Active Bayesian Causal Inference

Anonymous Author(s)
Affiliation
Address
email

Abstract

Causal discovery and causal reasoning are classically treated as separate and consecutive tasks: one first infers the causal graph, and then uses it to estimate causal effects of interventions. However, such a two-stage approach is uneconomical, especially in terms of actively collected interventional data, since the causal query of interest may not require a fully-specified causal model. From a Bayesian perspective, it is also unnatural, since a causal query (e.g., the causal graph or some causal effect) can be viewed as a latent quantity subject to posterior inference—quantities that are not of direct interest ought to be marginalized out in this process, thus contributing to our overall uncertainty. In this work, we propose Active Bayesian Causal Inference (ABCI), a *principled fully-Bayesian active learning framework for integrated causal discovery and reasoning*, which jointly infers a posterior over causal models and queries of interest. In our approach to ABCI, we focus on the class of causally-sufficient nonlinear additive Gaussian noise models, which we model using Gaussian processes. To capture the space of causal graphs, we use a continuous latent graph representation, allowing our approach to scale to practically relevant problem sizes. We sequentially design experiments that are maximally informative about our target causal query, collect the corresponding interventional data, update our beliefs, and repeat. Through simulations, we demonstrate that our approach is more data-efficient than existing methods that only focus on learning the full causal graph. This allows us to accurately learn downstream causal queries from fewer samples, while providing well-calibrated uncertainty estimates of the quantities of interest.

1 Introduction

Causal reasoning, that is, answering causal queries such as the effect of a particular intervention, is a fundamental scientific quest [3, 24, 27, 34]. A rigorous treatment of this quest requires a reference causal model, typically consisting at least of (i) a causal diagram, or directed acyclic graph (DAG), capturing the qualitative causal structure between a system's variables [38]; and (ii) a joint distribution which is Markovian w.r.t. this causal graph [52]. Other frameworks additionally model (iii) the functional dependence of each variable on its causal parents in the graph [39, 58]. If the graph is not known from domain expertise, *causal discovery* aims to infer it from data [33, 52]. However, given only observational (passively collected) data, causal discovery is fundamentally limited to recovering the Markov equivalence class (MEC) of DAGs implying the same conditional independences as the data [52]. Additional structural assumptions (e.g., linearity) can render the graph identifiable [25, 42, 49, 59] but are often hard to falsify, thus leading to risk of misspecification. These shortcomings motivate learning from experimental (interventional) data which suffices to uniquely recover the true graph [10, 11, 19]. Here, we are particularly interested in the *active* learning setting in which we can sequentially design and perform interventions that are most informative for the target causal query [1, 17, 19, 20, 35, 55].

Classically, causal discovery and reasoning are treated as separate, consecutive tasks that are studied by different communities. Prior work on experimental design has thus focused either purely on causal reasoning—how to best design experimental studies if the causal graph is known—or purely on causal discovery, whenever the graph is unknown. In contrast, we consider the arguably more common setting in which we are interested in performing causal reasoning but do not have access to a reference causal model a priori. In this case, causal discovery can be seen as a means to an end, rather than as the main objective. Nonetheless, existing experimental design approaches generally focus on learning

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

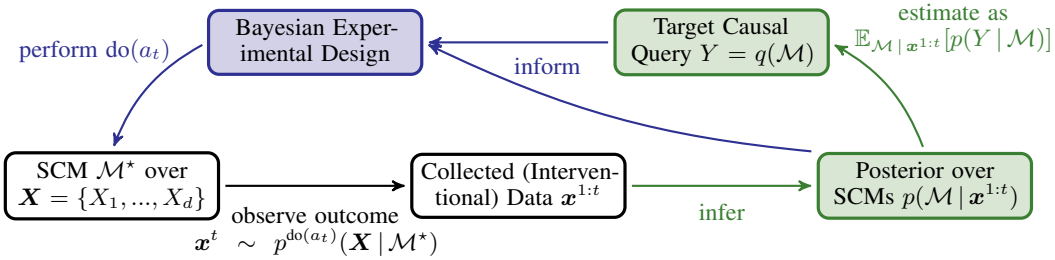

Figure 1: **Overview of the Active Bayesian Causal Inference (ABCI) framework.** At each time step t, we use Bayesian experimental design based on our current beliefs to choose a maximally informative intervention a_t to perform. We then collect a finite data sample from the interventional distribution induced by the environment, which we assume to be described by an unknown structural causal model (SCM) $\mathcal{M}^\star$ over a set of observable variables $\boldsymbol{X}$. Given (interventional) data $\boldsymbol{x}^{1:t}$ collected from the true SCM $\mathcal{M}^\star$, together with a prior distribution over the model class of consideration, we infer the posterior over a target causal query $Y = q(\mathcal{M})$ that can be expressed as a function of the causal model: for example, we may be interested in the graph (causal discovery), the presence of certain edges (partial causal discovery), the full SCM (causal model learning), a collection of interventional distributions or treatment effects (causal reasoning), or any combination thereof.

the graph, which is subsequently fixed for the causal reasoning phase. This can be disadvantageous for two reasons: first, wasting samples on learning the full causal graph is suboptimal when we are only interested in specific aspects of the causal model; and second, causal discovery from finite (especially small amounts of) data entails significant epistemic model uncertainty—e.g., from low statistical test power or multiple highly-scoring DAGs— which should be taken into account [2, 13].

In the present work, we propose *Active Bayesian Causal Inference (ABCI), a principled, fully-Bayesian framework for integrated causal discovery and reasoning with experimental design*. The basic approach is to put a Bayesian prior over the causal model class of choice, and to cast the learning problem as Bayesian inference over the model posterior. Moreover, we introduce the *target causal query* which is a function of the causal model that returns the (set of) causal quantities we are interested in. The model posterior together with the query function induce a *query posterior* which represents the result of our Bayesian learning procedure; it can be used, e.g., to derive a MAP solution or suitable expectation, or for down-stream decision tasks. The query posterior is incorporated in an active learning loop: we follow the Bayesian optimal experimental design approach [6, 28] and sequentially choose admissible interventions on the true causal model which are most informative about our target query w.r.t. our current beliefs. We then update our beliefs given the observed data by computing the posterior over causal models and queries, and use them to design the next experiment.

Since the general ABCI framework is computationally highly challenging, we implement it for the class of causally-sufficient nonlinear additive Gaussian noise models [25] which we model using Gaussian processes (GPs) [14, 57]. While this class is somewhat restrictive from a causal perspective, it is a flexible non-linear causal model which automates causal discovery in a wide range of scientific and engineering disciplines, as long as causal sufficiency can be reasonably assumed. To parameterize the combinatorial space of causal graphs, we use a recently proposed framework for differentiable Bayesian structure learning (DiBS) [30] that employs a continuous latent probabilistic graph representation to allow for tractable posterior inference. To approximately maximise information gain, we rely on Bayesian optimisation [31, 32, 51]. We highlight the following contributions:

- We propose ABCI as a flexible Bayesian active learning framework for efficiently inferring arbitrary sets of causal queries, subsuming causal discovery and reasoning as special cases (§ 3).
- We give a fully Bayesian treatment for the flexible class of nonlinear additive Gaussian noise models by leveraging GPs, continuous graph parametrisations, and Bayesian optimisation (§ 4).
- We demonstrate that our approach scales to relevant problem sizes and compares favourably to baselines in terms of efficiently learning the graph, full SCM, or interventional distributions (§ 5).

2 Related Work

Causal discovery and reasoning have been widely studied in machine learning and statistics [23, 42]. Given an already collected set of observations, there is a large body of literature on learning causal structure, both in the form of a point estimate [18, 41, 49, 52] and a Bayesian posterior [2, 8, 13, 21, 30]. Given a known causal graph, previous work studies how to estimate treatment effects or counterfactuals [39, 47, 48]. When interventional data is yet to be collected, existing work primarily focuses on the specific task of structure learning—without its downstream use. The concept

of (Bayesian) active causal discovery was first considered in discrete models with closed-form marginal likelihoods [35, 55] and later extended to nonlinear causal mechanisms [54, 56], multi-target interventions [53], and general models by using hypothesis testing [15]. Graph theoretic works give insights on the interventions required for full identifiability [10, 11, 19, 26].

Beyond learning the complete causal graph, few prior works have studied active causal inference. Concurrent work of Tigas et al. [54] considers experimental design for learning a full SCM parametrised by neural networks; there are significant differences to our approach: in particular, our framework (§ 3) is not limited to the information gain over the full model and provides a fully Bayesian treatment of the functions (§ 4). Agrawal et al. [1] consider actively learning a function of the causal graph under budget constraints, though not of the causal mechanisms and only for linear Gaussian models. Conversely, Rubenstein et al. [46] actively learn the causal mechanisms after the causal graph has been inferred. Thus, while prior work considers causal discovery and reasoning as a separate tasks, ABCI forms an integrated Bayesian approach for learning causal queries through interventions, reducing to previously studied settings in special cases. We further discuss related work in Appx. A.

3 Active Bayesian Causal Inference (ABCI) Framework

In this section, we first introduce the ABCI framework in generality, focusing on the main ideas and high-level ingredients, which are also illutrated in Fig. 1. In § 4 we then describe our particular implementation for the class of causally sufficient non-linear additive Gaussian noise models.

Notation. We use upper-case X and lower-case x to denote random variables and their realizations, respectively. Sets and vectors are written in bold face, $\boldsymbol{X}$ and $\boldsymbol{x}$. With a slight abuse of notation, we use $p(\cdot)$ to denote different distributions, or densities, which are distinguished by their arguments.

Causal Model. To treat causality in a rigorous way, we first need to postulate a mathematically well-defined causal model. Historically hard questions about causality can then be reduced to *epistemic questions*, that is, what and how much is known about the causal model. A prominent type of causal model is the *structural causal model* (SCM) [39]. From a Bayesian perspective, an SCM can be viewed as a hierarchical data-generating process involving latent random variables.

Definition 1 (SCM)**.** An SCM $\mathcal{M}$ over a set of *endogenous* (observed) variables $\boldsymbol{X} = \{X_1, \dots, X_d\}$ and *exogenous* (latent) variables $\boldsymbol{U} = \{U_1, \dots, U_d\}$ consists of structural equations, or *mechanisms*,

$$X_i := f_i(\mathbf{Pa}_i, U_i), \qquad \text{for} \quad i \in \{1, \dots, d\}, \tag{3.1}$$

which assign the value of each X_i as a *deterministic* function f_i of its direct causes, or *causal parents*, $\mathbf{Pa}_i \subseteq \boldsymbol{X} \setminus \{X_i\}$ and U_i; and a joint distribution $p(\mathbf{U})$ over the exogenous variables.

Associated with each SCM is a directed causal graph G with vertices $\boldsymbol{X}$ and edges $X_j \to X_i$ iff. $X_j \in \mathbf{Pa}_i$, which we assume to be *acyclic* (i.e., it is a DAG). Any acyclic SCM then induces a unique *observational distribution* $p(\boldsymbol{X} \mid \mathcal{M})$ over the endogenous variables $\boldsymbol{X}$, which is obtained as the pushforward measure of $p(\boldsymbol{U})$ through the causal mechanisms in Eq. (3.1).

Interventions. A crucial aspect of causal models such as SCMs is that they also model the effect of *interventions*—external manipulations to one or more of the causal mechanisms in Eq. (3.1)—which, in general, are denoted using Pearl's do-operator [39] as $\mathrm{do}(\{X_i = \tilde{f}_i(\mathbf{Pa}_i, U_i)\}_{i \in \mathcal{I}})$ with $\mathcal{I} \subseteq [d]$ and suitably chosen $\tilde{f}_i(\cdot)$. An intervention leads to a new SCM, the so-called *interventional SCM*, in which the relevant structural equations in Eq. (3.1) have been replaced by the new, manipulated ones. The interventional SCM thus induces a new distribution over the observed variables, the so-called *interventional distribution* which is denoted by $p^{\mathrm{do}(a)}(\boldsymbol{X} \mid \mathcal{M})$ with a denoting the (set of) intervention(s) $\{X_i = \tilde{f}_i(\mathbf{Pa}_i, U_i)\}_{i \in \mathcal{I}}$. *Causal effects*—expressions like $\mathbb{E}[X_j | \mathrm{do}(X_i = 3)]$—can then be derived from the corresponding interventional distribution via standard probabilistic inference.

Being Bayesian with Respect to Causal Models. The main epistemic challenge for causal reasoning stems from the fact that the *true causal model* $\mathcal{M}^\star$ is not (or not completely) known. The canonical response to such epistemic challenges is a *Bayesian approach*: put a prior $p(\mathcal{M})$ on causal models, collect data $\mathcal{D}$ from the true model $\mathcal{M}^\star$, and compute the posterior via Bayes rule:

$$p(\mathcal{M} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})}{p(\mathcal{D})} = \frac{p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})}{\int p(\mathcal{D} \mid \mathcal{M})\, p(\mathcal{M})\, \mathrm{d}\mathcal{M}}\,. \tag{3.2}$$

A full Bayesian treatment over $\mathcal{M}$ is computationally delicate, to say the least. First, we require a way to parametrise the class of models $\mathcal{M}$ we consider. Second, we need to be able to perform joint

posterior inference over this model class. In this paper, we present (one of) the first full Bayesian approaches which considers a flexible model class with nonlinear relationships (§ 4).

Bayesian Causal Inference. In the causal inference literature, the tasks of *causal discovery* (or, more generally, causal model learning) and *causal reasoning* are typically considered as separate problems. The former aims to learn (parts) of the causal model $\mathcal{M}^\star$ (typically the causal graph $G^\star$) while the latter, assuming that the relevant parts of $\mathcal{M}^\star$ are already known, aims to identify and estimate some query of interest, typically using only observational data. This separation essentially suggests a two-stage approach: causal discovery followed by causal reasoning. From a Bayesian perspective, however, this distinction is unnatural and there is no real conceptual difference between the two. Rather, we might define a *causal query function* q, which specifies a *target causal query* $Y = q(\mathcal{M})$ as a function of the causal model $\mathcal{M}$. This view thus subsumes and generalises causal discovery and reasoning. Concretely, possible causal queries are

Causal Discovery: $Y = q_{\text{CD}}(\mathcal{M}) = G$, that is, learning the full causal graph G;
Partial Causal Discovery: $Y = q_{\text{PCD}}(\mathcal{M}) = \phi(G)$, that is, learning some feature ϕ of the graph, such as the presence of a particular (set of) edge(s).
Causal Model Learning: $Y = q_{\text{CML}}(\mathcal{M}) = \mathcal{M}$, that is, learning the full SCM $\mathcal{M}$;
Causal Reasoning: $Y = q_{\text{CR}}(\mathcal{M}) = \{p^{\text{do}(\boldsymbol{X}_{\mathcal{I}(j)})}(X_j \mid \mathcal{M})\}_{j \in \mathcal{J}}$, that is, learning a set of interventional distributions induced by $\mathcal{M}$.[1]

Once we have fixed the causal query, Bayesian inference naturally extends to the *query posterior*:

$$p(Y \mid \mathcal{D}) = \int p(Y \mid \mathcal{M})\, p(\mathcal{M} \mid \mathcal{D})\, \mathrm{d}\mathcal{M} = \mathbb{E}_{\mathcal{M} \mid \mathcal{D}}[\, p(Y \mid \mathcal{M})]\,, \tag{3.3}$$

where $p(Y \mid \mathcal{M})$ is deterministically given by $q(\mathcal{M})$, i.e., a point mass. Evidently, computing Eq. (3.3) constitutes a hard computational problem in general, as we need to marginalise over all causal models. In § 4 we introduce a practical implementation for a restricted causal model class.

Identifiability of causal models and queries. A crucial concept is that of *identifiability* of a model class, which refers to the ability to uniquely recover the true model in the limit of infinitely many observations from it [16].[2] In the context of our setting, if the class of causal models $\mathcal{M}$ is identifiable, the model posterior $p(\mathcal{M} \mid \mathcal{D})$ in Eq. (3.2) and hence also the query posterior $p(Y \mid \mathcal{D})$ in Eq. (3.3) will collapse and converge to a point mass on their respective true values $\mathcal{M}^\star$ and $q(\mathcal{M}^\star)$, given infinite data and provided the true model has non-zero mass under our prior, $p(\mathcal{M}^\star) > 0$. Given only *observational* data, causal models are notoriously unidentifiable in general: without further assumptions on $p(\mathbf{U})$ and the structural form of Eq. (3.1), neither the graph nor the mechanisms can be recovered. In this case, $p(\mathcal{M} \mid \mathcal{D})$ may only converge to an equivalence class of models that cannot be further distinguished. Note, however, that even in this case, $p(Y \mid \mathcal{D})$ may still sometimes collapse, for example, if the Markov equivalence class (MEC) of graphs is identifiable (under causal sufficiency) and our query concerns the presence of a particular edge which is shared by all graphs in the MEC.

Active Learning with Sequential Interventions. Rather than collecting a *large, observational* dataset, we leverage observations from a *small* number of sequentially-performed *experiments*. The motivation for this is two-fold: first, experimental data can help resolve some of the non-identifiability issues discussed above; second, even if the model is identifiable (as for our approach in § 4), interventional data can still help learn our target causal query more quickly from *finite* data. Hence, at each time step t, we assume that we can perform an *experiment in the form of an intervention* a_t. The outcome of this experiment is a batch $\boldsymbol{x}^t$ of N_t i.i.d. observations from the *true* interventional distribution:

$$\boldsymbol{x}^t = \{\boldsymbol{x}^{t,n}\}_{n=1}^{N_t}, \qquad \boldsymbol{x}^{t,n} \overset{\text{i.i.d.}}{\sim} p^{\text{do}(a_t)}(\boldsymbol{X} \mid \mathcal{M}^\star) \tag{3.4}$$

Note that restricting to $a_t = \varnothing$—that is, sampling from the observational distribution—amounts to learning from observational data as a special case. Crucially, however, we design the experiment

[1] Here the set $\mathcal{J}$ can be uncountable, subsuming interventional distributions for a continuous set of interventions, possibly on different variables. Thus, in this case the return value of q is a set of density functions. In practice, these are implicitly represented in the learned Bayesian models, see § 5.

[2] It is worth pointing out that the term "identifiability" is sometimes used differently in the causal inference literature: within causal discovery, it typically refers to *structure identifiability*, that is, recovering only the causal graph; in the context of causal reasoning, on the other hand, it typically refers to whether an interventional (or counterfactual) query can be *expressed in terms of known quantities*, usually involving only the observational distribution. Here, we will use the term in its (original) statistical sense to refer to *identifiability of models*.

a_t so that it is *maximally informative* about our target causal query Y. In our Bayesian setting, this is naturally formulated by maximising the *information gain* between Y and the outcome $\boldsymbol{X}^t$ [6, 28]:

$$\max_{a_t} \quad \mathrm{I}(Y; \boldsymbol{X}^t \mid \boldsymbol{x}^{1:t-1}) \tag{3.5}$$

where $\boldsymbol{X}^t$ follows the *predictive interventional distribution* of the Bayesian causal model ensemble at time $t-1$ under intervention a_t, which is given by

$$\boldsymbol{X}^t \sim p^{\mathrm{do}(a_t)}(\boldsymbol{X} \mid \boldsymbol{x}^{1:t-1}) \propto \int p^{\mathrm{do}(a_t)}(\boldsymbol{X} \mid \mathcal{M})\, p(\mathcal{M} \mid \boldsymbol{x}^{1:t-1})\, \mathrm{d}\mathcal{M}. \tag{3.6}$$

By maximizing Eq. (3.5) we collect experimental data in a goal-oriented manner to learn our causal query Y as efficiently and quickly as possible.

4 Tractable ABCI for Nonlinear Additive Noise Models

Having discussed the general framework and conceptual ideas, we now present our concrete approach to ABCI. This requires specifying: (i) the class of causal models we consider in Eq. (3.1), including their parametrisation; (ii) the types of interventions a_t we consider at each step and the corresponding interventional likelihood in Eq. (3.4); (iii) our prior distribution $p(\mathcal{M})$ over models; (iv) how to do posterior inference, that is, how to compute the model posterior in Eq. (3.2); and finally (v) how to maximise the information gain in Eq. (3.5) for experimental design.

Model Class and Parametrisation. In our approach to ABCI, we consider SCMs of the form

$$X_i := f_i(\mathbf{Pa}_i) + U_i, \qquad \text{with} \qquad U_i \sim \mathcal{N}(0, \sigma_i^2), \qquad \text{for} \quad i \in \{1, \dots, d\}, \tag{4.1}$$

where the f_i are *smooth, nonlinear* functions and where the U_i are assumed to be mutually independent, corresponding to the assumption of *causal sufficiency* (no hidden confounding). That is, we consider the special case of *causally sufficient, non-linear, Gaussian additive noise models*. Any model $\mathcal{M}$ of this form can be described by a triple $\mathcal{M} = (G, \boldsymbol{f}, \boldsymbol{\sigma}^2)$, where G is a causal DAG, $\boldsymbol{f} = (f_1, \dots, f_d)$ is a vector of functions defined over the parent sets implied by G, and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_d^2)$ contains the Gaussian noise variances. Provided that the f_i are nonlinear and not constant in any of their arguments, the model is identifiable almost surely [25, 43].

Interventional Likelihood. We support the realistic setting where only a subset $\boldsymbol{W} \subseteq \boldsymbol{X}$ of all variables are *actionable*, i.e., only $\boldsymbol{W}$ can be the target of an intervention.[3] For simplicity, we consider *hard interventions* of the form $\mathrm{do}(a_t) = \mathrm{do}(\boldsymbol{X}_\mathcal{I} = \boldsymbol{x}_\mathcal{I})$ which fix a subset $\boldsymbol{X}_\mathcal{I} \subseteq \boldsymbol{W}$ to a constant $\boldsymbol{x}_\mathcal{I}$. Due to causal sufficiency, the interventional likelihood under such hard interventions a_t factorises over the causal graph G and is given by the *g-formula* [44] or *truncated factorisation* [52]:

$$p^{\mathrm{do}(a_t)}(\boldsymbol{X} \mid G, \boldsymbol{f}, \boldsymbol{\sigma}^2) = \mathbb{I}\{\boldsymbol{X}_\mathcal{I} = \boldsymbol{x}_\mathcal{I}\} \prod_{j \notin \mathcal{I}} p(X_j \mid f_j(\mathbf{Pa}_j^G), \sigma_j^2). \tag{4.2}$$

The last term in Eq. (4.2) is given by $\mathcal{N}(f_j(\mathbf{Pa}_j^G), \sigma_j^2)$ due to the Gaussian noise assumption. Let $\boldsymbol{x}^{1:t}$ be the entire dataset, collected up to time t. The likelihood of $\boldsymbol{x}^{1:t}$ is then given by

$$p(\boldsymbol{x}^{1:t} \mid G, \boldsymbol{f}, \boldsymbol{\sigma}^2) = \prod_{\tau=1}^t p^{\mathrm{do}(a_\tau)}(\boldsymbol{x}^\tau \mid G, \boldsymbol{f}, \boldsymbol{\sigma}^2) = \prod_{\tau=1}^t \prod_{n=1}^{N_t} p^{\mathrm{do}(a_\tau)}(\boldsymbol{x}^{\tau,n} \mid G, \boldsymbol{f}, \boldsymbol{\sigma}^2). \tag{4.3}$$

Structured Model Prior. To specify our model prior, we need to distinguish between *root nodes* X_i, for which $\mathbf{Pa}_i = \varnothing$ and thus $f_i = \text{const}$, and *non-root nodes* X_j. For a given G, denote by $\mathbf{R}(G) = \{i \in [d] : \mathbf{Pa}_i^G = \varnothing\}$ the index set of root nodes, and by $\mathbf{NR}(G) = [d] \setminus \mathbf{R}(G)$ that of non-root nodes. We then place the following structured prior over the class of models $\mathcal{M} = (G, \boldsymbol{f}, \boldsymbol{\sigma}^2)$:

$$p(\mathcal{M}) = p(G)\, p(\boldsymbol{f}, \boldsymbol{\sigma}^2 \mid G) = p(G) \prod_{i \in \mathbf{R}(G)} p(f_i, \sigma_i^2 \mid G) \prod_{j \in \mathbf{NR}(G)} p(f_j \mid G)\, p(\sigma_j^2 \mid G)\,. \tag{4.4}$$

Here, $p(G)$ is a prior over graphs, and $p(\boldsymbol{f}, \boldsymbol{\sigma}^2 \mid G)$ is a prior over the functions and noise variances in G. We factorise our prior conditional on G as in Eq. (4.4) not only to allow for a separate treatment of root nodes and non-root nodes but also to *share priors across similar graphs*: whenever $\mathbf{Pa}_i^{G_1} = \mathbf{Pa}_i^{G_2}$, we set $p(f_i, \sigma_i^2 \mid G_1) = p(f_i, \sigma_i^2 \mid G_2)$, and similarly for $p(f_j \mid G)$ and $p(\sigma_j^2 \mid G)$. As a consequence, the posteriors are also shared, which substantially reduces the computational burden. We also assume that $f_j \perp\!\!\!\perp f_{j'} \mid G$ and $\sigma_j^2 \perp\!\!\!\perp \sigma_{j'}^2 \mid G$ for all $j \neq j' \in \mathbf{NR}(G)$, which is motivated by the principle of independent causal mechanisms [42]. Our specific choices for $p(G)$, $p(f_i, \sigma_i^2 \mid G)$, $p(f_j \mid G)$, and $p(\sigma_j^2 \mid G)$ are guided by computational challenges and described in more detail below.

[3] In principle, the set of actionable variables might even change over time, in which case they are denoted $\boldsymbol{W}_t$.

Model Posterior. Given collected data $\boldsymbol{x}^{1:t}$, we can update our beliefs and quantify our uncertainty in $\mathcal{M}^\star$ by inferring a Bayesian posterior $p(\mathcal{M} \mid \boldsymbol{x}^{1:t})$ over SCMs $\mathcal{M} = (G, \boldsymbol{f}, \boldsymbol{\sigma}^2)$ as follows:[4]

$$p(\mathcal{M} \mid \boldsymbol{x}^{1:t}) = p(G \mid \boldsymbol{x}^{1:t}) \prod_{i \in \mathbf{R}(G)} p(f_i, \sigma_i^2 \mid \boldsymbol{x}^{1:t}, G) \prod_{j \in \mathbf{NR}(G)} p(f_j, \sigma_j^2 \mid \boldsymbol{x}^{1:t}, G)\,. \tag{4.5}$$

For root nodes $i \in \mathbf{R}(G)$, posterior inference is straight-forward: we have $f_i = \text{const}$, so f_i can be viewed as the mean of U_i, cf. Eq. (4.1). We thus place a conjugate Normal-Gamma$(\mu_i, \lambda_i, \alpha_i^{\text{R}}, \beta_i^{\text{R}})$ prior on $p(f_i, \sigma_i^2 \mid G)$, so that we can analytically compute the root node posterior $p(f_i, \sigma_i^2 \mid \boldsymbol{x}^{1:t}, G)$ in Eq. (4.5) in closed form [36]. We collect all the Normal-Gamma hyperparameters in $(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\alpha}^{\text{R}}, \boldsymbol{\beta}^{\text{R}})$.

The posteriors over graphs and non-root nodes $j \in \mathbf{NR}(G)$ are given as

$$p(G \mid \boldsymbol{x}^{1:t}) = \frac{p(\boldsymbol{x}^{1:t} \mid G)\, p(G)}{p(\boldsymbol{x}^{1:t})}\,, \qquad p(f_j, \sigma_j^2 \mid \boldsymbol{x}^{1:t}, G) = \frac{p(\boldsymbol{x}^{1:t} \mid G, f_j, \sigma_j^2)\, p(f_j, \sigma_j^2 \mid G)}{p(\boldsymbol{x}^{1:t} \mid G)}\,. \tag{4.6}$$

Computing these posteriors is more involved and discussed below.

4.1 Addressing Challenges for Posterior Inference with GPs and DiBS

The particular challenges in Eq. (4.6) are the terms $p(\boldsymbol{x}^{1:t} \mid G)$ and $p(\boldsymbol{x}^{1:t})$. In the following, we will address these by means of appropriate prior choices and approximations.

Challenge 1: Marginalising out Functions. The term $p(\boldsymbol{x}^{1:t} \mid G)$ in Eq. (4.6) reads

$$p(\boldsymbol{x}^{1:t} \mid G) = \int p(\boldsymbol{x}^{1:t} \mid G, f_j, \sigma_j^2)\, p(f_j \mid G)\, p(\sigma_j^2 \mid G)\, \mathrm{d}f_j\, \mathrm{d}\sigma_j^2 \tag{4.7}$$

and requires evaluating integrals over the function domain. We use *Gaussian processes* (GPs) [57] as an elegant way to solve this problem, as GPs can flexibly model *nonlinear* functions while offering convenient analytical properties. Specifically, we place a $\mathcal{GP}(0, k_j^G(\cdot,\cdot))$ prior on $p(f_j|G)$, where $k_j^G(\cdot,\cdot)$ is a covariance function over $\mathbf{Pa}_j^G$ with length scales $\boldsymbol{\kappa}_j$, which we collect in $\boldsymbol{\kappa}$. In line with the GP-literature, we refer to $(\boldsymbol{\kappa}_j, \sigma_j^2)$ as the *GP-hyperparameters*. We place Gamma$(\alpha_j^\sigma, \beta_j^\sigma)$ and Gamma$(\alpha_j^\kappa, \beta_j^\kappa)$ priors on $p(\sigma_i^2 \mid G)$ and $p(\kappa_i \mid G)$ and collect their parameters in $(\boldsymbol{\alpha}^{\text{GP}}, \boldsymbol{\beta}^{\text{GP}})$, see Fig. 2. For our model class, GPs then provide closed-form expressions for the "GP-marginal likelihood" $p(\boldsymbol{x}^{1:t} \mid G, \sigma_j^2, \boldsymbol{\kappa}_j)$, as well as for the "GP posteriors" $p(f_j \mid \boldsymbol{x}^{1:t}, G, \sigma_j^2, \boldsymbol{\kappa}_j)$, and the "predictive posteriors over observations" $p(\boldsymbol{X} \mid \boldsymbol{x}^{1:t}, G, \boldsymbol{\sigma}^2, \boldsymbol{\kappa})$ [57], see Appx. B for details.

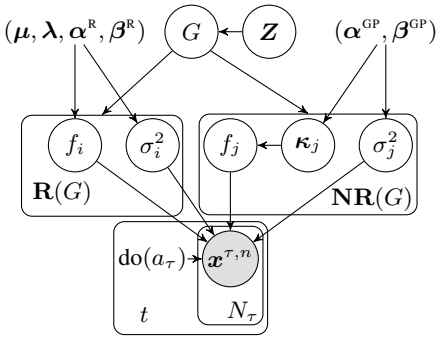

Figure 2: Graphical model representation of our GP-DiBS-ABCI approach.

Challenge 2: Marginalising out GP-Hyperparameters. While GPs allow for exact posterior inference *conditional on a fixed value of* $(\sigma_j^2, \boldsymbol{\kappa}_j)$, evaluating expressions such as $p(f_j \mid \boldsymbol{x}^{1:t}, G)$ requires marginalising out these GP-hyperparameters from the GP-posterior (see above). Unfortunately, this cannot, in general, be done exactly in connection with GPs as there is no closed-form expression for $p(\sigma_j^2, \kappa_j \mid \boldsymbol{x}^{1:t}, G)$. We therefore approximate such expectations with a maximum a posteriori (MAP) point estimate $(\hat{\sigma}_j^2, \hat{\boldsymbol{\kappa}}_j)$, obtained by performing gradient ascent on the unnormalized log posterior,

$$\nabla \log p(\sigma_j^2, \boldsymbol{\kappa}_j \mid \boldsymbol{x}^{1:t}, G) = \nabla \log p(\boldsymbol{x}^{1:t} \mid G, \sigma_j^2, \boldsymbol{\kappa}_j) + \nabla \log p(\sigma_j^2, \boldsymbol{\kappa}_j \mid G) \tag{4.8}$$

according to a predefined update schedule, cf. Alg. 1. That is, we use approximations of the form:

$$p(f_j \mid \boldsymbol{x}^{1:t}, G) = \int p(f_j \mid \boldsymbol{x}^{1:t}, G, \sigma_j^2, \boldsymbol{\kappa}_j) p(\sigma_j^2, \kappa_j \mid \boldsymbol{x}^{1:t}, G)\, \mathrm{d}\sigma_j^2\, \mathrm{d}\boldsymbol{\kappa}_j \approx p(f_j \mid \boldsymbol{x}^{1:t}, G, \hat{\sigma}_j^2, \hat{\boldsymbol{\kappa}}_j)$$

Challenge 3: Marginalising out Graphs. Further, the "evidence" $p(\boldsymbol{x}^{1:t})$ is given by

$$p(\boldsymbol{x}^{1:t}) = \textstyle\sum_G p(\boldsymbol{x}^{1:t} \mid G)\, p(G) \tag{4.9}$$

[4] To avoid further complicating the notation, we write all posteriors and likelihoods in terms of the full data $\boldsymbol{x}^{1:t}$. However, only observations of X_i and $X_j \mid \mathbf{Pa}_j^G$ matter for $i \in \mathbf{R}(G)$ and $j \in \mathbf{NR}(G)$.

Algorithm 1: GP-DiBS-ABCI for nonlinear additive Gaussian noise models

Input: no. of experiments T, batch sizes $\{N_t\}_{t=1}^T$, no. of latent particles M, no. of MC graphs K, particle resampling schedule $\{r_t\}_{t=1}^T$, hyperparameter update schedule $\{s_t\}_{t=1}^T$
Output: Posterior over target causal query $p(Y \mid \boldsymbol{x}^{1:T})$
for $t \leftarrow 1$ **to** T **do**
  $a_t \leftarrow \arg\max_{a=(\mathcal{I},\boldsymbol{x}_\mathcal{I})} U(a, \boldsymbol{x}^{1:t-1})$ ▷`design experiment: Eq. (4.11)`
  $\boldsymbol{x}^t \leftarrow \{\boldsymbol{x}^{(t,n)} \sim p^{\mathrm{do}(a_t)}(\boldsymbol{X} \mid \mathcal{M}^\star)\}_{n=1}^{N_t}$ ▷`perform experiment`
  if r_t **then**
    $\boldsymbol{z}^t \leftarrow$ `resample_particles` $(\boldsymbol{z}^t)$ ▷`see App.D`
  end
  $\boldsymbol{G} \leftarrow \{\{G^{(k,m)} \sim p(G \mid \boldsymbol{z}_m)\}_{k=1}^K\}_{m=1}^M$ ▷`sample graphs`
  $\boldsymbol{\kappa}, \boldsymbol{\sigma}^2 \leftarrow$ `estimate_hyperparameters`$(\boldsymbol{x}^{1:s_t}, \boldsymbol{G})$ ▷`see Eq. (4.8)`
  $\boldsymbol{z}^{t+1} \leftarrow \mathrm{SVGD}(\boldsymbol{z}^t, \boldsymbol{x}^{1:t})$ ▷`update latent particles`
end

and involves a summation over all possible DAGs G. This becomes intractable even for $d \geq 4$ variables as the number of DAGs grows super-exponentially in the number of variables [45]. To address this challenge, we employ the recently proposed DiBS framework [30]. By introducing a continuous prior $p(\boldsymbol{Z})$ that models G via $p(G \mid \boldsymbol{Z})$ and simultaneously enforces acyclicity of G, Lorch et al. [30] show that we can efficiently infer the discrete posterior $p(G \mid \boldsymbol{x}^{1:t})$ via $p(\boldsymbol{Z} \mid \boldsymbol{x}^{1:t})$ as

$$\mathbb{E}_{G \mid \boldsymbol{x}^{1:t}}\left[\phi(G)\right] = \mathbb{E}_{\boldsymbol{Z} \mid \boldsymbol{x}^{1:t}}\left[\frac{\mathbb{E}_{G \mid \boldsymbol{Z}}\left[p(\boldsymbol{x}^{1:t} \mid G)\,\phi(G)\right]}{\mathbb{E}_{G \mid \boldsymbol{Z}}\left[p(\boldsymbol{x}^{1:t} \mid G)\right]}\right] \tag{4.10}$$

where ϕ is some function of the graph. Since $p(\boldsymbol{Z} \mid \boldsymbol{x}^{1:t})$ is a continuous density with tractable gradient estimators, we can resort to efficient variational inference methods such as Stein Variational Gradient Descent (SVGD) for approximate inference [29], see Appx. D for additional details.

4.2 Approximate Bayesian Experimental Design with Bayesian Optimisation

As motivated in § 3, we aim to perform experiments a_t that are maximally informative about our target query $Y = q(\mathcal{M})$ by maximising the information gain from Eq. (3.5) given our current data $\mathcal{D} := \boldsymbol{x}^{1:t-1}$. In Appx. C we show that this is equivalent to maximising the following utility function:

$$\begin{aligned} U(a) =& H(\boldsymbol{X}^t \mid \mathcal{D}) + \mathbb{E}_{\mathcal{M} \mid \mathcal{D}}\left[\mathbb{E}_{\boldsymbol{X}^t, Y \mid \mathcal{M}}\left[\log \mathbb{E}_{\mathcal{M}' \mid \mathcal{D}}\left[p(\boldsymbol{X}^t \mid \mathcal{M}')\,p(Y \mid \mathcal{M}')\right]\right]\right], \\ \text{where} \quad & H(\boldsymbol{X}^t \mid \mathcal{D}) = \mathbb{E}_{\mathcal{M} \mid \mathcal{D}}\left[\mathbb{E}_{\boldsymbol{X}^t \mid \mathcal{M}}\left[\log \mathbb{E}_{\mathcal{M}' \mid \mathcal{D}}\left[p(\boldsymbol{X}^t \mid \mathcal{M}')\right]\right]\right] \end{aligned} \tag{4.11}$$

denotes the differential entropy of the experiment outcome, which depends on a and is distributed as in Eq. (3.6). This surrogate objective can be estimated using a nested Monte Carlo estimator, as long as we can sample from and compute $p(Y \mid \mathcal{M})$, see Appx. D for further details. For example, for $q_{\mathrm{CR}}(\mathcal{M}) = p^{\mathrm{do}(X_i=\psi)}(X_j \mid \mathcal{M})$ with $\psi \sim p(\psi)$ a distribution over intervention values, we get:

$$U_{\mathrm{CR}}(a) = H(\boldsymbol{X}^t \mid \mathcal{D}) + \mathbb{E}_{\boldsymbol{X}^t \mid \mathcal{D}}\,\mathbb{E}_{\psi}\,\mathbb{E}_{X_j}^{\mathrm{do}(X_i=\psi)}\left[\log \mathbb{E}_{\mathcal{M}' \mid \mathcal{D}}\left[p(\boldsymbol{X}^t \mid \mathcal{M}')\,p^{\mathrm{do}(X_i=\psi)}(X_j \mid \mathcal{M}')\right]\right].$$

Importantly, for specific instances of the query function $q(\cdot)$ discussed in § 3, we can derive simpler utilities than Eq. (4.11). For example, for $q_{\mathrm{CD}}(\mathcal{M}) = G$ and $q_{\mathrm{CML}}(\mathcal{M}) = \mathcal{M}$ we arrive at

$$U_{\mathrm{CD}}(a) = \mathbb{E}_{G \mid \mathcal{D}}\left[\mathbb{E}_{\boldsymbol{X}^t \mid G, \mathcal{D}}\left[\log p(\boldsymbol{X}^t \mid \mathcal{D}, G) - \log \mathbb{E}_{G' \mid \mathcal{D}}\left[p(\boldsymbol{X}^t \mid \mathcal{D}, G')\right]\right]\right], \tag{4.12}$$

$$U_{\mathrm{CML}}(a) = \mathbb{E}_{\mathcal{M} \mid \mathcal{D}}\left[\mathbb{E}_{\boldsymbol{X}^t \mid \mathcal{M}}\left[\log p(\boldsymbol{X}^t \mid \mathcal{M}) - \log \mathbb{E}_{G' \mid \mathcal{D}}\left[p(\boldsymbol{X}^t \mid \mathcal{D}, G')\right]\right]\right], \tag{4.13}$$

where the entropy $\mathbb{E}_{\boldsymbol{X}^t \mid \mathcal{M}}\left[\log p(\boldsymbol{X}^t \mid \mathcal{M})\right]$ can again be efficiently computed given our modelling choices. For the sake of brevity, we defer derivations and estimation details to Appxs. C and D.

Finding the optimal experiment $a_t^* = (\mathcal{I}^*, \boldsymbol{x}_\mathcal{I}^*)$ requires jointly optimising the utility function corresponding to our query with respect to (i) the set of intervention *targets* $\mathcal{I}$, and (ii) the corresponding intervention *values* $\boldsymbol{x}_\mathcal{I}$. This lends itself naturally to a nested, bi-level optimization scheme [56]:

$$\mathcal{I}^* \in \arg\max_\mathcal{I} U(\mathcal{I}, \boldsymbol{x}_\mathcal{I}^*), \quad \text{where} \quad \forall \mathcal{I}: \quad \boldsymbol{x}_\mathcal{I}^* \in \arg\max_{\boldsymbol{x}_\mathcal{I}} U(\mathcal{I}, \boldsymbol{x}_\mathcal{I}), \tag{4.14}$$

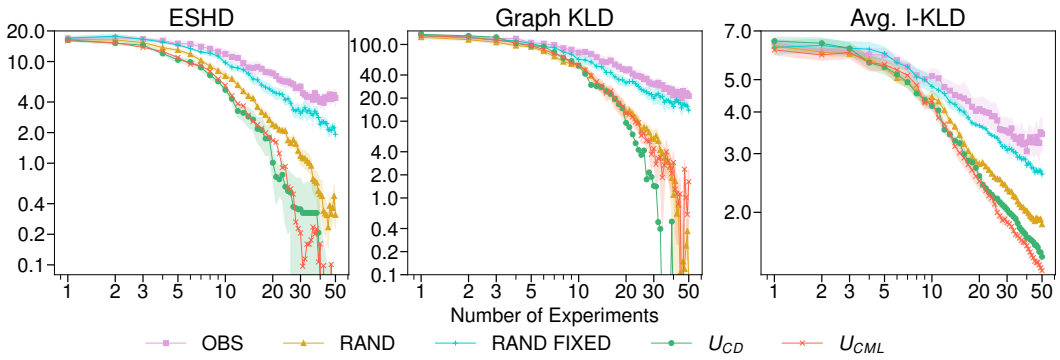

Figure 3: **Causal Discovery and SCM Learning.** Comparison of experimental design strategies for causal discovery (U_{CD}) and causal model learning (U_{CML}) with random and observational baselines on simulated ground truth models with 8 nodes. Lines and shaded areas show means ± 1 std. dev. across 30 runs (5 randomly sampled ground-truth SCMs with 6 restarts per SCM). **(a) ESHD.** Both our objectives significantly outperform the observational and random baselines. **(b) Graph-KLD.** U_{CD}, which optimises for this objective performs best as expected, but U_{CML} and the strong random baseline (RAND) perform competitively at learning the graph. **(c) Average I-KLD.** Both our strategies significantly outperform the baselines; U_{CML}, which aims to learn the full SCM, does slightly better than U_{CD} in terms of this proxy for causal model learning, as expected.

that is, we first estimate the optimal intervention values for all candidate intervention targets $\mathcal{I}$, and then select the intervention target that yields the highest utility. The intervention target $\mathcal{I}$ might contain multiple variables, which, however, yields a combinatorial problem. Thus, for simplicity, we consider only single-node interventions, i.e., $|\mathcal{I}| = 1$. To find $\boldsymbol{x}^*_{\mathcal{I}}$, we employ *Bayesian optimisation* [31, 32, 51] to efficiently estimate an optimal intervention value $\boldsymbol{x}^*_{\mathcal{I}}$, see Appx. D.

5 Experiments

Setup. We evaluate ABCI by inferring the query posterior on synthetic ground truth SCMs using several different experiment selection strategies. Specifically, we design experiments w.r.t. U_{CD} (causal discovery), U_{CML} (causal model learning), and U_{CR} (causal reasoning), see § 4.2. We compare against baselines which (i) only sample from the observational distribution (OBS) or (ii) pick an intervention target j uniformly at random from $[d] \cup \{\varnothing\}$ and set $X_j = 0$ (RAND FIXED, a weak random baseline used in prior work) or draw $X_j \sim \mathcal{U}(-7, 7)$ (RAND) if $X_j \neq \varnothing$. All methods follow our Bayesian GP-DiBS-ABCI approach from § 4. We sample ground truth SCMs over random scale-free graphs [4] of size $d = 8$, with mechanisms and noise variances drawn from our model prior Eq. (4.4). We initialise all methods with 5 observational samples, and then perform experiments with a batch size of 3. For specific prior choices and simulation details, see Appx. D.

Metrics. As ABCI infers a posterior over the target query Y, a natural evaluation choice is the Kullback-Leibler divergence (KLD) between the true query distribution and the inferred query posterior, $\mathrm{KL}(p(Y \mid \mathcal{M}^\star) || \, p(Y \mid \boldsymbol{x}^{1:t}))$. We report **Graph KLD**, a sample-based approximation of the KLD for posteriors over graphs (q_{CD}), and **Query KLD**, a KLD estimate for target interventional distributions (q_{CR}). As a proxy for the KLD of the SCM posterior (q_{CML}),[5] we report the average KLD across all single node interventional distributions $\{p^{\text{do}(X_i=\psi)}(\boldsymbol{X})\}_{i=1}^d$, with $\psi \sim \mathcal{U}(-7, 7)$ (**Average I-KLD**). We also report the *expected structural hamming distance* [9], **ESHD** $= \mathbb{E}_{G \mid \boldsymbol{x}^{1:t}}\left[\mathrm{SHD}(G, G^\star)\right]$, a commonly used causal discovery metric; see Appx. D for further details.

Causal Discovery and SCM Learning (Fig. 3). In our first experiment, we find that: (i) all our ABCI-based methods are able to meaningfully learn from small amounts of data, thus validating our Bayesian approach; further (ii) *performing targeted interventions using experimental design indeed yields improved performance over uninformed experimentation* (OBS, RAND FIXED, RAND). Notably, the stronger random baseline (RAND), which also randomises over intervention values, performs (surprisingly) well throughout—at least for the considered setting. As expected per the theoretical grounding of our information gain utilities, U_{CD} identifies the true graph the fastest (as measured by Graph KLD), whereas U_{CML} appears to most efficiently learn the full model, including the functions and noise variances, as measured by the Average I-KLD proxy, see the caption of Fig. 3 for further details.

Learning Interventional Distributions (Fig. 4). In our second experiment, we investigate ABCI's causal reasoning capabilities by randomly sampling ground truth SCMs (as described above) over the fixed graph shown in Fig. 4 (right)—which is *not* known to the methods—and treat the (uncountable)

[5] The SCM KLD is either zero, if the SCM posterior collapses onto the true SCM, or infinite, otherwise.

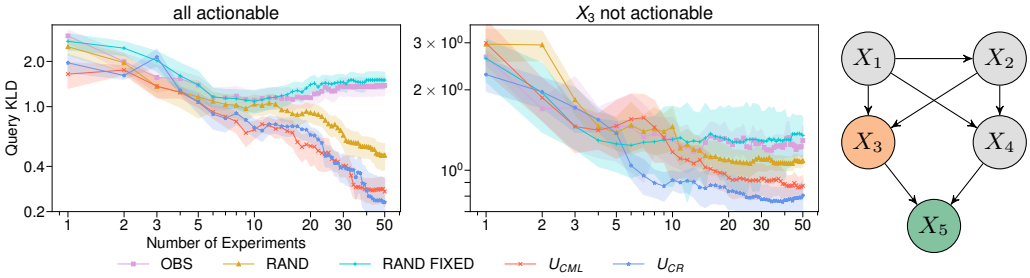

Figure 4: **Learning Interventional Distributions.** (left) Comparison of different methods w.r.t. learning a set of interventional distributions $p^{\text{do}(X_3=\psi)}(X_5 \mid \mathcal{M})$ with $\psi \sim \mathcal{U}[4, 7]$ on simulated ground truth models with fixed causal graph (right). Lines and shaded areas show mean ± 1 std. dev. across 25 runs (5 randomly sampled ground truth SCMs with 5 restarts each). **(a) All nodes actionable.** Our objectives significantly outperform the baselines; U_{CML} and U_{CR} perform similarly. In conjunction with results from Fig. 3, this suggests that U_{CML} yields a solid base model for performing downstream causal inference tasks. **(b)** X_3 **not actionable.** In this setting, where we cannot directly intervene on the treatment variable of interest, U_{CR} clearly outperforms all other methods for ≥ 5 experiments, suggesting that, in such a scenario, query-targeted experimental design is particularly helpful.

set of interventional distributions $p^{\text{do}(X_3=\psi)}(X_5 \mid \mathcal{M})$ with $\psi \sim \mathcal{U}[4, 7]$ as the target query. We find that *our informed experiment selection strategies significantly outperform the baselines at causal reasoning*, as measured by the Query KLD. In accord with the results from Fig. 3 and considering that, once we know the true SCM, we can compute any causal quantity of interest, U_{CML} thus seems to provide a reasonable experimental strategy in case the causal query of interest is *not* known a priori. However, our results indicate that if we *do* know our query of interest, then U_{CR} *provides an even faster way for its estimation, especially when the treatment variable of interest is not directly intervenable*. Note the different axis scales, indicating that the task is harder in this case, as expected.

6 Discussion

Assumptions, Limitations, and Extensions. In § 4, we have made several assumptions to facilitate tractable inference and showcase the ABCI framework in a relatively simple causal setting. In particular, our assumptions exclude heteroscedastic noise, unobserved confounding, and cyclic relationships. On the experimental design side, we only considered *hard* interventions, but for some applications *soft* interventions [12] are more plausible. On the query side, we only considered *interventional* distributions. However, SCMs also naturally lend themselves to *counterfactual* reasoning, so one could also consider counterfactual queries such as the effect of the treatment on the treated [22, 50]. *In principle*, the ABCI framework as presented in § 3 extends directly to such generalisations. *In practice*, however, these can be non-trivial to implement, especially with regard to model parametrisation and tractable inference. Since actively performed interventions allow for causal learning even under causal sufficiency violations, we consider this a promising avenue for future work and believe the ABCI framework to be particularly well-suited for exploring it. Extensions to other causal modelling frameworks, such as graphical causal models are also of interest.

Reflections on the ABCI Framework. The main conceptual advantages of the ABCI framework are that it is *flexible* and *principled*. By considering general target causal queries, we can precisely specify what aspects of the causal model we are interested in, thereby offering a fresh perspective on the classical divide between causal discovery and reasoning: sometimes, the main objective may be to foster scientific understanding by uncovering the qualitative causal structure underlying real-world systems; other times, causal discovery may only be a means to an end—to support causal reasoning. Of particular interest in the context of actively selecting interventions is the setting where we cannot directly intervene on variables whose causal effect on others we are interested in (see Fig. 4), which connects to concepts such as transportability and external validity [5, 40]. ABCI is also flexible in that it easily allows for incorporating available domain knowledge: if we know some aspects of the model a priori (as assumed in conventional causal reasoning) or have access to a large observational sample (from which we can infer the MEC of DAGs), we can encode this in our prior and only optimise over a smaller model class, which should boost efficiency. The principled Bayesian nature of ABCI evidently comes at a significant computational cost: most integrals are intractable, and approximating them with Monte-Carlo sampling is computationally expensive and can introduce bias when resources are limited. On the other hand, in many real-world applications, such as in the context of biological networks, active interventions are possible but only at a significant cost [7, 37]. Particularly in such cases, a careful and computationally-heavy experimental design approach as presented in the present work is warranted and might be easily amortised.

References

[1] Agrawal, R., Squires, C., Yang, K., Shanmugam, K., and Uhler, C. (2019). ABCD-strategy: Budgeted experimental design for targeted causal structure discovery. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3400–3409. PMLR.

[2] Agrawal, R., Uhler, C., and Broderick, T. (2018). Minimal I-MAP MCMC for scalable structure discovery in causal DAG models. In *International Conference on Machine Learning*, pages 89–98. PMLR.

[3] Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics*. Princeton University Press.

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

[5] Bareinboim, E. and Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352.

[6] Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, pages 273–304.

[7] Cho, H., Berger, B., and Peng, J. (2016). Reconstructing causal biological networks through active learning. *PloS one*, 11(3):e0150611.

[8] Cundy, C., Grover, A., and Ermon, S. (2021). BCD nets: Scalable variational approaches for Bayesian causal discovery. *Advances in Neural Information Processing Systems*, 34.

[9] de Jongh, M. and Druzdzel, M. J. (2009). A comparison of structural distance measures for causal bayesian network models. *Recent Advances in Intelligent Information Systems, Challenging Problems of Science, Computer Science series*, pages 443–456.

[10] Eberhardt, F. (2008). Almost optimal intervention sets for causal discovery. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 161–168. AUAI Press.

[11] Eberhardt, F., Glymour, C., and Scheines, R. (2006). N-1 experiments suffice to determine the causal relations among n variables. In *Innovations in machine learning*, pages 97–112. Springer.

[12] Eberhardt, F. and Scheines, R. (2007). Interventions and causal inference. *Philosophy of Science*, 74(5):981–995.

[13] Friedman, N. and Koller, D. (2003). Being Bayesian about network structure. a Bayesian approach to structure discovery in Bayesian networks. *Machine learning*, 50(1):95–125.

[14] Friedman, N. and Nachman, I. (2000). Gaussian process networks. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pages 211–219. Morgan Kaufmann Publishers Inc.

[15] Gamella, J. L. and Heinze-Deml, C. (2020). Active invariant causal prediction: Experiment selection through stability. *Advances in Neural Information Processing Systems*, 33:15464–15475.

[16] George Casella, R. L. B. (2002). *Statistical Inference*, volume 2. Duxbury.

[17] Ghassami, A., Salehkaleybar, S., Kiyavash, N., and Bareinboim, E. (2017). Budgeted experiment design for causal structure learning. *arXiv preprint arXiv:1709.03625*.

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

[19] Hauser, A. and Bühlmann, P. (2014). Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939.

[20] He, Y.-B. and Geng, Z. (2008). Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9(Nov):2523–2547.

[21] Heckerman, D. (1995). A Bayesian approach to learning causal networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 285–295. Morgan Kaufmann Publishers Inc.

[22] Heckman, J. J. (1992). Policy evaluation. *Evaluating welfare and training programs*, page 201.

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

[24] Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

[25] Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696.

[26] Hyttinen, A., Eberhardt, F., and Hoyer, P. O. (2013). Experiment selection for causal discovery. *The Journal of Machine Learning Research*, 14(1):3041–3071.

[27] Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[28] Lindley, D. V. et al. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005.

[29] Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose Bayesian inference algorithm. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

[30] Lorch, L., Rothfuss, J., Schölkopf, B., and Krause, A. (2021). DiBS: Differentiable Bayesian Structure Learning. *Advances in Neural Information Processing Systems*.

[31] Mockus, J. (1975). On Bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference*, pages 400–404. Springer.

[32] Mockus, J. (2012). *Bayesian Approach to Global Optimization: Theory and Applications*, volume 37. Springer Science & Business Media.

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

[34] Morgan, S. L. and Winship, C. (2014). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press.

[35] Murphy, K. P. (2001). Active learning of causal Bayes net structure.

[36] Murphy, K. P. (2007). Conjugate Bayesian analysis of the gaussian distribution. Technical report, University of British Columbia.

[37] Ness, R. O., Sachs, K., Mallick, P., and Vitek, O. (2017). A Bayesian active learning experimental design for inferring signaling networks. In *International Conference on Research in Computational Molecular Biology*, pages 134–156. Springer.

[38] Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

[39] Pearl, J. (2009). *Causality*. Cambridge University Press, 2nd edition.

[40] Pearl, J. and Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595.

[41] Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012.

[42] Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference - Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, MA, USA.

[43] Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053.

[44] Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512.

[45] Robinson, R. W. (1973). Counting labeled acyclic digraphs. *New Directions in the Theory of Graphs*, pages 239–273.

[46] Rubenstein, P. K., Tolstikhin, I., Hennig, P., and Schölkopf, B. (2017). Probabilistic active learning of functions in structural causal models. *arXiv preprint arXiv:1706.10234*.

[47] Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

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

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

[50] Shpitser, I. and Pearl, J. (2009). Effects of treatment on the treated: Identification and generalization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2009*, pages 514–521. AUAI Press.

[51] Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959.

[52] Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*. MIT press, 2nd edition.

[53] Sussex, S., Uhler, C., and Krause, A. (2021). Near-optimal multi-perturbation experimental design for causal structure learning. *Advances in Neural Information Processing Systems*, 34.

[54] Tigas, P., Annadani, Y., Jesson, A., Schölkopf, B., Gal, Y., and Bauer, S. (2022). Interventions, where and how? experimental design for causal models at scale. *arXiv preprint arXiv:2203.02016*.

[55] Tong, S. and Koller, D. (2001). Active learning for structure in Bayesian networks. In *International Joint Conference on Artificial Intelligence*, volume 17, pages 863–869.

[56] von Kügelgen, J., Rubenstein, P. K., Schölkopf, B., and Weller, A. (2019). Optimal experimental design via Bayesian optimization: active causal structure learning for Gaussian process networks. In *NeurIPS 2019 Workshop "Do the right thing": machine learning and causal inference for improved decision making*. *arXiv:1910.03962*.

[57] Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian Processes for Machine Learning*, volume 2. MIT Press Cambridge, MA.

[58] Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215.

[59] Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 647–655. AUAI Press.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Summarizing the abstract/introduction we claim (i) to introduce *a principled fully-Bayesian active learning framework for integrated causal discovery and reasoning* and to (ii) show the practicality of our approach through simulations. We lay out the former concisely in § 3 and § 4. We report the empirical evaluation in § 5.
   (b) Did you describe the limitations of your work? [Yes] See discussion in § 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We provide a short discussion in the Appendix.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We give a concise and rigorous treatment when formulating the general framework in § 3, as well as our approach and model specifics in § 4.
   (b) Did you include complete proofs of all theoretical results? [Yes] We provide the derivation of our utility functions in Appx. C.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Python code and instructions are provided in the supplement as source_code.zip
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We give a minimal set of details in § 5 and provide full information about our experiments in Appx. D.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figs. 3 and 4

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We give a brief summary in Appendix D.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We do not use external models or data. We use a set of Python packages that we list in Appendix D.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include our code base in the supplementary material and will make it publicly available via Github upon acceptance.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]