{8}$$

Algorithm 1 Adaptive multilevel splitting for conditional causal discovery

Require: Posterior $\pi(Z)$, score h, target level $\lambda_\star$, initial particles $\mathcal{P}$, particle size N, survival fraction ρ, mutation steps m, maximum levels $K_{\max}$.
Ensure: Tail-probability estimate $\widehat{p}_\star$ and constrained particles $\mathcal{P}$.
Initialize $\mathcal{P} = \{Z_n\}_{n=1}^N$ using an approximate posterior sampler or warm-start routine targeting $\pi(Z)$.
Set $\lambda_0 \leftarrow -\infty$ and $\widehat{p}_\star \leftarrow 1$.
for $k = 0, \dots, K_{\max} - 1$ **do**
  $\mathcal{P} \leftarrow \mathrm{Mutate}_{\lambda_k}(\mathcal{P}; m)$.
  Compute scores $s_n \leftarrow h(Z_n)$ for all $Z_n \in \mathcal{P}$.
  Set $\lambda_{k+1} \leftarrow \min\{Q_{1-\rho}(s_1, \dots, s_N), \lambda_\star\}$.
  Let $\mathcal{S} \leftarrow \{Z_n \in \mathcal{P} : s_n \geq \lambda_{k+1}\}$.
  Estimate $\widehat{\beta}_{k+1} \leftarrow |\mathcal{S}|/N$ and update $\widehat{p}_\star \leftarrow \widehat{p}_\star \widehat{\beta}_{k+1}$.
  if $|\mathcal{S}| = 0$ **then**
    return $0, \emptyset$.
  end if
  Resample N particles from $\mathcal{S}$ with replacement to form $\mathcal{P}$.
  if $\lambda_{k+1} = \lambda_\star$ **then**
    $\mathcal{P} \leftarrow \mathrm{Mutate}_{\lambda_\star}(\mathcal{P}; m)$.
    return $\widehat{p}_\star, \mathcal{P}$.
  end if
end for
return $\widehat{p}_\star, \mathcal{P}$.

Here $\mathrm{Mutate}_\lambda(\mathcal{P}; m)$ *denotes* m *Metropolis–Hastings steps per particle targeting* $\pi_\lambda(Z) \propto \pi(Z)\mathbf{1}\{h(Z) \geq \lambda\}$.

Here $\rho \in (0, 1)$ is the survival fraction and $Q_{1-\rho}$ is the empirical $(1-\rho)$-quantile. The estimated conditional factor at this level is the fraction of particles that survive the new threshold,

$$\widehat{\beta}_{k+1} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}\{h(Z_{k,n}) \geq \lambda_{k+1}\}. \tag{9}$$

The survivors are resampled with replacement to form the starting population for the next level. The final probability estimate is $\widehat{p}_\star = \prod_{k=0}^{K-1} \widehat{\beta}_{k+1}$, and after the final mutation step the particles approximate samples from $\pi_\star(Z)$.

4.3 PARTICLE INITIALIZATION

The splitting procedure should be viewed as a rare-event wrapper around Bayesian causal discovery rather than as a replacement for it. In principle, the initial population can be obtained from any posterior sampler that approximately targets $\pi(Z)$, including a long MCMC chain, an informed DAG-space sampler, or another Bayesian structure-learning method. The role of multilevel splitting is then to take these ordinary posterior samples and concentrate computation on the score-level region $\mathcal{A}_{\lambda_\star}$.

In our implementation, we use a simple warm-start scheme. We first draw sparse DAGs $G_{0,n}$ from the structural prior $p(G)$, using an Erdős–Rényi DAG prior in the experiments with edge inclusion probability p_{edge} chosen to match a target expected number of edges per node. Given each initial graph, we draw weights from the conjugate posterior $B_{0,n} \sim p(B \mid G_{0,n}, \mathcal{D})$ and set $Z_{0,n} = (G_{0,n}, B_{0,n})$.

Because these raw initial particles are not assumed to be exact samples from $\pi(Z)$, we use the inner MCMC kernel at the initial level $\lambda_0 = -\infty$ as a posterior warm-up before adapting the first nontrivial threshold. This makes the first population used in the splitting product approximate the unconstrained posterior, while retaining a simple and scalable initialization procedure.

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

At each score level λ_k, mutation must preserve the level-truncated target $\pi_{\lambda_k}(Z)$ from Eq. equation 7. We use a blocked Metropolis–Hastings kernel over the joint state $Z = (G, B)$, alternating between graph updates and coefficient updates. This is important because the rare-event constraint depends on both structure and weights: changing only G can leave the sampler stuck at a fixed set of coefficients, while changing only B cannot explore alternative pathways.

With probability p_{struct}, we propose a structure move. The new graph G' is drawn from a proposal $q_G(\cdot \mid G)$, instantiated either as Structure-MCMC, which adds, deletes, or reverses a single edge while rejecting cyclic graphs, or as PARNI-DAG [Liang et al., 2023], which uses locally informed adaptive neighborhoods guided by posterior edge information. After proposing G', we refresh only the coefficient blocks whose parent sets changed. For each affected node v, we sample

$$B'_{\mathrm{pa}_{G'}(v),v} \sim p\big(B_{\mathrm{pa}_{G'}(v),v} \mid G', \mathcal{D}\big), \tag{10}$$

set coefficients for absent edges to zero, and copy all unchanged blocks from B.

With probability $1 - p_{\mathrm{struct}}$, we propose a weight move at fixed structure. We keep $G' = G$, select a node v with at least one parent, and resample its incoming coefficient block,

$$B'_{\mathrm{pa}_G(v),v} \sim p\big(B_{\mathrm{pa}_G(v),v} \mid G, \mathcal{D}\big), \tag{11}$$

leaving all other entries unchanged. These blocked refreshes are inexpensive under the conjugate linear-Gaussian model because each conditional coefficient posterior is a multivariate t distribution.

Let $q(Z' \mid Z)$ denote the complete proposal density, including the selected move type, the graph proposal when applicable, and the coefficient-refresh density. Since the current state already satisfies $h(Z) \geq \lambda_k$, the level-λ_k acceptance

probability is

$$a_{\lambda_k}(Z, Z') = \mathbf{1}\{h(Z') \geq \lambda_k\} \times \min\left\{1, \frac{\pi(Z')q(Z \mid Z')}{\pi(Z)q(Z' \mid Z)}\right\}. \quad (12)$$

Thus, proposals that violate the current score-level constraint are rejected immediately, while feasible proposals are accepted according to the usual Metropolis–Hastings ratio for the unconstrained posterior and the proposal probabilities. Repeating these moves after each resampling step helps remove duplicate particles and produces a more representative approximation to the constrained posterior. Zs

5 EXPERIMENTS

5.1 EXPERIMENT SETUP

We evaluate the proposed conditional causal discovery procedure on synthetic linear-Gaussian datasets generated from random Erdős–Rényi DAGs with $d \in \{4, 8, 16, 32\}$. For each dimension, we first sample a data-generating DAG $G^\star$ using an Erdős–Rényi DAG generator with target sparsity approximately $2d$ edges. The generator samples an acyclic ordering, draws only order-compatible directed edges, and returns an adjacency matrix $G^\star$. The exact edge-sampling rule, including the edge probability or edge budget used for each d and the small-graph handling for $d = 4$, is provided in Appendix A.1.

Given $G^\star$, we draw raw edge weights independently as $\widetilde{B}_{ij} \sim \mathcal{N}(0, 1)$ and mask them by the sampled adjacency, so that $B^\star = \widetilde{B} \odot G^\star$. Equivalently, $B^\star_{ij} = \widetilde{B}_{ij}$ if $G^\star_{ij} = 1$ and $B^\star_{ij} = 0$ otherwise. We then generate an observational dataset $\mathcal{D}$ of size $n_{\text{obs}} = 1000$ from the linear-Gaussian SEM specified by $(G^\star, B^\star)$.

We evaluate our multilevel splitting framework instantiated with either the Structure-MCMC kernel or the PARNI-DAG kernel [Liang et al., 2023]. We compare against four baselines: (i) exhaustive enumeration, which is a gold-standard baseline feasible for $d = 4$; (ii) DiBS [Lorch et al., 2021]; (iii) OrderSPN [Wang et al., 2022]; and (iv) long single-chain MCMC using the same structural kernels as the multilevel splitting framework. Detailed sampler hyperparameters, thresholds, and evaluation settings are provided in Appendix A.1.

The experiments are organized around three questions: small-graph accuracy, multi-effect conditioning behavior, and scalability beyond the enumerable setting. We describe each question in the corresponding subsection below.

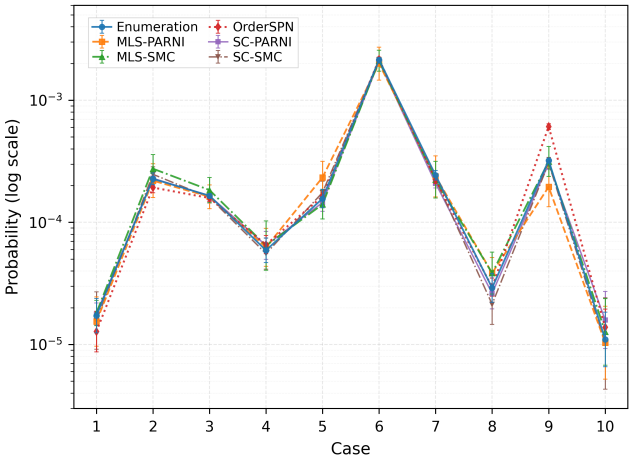


Figure 2: Single-effect one-sided tail-probability estimation on $d = 4$ across 10 test cases, with exhaustive enumeration as the reference posterior probability. Error bars show across-run variability over 10 independent runs. DiBS is omitted from the plot because its estimates are orders of magnitude larger on these tail events, which would obscure the comparison among the remaining methods; the corresponding numerical values are reported in Table 3.

5.2 CORRECTNESS OF CONDITIONAL CAUSAL DISCOVERY

The correctness experiments are designed to validate two complementary aspects of the method. The single-effect experiment checks the numerical accuracy of posterior tail-probability estimation against an exact enumerative reference. The multi-effect experiment checks the conditioning operator itself: when the conditioning event is constructed around ground-truth causal effects, the conditional posterior should assign more mass to graphs and weights close to the data-generating mechanism, and this concentration should increase as the constraint set becomes more informative.

5.2.1 Single-effect conditioning: validation on $d = 4$

We construct 10 test cases by sampling 10 independent $d = 4$ graphs and selecting one ordered node pair from each graph. For each pair, we evaluate a one-sided causal-effect tail query, represented in the score-level form $\mathcal{A}_{\lambda_\star} = \{Z : h(Z) \geq \lambda_\star\}$ from Sec. 4. For every test case, we run each stochastic method 10 times and report the mean estimate together with across-run variability.

Figure 2 compares all methods against exhaustive enumeration. We make two observations. First, all methods except DiBS are broadly consistent with the enumeration baseline across the $d = 4$ cases, with only minor deviations; the corresponding numerical values are reported in Table 3. Second, DiBS overestimates tail probabilities by orders of

Table 1: $d = 4$ **(PARNI)** structural accuracy under multi-effect conditioning.

Constraints	Runs	Samples/run	Mean SHD ↓	$\Pr(\text{SHD} = 0)$ ↑
Weak	10	20	0.325 ± 0.164	0.790 ± 0.117
Strong	10	20	0.115 ± 0.053	0.885 ± 0.053

magnitude on these tail events. We therefore exclude DiBS from subsequent experiments and focus on methods that give reliable small-graph estimates.

5.2.2 Multi-effect conditioning: validating the conditioning operator on $d = 4$ and $d = 8$

The previous experiment validates tail-probability estimation for a single causal-effect event. We next test whether the method correctly conditions the joint posterior over graphs and weights. To make this test interpretable, we construct constraints from the known data-generating state and ask whether the resulting conditional posterior concentrates toward that state.

Multi-constraint conditioning event. We select C ordered node pairs $\{(i_c, j_c)\}_{c=1}^{C}$. Let $e_c^\star = \text{CE}_{i_c j_c}(Z^\star)$ denote the ground-truth causal effect for pair (i_c, j_c) under the data-generating state $Z^\star = (G^\star, B^\star)$. For each pair, we impose a two-sided interval constraint around $e_c^\star$:

$$\text{CE}_{i_c j_c}(Z) \in [e_c^\star - \varepsilon,\, e_c^\star + \varepsilon], \quad c = 1, \ldots, C. \tag{13}$$

Each interval is represented by two one-sided inequalities, and the conjunction of all inequalities defines the conditioning event. Following Sec. 4, we aggregate these inequalities using the minimum normalized margin score $h_{\mathcal{C}}(Z)$, so that the target event is $\mathcal{E}_{\mathcal{C}} = \{Z : h_{\mathcal{C}}(Z) \geq 0\}$.

We compare a **weak** constraint set, with fewer constrained pairs and hence a looser event, against a **strong** constraint set, with more constrained pairs and hence a tighter event. All posterior summaries are computed from the final-stage multilevel-splitting particles and therefore condition on the target event by construction. To visualize the conditional posterior, we aggregate samples across runs to compute edge-frequency heatmaps, corresponding to posterior marginal edge probabilities, and average edge-weight heatmaps, corresponding to posterior mean weights. We also quantify structural recovery using the Structural Hamming Distance (SHD) between sampled DAGs and the ground-truth graph $G^\star$.

Multi-effect conditioning results. As demonstrated by Figs. 6 and 7 in Appendix A.7, across both $d = 4$ and $d = 8$, stronger, more informative multi-effect constraints consistently concentrate the conditional posterior toward the ground-truth mechanism. Quantitatively, Tables 1 and 2 show lower SHD under strong constraints than under weak constraints, together with a higher fraction of exact structural matches. Qualitatively, the corresponding edge-frequency and edge-weight heatmaps show the same pattern: posterior mass becomes sharper and closer to the ground truth as the constraint set becomes stronger. Together, these SHD improvements and posterior-summary concentration patterns provide evidence that the framework correctly conditions on multiple causal-effect constraints in the joint graph–parameter space.

Table 2: $d = 8$ **(PARNI)** structural accuracy under multi-effect conditioning. We report mean $\pm$ standard deviation of the per-run mean SHD, averaged over sampled DAGs within each run, and the fraction of samples with $\text{SHD} = 0$.

Constraints	Runs	Samples/run	Mean SHD ↓	$\Pr(\text{SHD} = 0)$ ↑
Weak	10	50	5.342 ± 0.442	0.040 ± 0.027
Strong	10	50	2.290 ± 0.199	0.108 ± 0.066

5.3 SCALABILITY

The scalability experiments examine whether the rare-event estimator remains stable when exact enumeration is no longer available. We focus on one-sided single-effect tail-probability estimation on $d \in \{8, 16, 32\}$ using multilevel splitting with the Structure-MCMC and PARNI-DAG kernels. These experiments test three aspects of scalability: whether estimated tail probabilities decrease smoothly as the target level $\lambda_\star$ becomes more stringent; whether independent runs give reproducible tail curves; and whether informed structure proposals become more important as the graph dimension increases.

Figure 3 summarizes one-sided single-effect tail-probability estimation on $d \in \{8, 16, 32\}$. On $d = 8$, we additionally include unconditional baselines, OrderSPN and two single-chain samplers, for comparison. Both multilevel-splitting variants produce smooth, monotone-decaying tail curves as the target level $\lambda_\star$ becomes more stringent. In contrast, the unconditional methods quickly fail to generate samples satisfying the more extreme targets, leading to degenerate probability estimates in the rare-event region. This highlights the practical advantage of multilevel splitting for conditional causal discovery on larger graphs, where rare-event conditioning makes unconditional sampling increasingly inefficient.

The same curves also provide a threshold-sensitivity check. The final threshold $\lambda_\star$ defines the scientific query, so the conditional posterior should change as the threshold changes. Numerically, however, the estimated posterior mass should vary smoothly and reproducibly as the threshold becomes more extreme. Across independent runs, both structure kernels yield stable and monotone tail curves in the moderate dimensions, supporting this expected behav-

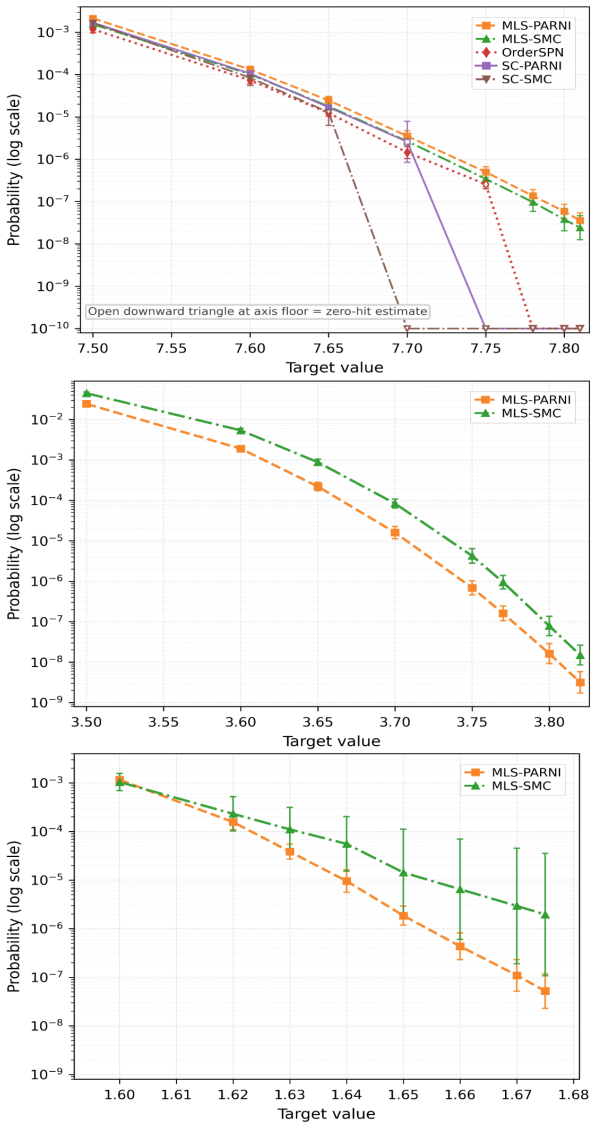


Figure 3: Single-effect one-sided tail-probability estimation on $d \in \{8, 16, 32\}$.

ior beyond the enumerable $d = 4$ setting.

Accordingly, for $d = 16$ and $d = 32$, we primarily assess robustness via internal consistency, because exact gold-standard posterior enumeration is unavailable at these scales. For $d = 16$, both multilevel-splitting instantiations produce stable and consistent probability estimates across runs, indicating that the framework remains well behaved as graph size grows. At $d = 32$, the two instantiations separate more clearly: PARNI-DAG retains relatively stable run-to-run behavior, whereas Structure-MCMC exhibits noticeably larger variability in the extreme tail. This suggests that reliable deep-tail estimation in larger graphs benefits from more informed structure proposals, as well as appropriate hyperparameter settings.

6 CASE STUDY

We now study an application of our conditional causal discovery framework to a real world protein-signalling dataset [Sachs et al., 2005] commonly used in causal discovery benchmarks. In particular, we will ask the following question: *Which causal structures and directed pathways most plausibly realize a specified extreme-effect event?*

Rather than claiming new biological insights, our goal is to show that conditional sampling can identify and summarize plausible mechanisms in an extreme-effect region that is rare under the unconditional posterior.

Data, reference network, and target pairs. We use the Sachs protein-signaling dataset, containing $n = 7466$ measurements of $d = 11$ proteins [Sachs et al., 2005]. As a qualitative reference, we use the 20-edge consensus network in Fig. 4 [Koch et al., 2009, Sachs et al., 2005]. We study two target pairs:

(i) PIP3 $\to$ PIP2 (indices $6 \to 5$), which admits a clean SEM decomposition into a direct route PIP3 $\to$ PIP2 and a mediated route PIP3 $\to$ Plcg $\to$ PIP2;

(ii) Erk $\to$ Akt (indices $1 \to 10$), a widely discussed cross-module influence in the Sachs signaling system. [1]

A published SEM mediation example on the same dataset reports a total effect of approximately 0.5665 for PIP3 $\to$ PIP2, with an explicit direct/indirect decomposition via Plcg, which we use as an external numerical anchor for the effect scale [Madhanagopal and Amrhein, 2019]. We treat such literature values as sanity checks on sign and magnitude rather than strict targets, since estimators and preprocessing differ.

Conditioning events. Using the notation from Sec. 3, let $Z = (G, B)$ denote a graph–weight state sampled from the posterior $\pi(Z) = p(G, B \mid \mathcal{D})$, and let $\mathrm{CE}_{ij}(Z)$ denote the signed linear-SEM total effect. For the two target pairs, define the events $\mathcal{E}_{\mathrm{PIP}}(t) = \{Z : \mathrm{CE}_{\mathrm{PIP3,PIP2}}(Z) \geq t\}$, and$\mathcal{E}_{\mathrm{ERK}}(t) = \{Z : \mathrm{CE}_{\mathrm{Erk,Akt}}(Z) \geq t\}$. We compare four posterior conditions:

- **Unconditioned:** $Z \sim \pi(Z) = p(G, B \mid \mathcal{D})$.
- **Cond-PIP:** $Z \sim p(Z \mid \mathcal{D}, \mathcal{E}_{\mathrm{PIP}}(t_{\mathrm{PIP}}))$, with $t_{\mathrm{PIP}} = 0.77$.
- **Cond-ERK:** $Z \sim p(Z \mid \mathcal{D}, \mathcal{E}_{\mathrm{ERK}}(t_{\mathrm{ERK}}))$, with $t_{\mathrm{ERK}} = 0.67$.
- **Cond-Joint:** $Z \sim p(Z \mid \mathcal{D}, \mathcal{E}_{\mathrm{PIP}}(t'_{\mathrm{PIP}}) \cap \mathcal{E}_{\mathrm{ERK}}(t'_{\mathrm{ERK}}))$, with

[1] Node order used throughout: [`Raf`, `Erk`, `Plcg`, `PKC`, `PKA`, `PIP2`, `PIP3`, `Mek`, `P38`, `Jnk`, `Akt`]. In particular, PIP3 is index 6, PIP2 is index 5, Plcg is index 2, Erk is index 1, and Akt is index 10.

$(t'_{\mathrm{PIP}}, t'_{\mathrm{ERK}}) = (0.74, 0.65)$, targeting a similarly rare posterior region.

The estimated posterior masses of these events are reported in Appendix A.6 and Table 6. These conditioning events do *not* assert that the dataset corresponds to a single "ground-truth" extreme state. Instead, they define posterior queries of the form $p(Z \mid \mathcal{D}, \mathcal{E})$, asking which graph–weight states and pathways remain plausible under the observed data when a specified causal effect, or a pair of causal effects, is unusually large.

Why unconditioned summaries can disagree with the consensus graph. We run experiments on each condition with settings listed in Sec. A.1. Table 7 shows substantial unconditioned mass on zero total effect, corresponding to sampled DAGs with no directed path from the source to the target, even though the consensus network suggests nonzero coupling. This is expected: the consensus graph is a qualitative reference rather than a uniquely identified ground truth, and the pooled perturbation data do not uniquely determine reachability under our model, so many near-equivalent posterior graphs omit these paths [Sachs et al., 2005, Koch et al., 2009, Friedman and Koller, 2003].

Mechanistic analysis under conditioning.

- **Cond-PIP:** The conditional posterior yields a highly concentrated explanation for $\mathrm{PIP3} \rightarrow \mathrm{PIP2}$: the direct route and the Plcg-mediated route appear in essentially all conditional samples, and the conditional mean effect is approximately 0.776 (calculated in Sec. A.6), with an increased Plcg-mediated share (Table 8).
- **Cond-ERK:** $\mathrm{Erk} \rightarrow \mathrm{Akt}$ remains direct-dominated, while $\mathrm{PIP3} \rightarrow \mathrm{PIP2}$ becomes almost always reachable but is supported by many alternative, partly cancelling paths. This contrasts with the two-path concentration under Cond-PIP (Tables 7 and 8).
- **Non-symmetry:** Conditioning on $\mathrm{PIP3} \rightarrow \mathrm{PIP2}$ does not force $\mathrm{Erk} \rightarrow \mathrm{Akt}$ to become extreme; for example, $\Pr(\mathrm{CE}_{\mathrm{Erk},\mathrm{Akt}}(Z) > 0 \mid \mathcal{D}, \mathcal{E}_{\mathrm{PIP}}(t_{\mathrm{PIP}})) \approx 0.498$.
- **Cond-Joint:** Joint conditioning increases the frequency of the mediated $\mathrm{Erk} \rightarrow \mathrm{Plcg} \rightarrow \mathrm{Akt}$ route relative to Cond-ERK, highlighting Plcg as a shared mediator under co-extreme coupling, beyond what either single-condition run reveals alone (Table 8).

Graph-level shifts under conditioning. Figure 5 illustrates these structural shifts, summarizing the edge frequencies and mean edge weights under the unconditioned posterior and the three conditional posteriors.

Takeaway. Without conditioning, posterior samples often imply zero or near-zero total effect, or distribute the effect across many paths that partially cancel each other, so pathway summaries are hard to determine. Single-effect constraints make the dominant routes clearer: **Cond-PIP** explains $\mathrm{PIP3} \rightarrow \mathrm{PIP2}$ mainly through two routes, whereas **Cond-ERK** keeps $\mathrm{Erk} \rightarrow \mathrm{Akt}$ mostly direct but yields a broader set of $\mathrm{PIP3} \rightarrow \mathrm{PIP2}$ paths with partial cancellation (Table 8). Joint conditioning is not just the overlap of the two single-effect results: compared to **Cond-ERK**, it more often highlights the mediated route $\mathrm{Erk} \rightarrow \mathrm{Plcg} \rightarrow \mathrm{Akt}$, pointing to Plcg as a shared mediator that emerges under the joint extreme-effect query (Table 8).

7 CONCLUSIONS

We introduced *conditional causal discovery*, a framework for posterior inference over causal graph structures and edge weights under user-specified causal-effect constraints. The framework is designed to answer two questions simultaneously: how likely a specified extreme-effect event is under the posterior, and which graph–weight configurations remain plausible when that event occurs.

To make such inference practical in rare-event regions, we developed an adaptive multilevel splitting framework with an MCMC kernel over the joint graph–weight space. By combining structure proposals, such as PARNI-DAG or Structure-MCMC, with blocked weight moves, the method estimates one-sided signed causal-effect tail probabilities and produces representative samples from the corresponding constrained posterior.

Empirically, we validated the estimator against exhaustive enumeration in a four-node setting and showed that the approach remains effective as problem size grows, where unconditional baselines often degenerate in the rare-event region. In a case study with the Sachs dataset, conditional samples provided compact pathway-level explanations and highlighted coherent mechanisms under both single-effect and joint-effect queries.

Limitations and future work. Our current implementation assumes linear-Gaussian SEMs under causal sufficiency. Extending the framework to nonlinear mechanisms, latent confounding, and interventional data remains an important direction. At the algorithmic level, the multilevel splitting procedure only requires a scalar score function defining the event of interest. For nonlinear structural causal models, the main changes would be to replace the BGe and conjugate posterior components with suitable nonlinear posterior inference modules, and to replace the closed-form linear causal-effect evaluator with an appropriate effect-estimation procedure. More broadly, incorporating richer constraint families, such as path-specific effects or qualitative monotonicity constraints, could enable deeper mechanistic analysis of complex scientific systems.

References

Tara V Anand, Adele H Ribeiro, Jin Tian, and Elias Bareinboim. Causal effect identification in cluster dags. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

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.

Juraj Bodik, Milan Paluš, and Zbyněk Pawlas. Causality in extremes of time series. *Extremes*, 27(1):67–121, 2024.

Giorgos Borboudakis and Ioannis Tsamardinos. Incorporating causal prior knowledge as path-constraints in bayesian networks and maximal ancestral graphs. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 427–434, 2012.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33: 21865–21877, 2020.

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

Eunice Yuh-Jie Chen, Yujia Shen, Arthur Choi, and Adnan Darwiche. Learning bayesian networks with ancestral constraints. *Advances in Neural Information Processing Systems*, 29, 2016.

Yizuo Chen and Adnan Darwiche. Identifying causal effects under functional dependencies. *Advances in Neural Information Processing Systems*, 37:87013–87041, 2024.

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.

Chris Chinenye Emezue, Alexandre Drouin, Tristan Deleu, Stefan Bauer, and Yoshua Bengio. Benchmarking bayesian causal discovery methods for downstream treatment effect estimation. *arXiv preprint arXiv:2307.04988*, 2023.

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.

Nicola Gnecco, Nicolai Meinshausen, Jonas Peters, and Sebastian Engelke. Causal discovery in heavy-tailed models. *The Annals of Statistics*, 49(3):1755–1778, 2021.

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.

Alain Hauser and Peter Bühlmann. Jointly interventional and observational data: estimation of interventional markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(1):291–318, 2015.

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

Mitchell Koch, Bradley M Broom, and Devika Subramanian. Learning robust cell signalling models from high throughput proteomic data. *International journal of bioinformatics research and applications*, 5(3):241–253, 2009.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. 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.

Xitong Liang, Alberto Caron, Samuel Livingstone, and Jim Griffin. Structure learning with adaptive random neighborhood informed mcmc. *Advances in Neural Information Processing Systems*, 36:40760–40772, 2023.

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

Banoo Madhanagopal and John Amrhein. Analyzing structural causal models using the calis procedure. In *SAS Global Forum*, pages 3240–2019, 2019.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 403–410, 1995.

Olivier C Pasche, Valérie Chavez-Demoulin, and Anthony C Davison. Causal modelling of heavy-tailed variables and confounders with application to river flow. *Extremes*, 26(3):573–594, 2023.

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.

Michael E Sobel. Effect analysis and causation in linear structural equation models. *Psychometrika*, 55(3):495–515, 1990.

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 EXPERIMENT SETTING

For all synthetic experiments, we generate the ground-truth DAG $G^\star$ using an ordered Erdős–Rényi construction. Given dimension d and edge budget parameter $\texttt{edges\_per\_node} = 2$, we set

$$p_{\text{edge}} = \min\left\{\frac{\texttt{edges\_per\_node} \cdot d}{d(d-1)/2},\, 0.5\right\}.$$

We then sample a Bernoulli adjacency matrix with edge probability p_{edge}, keep only the strict lower-triangular part to ensure acyclicity under the current ordering, and finally apply a random permutation of the node labels. The effective edge probabilities are therefore $p_{\text{edge}} = 0.5$ for $d = 4$, $p_{\text{edge}} = 0.5$ for $d = 8$, $p_{\text{edge}} = 4/15$ for $d = 16$, and $p_{\text{edge}} = 4/31$ for $d = 32$. The cap at 0.5 is the small-graph handling used by the generator: for $d = 4$, the nominal value $4/3$ is clipped to 0.5, giving an expected 3 edges instead of an overly dense four-node DAG. The corresponding expected edge counts are 3, 14, 32, and 64 for $d = 4, 8, 16, 32$, respectively.

Conditional on $G^\star$, absent edges are assigned weight zero and nonzero edge weights are sampled independently from a Gaussian distribution, as in the data generation script. Observations are then generated from the linear-Gaussian SEM using $n_{\text{obs}} = 1000$ training samples.

Below are the experiment settings for each method:

1. *Exhaustive enumeration ($d = 4$ only):* enumerate all DAGs ($|\mathcal{G}_{d=4}| = 543$) and draw 20000 weight samples per graph.
2. *OrderSPN [Wang et al., 2022]:* sample 5000 graphs from the learned circuit and draw 100 weight samples per graph.
3. *DiBS [Lorch et al., 2021]:* draw 500000 joint graph–weight samples in total.
4. *PARNI-DAG single chain [Liang et al., 2023]:* run a single long MCMC chain over (G, B) using PARNI-DAG as the structure proposal kernel for 500000 iterations with 10% burn-in.
5. *Structure-MCMC single chain:* run a single long MCMC chain over (G, B) using Structure-MCMC as the structure proposal kernel for 500000 iterations with 10% burn-in.
6. *MLS-PARNI-DAG [Liang et al., 2023]:* use our MLS framework with PARNI-DAG as the structure proposal kernel. We use $N = 200$ particles, $m = 2000$ MCMC iterations per level, and $K_{\max} = 10$ levels for all dimensions.
7. *MLS-Structure-MCMC:* use our MLS framework with Structure-MCMC as the structure proposal kernel, with the same MLS hyperparameters as above.

A.2 TARGET-THRESHOLD CONSTRUCTION

We construct target score levels to evaluate each method's ability to estimate *deep-tail* probabilities under increasingly extreme one-sided causal-effect constraints. Let

$$e_0 = \mathrm{CE}_{ij}(Z^\star) = \mathrm{CE}_{ij}(G^\star, B^\star)$$

denote the ground-truth signed total causal effect for a queried ordered pair (i, j), where $Z^\star = (G^\star, B^\star)$ is the data-generating graph–weight state.

For synthetic experiments, we choose the tail direction using the sign of e_0. Equivalently, define $s = \mathrm{sign}(e_0)$ and use the scalar score

$$h(Z) = s\,\mathrm{CE}_{ij}(Z).$$

Then larger values of $h(Z)$ always correspond to more extreme effects in the selected direction. A target score level λ defines the event

$$\mathcal{A}_\lambda = \{Z : h(Z) \geq \lambda\}.$$

When $e_0 > 0$, this is the right-tail event $\mathcal{E}_{ij}^{+}(\lambda) = \{Z : \mathrm{CE}_{ij}(Z) \geq \lambda\}$. When $e_0 < 0$, it is the left-tail event $\mathcal{E}_{ij}^{-}(\lambda) = \{Z : \mathrm{CE}_{ij}(Z) \leq -\lambda\}$. This is the same score-level representation used by the MLS sampler in Sec. 4.

1. **Initial score level.** We initialize the threshold grid at the magnitude of the ground-truth effect, $\lambda_1 = |e_0|$.
2. **Pilot run.** We run a short, low-budget adaptive multilevel-splitting pilot to identify a more extreme target score level λ_T such that the corresponding tail probability is already in the rare-event regime, approximately 10^{-6} or smaller.
3. **Threshold grid.** We form a monotone increasing sequence of T target score levels $\lambda_1 < \lambda_2 < \cdots < \lambda_T$. Larger λ always corresponds to a more extreme one-sided effect in the selected direction.

A.3 DETAILED RESULT FOR D=4 BASELINE EXPERIMENT

Table 3 lists statistics for the $d = 4$ baseline experiments across seven methods.

A.4 TIME COMPLEXITY

We separate the cost of the outer adaptive multilevel-splitting loop from the cost of the inner MCMC mutation kernel. Let N be the number of particles, m the number of

Table 3: $d = 4$ per-case probability estimates across runs (mean $\pm$ SD). MLS refers to multilevel splitting, and SC refers to single-chain sampling.

Case	**Enumeration** $\hat{p} \pm \mathrm{SD}$	**PARNI-MLS** $\hat{p} \pm \mathrm{SD}$	**Structure-MCMC-MLS** $\hat{p} \pm \mathrm{SD}$	**SPN** $\hat{p} \pm \mathrm{SD}$	**PARNI-SC** $\hat{p} \pm \mathrm{SD}$	**Structure-MCMC-SC** $\hat{p} \pm \mathrm{SD}$	**DIBS** $\hat{p} \pm \mathrm{SD}$
1	1.87×10^{-5} $\pm 7.25 \times 10^{-6}$	1.83×10^{-5} $\pm 1.07 \times 10^{-5}$	1.9×10^{-5} $\pm 6.09 \times 10^{-6}$	1.42×10^{-5} $\pm 6 \times 10^{-6}$	1.84×10^{-5} $\pm 5.32 \times 10^{-6}$	1.4×10^{-5} $\pm 1.01 \times 10^{-5}$	2.32×10^{-2} $\pm 1.35 \times 10^{-3}$
2	2.3×10^{-4} $\pm 1.13 \times 10^{-4}$	2.5×10^{-4} $\pm 1.07 \times 10^{-4}$	2.41×10^{-4} $\pm 8.04 \times 10^{-5}$	1.37×10^{-2} $\pm 8.62 \times 10^{-4}$	2.28×10^{-4} $\pm 7.49 \times 10^{-5}$	2.53×10^{-4} $\pm 5.78 \times 10^{-5}$	1.35×10^{-2} $\pm 7.72 \times 10^{-4}$
3	1.65×10^{-4} $\pm 5.96 \times 10^{-5}$	1.7×10^{-4} $\pm 5.47 \times 10^{-5}$	1.61×10^{-4} $\pm 1.41 \times 10^{-5}$	1.83×10^{-1} $\pm 7.92 \times 10^{-3}$	1.67×10^{-4} $\pm 2.42 \times 10^{-5}$	1.79×10^{-4} $\pm 2.31 \times 10^{-5}$	1.81×10^{-1} $\pm 8.38 \times 10^{-3}$
4	6.17×10^{-5} $\pm 4.07 \times 10^{-5}$	5.89×10^{-5} $\pm 3.67 \times 10^{-5}$	6.79×10^{-5} $\pm 2.13 \times 10^{-5}$	2.67×10^{-2} $\pm 1.76 \times 10^{-3}$	7.06×10^{-5} $\pm 2.85 \times 10^{-5}$	7.21×10^{-5} $\pm 3.06 \times 10^{-5}$	2.65×10^{-2} $\pm 1.66 \times 10^{-3}$
5	1.6×10^{-4} $\pm 5.56 \times 10^{-5}$	1.71×10^{-4} $\pm 5.9 \times 10^{-5}$	1.62×10^{-4} $\pm 1.74 \times 10^{-5}$	3.34×10^{-2} $\pm 1.88 \times 10^{-3}$	1.55×10^{-4} $\pm 4.54 \times 10^{-5}$	1.69×10^{-4} $\pm 4.88 \times 10^{-5}$	3.34×10^{-2} $\pm 1.63 \times 10^{-3}$
6	2.15×10^{-3} $\pm 7.08 \times 10^{-4}$	2.21×10^{-3} $\pm 9.14 \times 10^{-4}$	2.16×10^{-3} $\pm 7.4 \times 10^{-4}$	2.27×10^{-2} $\pm 1.61 \times 10^{-3}$	2.09×10^{-3} $\pm 1.38 \times 10^{-4}$	2.15×10^{-3} $\pm 1.78 \times 10^{-4}$	2.3×10^{-2} $\pm 1.55 \times 10^{-3}$
7	2.44×10^{-4} $\pm 1.13 \times 10^{-4}$	2.4×10^{-4} $\pm 1.06 \times 10^{-4}$	2.19×10^{-4} $\pm 2.48 \times 10^{-5}$	9.71×10^{-2} $\pm 6.54 \times 10^{-3}$	2.22×10^{-4} $\pm 2.25 \times 10^{-5}$	2.42×10^{-4} $\pm 2.5 \times 10^{-5}$	9.65×10^{-2} $\pm 6.06 \times 10^{-3}$
8	3.09×10^{-5} $\pm 2.84 \times 10^{-5}$	4.48×10^{-5} $\pm 2.71 \times 10^{-5}$	3.1×10^{-5} $\pm 6.72 \times 10^{-6}$	5.76×10^{-3} $\pm 6.65 \times 10^{-4}$	2.82×10^{-5} $\pm 8.1 \times 10^{-6}$	3.4×10^{-5} $\pm 1.35 \times 10^{-5}$	5.86×10^{-3} $\pm 5.4 \times 10^{-4}$
9	3.22×10^{-4} $\pm 1.34 \times 10^{-4}$	3.25×10^{-4} $\pm 1.33 \times 10^{-4}$	3.22×10^{-4} $\pm 5.19 \times 10^{-5}$	1.32×10^{-2} $\pm 1.46 \times 10^{-3}$	3.09×10^{-4} $\pm 6.17 \times 10^{-5}$	3.3×10^{-4} $\pm 4.04 \times 10^{-5}$	1.33×10^{-2} $\pm 1.45 \times 10^{-3}$
10	1.33×10^{-5} $\pm 1.02 \times 10^{-5}$	1.98×10^{-5} $\pm 1.78 \times 10^{-5}$	1.37×10^{-5} $\pm 1.11 \times 10^{-5}$	1.82×10^{-2} $\pm 1.55 \times 10^{-3}$	1.68×10^{-5} $\pm 1.52 \times 10^{-5}$	1.21×10^{-5} $\pm 1.58 \times 10^{-5}$	1.8×10^{-2} $\pm 1.72 \times 10^{-3}$

MCMC mutation steps per particle per level, K the number of splitting levels, and C_{MH} the average cost of one Metropolis– Hastings proposal, including proposal generation, posterior-ratio evaluation, and causal-effect score evaluation. At each level, the algorithm mutates N particles for m steps, computes N scores, sorts or partially sorts the scores to choose the next empirical quantile, and resamples the survivors. The overall cost is therefore

$$O(KNmC_{\mathrm{MH}} + KN \log N)\,,$$

where the $KN \log N$ term comes from quantile selection. If selection is implemented by a linear-time order-statistic routine, this sorting term can be reduced to $O(KN)$ and the mutation cost dominates.

The average proposal cost depends on the mixture of structure and weight moves. If p_{struct} is the probability of proposing a structure move, then

$$C_{\mathrm{MH}} \approx p_{\mathrm{struct}} C_{\mathrm{struct}} + (1 - p_{\mathrm{struct}}) C_{\mathrm{weight}} + C_h,$$

where C_{struct} is the cost of proposing and scoring a graph update, C_{weight} is the cost of a blocked coefficient refresh, and C_h is the cost of evaluating the scalar score $h(Z)$. In the linear-Gaussian implementation, the coefficient refreshes use node-wise conjugate posterior updates, while $h(Z)$ is computed from the linear total-effect matrix. A direct matrix inverse gives $C_h = O(d^3)$, although for a DAG and a single target pair one can exploit the topological structure or solve a triangular system to reduce this cost in practice.

The key rare-event advantage is the dependence on the event probability $p_\star$. A direct posterior sampler needs about $1/p_\star$ samples to see one sample from an event of posterior probability $p_\star$, and $O(1/(r^2 p_\star))$ samples to estimate that probability with fixed relative error r. By contrast, adaptive multilevel splitting replaces the single rare event with a product of moderate conditional survival probabilities. If the empirical survival fraction is approximately ρ at each level, then K grows roughly like $\log(p_\star)/\log(\rho)$, so the leading cost grows approximately logarithmically in $1/p_\star$ rather than linearly in $1/p_\star$. This is why multilevel splitting remains useful in the deep-tail regimes where unconditional sampling degenerates.

Runtime measurements. Table 4 reports wall-clock running times for the $d = 4$ case 1 experiment. Table 5 reports the average time per outer MLS loop for the two structure kernels across dimensions. PARNI-DAG is more expensive per loop because its locally informed proposal requires additional neighborhood construction and scoring, while Structure-MCMC has cheaper local edge proposals. The benefit of PARNI-DAG is not lower per-iteration cost, but more stable deep-tail behavior in larger graphs, as shown in the scalability experiments.

A.5 EXTENSION TO NONLINEAR GAUSSIAN MECHANISMS

The outer MLS framework is not specific to linear-Gaussian SEMs. It only requires three ingredients: an unconstrained posterior target $\pi(Z)$, a scalar score $h(Z)$ whose super-level set defines the event of interest, and an MCMC kernel that approximately preserves the level-truncated posterior $\pi(Z)\mathbf{1}\{h(Z) \geq \lambda\}$. The linear-Gaussian assumptions used in the main experiments provide convenient closed forms for the BGe marginal likelihood, the node-wise coefficient posterior, and the total

Table 4: Average running time for the $d = 4$ case 1 experiment over 10 runs. We report mean $\pm$ standard deviation in seconds.

Method	Time (s)
MLS-PARNI-DAG	192.3 ± 4.8
MLS-Structure-MCMC	56.1 ± 2.1
PARNI-DAG	268.2 ± 5.3
Structure-MCMC	105.5 ± 3.0
Enumeration-Baseline	301.6 ± 4.6
OrderSPN	166.8 ± 3.9
DiBS	434.2 ± 6.7

Table 5: Average running time per outer MLS loop in seconds. Standard deviations over 10 runs are shown in parentheses.

Kernel	$d = 4$	$d = 8$	$d = 16$	$d = 32$
PARNI-DAG	38.4 (1.1)	56.8 (1.6)	92.4 (2.8)	172.5 (4.9)
Structure-MCMC	11.2 (0.5)	19.5 (0.8)	34.8 (1.2)	58.5 (2.1)

causal effect, but they are not required by the splitting principle itself.

For a nonlinear Gaussian structural causal model, one could write

$$X_j = f_j\big(X_{\mathrm{pa}_G(j)}; \phi_j\big) + \epsilon_j, \qquad \epsilon_j \sim \mathcal{N}(0, \sigma_j^2),$$

and replace the graph–weight state $Z = (G, B)$ by a graph–mechanism state $Z = (G, \phi, \sigma)$. The posterior target would become $p(G, \phi, \sigma \mid \mathcal{D})$, obtained using an appropriate nonlinear mechanism class such as splines, Gaussian processes, or neural networks. The inner mutation kernel would then update graph structure and mechanism parameters instead of graph structure and linear coefficients.

The causal-effect score would also be replaced by a nonlinear effect evaluator. For example, for a scalar intervention one may define an average interventional contrast,

$$h(Z) = \mathbb{E}_Z[X_j \mid \mathrm{do}(X_i = x + \Delta)] - \mathbb{E}_Z[X_j \mid \mathrm{do}(X_i = x)]\,,$$

or an averaged derivative when such derivatives are well-defined. These expectations can be estimated by ancestral simulation under the proposed nonlinear SCM. The outer splitting loop would remain unchanged; the additional computational burden would come from evaluating nonlinear posterior ratios and interventional scores. Therefore, the main technical requirement for such an extension is not a new rare-event algorithm, but reliable posterior inference and causal-effect evaluation for the chosen nonlinear mechanism class.

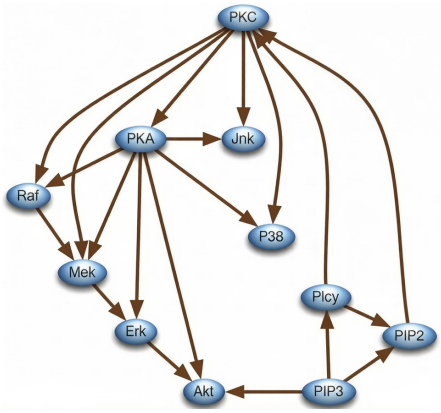


Figure 4: Consensus signaling network (Sachs) used as a qualitative reference.

A.6 SACHS EXPERIMENT CALCULATIONS AND TABLES

This section provides supplementary numerical details for the Sachs case study in Sec. 6, including (i) the qualitative reference network, (ii) estimated posterior masses of the tail events, and (iii) pathway-level decompositions of conditional effects.

Reference network. Figure 4 shows the 11-node, 20-edge consensus signaling network used as a qualitative reference in Sec. 6.

Posterior mass of tail events. We run MLS with 10 independent repetitions for each conditioning regime. Table 6 reports (i) the *geometric mean* estimate of $\Pr(\mathcal{E} \mid \mathcal{D})$ (i.e., $\exp(\mathbb{E}[\log \hat{p}])$ across runs) and (ii) the mean $\pm$ sd of $-\log \hat{p}$ across runs.

Table 6: Estimated posterior mass of Sachs tail events over 10 MLS runs.

Condition	$\Pr(\mathcal{E} \mid \mathcal{D})$ (geo-mean)	$-\log \hat{p}$ (mean $\pm$ sd)
Cond-PIP	2.2×10^{-7}	15.32 ± 0.44
Cond-ERK	1.3×10^{-7}	15.84 ± 0.53
Cond-Joint	5.2×10^{-7}	14.47 ± 0.67

Path decomposition under Cond-PIP. Across samples drawn from the conditional posterior

$$p(Z \mid \mathcal{D}, \mathcal{E}_{\mathrm{PIP}}(t_{\mathrm{PIP}}))\,,$$

the two canonical pathways PIP3 $\to$ PIP2 and PIP3 $\to$ Plcg $\to$ PIP2 are present in *100%* of samples. These two routes account for essentially the entire conditional mean

effect:

$$\mathbb{E}[\mathrm{CE}_{\mathrm{PIP3,PIP2}}(Z) \mid \mathcal{D}, \mathcal{E}_{\mathrm{PIP}}(t_{\mathrm{PIP}})] \approx \underbrace{0.6142}_{\text{direct}} + \underbrace{0.1614}_{\text{via Plcg}} + \underbrace{2 \cdot 10^{-4}}_{\text{other paths}} \approx 0.7758. \quad (14)$$

The contribution from all remaining directed paths is numerically negligible, confirming that conditioning isolates a highly concentrated two-path mechanism.

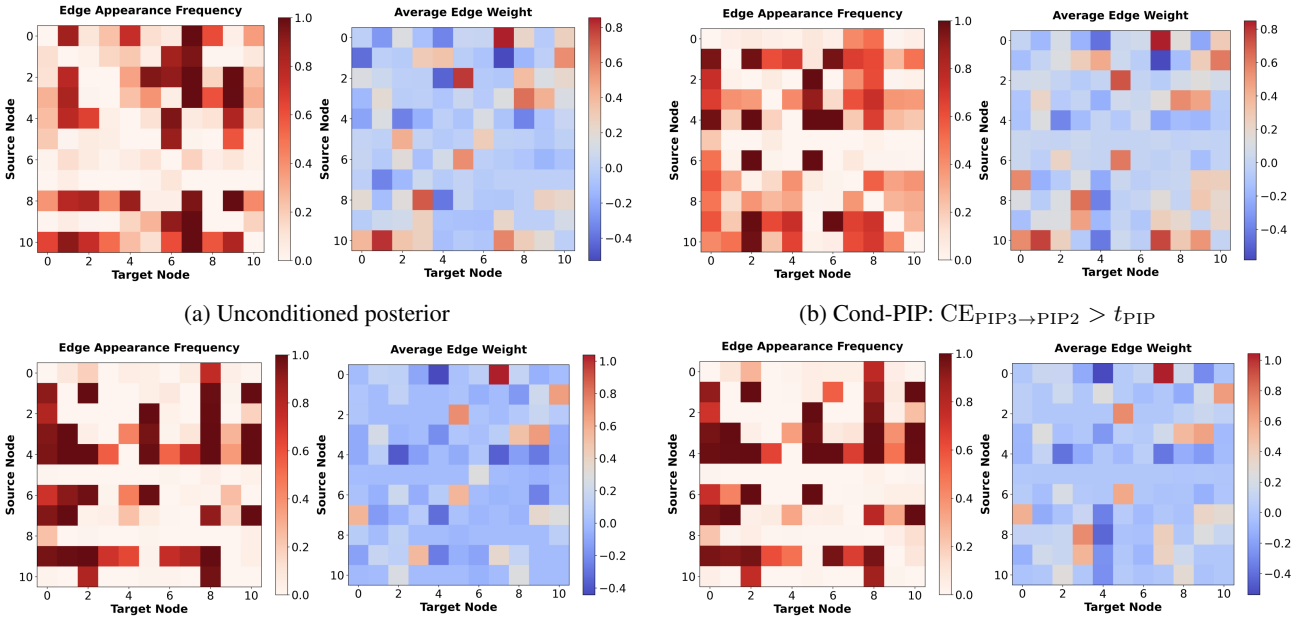
Summary tables. Table 7 reports posterior effect statistics (e.g., $\Pr(\mathrm{CE} > 0)$, conditional means, and path counts) across the four regimes. Table 8 further decomposes the conditional mean effects into direct and mediated contributions for the two target pairs (PIP3→PIP2 and Erk→Akt), together with the frequency of the canonical mediated paths.

A.7 SYNTHETIC MULTI-EFFECT HEAT MAPS

Figures 6 and 7 provide the detailed heat-map visualizations for the multi-effect conditioning experiments discussed in Sec. 5.2.2. Each row displays the aggregated edge frequency and average edge weight summaries for one posterior condition. These figures are included in the appendix because they are visually dense; the main text reports the more compact SHD summaries.

A.8 PSEUDO-CODE

Algorithm 2 summarizes the adaptive multilevel-splitting loop, and Algorithm 3 summarizes the level-truncated MCMC mutation kernel. The notation matches Sec. 4: h is the scalar score, $\lambda_\star$ is the target level, and $\mathcal{A}_\lambda = \{Z : h(Z) \geq \lambda\}$ is the corresponding score-level event.



(a) Unconditioned posterior

(b) Cond-PIP: $\mathrm{CE}_{\mathrm{PIP3}\to\mathrm{PIP2}} > t_{\mathrm{PIP}}$

(c) Cond-ERK: $\mathrm{CE}_{\mathrm{Erk}\to\mathrm{Akt}} > t_{\mathrm{ERK}}$

(d) Cond-Joint: both constraints

Figure 5: Heatmaps summarizing posterior graph features under four conditioning settings. Each panel contains two heatmaps: **edge frequency** (probability an edge appears in sampled graphs; left) and **mean edge weight** (conditional mean of B_{uv} given the edge is present; right).

Table 7: Comparison across four conditioning settings for two target pairs. $\Pr(\mathrm{CE} > 0)$ is the posterior mass with a directed influence (i.e., at least one directed path), $\mathbb{E}[\mathrm{CE} \mid \mathrm{CE} > 0]$ and quantiles are computed on the nonzero subset, and #paths is the median number of directed paths among samples with $\mathrm{CE} > 0$. We run PARNI-DAG single chain with $200,000$ iterations and collect 180000 samples for the unconditioned case. For each of the other conditions, we collect 500 survivor samples from 10 runs with 50 samples for each run.

	PIP3→PIP2			**Erk→Akt**		
Condition	$\Pr(\mathrm{CE} > 0)$	$\mathbb{E}[\mathrm{CE} \mid \mathrm{CE} > 0]$	$\#\mathrm{paths}_{\mathrm{med}}$	$\Pr(\mathrm{CE} > 0)$	$\mathbb{E}[\mathrm{CE} \mid \mathrm{CE} > 0]$	$\#\mathrm{paths}_{\mathrm{med}}$
Unconditioned	0.109	0.600 [0.543, 0.681]	4	0.085	0.569 [0.532, 0.617]	1
Cond-PIP ($\mathrm{CE}_{\mathrm{PIP}} > t_{\mathrm{PIP}}$)	1.000	0.776 [0.771, 0.782]	2	0.498	0.596 [0.538, 0.634]	1
Cond-ERK ($\mathrm{CE}_{\mathrm{ERK}} > t_{\mathrm{ERK}}$)	0.984	0.655 [0.562, 0.710]	10	1.000	0.673 [0.670, 0.676]	1
Cond-Joint ($\mathrm{CE}_{\mathrm{PIP}} > t'_{\mathrm{PIP}} \wedge \mathrm{CE}_{\mathrm{ERK}} > t'_{\mathrm{ERK}}$)	1.000	0.747 [0.742, 0.758]	3	1.000	0.656 [0.652, 0.662]	1

Table 8: Mechanism decomposition across conditions, using the two pathways most relevant for interpretation: for PIP3→PIP2 we use the direct path (PIP3→PIP2) and the Plcg-mediated path (PIP3→Plcg→PIP2); for Erk→Akt we use the direct path (Erk→Akt) and the Plcg-mediated path (Erk→Plcg→Akt). For unconditioned samples, we select those with $\mathrm{CE} > 0$, and calculate $\mathbb{E}[\mathrm{CE} \mid \mathrm{CE} > 0]$. The last column reports the frequency that the canonical mediated path edges are present (over all samples under the condition).

Pair	**Condition**	$\mathbb{E}[\mathrm{CE} \mid \mathrm{CE} > 0]$	Direct contrib	Via-Plcg contrib	$\Pr$(Via-Plcg path)
PIP3→PIP2	Unconditioned	0.600	0.560 (93.4%)	0.076 (12.7%)	0.064
	Cond-PIP	0.776	0.614 (79.2%)	0.161 (20.8%)	1.000
	Cond-ERK	0.655	0.559 (85.2%)	0.140 (21.4%)	0.975
	Cond-Joint	0.747	0.598 (79.9%)	0.154 (20.5%)	1.000
Erk→Akt	Unconditioned	0.569	0.563 (98.9%)	0.000 (0.0%)	0.0004
	Cond-PIP	0.596	0.606 (101.6%)	0.001 (0.2%)	0.026
	Cond-ERK	0.673	0.668 (99.3%)	0.004 (0.6%)	0.177
	Cond-Joint	0.656	0.650 (99.1%)	0.005 (0.8%)	0.244

Note: Percentages are (path contribution)/(mean effect). They may exceed 100% or be negative due to path cancellations.

Algorithm 2 Adaptive Multilevel Splitting for Conditional Causal Discovery

Require: Data $\mathcal{D}$; posterior density $\pi(Z) = p(G, B \mid \mathcal{D})$; score h; target level $\lambda_\star$; particle size N; survival fraction ρ; mutation steps m per level; maximum number of levels $K_{\max}$; structure kernel Kernel.

Ensure: Tail-probability estimate $\widehat{p}_\star$ and approximately constrained particles $\mathcal{P}$.

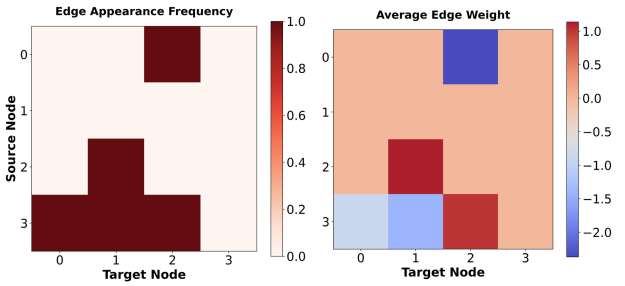
Initialize particles $\mathcal{P} = \{Z_n = (G_n, B_n)\}_{n=1}^N$ by drawing $G_n \sim p(G)$ and $B_n \sim p(B \mid G_n, \mathcal{D})$.
Set $\lambda_0 \leftarrow -\infty$, $\widehat{p}_\star \leftarrow 1$, and $k \leftarrow 0$.
$\mathcal{P} \leftarrow \textsc{MCMCMutation}(\mathcal{P}, \mathcal{D}, h, \lambda_0, m, \mathsf{Kernel})$.
while $k < K_{\max}$ **do**
  Compute scores $s_n \leftarrow h(Z_n)$ for all $Z_n \in \mathcal{P}$.
  $\widetilde{\lambda}_{k+1} \leftarrow Q_{1-\rho}(\{s_n\}_{n=1}^N)$.
  $\lambda_{k+1} \leftarrow \min\{\widetilde{\lambda}_{k+1}, \lambda_\star\}$.
  $\mathcal{S} \leftarrow \{Z_n \in \mathcal{P} : h(Z_n) \geq \lambda_{k+1}\}$.
  $\widehat{\beta}_{k+1} \leftarrow |\mathcal{S}|/N$.
  $\widehat{p}_\star \leftarrow \widehat{p}_\star \, \widehat{\beta}_{k+1}$.
  if $|\mathcal{S}| = 0$ **then**
    return $(0, \emptyset)$.
  end if
  Resample N particles from $\mathcal{S}$ with replacement to form $\mathcal{P}$.
  $\mathcal{P} \leftarrow \textsc{MCMCMutation}(\mathcal{P}, \mathcal{D}, h, \lambda_{k+1}, m, \mathsf{Kernel})$.
  if $\lambda_{k+1} = \lambda_\star$ **then**
    return $(\widehat{p}_\star, \mathcal{P})$.
  end if
  $k \leftarrow k + 1$.
end while
return $(\widehat{p}_\star, \mathcal{P})$.

Algorithm 3 Level-Truncated MCMC Mutation over Graph–Weight States

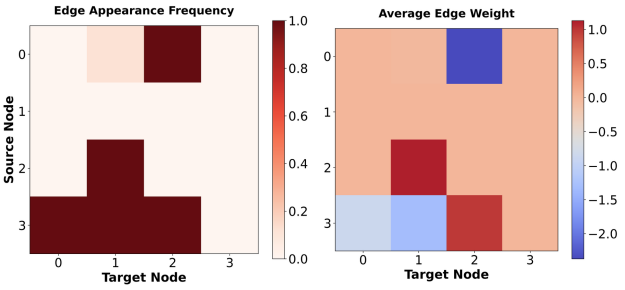
Require: Population $\mathcal{P}$; data $\mathcal{D}$; score h; level λ; mutation steps m; structure kernel Kernel; structure-move probability p_{struct}.

Ensure: Mutated population $\mathcal{P}'$.

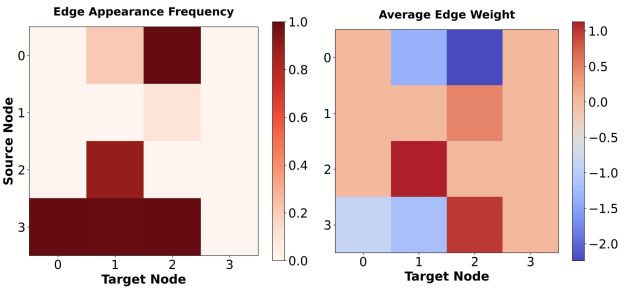
$\mathcal{P}' \leftarrow \emptyset$.
for all $Z = (G, B) \in \mathcal{P}$ **do**
  for $r = 1$ **to** m **do**
    Draw $u \sim \mathrm{Uniform}(0, 1)$.
    if $u < p_{\text{struct}}$ **then**
      Propose a graph $G' \sim q_G(\cdot \mid G)$ using Kernel.
      Refresh affected coefficient blocks from $p(B \mid G', \mathcal{D})$ and copy unchanged blocks to obtain B'.
    else
      Set $G' \leftarrow G$ and refresh one coefficient block from $p(B \mid G, \mathcal{D})$ to obtain B'.
    end if
    Set $Z' \leftarrow (G', B')$ and let $q(Z' \mid Z)$ denote the complete proposal density.
    if $h(Z') < \lambda$ **then**
      Reject Z' and continue.
    else
      $\alpha \leftarrow \min\left\{1, \dfrac{\pi(Z')q(Z \mid Z')}{\pi(Z)q(Z' \mid Z)}\right\}$.
      Draw $a \sim \mathrm{Uniform}(0, 1)$.
      if $a \leq \alpha$ **then**
        $Z \leftarrow Z'$.
      end if
    end if
  end for
  $\mathcal{P}' \leftarrow \mathcal{P}' \cup \{Z\}$.
end for
return $\mathcal{P}'$.



(a) Ground-truth graph.

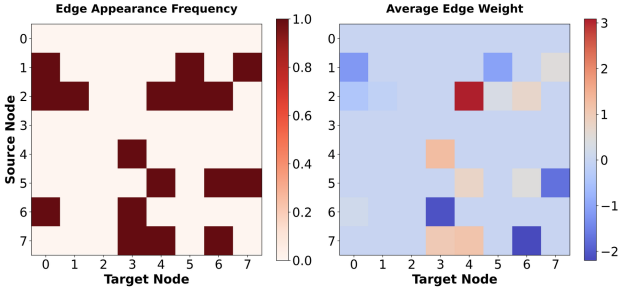


(b) Conditional posterior under strong multi-effect constraints.

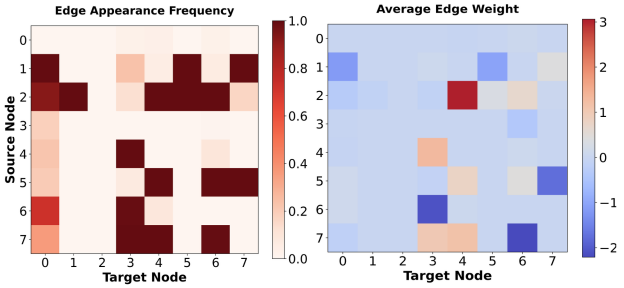


(c) Conditional posterior under weak multi-effect constraints.

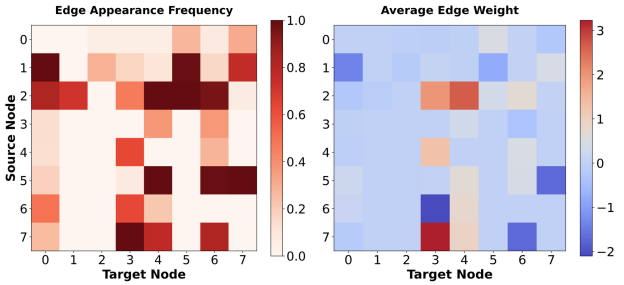
Figure 6: $d = 4$ multi-effect conditioning results. Each panel summarizes edge frequencies and average edge weights. Stronger constraints concentrate posterior mass more tightly around the ground-truth mechanism.



(a) Ground-truth graph.



(b) Conditional posterior under strong multi-effect constraints.



(c) Conditional posterior under weak multi-effect constraints.

Figure 7: $d = 8$ multi-effect conditioning results, with the same layout as Fig. 6. The strong constraints produce sharper edge and weight summaries than the weak constraints.