Bioinformatics, 38, 2022, i350–i358
https://doi.org/10.1093/bioinformatics/btac251
ISCB/ISMB 2022

Do-calculus enables estimation of causal effects in partially observed biomolecular pathways

Sara Mohammad-Taheri[1,*], Jeremy Zucker[2], Charles Tapley Hoyt[3], Karen Sachs[4,5], Vartika Tewari[1], Robert Ness[6] and Olga Vitek[1,*]

[1]Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, USA, [2]Computational Biology, Pacific Northwest National Laboratory, Richland, Washington, DC 99354, USA, [3]Laboratory of Systems Pharmacology, Harvard Medical School, Boston, MA 02115, USA, [4]Next Generation Analytics, Palo Alto, CA 94301, USA, [5]Answer ALS Consortium, LA, CA 70184, USA and [6]Microsoft Research, Redmond, WA 98052, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Estimating causal queries, such as changes in protein abundance in response to a perturbation, is a fundamental task in the analysis of biomolecular pathways. The estimation requires experimental measurements on the pathway components. However, in practice many pathway components are left unobserved (latent) because they are either unknown, or difficult to measure. Latent variable models (LVMs) are well-suited for such estimation. Unfortunately, LVM-based estimation of causal queries can be inaccurate when parameters of the latent variables are not uniquely identified, or when the number of latent variables is misspecified. This has limited the use of LVMs for causal inference in biomolecular pathways.

Results: In this article, we propose a general and practical approach for LVM-based estimation of causal queries. We prove that, despite the challenges above, LVM-based estimators of causal queries are accurate if the queries are identifiable according to Pearl's do-calculus and describe an algorithm for its estimation. We illustrate the breadth and the practical utility of this approach for estimating causal queries in four synthetic and two experimental case studies, where structures of biomolecular pathways challenge the existing methods for causal query estimation.

Availability and implementation: The code and the data documenting all the case studies are available at https://github.com/srtaheri/LVMwithDoCalculus.

Contact: mohammadtaheri.s@northeastern.edu or o.vitek@northeastern.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Biomolecular pathways are governed by complex patterns of controls such as signaling, gene regulation and metabolic reactions. Biomolecular pathways are often represented as graphs, where nodes are signaling proteins, genes, transcripts or metabolites, and directed edges are causal regulatory relationships. The graph-based representations are useful for simulating experimental perturbations, and answering, *in silico*, causal queries of the form 'When we perturb X, what is the effect on its descendent Y?'. However, the estimation of causal queries requires more than a qualitative topology of the graph. It also requires experimental measurements on the nodes of the graph, in order to quantitatively characterize the causal relationships (Pearl, 2009).

Unfortunately, no measurement modality can currently capture all the molecular components of a pathway. The incomplete data arise in at least two general, ubiquitous scenarios. The first occurs when components of a biomolecular pathway are not fully known. For example, there may be empirical evidence for the regulation of an enzyme, but the identity of the molecule or protein that regulates the enzyme may be unknown (Cannon *et al.*, 2021). The second scenario occurs when, due to limitations of the measurement techniques, some pathway components are unobserved. For example, antibodies for a protein may not be available. Alternatively, while RNA abundances may be characterized, levels of the corresponding protein or the state of its post-translational modifications may be unknown (McNaughton *et al.*, 2021).

Latent variable models (LVMs) are particularly useful for representing biological pathways with partially known topology or missing measurements of pathway components (Durbin *et al.*, 1998; Ernst *et al.*, 2007; Kondofersky *et al.*, 2015; Shojaie and Michailidis, 2009; St John *et al.*, 2019). LVMs are probabilistic models of a joint distribution on a set of observed and unobserved variables. A broad class of LVMs has a directed acyclic graphical (DAG) structure. LVM-based estimation of a causal query proceeds by removing edges in the DAG that point to the target of intervention. Trained on observational data once, an LVM can estimate multiple causal queries corresponding to multiple mutilated versions of the original DAG.

There currently exists some controversy as to whether LVM-based estimation of causal queries is accurate. One argument against this approach is that the parameters of the LVM may not be uniquely identified from the observed data (Shpitser *et al.*, 2014). Another argument is that the number of latent variables may be misspecified (Shpitser *et al.*, 2012). As a result, currently accepted approaches to

LVM-based causal query estimation are limited to LVMs with specialized structural properties, such as the existence of proxy variables (Kuroki and Pearl, 2014; Louizos *et al.*, 2017), or the presence of multiple causes (Wang and Blei, 2019). The latter approach, although scalable to a large number of variables, is not correct in general and requires strong parametric assumptions (D'Amour, 2019). Since biomolecular pathways have complex and diverse topology, are frequently large-scale, and have many (possibly unknown) latent variables, the controversy has so far limited the use of LVM for causal inference in this context.

In this article, we argue that LVM-based estimators of causal queries are in fact accurate when the queries are identifiable according to Pearls do-calculus, and describe a simple and practical algorithm for its estimation. We show that the estimated probability distribution associated with the causal query converges to the true distribution, and that the estimate of its expected value is consistent. This holds even when the parameters of the model are not uniquely identified, or when the true number of the latent variables is unknown.

We showcase the breadth of applicability, and the practical utility of LVM-based estimation of identifiable causal queries in four synthetic and two experimental case studies of biomolecular pathways. The case studies demonstrate the accuracy of the estimated causal effects, even when some pathway components are not experimentally quantified or unknown, and when the parametric assumptions only approximately represent the true data generating process. The case studies also demonstrate that the proposed approach expands the use of causal inference to pathways where the existing alternative approaches do not apply, and enables the estimation of multiple causal queries from a single trained model.

2 Background

2.1 Notation

Let $\mathbf{V} = \{V_1, \ldots, V_J\}$ be a set of observed random variables, and $\mathbf{U} = \{U_1, \ldots, U_L\}$ be a set of latent variables. Let v_i be an instance of V_i, and $\mathbf{v} = \{v_1, \ldots, v_J\}$ an instance of $\mathbf{V}$. Let $P(v_1, \ldots, v_J)$ be the joint probability distribution of the event $\mathbf{V} = \mathbf{v}$, and let $P(V_i = v_i | V_j = v_j)$ be the conditional probability distribution for the event $V_i = v_i$ given $V_j = v_j$. Denote $P(\mathbf{U})$ the prior distribution over all the latent variables, and $P(\mathbf{U}|\{\mathbf{v}_i\}_{i=1}^N)$ the posterior distribution over all latent variables $\mathbf{U}$ given N observations of $\mathbf{V}$. In this article, we simplify the notation for the marginalized joint distribution $\int_{\mathbf{u}} P(\mathbf{U}, \mathbf{V}) d\mathbf{u}$ as $P(\mathbf{V})$. Let G be a DAG with nodes $\mathbf{V} \cup \mathbf{U}$, where $\mathrm{Pa}(V_j)$ denotes the parents of a node V_j in G. The joint distribution between variables $\mathbf{V} \cup \mathbf{U}$ in DAG G is formulated as, $P(\mathbf{U}, \mathbf{V}) = \prod_{j=1}^J P(V_j | Pa(V_j)) \prod_{l=1}^L P(U_l | Pa(U_l))$.

2.2 Latent variable models

A *latent variable model* (LVM) is a probability distribution over two sets of variables $\mathbf{V}$, $\mathbf{U}$, where $\mathbf{V}$ are observed at the learning time, and $\mathbf{U}$ are not observed. LVMs are generative, in the sense that they allow us to sample from the joint distribution of all the variables.

A *causal LVM* $\mathcal{M}$ is an LVM with DAG structure where $\mathrm{Pa}(V_i)$ are interpreted as *direct causes* of V_i. In Bayesian framework, parameter vector θ of the causal LVM are assigned prior probability distributions, and are absorbed into the set of latent variables denoted by $\theta \subseteq \mathbf{U}$.

Given a causal LVM with a DAG G, observed variables $\mathbf{V}$, and latent variables $\mathbf{U}$ (Evans, 2016) compactly represents LVMs with many latent variables by an LVM with a single latent variable between each pair of observed variables, according to the following rules:

1. Remove latent variables with no children.
2. Remove a latent variable U with observable parents by connecting all the parents of U to its children.
3. If U, W are latent variables with $children(W) \subseteq children(U)$, then remove W.

Figure 1a illustrates a causal LVM with many latent variables, and Figure 1b, a causal LVM obtained from (a) by applying the simplification. Figure 1c is an *acyclic directed mixed graph (ADMG)* (Richardson *et al.*, 2017) representing both Figure 1a and b. It shows the existence of latent variables between X_1 and X_2 by a dashed bi-directed edge.

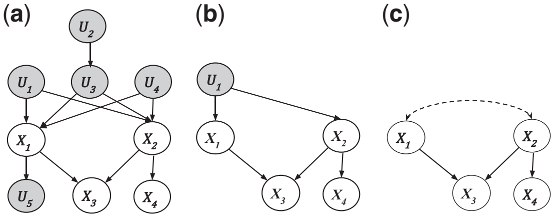


Fig. 1. (**a**) An LVM with four observed (white) and five latent (dark grey) variables. (**b**) A different LVM with 1 latent variable. (**c**) An ADMG representing both (a) and (b)

Inference algorithms (Bishop, 2006) sample from the posterior distribution $P(\mathbf{U}|\{\mathbf{v}_i\}_{i=1}^N)$ of latent variables in the ADMG, including the parameters θ, given N observations of $\mathbf{V}$. In particular, exact algorithms such as Hamiltonian Monte Carlo (HMC) (Girolami and Calderhead, 2011) guarantee asymptotically exact samples but are computationally expensive (Robert and Casella, 2013). Approximate probabilistic inference algorithms such as variational inference (Blei *et al.*, 2017) trade off accuracy for speed by searching with gradient descent a parameterized family of functions that approximate the target distribution. A trained causal LVM $\hat{\mathcal{M}}$ is an LVM where posterior distributions of the parameters are learned with an inference algorithm. Many packages such as PyStan (Van Hoey *et al.*, 2013) or pyro (Bingham *et al.*, 2019) in Python, or RStan (2020) in R take as input an LVM and output a trained LVM.

2.3 Causal query identification

Frequently, we are interested in an *intervention* on a set of target variables $\mathbf{X} \subseteq \mathbf{V}$ which fixes a set of variables $\mathbf{X}$ to constant values $\mathbf{x}'$ (denoted $do(\mathbf{X} = \mathbf{x}')$, shortened to $do(\mathbf{x}')$), and makes it independent of its causes (Eberhardt and Scheines, 2007; Spirtes *et al.*, 2000). *Graph mutilation* in a causal LVM simulates an intervention. It severs the edges incoming to the target nodes and fixes each node $X \in \mathbf{X}$ to its intervention value $x' \in \mathbf{x}'$ (Koller and Friedman, 2009), producing a graph that we denote $G_{\overline{\mathbf{X}}}$. Denote $P_{G_{\overline{\mathbf{X}}}}(\mathbf{v})$ the probability distribution encoded by $G_{\overline{\mathbf{X}}}$. Denote $\mathcal{M}_{\overline{\mathbf{X}}}$ the causal LVM with structure $G_{\overline{\mathbf{X}}}$ (the subscript $\overline{\mathbf{X}}$ in this notation distinguishes the intervened model from the original model). Denote $P_{\mathcal{M}_{\overline{\mathbf{X}}}}(\mathbf{v})$ the probability distribution, and $E_{\mathcal{M}_{\overline{\mathbf{X}}}}[\mathbf{v}]$ the expected value of the variables $\mathbf{v}$ in the intervened model $\mathcal{M}_{\overline{\mathbf{X}}}$.

A *causal query* $Q_{\mathbf{X}}$ with respect to a causal LVM $\mathcal{M}$ is a probabilistic query that conditions a set of outcomes $\mathbf{Y} \subseteq \mathbf{V}\ \mathbf{X}$ on a set of interventions, such as $Q_{\mathbf{X}} = P_{\mathcal{M}_{\overline{\mathbf{X}}}}(\mathbf{Y}|do(\mathbf{x}'))$ or $Q_{\mathbf{X}} = E_{\mathcal{M}_{\overline{\mathbf{X}}}}[\mathbf{Y}|do(\mathbf{x}')]$. To denote the distribution of the outcome variable obtained from a mutilated model that was trained on pre-interventional data, we use counterfactual subscript notation $\mathbf{Y}_{do(\mathbf{x}')} \sim P(\mathbf{Y}_{do(\mathbf{x}')}|\{x_i, y_i\}_{i=1}^N)$.

A causal query $Q_{\mathbf{X}}$ is *identifiable* with respect to $P(\mathbf{V})$ and an ADMG A, if all LVMs that project onto A and agree on $P(\mathbf{v})$ also agree on the value of $Q_{\mathbf{X}}$ (Shpitser and Pearl, 2008). A causal query is identifiable if it satisfies the back-door or the front-door criteria (Pearl, 2009). The back-door and the front-door criteria rely on the following concepts of graphical modeling. In a DAG G, there is a *path* between V_i and V_j, if there is a sequence of edges connecting V_i to V_j. A variable is a *collider* when both edges adjacent to the variable on the path point into it. A path is *blocked* if we observe the value of a non-collider on that path or we do not observe the value of a collider.

The *back-door criterion* (Pearl, 2009) holds for $X, Y \in \mathbf{V}$ in ADMG A if there is no path from X to Y consisting of bi-directed edges, and there exists a set $\mathbf{Z} \subseteq \mathbf{V}\ \{X, Y\}$ such that no node is a descendant of X, and $\mathbf{Z}$ blocks every path between X and Y that contains an arrow into X (Pearl, 2009). If a set of variables $\mathbf{Z}$ satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is identifiable and is given by $P(Y|do(x')) = \sum_z P(Y|x', z)P(z)$. The

Algorithm 1 Estimation of an identifiable causal query

Input $\hat{\mathcal{M}}$, a causal LVM trained on observational data with an exact inference algorithm
$\mathbf{x}' \subseteq \mathbf{v}$, target values of the intervention
$\mathbf{Y} \subseteq \mathbf{V}$, effects of the intervention
$Q_{\mathbf{X}} = P_{\mathcal{M}_{\overline{\mathbf{X}}}}(\mathbf{Y}|do(\mathbf{x}'))$ or $E_{\mathcal{M}_{\overline{\mathbf{X}}}}[\mathbf{Y}|do(\mathbf{x}')]$ query
Param S, # of samples from the posterior distribution
L, # of samples for each variable
Output $\hat{P}_{\hat{\mathcal{M}}_{\overline{\mathbf{X}}}}(\mathbf{Y}|do(\mathbf{x}'))$ or $\hat{E}_{\hat{\mathcal{M}}_{\overline{\mathbf{X}}}}[\mathbf{Y}|do(\mathbf{x}')]$

```
Check identifiability of $Q_{\mathbf{X}}$
if $Q_{\mathbf{X}}$ is not identifiable then
    break
else
    Set $\mathbf{X} = \mathbf{x}'$
    Create $\hat{\mathcal{M}}_{\overline{\mathbf{X}}}$, the mutilated model
    for $s$ in 1: $S$ do
        Sample $\theta_s \sim P_{\hat{\mathcal{M}}_{\overline{\mathbf{X}}}}(\theta|\{\mathbf{v}_i\}_{i=1}^{N})$
        for $W$ in topological-sort($\{\mathbf{U} \cup \mathbf{V}\}$) do
            Sample $w_s \sim P_{\hat{\mathcal{M}}_{\overline{\mathbf{X}}}}(W|Pa(W); \theta_s)$ $L$ times
        end for
        Collect $\mathbf{y}_s \subseteq \mathbf{w}_s$
    end for
    Return density($\{\mathbf{y}_s\}_{s=1}^{S}$) or $\frac{1}{S}\sum_{s=1}^{S} \mathbf{y}_s$
end if
```

front-door criterion (Pearl, 2009) holds when there is an unobserved confounder, but there exists a mediator between cause and effect that is shielded from confounding (Pearl, 1993, 1995, 2009). If a set of variables $\mathbf{Z}$ satisfies the front-door criterion relative to (X, Y), and if $P(x, z) > 0$, then the causal effect of X on Y is identifiable and is given by the formula, $P(Y|do(x')) = \sum_z P(z|x') \sum_x P(Y|x, z)P(x)$. For example, neither the back-door nor the front-door criterion hold in Figure 2a, but the front-door criterion holds in Figure 2b. The back-door and front-door criteria are sufficient but not necessary for causal identifiability.

The *do-calculus* is comprised of three rules for symbolically manipulating interventional and observational joint distributions. Let $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ and $\mathbf{W}$ be disjoint sets of variables in the joint distribution entailed by ADMG G. Let $G_{\overline{\mathbf{X}}}$ denote the graph produced by mutilating G such that all incoming edges to $\mathbf{X}$ are removed. Similarly, $G_{\underline{\mathbf{Z}}}$ is the graph created when G is mutilated by removing all outgoing edges from $\mathbf{Z}$. The three rules of do-calculus are as follows:

1: $P(\mathbf{Y}|do(\mathbf{x}), \mathbf{z}, \mathbf{w}) = P(\mathbf{Y}|do(\mathbf{x}), \mathbf{w})$ if $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}, \mathbf{W})G_{\overline{\mathbf{X}}}$
2: $P(\mathbf{Y}|do(\mathbf{x}, \mathbf{z}), \mathbf{w}) = P(\mathbf{Y}|do(\mathbf{x}), \mathbf{z}, \mathbf{w})$ if $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}, \mathbf{W})G_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}$
3: $P(\mathbf{Y}|do(\mathbf{x}, \mathbf{z}), \mathbf{w}) = P(\mathbf{Y}|do(\mathbf{x}), \mathbf{w})$ if $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}, \mathbf{W})G_{\overline{\mathbf{X}}, \overline{\mathbf{Z}(\mathbf{W})}}$

Here, $\mathbf{Z}(\mathbf{W})$ is the subset of nodes in $\mathbf{Z}$ that are not ancestors of any node in $\mathbf{W}$. The do-calculus rules are complete (Huang and Valtorta, 2006; Shpitser and Pearl, 2006), meaning if a causal query is identifiable, then it can be derived using these three rules.

A causal query containing a $do()$ operator is identifiable in a given ADMG if the do-calculus transforms it into an equivalent *do*-free estimand. The do-calculus estimands are non-parametric, in the sense that they do not impose constraints on $P(\mathbf{x})$. Any causal query in an ADMG identifiable by the do-calculus is also identifiable in every causal LVM that projects onto that ADMG (Richardson *et al.*, 2017).

Several sound and complete algorithms take as input an ADMG and a causal query and determine whether the query is identifiable according to the do-calculus (Richardson *et al.*, 2017; Shpitser and Pearl, 2008). These algorithms have polynomial time complexity (Galles and Pearl, 2013).

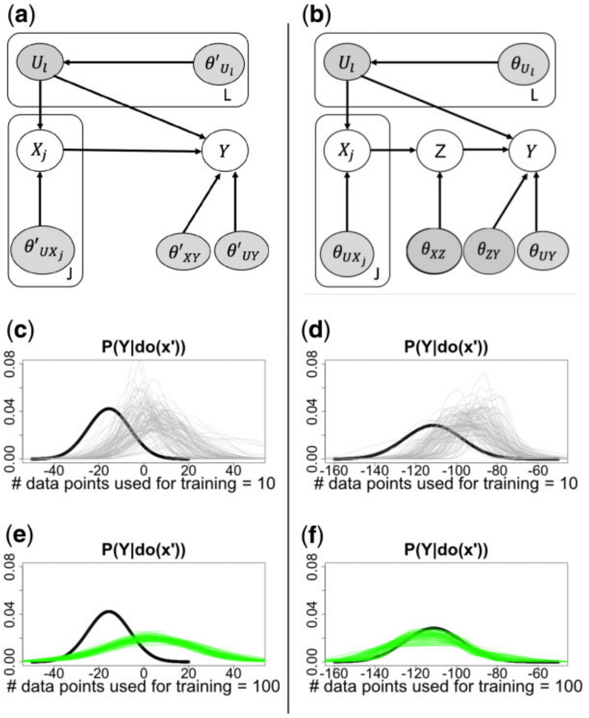


Fig. 2. The estimates of a non-identifiable causal query $P_{\mathcal{M}_{\overline{X}}}(Y|do(\mathbf{x}'))$ fail to converge to the true distribution as number of data points used to train the LVM increases (left column). The estimates of an identifiable causal query converges to the true distribution (right column). (**a**) An LVM where $P_{\mathcal{M}_{\overline{X}}}(Y|do(\mathbf{x}'))$ is not non-parametrically identified. Boxes indicate sets of variables with the same structure. Circular white/gray nodes are observed/latent variables. θ' are model parameters. Each parameter such as θ'_U has a prior distribution, e.g., $\theta'_U \sim P(q_{\theta'_U})$, where $q_{\theta'_U}$ is a hyperparameter. (**b**) As in (a), but in this case $P_{\mathcal{M}_{\overline{X}}}(Y|do(\mathbf{x}'))$ is non-parametrically identified. (**c, e**) relate to (a). Thick curve estimates the true distribution $P_{\mathcal{M}_{\overline{X}}}(Y|do(\mathbf{x}'); \theta)$, with θ used to generate interventional data. After training the LVM on $N = 10$ 100 observational data points, each thin curve estimates $P_{\hat{\mathcal{M}}_{\overline{X}}}(Y_{do(\mathbf{x}')}; \{x_i, y_i\}_{i=1}^{N}, \theta)$ for each sampled θ. The curves do not approach the true distribution as number of data points increases. (**d, f**) relate to (b). The curves converge to the true distribution as the number of data points increases

2.4 Causal query estimation

For queries of a form of $P_{\mathcal{M}_{\overline{X}}}(\mathbf{Y}|do(\mathbf{x}'))$, a desirable property of the estimator is the convergence of the estimated probability distribution to the true probability distribution. For queries of a form of $E_{\mathcal{M}_{\overline{X}}}[\mathbf{Y}|do(\mathbf{x}')]$, a desirable property of the estimator is consistency. An estimator of $E_{\mathcal{M}_{\overline{X}}}[\mathbf{Y}|do(\mathbf{x}')]$ is *consistent* if, as the number of data points used to estimate the query tends to infinity, the sequence of the estimates converges in probability to its expected value.

Several non-LVM approaches for estimating causal queries with these desirable properties exist such as semi-parametric primal IPW (PIPW), dual IPW (DIPW), nested IPW and augmented nested IPW (Bhattacharya *et al.*, 2020). They are all implemented and well-documented in Ananke (Bhattacharya *et al.*, 2020). Unfortunately, these approaches derive a separate statistical estimand for each causal query anew (Pearl, 2019). In addition, they are limited to causal queries with one cause and one effect, and the cause must be binary-valued. This has limited, the scope of their applicability in systems biology where one is often interested in the simultaneous effect of multiple cause on one or multiple effects and the variables are not always discrete. Other approaches such as (WERM-ID) (Jung *et al.*, 2020) and double/debiased machine learning (DML) (Jung *et al.*, 2021) proposed estimators for any identifiable query but are inadequate in large data regimes where it is computationally expensive to train a new estimator for each query of interest. The implementations for these approaches are unavailable for the public.

In this article, we advocate for the explicit use of LVMs for causal query estimation in presence of latent variables when causal queries contain multiple-causes, non-discrete cause(s) or multiple effects, as these are common in biology. We demonstrate that if the graph topology of an LVM correctly reflects the true underlying causal structure of the observed variables, and if the causal query of interest is identifiable according to Pearl's do-calculus, then LVM-based estimators have the desired properties.

3 Methods

3.1 Contribution of this work

In this article, we propose a simple and practical algorithm (Algorithm 1) for LVM-based causal query estimation. The algorithm takes as inputs a causal query of interest in the form of the distribution over the effect Y given an intervention on the cause X, i.e., $\mathbf{Y}_{do(\mathbf{x}')} \sim P_{\mathcal{M}_{\overline{X}}}(\mathbf{Y}|do(\mathbf{x}'))$, or in the form of the expected value of this distribution, i.e., $E_{\mathcal{M}_{\overline{X}}}[\mathbf{Y}|do(\mathbf{x}')]$, target values of the intervention, effects of the intervention, and an LVM with known DAG or ADMG structure that is trained on observational data. The output of the algorithm is the estimate of the causal query of interest, i.e., $\hat{P}_{\hat{\mathcal{M}}_{\overline{X}}}(\mathbf{Y}|do(\mathbf{x}'))$ or $\hat{E}_{\hat{\mathcal{M}}_{\overline{X}}}[\mathbf{Y}|do(\mathbf{x}')]$.

The algorithm first determines whether the causal query of interest is identifiable according to Pearl's do-calculus (line 1). If the query is identifiable, Algorithm 1 proceeds with its estimation. We take a Bayesian viewpoint (Lattimore and Rohde, 2019a, b), and follow the abduction, action, prediction paradigm (Pearl, 2009). Abduction estimates the posterior distribution over the latent variables (including the model parameters) given the training data. A trained LVM, including these posterior distributions, is an input to Algorithm 1. Action fixes the values of the intervened variables (line 5) and breaks the relationship of the intervened variables to their parents (line 6). Prediction samples the parameters from their posterior distributions (line 8) and then samples from each variable given its parents (line 10) until we are ready to estimate the causal query (line 14). Thus, the estimator can be thought of as a posterior predictive statistic over the marginal of the parameters.

The algorithm takes as input a trained LVM. In particular, it can take a trained LVM with continuous distributions, and multiple causes and effects, where non-parametric or current parametric approaches are limited. While training an LVM is NP-complete (and in practice depends on the specific LVM and on the choice of inference algorithm), it amortizes most of the computational work into this single training step. Given a single trained model, it can estimate an arbitrary number of queries.

3.2 Convergence and consistency of the estimator in Algorithm 1 in correctly specified LVMs

Motivating examples. We illustrate the practical application of Algorithm 1 in the special case of the LVM in Figure 2a. It represents a situation where, e.g., a protein product of gene X affects gene Y, while both are under regulation of the same transcription factor(s) and/or enhancer(s). The causal query $P_{\mathcal{M}_{\overline{X}}}(Y|do(x'))$ is not identifiable, and we show empirically that its LVM-based estimator is biased (Fig. 2c and e).

Extending the causal LVM with a mediator Z in Figure 2b makes the query identifiable according to the front-door criterion. This pattern occurs frequently in transcriptional cascades which involve multiple steps, or signaling pathways in which Y is not a direct substrate of X. We show empirically that the estimate of $P_{\mathcal{M}_{\overline{X}}}(Y|do(x'))$ converges to the true distribution (Fig. 2d and f).

Empirical example 1: Figure 2a Assume a model $\mathcal{M}$: $U := \theta'_U; X := U\theta'_{UX} + \theta'_X; Y := X\theta'_{XY} + U\theta'_{UY} + \theta'_Y$ where $\theta'_X \sim N(\mu'_X, \sigma'_X), \theta'_Y \sim N(\mu'_Y, \sigma'_Y), \theta'_U \sim N(\mu'_U, \sigma'_U)$ and a non-identifiable causal query of $P_{\mathcal{M}_{\overline{X}}}(Y|do(x'))$. We generated observational data with $N = 10$, 100 samples from the likelihood with a randomly chosen vector of true values of θ. The true $P_{\mathcal{M}_{\overline{X}}}(Y|do(x'); \theta)$ was estimated with Algorithm 1, where line 8 as substituted by the true values of θ (black curves in Fig. 2c and e).

To learn a model $\hat{\mathcal{M}}$ from this training data, we assumed a Gaussian prior on the parameters: $\mu'_U, \mu'_X, \mu'_Y, \sigma'_U, \sigma'_X, \sigma'_Y, \theta'_{UX} \sim N(0,1)$ and $\theta'_{XY}, \theta'_{UY} \sim N(0, 10)$, and trained the model with HMC. Thin lines in Figure 2c and e estimate $P_{\hat{\mathcal{M}}_{x'}}(Y|do(x'), \{x_i, y_i\}_{i=1}^N, \theta)$ for each sampled θ (line 10). As N increases, the distributions became less diverse, but did not approach the ground truth.

Empirical example 2: Figure 2b Expanding the previous example with a mediator Z, we assume a model $U := \theta_U$, $X := U\theta_{UX} + \theta_X$, $Z := X\theta_{XZ} + \theta_Z$, $Y := Z\theta_{ZY} + U\theta_{UY} + \theta_Y$ where, $\theta_U \sim N(\mu_U, \sigma_U)$, $\theta_X \sim N(\mu_X, \sigma_X), \theta_Y \sim N(\mu_Y, \sigma_Y), \theta_Z \sim N(\mu_Z, \sigma_Z)$.

With this expansion, the causal query $P_{\mathcal{M}_{\overline{X}}}(Y|do(x'))$ becomes identifiable. Repeating the same analysis, Figure 2d and f show that, as N increased, the distributions converged to the ground truth. The analytical proof of this empirical result for multivariate U and X can be found in Supplementary Materials.

The following Lemma 1 proves the empirical results for any arbitrary distribution.

Lemma 1 *Consider the LVM in Figure 2b with a DAG G.* $\mathbf{X}$, Z, *and* Y *are observed and* $\mathbf{U}$ *are latent. The front-door adjustment estimand of the query* $P(Y|do(\mathbf{x}'))$ *is equivalent to the estimand of that query in the mutilated LVM.*

Proof. Consider a mutilated version of G, $G_{\overline{\mathbf{X}}}$, where all the incoming edges to $\mathbf{X}$ are removed. A causal query $P(Y|do(\mathbf{x}'))$ transforms $P(.)$ into a distribution $P_{\overline{\mathbf{X}}}(.)$, and $P(Y|do(\mathbf{x}')) = P_{\overline{\mathbf{X}}}(Y|\mathbf{x}')$. Hence,

$$\begin{aligned} P(Y|do(\mathbf{x}')) = P_{\overline{\mathbf{X}}}(Y|\mathbf{x}') &= \int_{\mathbf{u},z} P_{\overline{\mathbf{X}}}(Y, \mathbf{u}, z|\mathbf{x}')d\mathbf{u}dz \\ &= \int_z \left(\int_{\mathbf{u}} P_{\overline{\mathbf{X}}}(Y|\mathbf{u}, z, \mathbf{x}') P_{\overline{\mathbf{X}}}(\mathbf{u}|z, \mathbf{x}')d\mathbf{u} \right) P_{\overline{\mathbf{X}}}(z|\mathbf{x}')dz \\ &= \int_z P_{\overline{\mathbf{X}}}(Y|z) P_{\overline{\mathbf{X}}}(z|\mathbf{x}')dz \\ &= \int_z P(Y|do(z))P(z|\mathbf{x}')dz \end{aligned} \tag{1}$$

$$= \int_z \left(\int_{\mathbf{x}} P(Y|z, \mathbf{x}) P(\mathbf{x}) d\mathbf{x} \right) P(z|\mathbf{x}')dz \tag{2}$$

Equation (1) holds because in $G_{\overline{\mathbf{X}}}$, Y is independent from $\mathbf{X}$ given Z. Since $P_{\overline{\mathbf{X}}}(z|\mathbf{x}')$ is unaffected by the mutilation of G, $P_{\overline{\mathbf{X}}}(z|\mathbf{x}') = P_G(z|\mathbf{x}')$. Equation (2) follows from the back-door path between Y and Z in G. The expression on the right-hand side of (2) is the estimand for $P(Y|do(\mathbf{x}'))$ derived from the do-calculus front-door adjustment formula. □

The following theorem proves that in general, for any LVM topology, any set of parametric distributions, and any identifiable causal query, Algorithm 1 accurately estimates causal queries in an LVM that correctly reflects the true underlying causal structure.

Theorem 1 *Consider a causal LVM* $\mathcal{M}$, *which includes the true likelihood that generated the observational data. Consider a causal query* $Q_{\mathbf{X}} = P_{\mathcal{M}_{\overline{X}}}(\mathbf{Y}|do(\mathbf{x}'))$ *or* $Q_{\mathbf{X}} = E_{\mathcal{M}_{\overline{X}}}[\mathbf{Y}|do(\mathbf{x}')]$, *identifiable according to the do-calculus with respect to* $\mathcal{M}$. *When estimating the causal query as in Algorithm 1, the estimate* $\hat{P}_{\hat{\mathcal{M}}_{\overline{X}}}(\mathbf{Y}|do(\mathbf{x}'))$ *converges to the true distribution, and the estimator* $\hat{E}_{\hat{\mathcal{M}}_{\overline{X}}}[\mathbf{Y}|do(\mathbf{x}')]$ *is consistent.*

Proof. When the ground truth parameters θ are known, samples from the likelihood $v_s \sim P(V|Pa(V), \theta)$ for all $V \in \mathbf{V}$ converge to the true joint observational distribution $\prod_{V \in \mathbf{V}} P(V|Pa(V), \theta)$ as $N \to \infty$. N is the number of data points.

In practice, parameters of the LVM are trained on observational data. If the parameters are not identifiable during training, their posterior distribution $\theta_r \sim P(\theta|\{\mathbf{v}_i\}_{i=1}^N)$ is not guaranteed to converge to the true value. Nonetheless, samples from the observed variables $v_s \sim P(V|Pa(V), \theta_r)$, $V \in \mathbf{V}$, converge to the same true joint observational distribution $\prod_{V \in \mathbf{V}} P(V|Pa(V), \theta)$. For identifiable causal queries, all parametrizations that agree on the joint observational distribution agree on the queries (Shpitser and Pearl, 2008). Therefore, since under stability conditions exact inference algorithms provide guarantees of asymptotically exact samples, the posterior predictive distribution $P(\mathbf{Y}_{do(\mathbf{x}')}|\{\mathbf{v}_i\}_{i=1}^{N})$ converges to the true distribution, and its expected value $E[\mathbf{Y}_{do(\mathbf{x}')}|\{\mathbf{v}_i\}_{i=1}^{N}]$ is consistent (Gelman *et al.*, 2014; Robert and Casella, 2013). □

3.3 Convergence and consistency of the estimator in Algorithm 1 in presence of miss-specified number of latent variables

The following corollary proves that queries of the form $E_{\mathcal{M}_{\overline{X}}}(\mathbf{Y}|do(\mathbf{x}))$ or $P_{\mathcal{M}_{\overline{X}}}(\mathbf{Y}|do(\mathbf{x}))$ can be accurately estimated even when the true number of latent variables is unknown.

Corollary 1 Consider a causal LVM $\mathcal{M}$, which includes the true likelihood that generated the observational data. Consider a class of LVMs $\mathbb{M}$ that projects on the same ADMG as $\mathcal{M}$. Consider a causal query $Q_{\mathbf{X}} = P_{\mathcal{M}_{\overline{X}}}(\mathbf{Y}|do(\mathbf{x}'))$ or $Q_{\mathbf{X}} = E_{\mathcal{M}_{\overline{X}}}[\mathbf{Y}|do(\mathbf{x}')]$, identifiable according to the do-calculus with respect to $\mathbb{M}$. When estimating the causal query as in Algorithm 1, the estimate $\hat{P}_{\hat{\mathcal{M}}_{\overline{X}}}(\mathbf{Y}|do(\mathbf{x}'))$ converges to the true distribution, and the estimate $\hat{E}_{\hat{\mathcal{M}}_{\overline{X}}}[\mathbf{Y}|do(\mathbf{x}')]$ is consistent.

Proof. Let θ' be the parameters of $\mathcal{M}' \in \mathbb{M}$. Following the same logic as in proof of Theorem, the samples $v'_s \sim P(V|Pa(V), \theta'_r)$, $V \in \mathbf{V}$, converge to the same true joint observational distribution $\prod_{V \in \mathbf{V}} P(V|Pa(V), \theta)$ as for the correctly specified model $\mathcal{M}$. Therefore, the posterior predictive distribution converges to the true distribution, and its expected value is consistent. □

This result is useful in practical applications, as choosing an instance from the right set of LVMs is less challenging than choosing the exactly right LVM. Therefore, given several candidate LVMs projecting on the same ADMG, we can rely on Occam's Razor (Balasubramanian, 1997; Rasmussen and Ghahramani, 2001) and favor the LVM with the simplest DAG structure.

4 Case studies

4.1 Overview

We illustrated the breadth and the practical utility of the proposed LVM-based estimation of identifiable causal queries in four synthetic and two experimental studies of biomolecular pathways, with topologies that challenge the existing methods for causal query estimation. We considered both the LVMs with the correct topology, and the LVMs with the correct topology for the observed variables but with a misspecified number of the latent variables. Posterior distributions of the parameters were inferred with HMC in Stan (2018).

The synthetic case studies illustrated the consistency of the causal queries of the form $Q_{\mathbf{X}} = E_{\mathcal{M}_{\overline{X}}}[\mathbf{Y}|do(\mathbf{X} = \mathbf{x}')]$ (in the following, we omit the subscript $\mathcal{M}_{\overline{X}}$, and state the value of $\mathbf{x}'$). The case studies incorporated a mix of probability distributions and a mix of informative and non-informative priors. Case Studies 1, 2 and 4 simulated observational data from parametric distributions with randomly selected values of parameters θ. Case Study 3 generated data using stochastic differential equations. The interventional datasets were obtained by sampling from the distribution with the true θ and the fixed targets of the interventions. To evaluate the performance of the proposed approach, the true values of $Q_{\mathbf{X}}$ were obtained by averaging 10 000 samples from the interventional datasets.

The experimental case studies illustrated the accuracy of the causal queries of a different form, namely $Q_{\mathbf{X}} = P_{\mathcal{M}_{\overline{X}}}(\mathbf{Y}|do(\mathbf{X} = \mathbf{x}'))$. The experimental data were downloaded from Precision RNA-seq Expression Compendium for Independent Signal Exploration (PRECISE, Sastry *et al.*, 2019). They contained 278 RNA-seq normalized expression profiles of *Escherichia coli* K-12 MG1655 and BW25113 across 154 unique experimental conditions. This manuscript focuses on pathways for which both observational and interventional data were available. To evaluate the performance of the proposed approach, experimentally observed instances from $P(\mathbf{Y}|do(\mathbf{X} = \mathbf{x}'))$ were plotted against the estimated distributions.

Each case study was run on a single standard virtual machine on Google Cloud Platform with 2 vCPUs and 8 GB memory. Several virtual machine instances were used to run the case studies in parallel. The case studies took between 1.5 min and 1.8 h.

4.2 Synthetic Case Study 1: the multi-cause feed-forward transcriptional regulatory network motif

The system in Figure 3a is an example of a common feed-forward network motif in *E.coli* and many other prokaryotes (Alon, 2019). The network was obtained by querying the EcoCyc database (Keseler *et al.*, 2021) to discover which front door motifs with one or more confounders and one or more causes exist in *E.coli*. A total of 1945 different cases were found. For this case study, we randomly selected one case. Despite being ubiquitous, the case study is challenging because it has multiple causes.

Query of interest $Q_{\mathbf{X}} = E[cas2|do(\mathbf{X} = 0)]$, where $\mathbf{X} = \{dsrA, gadX, fis\}$. In this query ,the back-door criterion does not hold but the front-door criterion holds.

Data of the latent variables followed a Normal distribution, and the remaining variables a Bernoulli distribution with logit parameterization.

LVM with correct topology assumed the correct data generation process with non-informative $\mathcal{N}(0, 10)$ priors over all the parameters.

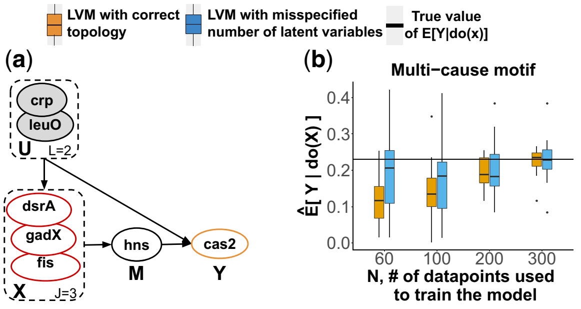
LVM with misspecified number of latent variables wrongly assumed only one latent variable.

4.3 Synthetic Case Study 2: the Napkin motif

The system in Figure 4b is called the second Napkin problem in Pearl and Mackenzie (2018). The network was obtained by querying the EcoCyc database (Keseler *et al.*, 2021) to discover all napkin motifs with two or more confounders in *E.coli*. 911 different cases were found. For this case study we randomly selected one case.

Causal query of interest $Q_{lrp} = E[topA|do(lrp = 1)]$. The system requires a non-trivial application of the do-calculus, because we cannot block the back-door path from *lrp* to *topA* (*hns* is a collider and *gadE* is an ancestor of a collider), and because the front-door criterion does not hold (there is no mediator between *lrp* and *topA*) (Helske *et al.*, 2021; Hughes *et al.*, 1998; Jung *et al.*, 2020; Pearl and Mackenzie, 2018).

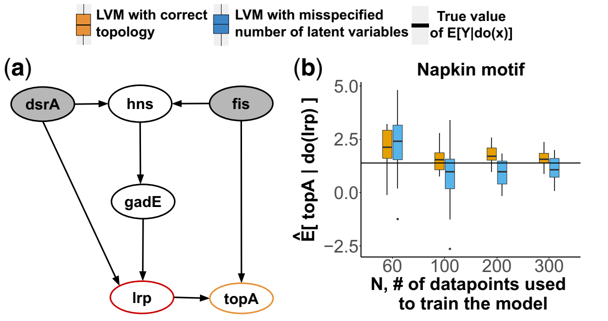
Data of *hns* was modeled with a gamma distribution (representative of expression measurements with a fluorescent reporter). The expression of all the other genes *dsr*, *fis*, *gadE* and *topA* were modeled with Gaussian distributions (representative of measurements or relative expression, such as with RT-PCR). *lrp* followed a Bernoulli distribution with logit parametrization.



Case study 1 : Multi-cause motif

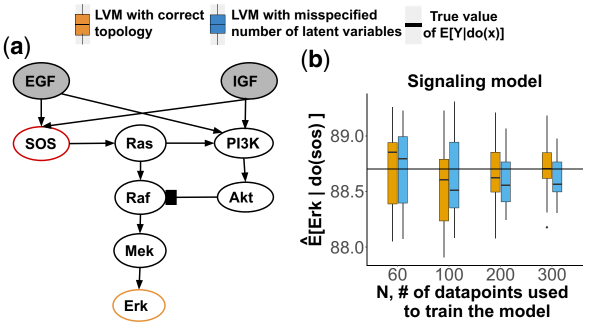
ˆE[cas2|do(dsrA, gadX, fis)]

Fig. 3. Synthetic Case Study 1. Red nodes are targets of the intervention, orange nodes are the effect. gray nodes are latent. (**a**) The multi-cause feed-forward transcriptional regulatory network motif. (**b**) Sampling distribution of $\hat{Q}_{\mathbf{x}} = \hat{E}[cas2|do(dsrA, gadX, fis = 0)]$ over 20 observational datasets



Case study 2 : Napkin motif $\hat{E}[topA|do(lrp=0)]$

Fig. 4. Synthetic Case Study 2. DAG labeled as in Figure 3. (**a**) The Napkin network motif. (**b**) Sampling distribution of $\hat{Q}_x = \hat{E}[topA|do(lrp = 0)]$ over 20 observational datasets



Case study 3 : Signaling model $\hat{E}[Erk|do(SOS{=}70)]$

Fig. 5. Synthetic Case Study 3. DAG labeled as in Figure 3. (**a**) The signaling model. Nodes are proteins, pointed/flat-headed edges are relationships of typge *increase/decrease*. (**b**) Sampling distribution of $\hat{Q}_x = \hat{E}[Erk|do(SOS = 70)]$ over 20 observational datasets

LVM with correct topology assumed the correct data generation process with non-informative $\mathcal{N}(0, 10)$ priors over all the parameters.

LVM with misspecified number of latent variables wrongly assumed two latent variables between *hns* and *topA*.

4.4 Synthetic Case Study 3: the signaling model

The system in Figure 5a is a well-studied insulin-like growth factor signaling system regulating growth and energy metabolism of a cell (Zucker *et al.*, 2021). IGF and EGF are latent.

Causal query of interest $Q_{SOS} = E[Erk|do(SOS = 70)]$. Similar to Case Study 2, Q_{SOS} does not satisfy the back-door or the front-door criteria.

Data mimicked the experimental process of collecting observational and interventional data. Since dynamics of this system are well characterized in form of stochastic differential equations (SDE) (Bianconi *et al.*, 2012), we generated observational data by simulating from the SDE. We set the initial amount of each protein molecule to 100, and generated subsequent observations via the Gillespie algorithm (Gillespie, 1977) in the *smfsb* (Wilkinson, 2018) R package. Replicates were generated by randomly initializing EGF and IGF. Interventional data were generated similarly, while fixing SOS = 70.

LVM with correct topology Unlike in the previous case studies, the variables were not modeled following the data generation process, but only approximated it. The exogenous variables were modeled with a Gaussian distribution. The rest of the variables were modeled by representing the biomolecular reactions with a Hill function, as common in the biological practice (Alon, 2019), and were approximated with a sigmoid function as follows, $\mathcal{N}\left(\frac{100}{1+\exp(\theta^T Pa(X)+\theta_0)}, \sigma_X\right)$. For a node X with q parents, $Pa(X)$ was a $q \times 1$ vector of measurements on the parent nodes, θ^T was a $1 \times q$ vector of unknown parameters, and θ_0 was an unknown scalar parameter. The non-informative $\mathcal{N}(0, 10)$ priors of the parameters θ in the sigmoid had a constraint of being positive for the relationships of type increase and negative for relationships of type decrease.

LVM with misspecified number of latent variables only included EGF as latent, and omitted IGF.

4.5 Synthetic Case Study 4: the SARS-CoV-2 model

The system in Figure 6a models activation of Cytokine Release Syndrome (Cytokine Storm), known to cause tissue damage in severely ill SARSCoV-2 patients (Ulhaq and Soraya, 2020). The simultaneous activation of the NF-κB and IL6-STAT3 activates IL6-AMP, which in turn activates Cytokine Storm (Hirano and Murakami, 2020). The system showcases the ability of a causal LVM to estimate multiple causal queries after a single instance of training.

The network was extracted from COVID-19 Open Research Dataset (CORD-19) (13) document corpus using the Integrated Dynamical Reasoner and Assembler (INDRA) (Gyori *et al.*, 2017) workflow (Zucker *et al.*, 2021), and by quering and expressing the corresponding causal statements in the Biological Expression Language (BEL) (Slater, 2014) using PyBEL (Hoyt *et al.*, 2018). Presence of latent variables was determined by querying pairs of entities in the network for common causes in the corpus.

Causal queries of interest examine the ability of two different drugs to prevent Cytokine Storm. Tocilizumab (Toci) is an immunosuppressive drug that targets sIL6Rα and blocks the IL6 signal transduction pathway (Zhang *et al.*, 2020). The first causal query examined the effect of Toci by setting its target sIL6Rα = 20 (low value), i.e., $Q_{sIL6R\alpha} = E[Cytokine|do(sIL6R\alpha) = 20)]$. The query is identifiable using the backdoor criterion. The drug Gefitinib (Gefi) blocks *EGFR*. The second causal query examined the effect of Gefi, i.e., $Q_{EGFR} = E[Cytokine|do(EGFR) = 20)]$. The query is not identifiable via either the backdoor or the front-door criterion, but is identified via the do-calculus.

Data of the latent variables had Gaussian distributions, Cytokine storm had a Bernoulli distribution with logit parameterization, and the remaining variables were simulated with a Hill function as in Case Study 3.

LVM with the correct topology assumed the correct data generation process where it contained two latent variables between (SARS-CoV-2 and Angiotensin II), (ADAM17 and sIL6Rα) and (PRR and NF-κB), and one latent variable for each remaining dotted edge. A mixture of Non-informative $\mathcal{N}(0, 10)$ and informative priors $\mathcal{N}(E[\theta], 1)$ where $E[\theta]$ were between 20 and 45 was used.

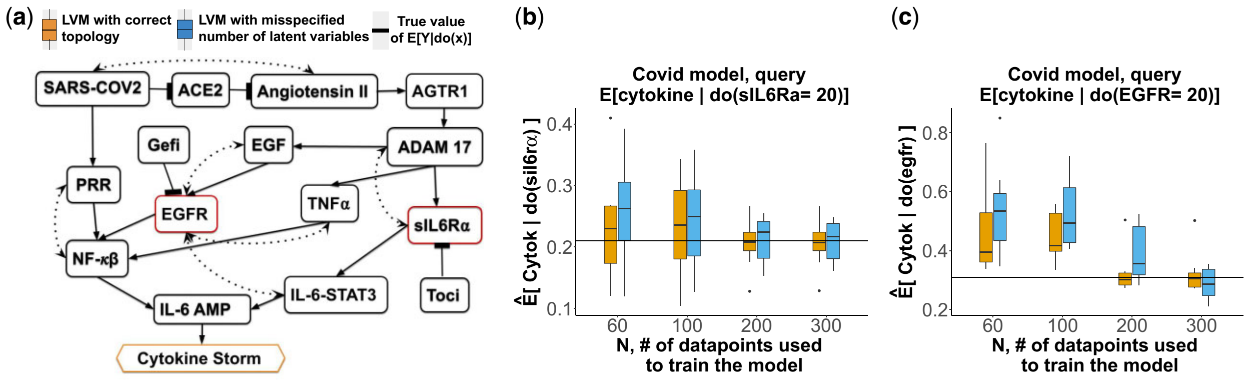
LVM with misspecified number of latent variables wrongly assumed only one latent variable for each dotted edge.

4.6 Experimental Case Study 5: the single-cause feed-forward transcriptional regulatory network motif

The system in Figure 7a is a common feed-forward network motif in the transcriptional regulatory network of *E.coli*, where the effect variable is not a direct effect of the cause variable. The network was obtained by querying the EcoCyc database (Keseler *et al.*, 2021) with the Pathway-tools lisp api (Karp *et al.*, 2021) for all 3-hop ancestors of all 2-hop descendants of the genes with available experimental interventional data. 5800 such cases were found. We randomly selected the pathway in Figure 7a. This system is similar to Case Study 1, but with a single cause and a single latent variable.

Causal query of interest $Q_{soxS} = P(ybiT|do(soxS = 0))$. Although *lrp* is latent, the mediator *rob* made the query identifiable according to the front-door criterion.

Experimental data contained 278 RNA-seq normalized expression profiles of *E.coli* K-12 MG1655 and BW25113 across 154 unique experimental conditions. Interventional data corresponded to the query of interest, i.e., $soxS = 0$. We used this data to evaluate



Case study 4 : Covid model

Fig. 6. Synthetic Case Study 4. DAG labeled as in Figure 3. (**a**) The SARS-CoV-2 model. Dotted edges indicate presence of latent variables. sIL6Rα and EGFR are targets of intervention, Cytokine Storm is the effect. (**b**) Sampling distribution of $\hat{Q}_{\mathbf{x}} = \hat{E}[Cytokine|do(sIL6R\alpha = 20)]$ over 20 observational datasets. (**c**) As in (b), for $\hat{Q}_{\mathbf{x}} = \hat{E}[Cytokine|do(EGFR = 20)]$

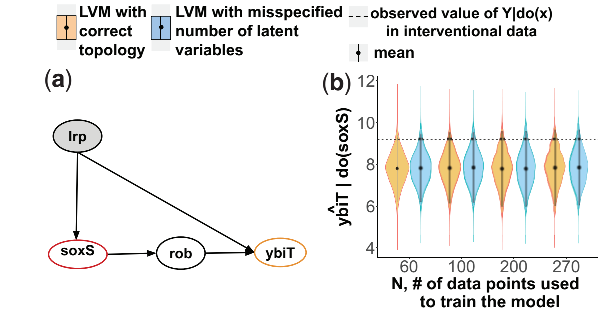


Fig. 7. Experimental Case Study 5. (**a**) The transcriptional regulatory network with the single-cause feed-forward motif. (**b**) The causal query in form of a probability distribution $\hat{Q}_{soxS} = \hat{P}(ybiT|do(soxS) = 0)$. The two (overlapping) horizontal dashed lines indicate two observed values of $ybiT|soxS = 0$ in the interventional experiments

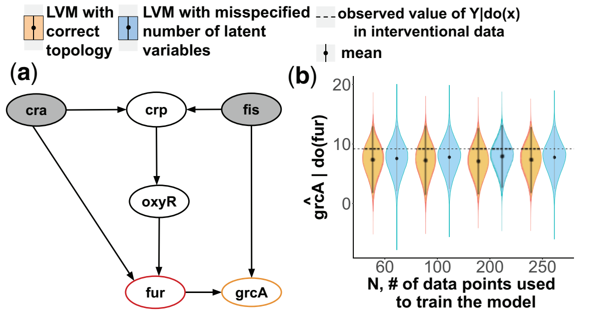


Fig. 8. Experimental Case Study 6. (**a**) The transcriptional regulatory network with the Napkin motif. (**b**) The causal query in form of a probability distribution $\hat{Q}_{fur} = \hat{P}(grcA|do(fur = 0))$. The two (overlapping) horizontal dashed lines indicate two observed values of $grcA|fur = 0$ in the interventional experiments

the performance of the proposed approach. The observational and interventional data was obtained from the PRECISE database (Sastry *et al.*, 2019).

LVM with correct topology specified Gaussian distributions with non-informative priors $N(0, 10)$.

LVM with misspecified number of latent variables wrongly assumed two latent variables.

4.7 Experimental Case Study 6: the napkin motif

The system in Figure 4.7 is the same system as in case study 2. The network was obtained by querying the EcoCyc database (Keseler *et al.*, 2021) with the Pathway-tools lisp api (Karp *et al.*, 2021) for all 3-hop ancestors of all 2-hop descendants of the genes with available experimental interventional data. 5500 such cases were found. We randomly selected the pathway in Figure 8a.

Causal query of interest $Q_{fur} = P(grcA|do(fur = 0))$.

Experimental data were as in Case Study 5. Interventional data corresponded to the query of interest, i.e. $fur = 0$. We used this data to evaluate the performance of the proposed approach.

LVM with correct topology assumed a Gaussian distribution over all the variables with non-informative priors $N(0, 10)$.

LVM with misspecified number of latent variables wrongly assumed two latent variables between *crp* and *grcA*.

5 Results

In the synthetic case studies with correct LVM topologies, the estimates $\hat{E}[Y|do(X = x')]$ were consistent Figures 3b, 4b, 5b and 6b and c show sampling distributions of 20 $\hat{E}[Y|do(X = x')]$, summarized from 20 repetitions of generating observational data with N replicates and estimating the causal query (orange boxes). Although the expected values and the variances of the sampling distributions depended on the data and on the system, all the estimates approached the true value with reduced variability as N increased. This was the case despite the diverse topologies of the networks, and despite the diverse distributional assumptions.

In the synthetic case studies with LVM with misspecified number of latent variables, the estimates $\hat{E}[Y|do(X = x')]$ *were consistent.* Figures 3b, 4b, 5b, and 6b and c show sampling distributions of 20 $\hat{E}[Y|do(X = x')]$, summarized from 20 repetitions of generating observational data and estimating the causal query with LVMs specifying a wrong number of latent variables (blue boxes). While these sampling distributions had more bias and variance than the distributions from the correctly specified LVMs for small N, they approached the true values with reduced variation as N increased.

In the experimental case studies, the estimates $\hat{P}(Y|do(X = x'))$ *accurately represented the observed interventional data.* Figures 7b and 8b display the estimated query specified in a different form, namely posterior interventional distributions $\hat{P}(Y|do(X = x'))$, as function of the number of observations used to train the LVM. The horizontal lines correspond to two experimental interventional measurements. As the values were very similar, the lines overlap. Despite real-life experimental artifacts, such as dynamic range compression and measurement errors that affected the observational data, and despite the approximate nature of the modeling assumptions, the estimated distributions covered the values observed from the experimental interventional data.

In both synthetic and experimental case studies, the estimates were accurate despite the approximate nature of the parametric assumptions. A common criticism of LVMs is the requirement of parametric distributional assumptions. Despite the criticism, in Case Study 3 the estimate $\hat{E}[Y|do(X = x')]$ was consistent, and in Case Studies 5 and 6 the estimates $\hat{P}(Y|do(X = x'))$ were accurate, even though the LVMs imperfectly approximated the unknown data generating distributions.

The proposed approach expanded the current use of LVMs for estimating causal queries. All the case studies showed the accuracy of LVM-based estimation in situations where LVM-based estimators have so far not been traditionally applied, i.e., pathways without proxy variables (Kuroki and Pearl, 2014; Louizos *et al.*, 2017), or with multiple causes (Wang and Blei, 2019).

Non-LVM-based approaches could not be applied to any case study in this manuscript. An alternative to LVM-based estimators are non-parametric or semi-parametric estimators. Unfortunately, we could not apply any of these methods to the case studies in this manuscript, as their implementations were either not publicly available (Jung *et al.*, 2020) or could not handle continuous or multi-cause queries (Bhattacharya *et al.*, 2020). However, in biomolecular systems multiple causes and effects are common, and the variables are not always discrete. All the case studies in this manuscript had either multiple causes or a continuous cause. They demonstrated the utility and the accuracy of LVM-based estimation in these situations.

Synthetic case studies 1, 2 and experimental case studies 5, and 6 represented commonly occurring patterns in biomolecular pathways. To demonstrate the ubiquity of the feed-forward motif presented in case studies 1 and 5, we queried the EcoCyc database (Keseler *et al.*, 2021) for all *E.coli* front-door motifs with one or more confounders and one or more causes, and found over 1000 such motifs. To demonstrate the ubiquity of the motif in case studies 2 and 6, we queried the EcoCyc database for all *E.coli* Napkin motifs with two or more confounders, and found over 1000 such motifs.

To further illustrate the ubiquity of network motifs that satisfy the front-door criterion of studies 1 and 5 in other organisms, we queried the repository INDRA (Gyori *et al.*, 2017) for all such motifs in humans. We applied strict quality filters to ensure each causal relationship was supported by at least 5 publications. We found 90 such instances. Reference Mangan and Alon (2003) provides additional examples of the feed-forward motif in yeast. Overall, the case studies illustrate the ability of the proposed approach to quantitatively answer potentially large numbers of important biological questions, limited only by qualitative prior knowledge of the organism's regulatory network and the availability of experimental data.

In synthetic case study 3, LVM-based estimators estimated multiple queries from a single trained model. After training the LVM once, the proposed approach enabled the estimation of two distinct causal queries. This is particularly valuable for probabilistic reasoning systems, e.g., in systems biology or medical diagnosis, where large-scale models are expensive to train and maintain. This is not possible with non-LVM-based estimators, which require us to derive a new statistical estimand for each causal query anew.

6 Discussion

A major criticism of traditional pathway modeling is its inability to account for external influences on pathway components. This is particularly relevant to causal inference, as ignoring the effect of unobserved confounding can undermine the inaccuracy of the results. In this article, we advocate for the explicit use of LVMs, and for applying Pearl's do-calculus to determine whether the causal effect can be identified.

We proposed training the LVMs with exact inference algorithms as they guarantee asymptotically exact samples. These algorithms are computationally expensive. However, trained on observational data once, an LVM can estimate multiple causal queries corresponding to multiple mutilated versions of the original DAG. After training, the estimation of causal query is instantaneous. This indicates that with enough experimental replicates and computational resources the proposed approach can be scaled to larger networks.

We showed that LVM-based estimation of identifiable causal queries is successful in situations that challenge other statistical estimators, e.g., in presence of interventions on continuous variables, and queries with multiple causes and effects. The estimation is robust to latent variable misspecification, and to parametric approximations of complex processes of data generation. As all these situations are very common, the proposed approach expands the feasibility and scope of causal inference in biomolecular pathways.

Funding

J.Z's contribution was funded by the Data Model Convergence Initiative at Pacific Northwest National Laboratory. It was conducted under the Laboratory Directed Research and Development Program at PNNL, a multi-program national laboratory operated by Battelle for the U.S. Department of Energy. C.T.H's contribution was supported by the DARPA Young Faculty Award W911NF2010255 (PI: Benjamin M. Gyori). K.S's contribution was supported by the Muscular dystrophy association grant # 574137 and Answer ALS consortium. O.V. acknowledges the support of NSF-BIO/DBI 1759736, NSF-BIO/DBI 1950412, NIH-NLM-R01 1R01LM013115 and of the Chan-Zuckerberg foundation.

Conflict of Interest: none declared.

References

Alon,U. (2019) *An Introduction to Systems Biology: Design Principles of Biological Circuits*. CRC Press, Boca Raton, Florida.

Balasubramanian,V. (1997) Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Comput.*, **9**, 349–368.

Bhattacharya,R. *et al.* (2020) Semiparametric inference for causal effects in graphical models with hidden variables. arXiv preprint arXiv:2003.12659.

Bianconi,F. *et al.* (2012) Computational model of EGFR and IGF1R pathways in lung cancer: a systems biology approach for translational oncology. *Biotechnol. Adv.*, **30**, 142–153.

Bingham,E. *et al.* (2019) Pyro: deep universal probabilistic programming. *J. Mach. Learn. Res.*, **20**, 973–978.

Bishop,C.M. (2006) *Pattern Recognition and Machine Learning*. Springer.

Blei,D.M. *et al.* (2017) Variational inference: a review for statisticians. *J. Am. Stat. Assoc.*, **112**, 859–877.

Cannon,W.R. *et al.* (2021) Cracking the code of metabolic regulation in biology using maximum entropy/caliber and reinforcement learning. In: MDPI. pp. 9867.

D'Amour,A. (2019) On multi-cause causal inference with unobserved confounding: counterexamples, impossibility, and alternatives. arXiv preprint arXiv:1902.10286.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, England.

Eberhardt,F. and Scheines,R. (2007) Interventions and causal inference. *Philos. Sci.*, **74**, 981–995.

Ernst,J. *et al.* (2007) Reconstructing dynamic regulatory maps. *Mol. Syst. Biol.*, **3**, 74.

Evans,R.J. (2016) Graphs for margins of Bayesian networks. *Scand. J. Statist.*, **43**, 625–648.

Galles,D. and Pearl,J. (2013) Testing identifiability of causal effects. *arXiv preprint arXiv:1302.4948*.

Gelman,A. *et al.* (2014) *Bayesian Data Analysis*. 3rd edn. CRC Press, Boca Raton, Florida.

Gillespie,D.T. (1977) Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, **81**, 2340–2361.

Girolami,M. and Calderhead,B. (2011) Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. B*, **73**, 123–214.

Gyori,B.M. *et al.* (2017) From word models to executable models of signaling networks using automated assembly. *Mol. Syst. Biol.*, **13**, 954.

Helske,J. *et al.* (2021) Estimation of causal effects with small data in the presence of trapdoor variables. *J. R. Stat. Soc., A: Stat. Soc.*, **184**, 1030–1051.

Hirano,T. and Murakami,M. (2020) COVID-19: a new virus, but a familiar receptor and cytokine release syndrome. *Immunity*, **52**, 731–733.

Hoyt,C.T. *et al.* (2018) PyBEL: a computational framework for biological expression language. *Bioinformatics*, **34**, 703–704.

Huang,Y. and Valtorta,M. (2006) Pearl's calculus of intervention is complete. In: *Proceedings of Uncertainty in Artificial Intelligence*, UAI'06, Boston, MA.

Hughes,M.D. *et al.* (1998) CD4 cell count as a surrogate endpoint in HIV clinical trials: a meta-analysis of studies of the aids clinical trials group. *Aids*, **12**, 1823–1832.

Jung,Y. *et al.* (2020) Learning causal effects via weighted empirical risk minimization. In: *Advances in Neural Information Processing Systems*, New Orleans, LA, vol. 33.

Jung,Y. *et al.* (2021) Estimating identifiable causal effects through double machine learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, held virtually.

Karp,P.D. *et al.* (2021) Pathway tools version 23.0 update: software for pathway/-genome informatics and systems biology. *Brief. Bioinform.*, **22**, 109–126.

Keseler,I.M. *et al.* (2021) The ECOCYC database in 2021. *Front. Microbiol.*, **12**, 2098.

Koller,D. and Friedman,N. (2009) *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA.

Kondofersky,I. *et al.* (2015) Identifying latent dynamic components in biological systems. *IET Syst. Biol.*, **9**, 193–203.

Kuroki,M. and Pearl,J. (2014) Measurement bias and effect restoration in causal inference. *Biometrika*, **101**, 423–437.

Lattimore,F. and Rohde,D. (2019a) Causal inference with Bayes rule. arXiv preprint arXiv:1910.01510.

Lattimore,F. and Rohde,D. (2019b) Replacing the do-calculus with Bayes rule. arXiv preprint arXiv:1906.07125.

Louizos,C. *et al.* (2017) Causal effect inference with deep latent-variable models. In: *Advances in Neural Information Processing Systems, Long Beach, CA*. pp. 6446.

Mangan,S. and Alon,U. (2003) Structure and function of the feed-forward loop network motif. *Proc. Natl. Acad. Sci. USA*, **100**, 11980–11985.

McNaughton,A.D. *et al.* (2021) Bayesian inference for integrating *Yarrowia lipolytica* multiomics datasets with metabolic modeling. *ACS Synth Biol.*, **10**, 2968–2981.

Pearl,J. (1993) Bayesian analysis in expert systems: comment: graphical models, causality and intervention. *Stat. Sci.*, **8**, 266.

Pearl,J. (1995) Causal diagrams for empirical research. *Biometrika*, **82**, 669–688.

Pearl,J. (2009) *Causality*. Cambridge University Press, Cambridge, England.

Pearl,J. (2019) The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, **62**, 54–60.

Pearl,J. and Mackenzie,D. (2018) *The Book of Why: The New Science of Cause and Effect*. Basic Books, Ney York, NY.

Rasmussen,C.E. and Ghahramani,Z. (2001) Occam's razor. In: *Advances in Neural Information Processing Systems, Vancouver, British Columbia, Canada*. pp. 294.

Richardson,T.S. *et al.* (2017) Nested Markov properties for acyclic directed mixed graphs. arXiv preprint arXiv:1701.06686.

Robert,C. and Casella,G. (2013) *Monte Carlo Statistical Methods*. Springer Science & Business Media, Berlin, Germany.

Sastry,A.V. *et al.* (2019) The *Escherichia coli* transcriptome mostly consists of independently regulated modules. *Nat. Commun.*, **10**, 1–14.

Shojaie,A. and Michailidis,G. (2009) Analysis of gene sets based on the underlying regulatory network. *J. Comput. Biol.*, **16**, 407–426.

Shpitser,I. and Pearl,J. (2006) Identification of joint interventional distributions in recursive semi-Markovian causal models. In: *Proceedings of the National Conference on Artificial Intelligence, Boston, MA, USA*, vol. 21. pp. 1219.

Shpitser,I. and Pearl,J. (2008) Complete identification methods for the causal hierarchy. *J. Mach. Learn. Res.*, **9**, 1941.

Shpitser,I. *et al.* (2012) Parameter and structure learning in Nested Markov Models. arXiv preprint arXiv:1207.5058.

Shpitser,I. *et al.* (2014) Introduction to nested Markov models. *Behaviormetrika*, **41**, 3–39.

Slater,T. (2014) Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discov. Today*, **19**, 193–198.

Spirtes,P. *et al.* (2000) *Causation, Prediction, and Search*. MIT Press, Cambridge, MA.

St John,P.C. *et al.* (2019) Bayesian inference of metabolic kinetics from genome-scale multiomics data. *PLoS Comput. Biol.*, **15**, e1007424.

Stan Development Team (2018) RStan: the R interface to Stan. R Package Version 2.17.3.

Stan Development Team. (2020) RStan: the R interface to Stan, 2020. R package version 2.21.2. https://mc-stan.org/.

Ulhaq,Z.S. and Soraya,G.V. (2020) Interleukin-6 as a potential biomarker of COVID-19 progression. *Med. Mal. Infect.*, **50**, 382–383.

Van Hoey,S. *et al.* (2013) Python package for model structure analysis (pySTAN). In: EGU General Assembly Conference Abstracts, Vienna, Austria. pp. EGU2013–10059.

Wang, L.W. *et al.* (2020) CORD-19: The COVID-19 Open Research Dataset. arXiv preprint arXiv:2004.10706v2.

Wang,Y. and Blei,D.M. (2019) The blessings of multiple causes. *J. Am. Stat. Assoc.*, **114**, 1574–1596.

Wilkinson,D (2018). *Package smfsb, Smfsb-stochastic modeling for systems biology. R Package Version, 1*. https://cran.microsoft.com.

Zhang,C. *et al.* (2020) Cytokine release syndrome in severe COVID-19: interleukin-6 receptor antagonist tocilizumab may be the key to reduce mortality. *Int. J. Antimicrob. Agents*, **55**, 105954.

Zucker,J. *et al.* (2021) Leveraging structured biological knowledge for counterfactual inference: a case study of viral pathogenesis. *IEEE Trans. Big Data*, **7**, 25–37.