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] **Olga Vitek** [1]

Abstract

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 modeling such partially observed pathways. Unfortunately, LVM-based estimation of causal queries can be inaccurate when parameters of the latent variables are not uniquely identified during training, or when the number of true latent variables is misspecified. This so far has limited the use of LVMs for causal inference from biomolecular pathways. In this manuscript 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. We further provide an open-source implementation evaluating whether a causal query is identifiable, and describe an algorithm for its estimation. The proposed approach opens the door for causal inference in a broad variety of biomolecular pathways. We illustrate the breadth and the practical utility of this approach for estimating causal queries in four case studies with varying complexity.

[1]Khoury College of Computer Sciences, Northeastern University, Boston, Massachusetts, USA [2]Pacific Northwest National Laboratory, Richland, Washington, USA [3]Laboratory of Systems Pharmacology, Harvard Medical School, Boston, Massachusetts, USA [4]Next Generation Analytics, Palo Alto California, USA [5]Answer ALS Consortium, USA [6]Microsoft Research, Redmond, Washington, USA. Correspondence to: Sara Mohammad-Taheri <mohammadtaheri.s@northeastern.edu>, Olga Vitek <o.vitek@northeastern.edu>.

1. Introduction

Biomolecular pathways are governed by intricate 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 wet lab perturbations, and answering, *in silico*, causal queries of the form "when we perturb X, what is the effect on its descendent Y?". However, 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 and estimate their parameters (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.). 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.).

Latent variable models (LVMs) are particularly useful for representing biological pathways with partially known topology or missing measurements of pathway components (Durbin et al.; Kondofersky et al.; Ernst et al.; St John et al.; Shojaie & Michailidis). LVMs are probabilistic models of a joint distribution on a set of observed and unobserved variables. A broad class of LVMs have 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 (Louizos et al., 2017; Kuroki & Pearl, 2014), or the presence of multiple causes and no unobserved confounders, (i.e., no hidden regulators of cause and effect) (Wang & Blei, 2019). The latter approach 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 manuscript, we argue that LVM-based estimators of causal queries are in fact accurate when the queries are identifiable according to Pearl's do-calculus. 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 provide an open-source implementation for evaluating whether a causal query is identifiable, and describe a simple and practical algorithm for its estimation.

We demonstrate the breadth of applicability, and the practical utility of LVM-based estimation of identifiable causal queries of biomolecular pathways in four case studies of varying complexity. The first two case studies consider network motifs frequently occurring in the transcriptional regulatory network of *Escherichia coli*. The second two case studies focus on human signaling pathways, where information about a stimulus at the cell surface is transmitted via series of protein-protein interactions, in order to activate or repress a set of transcription factors in the nucleus. The case studies demonstrate the accurate and consistent estimation of causal effects, even when some components of the network motif are unknown or cannot be experimentally quantified.

2. Background

2.1. Notation

Let $\mathbf{V} = \{V_1, ..., V_J\}$ be a set of observed random variables, and $\mathbf{U} = \{U_1, ..., U_L\}$ 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 $P(v_1, ..., v_j)$ 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 manuscript, 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 $\text{Pa}(V_j)$ 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 broad class of LVMs have a directed acyclic graphical (DAG) structure. Canonical examples of them include topic models, hidden Markov models, Gaussian mixture models (Blei, 2014), and deep generative latent variable models such as variational autoencoders (Kingma & Welling, 2014).

A **causal LVM** is an LVM with DAG structure where $\text{Pa}(V_i)$ 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 structure G, observed variables $\mathbf{V}$, and latent variables $\mathbf{U}$, (Evans, 2016) offers the following simplification rules to compactly represent LVMs with many latent variables by LVMs that only include a single latent variable between each pair of observed variables.

- We can remove latent variables with no children.
- We can remove a latent variable U with observable parents by connecting all the parents of U to its children.
- If U, W are latent variables with $children(W) \subseteq children(U)$, then we can remove W.

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

Inference algorithms such as belief propagation (Pearl, 2014; Lauritzen & Spiegelhalter, 1988) and variable elimination(Shpitser et al., 2012), exact sampling techniques such

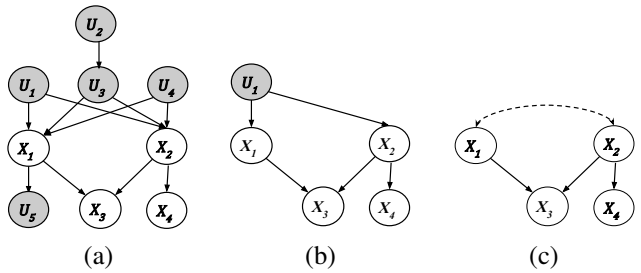


Figure 1. (a) An LVM with 4 observed (white) and 5 latent (dark grey) variables. (b) A different LVM with 1 latent variable. (c) An ADMG representing both (a) and (b).

as Hamiltonian Monte Carlo (HMC) (Girolami & Calderhead, 2011; Duane et al., 1987), and approximate gradient-based optimization methods such as stochastic variational inference (SVI) (Hoffman et al., 2013) 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 $\mathbf{V}$. Exact sampling-based algorithms such as HMC guarantee asymptotically exact samples, but are computationally expensive (Robert & Casella, 2004). Approximate probabilistic inference algorithms such as variational inference (Wainwright & Jordan, 2008; Bishop, 2006; Hoffman et al., 2013; Blei et al., 2017) trade off accuracy for speed by searching with gradient descent a parameterized family of functions that approximate the target distribution.

2.3. Causal inference

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 $do(\mathbf{x})$ by (Pearl, 1995)), and makes it independent of its causes (Spirtes et al., 2000; Eberhardt & Scheines, 2007). **Graph mutilation** in a causal LVM simulates an intervention. It severs the edges incoming to the target nodes, and fixes each node $X_i \in \mathbf{X}$ to its intervention value $x_i \in \mathbf{x}$ (Koller & Friedman, 2009). We denote $G_{\bar{\mathbf{X}}}$ as the graph resulting from removing all incoming edges to nodes $\mathbf{X}$ and $P_{G_{\bar{\mathbf{X}}}}(\mathbf{v})$ the probability distribution encoded by $G_{\bar{\mathbf{X}}}$.

A **causal query** $Q_{\mathbf{x}}$ is a probabilistic query that conditions a set of outcomes $\mathbf{Y} \subseteq \mathbf{V}\backslash\mathbf{X}$ on a set of interventions, such as $Q_{\mathbf{x}} = P(\mathbf{Y}|do(\mathbf{x}))$ or $Q_{\mathbf{x}} = E[\mathbf{Y}|do(\mathbf{x})]$. Sampling from $P(\mathbf{Y}|do(\mathbf{x}))$ is achieved by applying algorithmic inference to $G_{\bar{\mathbf{X}}}$ and sampling from $P_{G_{\bar{\mathbf{X}}}}(\mathbf{Y}|\mathbf{x})$. To denote the value of the outcome variable obtained from a mutilated model that was trained on observational data, we must use counterfactual subscript notation $\mathbf{Y}_{do(\mathbf{x}')} \sim P(\mathbf{Y}_{do(\mathbf{x}')}|\{x_i, y_i\}_{i=1}^N)$ to distinguish the value $\mathbf{y}_{do(\mathbf{x}')}$ of an outcome variable given an intervention $do(\mathbf{x}')$ from an observation of $\mathbf{x}$ and $\mathbf{y}$ in the training data.

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 & Pearl, 2008). A causal query is identifiable if it satisfies the back-door or the front-door criteria. 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 bidirected edges, and there exists a set $\mathbf{Z} \subseteq \mathbf{V}\backslash\{X, Y\}$ 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). The **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, 2009; 1993; 1995). For example, the back-door criterion or the front-door criterion does not hold in Fig. 2 (a) or (b), but the front-door criterion holds in Fig. 2(b).

The back-door and front-door criteria are sufficient but not necessary for causal identifiability. The **do-calculus**, comprised of three graph-mutilation-based rules (Pearl, 2009), is necessary and sufficient for causal identifiability. A causal query containing a $do()$ operator is identifiable 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 $P(\mathbf{x})$.

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 & Pearl, 2008). For example, the causal query $P(y|do(\mathbf{x}))$ in the causal LVM in Fig. 2(a) is not identifiable according to the do-calculus but is identifiable in Fig. 2(b). If the query is identifiable, these algorithms generate an estimand computable from observational data (Huang & Valtorta, 2006; Shpitser & Pearl, 2006). 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).

There exist several implementations of the identification algorithm, however each has a different limitation that hampers its utility. Causal Fusion (Bareinboim & Pearl, 2016) generates a symbolic representation of the probabilistic estimand, and the means to estimate the estimand from data, but it is closed source. CausalEffect (Tikka & Karvanen, 2017) generates a symbolic representation of the probabilistic estimand, but it does not provide the means to learn the estimand from data. Ananke (Bhattacharya et al., 2020) generates an influence function that can be used to learn the estimand from data, but it does not generate a symbolic representation of the probabilistic estimand.

2.4. Causal query estimation

For queries of a form of $P(\mathbf{Y}|do(\mathbf{x}))$, 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 $E(\mathbf{Y}|do(\mathbf{x})$, a desirable property of the estimator is consistency. An estimator of $E[\mathbf{Y}|do(\mathbf{x})]$ is consistent if, as the number of data points used to estimate the query tends to infinity, the sequence of the estimates of the causal query converges in probability to its expected value.

Several non-LVM approaches for estimating causal queries with these desirable properties exist. These approaches derive a separate statistical estimand for each causal query anew (Pearl, 2019). Unfortunately, the scope of their applicability is somewhat limited. Some of the approaches are restricted to causal queries with one cause and one effect, and the cause must be binary-valued (Bhattacharya et al., 2020). Others are inadequate in large data regimes where it is computationally expensive to train a new estimator for each query of interest (Jung et al., 2020; 2021).

Below 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. Since LVM-based approach for estimating causal queries is not limited in the number of cause and effects and the type of their distributions, and since it can estimate multiple causal queries from a single LVM training, it is applicable to a much broader set of circumstances. It is particularly useful for probabilistic causal reasoning in larger biomolecular pathways, where the true latent structure may be unknown, and where models are expensive to train and maintain.

3. Methods

3.1. LVM-based estimation of identifiable causal queries

The proposed approach (Algorithm 1) takes as inputs a causal query of interest, target values of the intervention, effects of the intervention, and an LVM trained with an exact sampling based inference algorithm. It first determines whether the causal query of interest is identifiable according to Pearl's do-calculus (line 1). This is done with an identification algorithm implementation such as in Causal Fusion, CausalEffect, or Ananke. We advocate the use of the Y_0 causal reasoning engine introduced in this manuscript (Sec. 3.4). If the query is not identifiable, we do not proceed (line 2).

If the query is identifiable, Algorithm 1 proceeds with its estimation. We take a Bayesian viewpoint (Lattimore & 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)

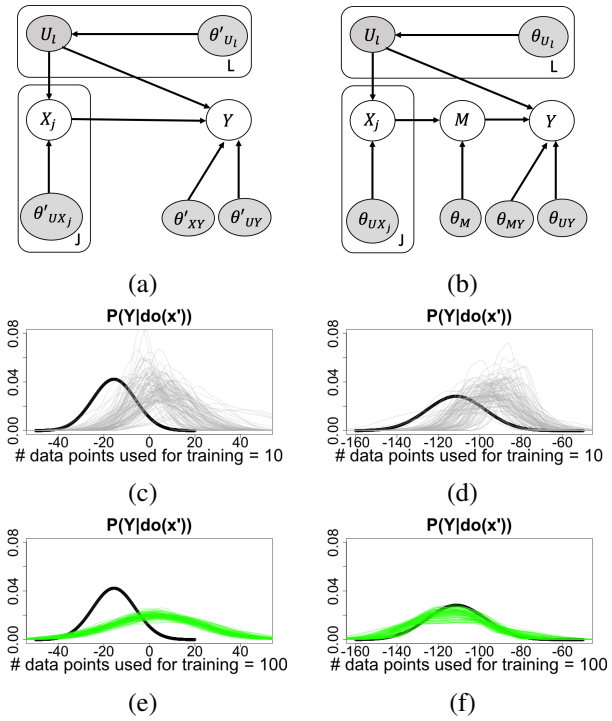


Figure 2. **The estimates of distribution of an identifiable causal query $P(Y|do(\mathbf{x}'))$ (Figures to the right) converges to the true distribution as number of data points increases but it fails to to do so for a not identifiable causal query (Figures to the left)** (a) An LVM where $P(Y|do(\mathbf{x}'))$ 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 $P(Y|do(\mathbf{x}'))$ is non-parametrically identified. (c,e) relate to (a). Black curve estimates the true distribution $P(Y|do(\mathbf{x}'); \theta)$, with θ used to generate interventional data. After training the LVM on $N = 10,100$ observational datapoints, each gray/green curve estimate $P_{\hat{M}_{\bar{x}}}(Y_{do(\mathbf{x}')}; \{x_i, y_i\}_{i=1}^N, \theta)$ 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.

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.

3.2. Motivating examples

We illustrate the practical application of this method in the special case of the LVM in Fig. 2(a) where protein product of gene X affects gene Y, while both are under regulation

Algorithm 1 Estimation of an identifiable causal query

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

1: Check identifiability of $Q_{\mathbf{x}}$, e.g. with Y_0
2: **if** $Q_{\mathbf{x}}$ is not identifiable **then**
3: break
4: **else**
5: Set $\mathbf{X} = \mathbf{x}'$
6: Create $\hat{\mathcal{M}}_{\bar{\mathbf{x}}}$, the mutilated model
7: **for** s in 1:S **do**
8: Sample $\theta_s \sim P_{\hat{\mathcal{M}}_{\bar{\mathbf{x}}}}(\theta|\{\mathbf{v}_i\}_{i=1}^N)$
9: **for** W in topological-sort($\{\mathbf{U} \cup \mathbf{V}\}$) **do**
10: Sample $w_s \sim P_{\hat{\mathcal{M}}_{\bar{\mathbf{x}}}}(W|Pa(W);\theta_s)$
11: **end for**
12: Collect $\mathbf{y}_s \subseteq \mathbf{w}_s$
13: **end for**
14: **Return** density$\left(\{\mathbf{y}_s\}_{s=1}^S\right)$ or $\frac{1}{S}\sum_{s=1}^S \mathbf{y}_s$
15: **end if**

of the same transcription factor(s) and/or enhancer(s). The causal query $P(Y|do(X = x'))$ is not identifiable, and we show empirically that its LVM-based estimator is biased. We then extend the causal LVM with a mediator Z in Fig. 2(b), such that the query becomes identifiable according to the front-door criterion. This 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 $P(Y|do(X = x'))$ converges to the true distribution.

Empirical example 1: Fig. 2(a) 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 $P(Y|do(X = x'))$. We generated observational data with $N = 10,100$ samples from the likelihood with a randomly chosen vector of true values of θ. The true $P(Y|do(X = x'); \theta)$ was estimated with Algorithm 1, where line 8 was substituted by the true values of θ (black curves in Fig. 2(c-f)).

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

Empirical example 2: Fig. 2(b) 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 $P(Y|do(X = x'))$ becomes identifiable. Repeating the same analysis, Fig. 2(d,h) show that, as N increased, the distributions converged to the ground truth.

3.3. Proofs

In this part, Lemma 1 proves the empirical results of examples 1 and 2 for the same LVM in Fig. 2(a) and (b) but with arbitrary distributions. Then, Theorem 1 proves that in general for **any identifiable causal query**, Algorithm 1 accurately estimates causal queries in an LVM that correctly reflects the true underlying causal structure of the observed variables. Finally, Corollary 1 proves that the results of Theorem 1 holds for LVMs with misspecified number of latent variables.

Lemma 1 *Consider the LVM in Fig. 2 (b) with a DAG G. $\mathbf{X}$, M, 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, $G_{\bar{\mathbf{X}}}$, where all the incoming edges to $\mathbf{X}$ are removed. A causal query $P(Y|do(\mathbf{x}'))$ transforms $P(.)$ into a distribution $P_{\bar{\mathbf{X}}}(.)$, and $P(Y|do(\mathbf{x}')) = P_{\bar{\mathbf{X}}}(Y|\mathbf{x}')$. Hence,

$$P(Y|do(\mathbf{x}')) = P_{\bar{X}}(Y|\mathbf{x}') = \int_{\mathbf{u},z} P_{\bar{\mathbf{X}}}(Y,\mathbf{u},z|\mathbf{x}')d\mathbf{u}dz$$

$$= \int_z \left(\int_{\mathbf{u}} P_{\bar{\mathbf{X}}}(Y|\mathbf{u},z,\mathbf{x}')P_{\bar{\mathbf{X}}}(\mathbf{u}|z,\mathbf{x}')d\mathbf{u}\right) P_{\bar{\mathbf{X}}}(z|\mathbf{x}')dz$$

$$= \int_z P_{\bar{\mathbf{X}}}(Y|z)P_{\bar{\mathbf{X}}}(z|\mathbf{x}')dz$$

$$= \int_z P(Y|do(z))P(z|\mathbf{x}')dz \quad (1)$$

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

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

Next we demonstrate that this result holds in all generality for any identifiable causal query.

Theorem 1 *Consider a causal LVM $\mathcal{M}$, which includes the true likelihood that generated the observational data.*

Consider a causal query $Q_{\mathbf{x}} = P(\mathbf{Y}|do(\mathbf{x}))$ or $Q_{\mathbf{x}} = E[\mathbf{Y}|do(\mathbf{x})]$, identifiable according to the do-calculus with respect to $\mathcal{M}$. When estimating the causal query as in Algorithm 1, the estimate $\hat{P}(\mathbf{Y}|do(\mathbf{x}))$ converges to the true distribution, and the estimator $\hat{E}[\mathbf{Y}|do(\mathbf{x})]$ 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 & 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). □

Finally, we show that the LVM does not need to include a precise specification of the latent variables, as long as the misspecified LVM and the true LVM project to the same ADMG and agree on the joint observational distribution.

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

Proof. Let θ' be the parameters of $\mathcal{M}' \in M$. 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. □

3.4. Implementation and computational complexity

We implemented the identification algorithm (Shpitser & Pearl, 2006) in the Y_0 causal reasoning engine, an open-source Python software package available under the permissive BSD license, using modern software engineering practices such as unit testing, linting, and continuous integration. Y_0 is available at `https://github.com/y0-causal-inference/y0`. It overcomes each of the previously described limitations of the previous implementations. Y_0 takes as input a causal LVM and a query, and determines whether the query is identifiable according to Pearl's do-calculus. Documentation is available through ReadTheDocs and demos as Jupyter notebooks. Determining the identifiability was nearly-instant for all the case studies in this manuscript. To further examine the scalability of Y_0 to larger systems, we ran it on over 2,100 larger networks with 77 nodes each. The whole experiment took 13 minutes and 26 seconds, or about 0.23 seconds for each identification testing.

The proposed approach is based on *ad hoc* training of each individual LVM. While training an LVM is NP-complete (and in practice depends on the specific LVM and on the choice of inference algorithm), the proposed approach amortizes most of the computational work into this single training step. Given a single trained model, Algorithm 1 can estimate an arbitrary number of queries. All the experiments in this manuscript, including the *ad-hoc* implementations of Algorithm 1, are in `https://github.com/srtaheri/LVMwithDoCalculus` . The case studies took between 1.5 minutes and 1.8 hours on a Google Cloud Platform.

4. Case studies

4.1. Overview

We illustrate the practical utility and the accuracy of the LVM-based estimation of identifiable causal queries in four case studies of biomolecular pathways with varying complexity. The first case study considers a common network motif in the transcriptional regulatory network of *Escherichia coli*. Despite the commonality of its structure, its causal query cannot be estimated with the existing non-LVM approaches due to the presence of multiple causes. The second case study is another transcriptional regulatory network of *Escherichia coli*, where it is not possible to block the back-door path between the cause and effect, and the front-door criteria does not hold. As the result, the causal query cannot be estimated with most non-LVM estimators. The third case study represents an insulin-like growth factor(IGFR) signaling network, where data are generated from a stochastic process with uncharacterized distributions, yet can be approximated by the LVM. The last case study is a larger-scale molecular biology expert system of host response to viral infection of SARS-CoV-2, where the LVM is trained once, and used to estimate two distinct causal queries of interest.

To reflect the heterogeneity of distributions observed in biomolecular datasets, we avoid making specific assumptions regarding the distribution of each variable in the model, and instead incorporate a mix of distributions in each of our case studies. Each case study (except study 3) specified randomly selected true values of θ, and simulated 20 observational and 20 interventional datasets. All the parameters had non-informative $\mathcal{N}(0, 10)$ priors. The parameters regarding the mean of latent variables in case study 4 have $\mathcal{N}(\mu, 1)$ priors where μ is between 20-45. Posterior distributions of the parameters were inferred with HMC in `Stan` (Team, 2018).

The causal queries were of the form $Q_{\mathbf{x}} = E[\mathbf{Y}|do(\mathbf{X})]$. The true value $Q_{\mathbf{x}}$ was obtained by averaging samples from the interventional datasets with the true θ. For each observational dataset, $\hat{Q}_{\mathbf{x}}$ was estimated as in Algorithm 1. To evaluate the robustness of $\hat{Q}_{\mathbf{x}}$ to model misspecification, we also considered LVMs with a wrong number of latent variables but same ADMG.

For each case study, we simultaneously checked the identifiability of the causal queries and generated their corresponding estimand using our implementation of the identification algorithm in Y_0.

4.2. Case study 1: The feed-forward transcriptional regulatory network motif

The system The famous feed-forward loop is an example of a common network motif in *E. coli* and many other prokaryotes (Alon, 2019). For this case study, we consider the causal effect of $marA$, $soxS$ and $ompR$ on $ypiT$ in the multi-node generalization of the feed-forward network motif shown in Fig. 3(a). This network satisfies the back-door criterion when rob is latent, and satisfies the front-door criterion when the variables lrp and crp are latent, but cannot be identified when lrp, crp and rob are latent. To demonstrate the ubiquity of this motif, we queried the EcoCyc database (Keseler et al., 2013; 2017) to discover which front door motifs with one or more confounders and one or more causes exist in *E. coli*, all 1945 of which are available at `https://ecocyc.org/group?id=biocyc14-15682-3843672784`.

LVM To demonstrate that causal effects on the front-door network motif can be identified even when the number of latent confounders $\mathbf{U}$ are misspecified, we generated data from an LVM with 3 causes ($marA, soxS$ and $ompR$) and two latent variables (lrp and crp). We then used that data to train an LVM with the correct number of causes and latent variables, and a misspecified LVM with the correct number of causes, but only one latent variable.

Data were generated from the true model where the latent variables were Gaussian and the remaining variables followed a Bernoulli distribution with logit parameterization.

Estimates $\hat{Q}_{marA,sosX,ompR}$ (Fig. 3(d)) based on mutilating the trained LVM with the correct topology had less variance than the estimates from the misspecified mutilated LVM. However both estimates converged to the true value as N increased.

4.3. Case study 2 : The Napkin motif

The system in Fig. 3(b), called the second Napkin problem by (Pearl & Mackenzie, 2018), requires a non-trivial application of the do-calculus, as we cannot block the back-door path from lrp to $topA$ because hns is a collider and $gadE$ is an ancestor of a collider, and the front-door criterion does not hold because there is no mediator between lrp and $topA$ (Helske et al., 2021; Hughes et al., 1998; Pearl & Mackenzie, 2018; Jung et al., 2020). The causal estimands that these kinds of queries generate are difficult to estimate using statistical methods (Bhattacharya et al., 2020; Schulman & Srivastava, 2016). To investigate the ubiquity of this motif, we queried the EcoCyc database (Keseler et al., 2013; 2017) to discover all napkin motifs with two or more confounders, all 911 of which are available at `https://ecocyc.org/group?id=biocyc14-15682-3844537443`. We are interested in the causal effect of knocking out lrp on $topA$, so we model lrp as a binary variable. We also assume that hns is measured with a fluorescent reporter, so we model the expression of hns with a gamma distribution. Lastly, we assume that all the other genes are measured using relative expression, such as RT-PCR, so we model them with Gaussian distributions.

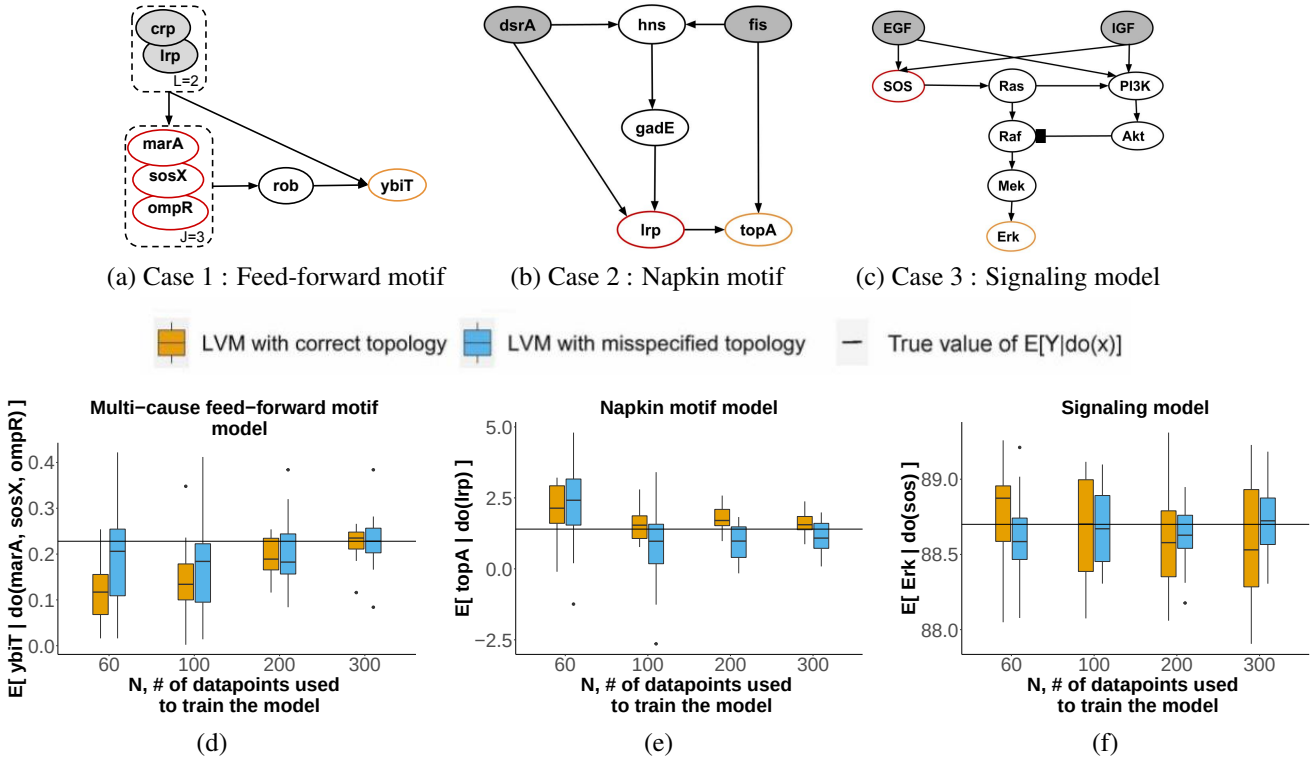
LVM with the correct topology had one latent variable between hns and lrp and one latent variable between hns and $topA$. The LVM with misspecified topology had one latent variable between hns and lrp, but two latent variables between hns and $topA$.

Data were generated from the model where dsr, fis, $gadE$ and $topA$ were simulated from a Gaussian, and hns from a Gamma distribution. lrp was simulated from a Bernoulli distribution with logit parametrization.

Estimates $\hat{Q}_{\mathbf{lrp}} = E[topA|do(lrp)]$ performed as in case study 1. In this case, the misspecified model did better than the correctly specified model, likely due to the additional degrees of freedom afforded by the second latent variable.

4.4. Case study 3: The signaling model

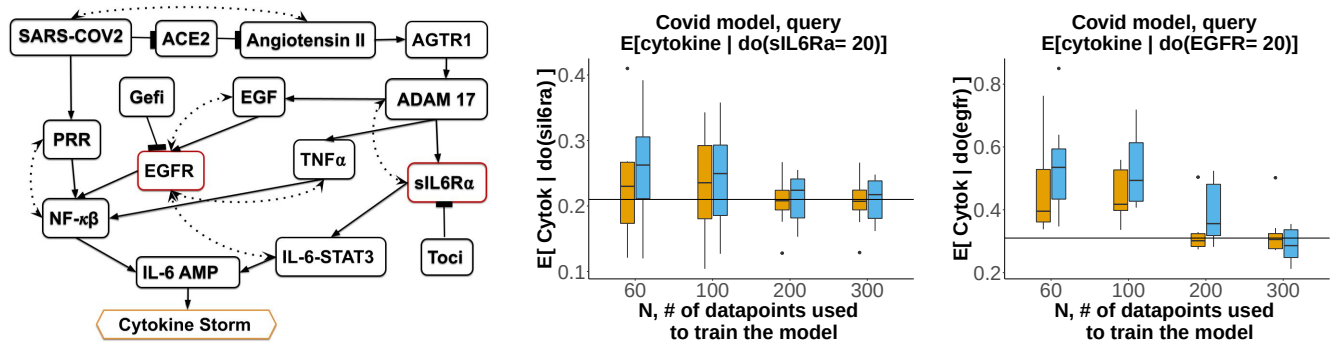
The system in Fig. 3(c) is a well-studied insulin-like growth factor signaling system regulating growth and energy metabolism of a cell (Zucker et al., 2021). It is activated by external stimuli IGF and EGF. Nodes in the system are



(a) Case 1 : Feed-forward motif (b) Case 2 : Napkin motif (c) Case 3 : Signaling model

(d) (e) (f)

Figure 3. Case studies 1-3. DAGs labeled as in Fig. 2. Red nodes are targets of the intervention, orange nodes are the effect. (a) The multi-cause feed-forward transcriptional regulatory network motif. (b) The Napkin network motif. (c) The signaling model. Nodes are proteins, pointed/flat-headed edges are relationships of type *increase/decrease*. (d) Sampling distribution of $\hat{Q}_{\mathbf{x}} = \hat{E}[ybiT|do(marA, sosX, ompR = 0)]$ over 20 observational datasets. (e) As in (d), $\hat{Q}_{\mathbf{x}} = \hat{E}[topA|do(lrp = 0)]$. (f) As in (d), $\hat{Q}_{\mathbf{x}} = \hat{E}[Erk|do(SOS = 70)]$.



(a) Case 4 : Covid model (b) (c)

Figure 4. Case study 4. Legends are as in Fig. 3. (a) The SARS-CoV-2 model. Dotted edges indicate presence of latent variables. sIL6Rα and EGFR are targets, Cytokine Storm is the effect. (b) $\hat{Q}_{\mathbf{x}} = \hat{E}[Cytokine|do(sIL6R\alpha = 20)]$. (c) $\hat{Q}_{\mathbf{x}} = \hat{E}[Cytokine|do(EGFR = 20)]$.

proteins, and edges are the effect of the upstream protein on the downstream protein's activity. IGF and EGF are latent. $Q_{\mathbf{x}} = E[Erk|do(SOS = 70)]$. This query does not satisfy the back-door or the front-door criteria.

LVM represented the biomolecular reactions with a Hill function, as common in the biological practice (Alon, 2019), and approximated them with a sigmoid. We modeled the root nodes with a Gaussian distribution and the non-root nodes with $\mathcal{N}(\frac{100}{1+exp(\theta^T Pa(X)+\theta_0)}, \sigma_X)$. 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 pa-

rameters, 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. The LVM with misspecified topology had a similar structure, but only including EGF as latent and omitting IGF.

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. Therefore, unlike in the other case studies, the LVM did not exactly represent the data generation process, but only approximated it.

Estimates $\hat{Q}_{\mathbf{x}}$ performed performed similarly to the previous case studies.

4.5. Case study 4: The SARS-CoV-2 model

The system (Fig. 4(a)) showcases the ability of a causal LVM to estimate multiple causal queries after a single instance of training. It models activation of Cytokine Release Syndrome (Cytokine Storm), known to cause tissue damage in severely ill SARS-CoV-2 patients (Ulhaq & Soraya, 2020). The simultaneous activation of the NF-κB and IL6-STAT3 activates IL6-AMP, which in turn activates Cytokine Storm (Hirano & Murakami, 2020).

The network was extracted from COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020) 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 examined 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. $Q_{\mathbf{x}} = E[Cytokine|do(sIL6R\alpha) = 20)]$. 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. $Q_{\mathbf{x}} = E[Cytokine|do(EGFR) = 20)]$. The query is not identifiable via either the backdoor or the front-door criterion, but is identified via the do-calculus.

LVM with the correct topology 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 in Fig. 4(a). Relationships between the nodes were modeled as in case study 3. The LVM with misspecified topology had only one latent variable for each dotted edge.

Data were generated from the model with a true θ, and the nodes were simulated with a Hill function as discussed in case study 3. Cytokine storm had a Bernoulli distribution with logit parameterization.

Estimates $\hat{Q}_{\mathbf{x}}$ (Fig. 4 (b-c)) performed as in the other case studies.

5. 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 produce inaccurate results. This manuscript offers an alternative to measuring every molecular component of the cell. By acknowledging the existence of latent variables, and applying Pearl's do-calculus, we can determine whether the causal effect can be identified. We further show that LVM-based estimation of identifiable causal queries is successful even 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 scope of causal inference in biomolecular pathways.

Real biological experiments present many challenges. The underlying data generating process may contain cycles, data may be missing not at random, there may be selection bias and other batch effects, and *in vitro* data may differ significantly from *in vivo* measurements (Bareinboim & Pearl, 2016; Sherman et al., 2020; Nabi et al., 2020; Forré & Mooij, 2019). While future work will address these threats to validity, we have systematically taken several steps towards practical use. We evaluated the accuracy of our approach on real biological pathways (Gyori et al., 2017; Ostaszewski & Niarakis, 2021) in the context of a complex environment that contains many unobserved confounders, and proposed the open-source Y_0 implementation to determine when a causal query is still identifiable according to the do-calculus. The LVMs are compatible with realistic probability distributions describing diverse regulatory events. Parameters

of LVMs must be estimated from experimental data with a sufficient number of replicates, and modern single-cell technologies make these data types increasingly available.

The proposed approach opens the door to many directions of future methodological research, and to many applications. For example, LVMs can directly incorporate informative priors regarding latent variables, or regarding the processes that govern the regulatory events, thus improving model accuracy. The estimation of parameters of latent variables can help us characterize pathway components even when they are unmeasured. Further implementation improvements can help address the remaining limitations of LVM-based causal query estimation. In particular, Y_0 can be more tightly integrated with the LVM-based estimation, enabling automated estimation of general classes of LVM. For LVMs, a limitation is the requirement for parametric assumptions, which can introduce a bias when the assumptions are not justified. These difficulties may be navigated with traditional model evaluation techniques, such as posterior predictive checks, model selection, and relying on Occam's Razor to favor the simplest LVM.

Overall, although molecular biology is in a golden age of intense data accumulation, the problem of unmeasured confounders remains. In this context, we believe that LVM-based estimation of causal queries and its future extensions will have a strong impact.

6. Acknowledgments

JZ is supported by the PNNL Directed R&D Initiative. PNNL is operated for the DOE by Battelle Memorial Institute under Contract DE-AC05-76RLO1830. CTH is supported by the DARPA Young Faculty Award W911NF2010255 (PI: Benjamin M. Gyori). KS is supported by Muscular dystrophy association grant # 574137 and Answer ALS consortium. OV acknowledges the support of NSF-BIO/DBI 1759736, NSF-BIO/DBI 1950412, NIH-NLM-R01 1R01LM013115 and of the Chan-Zuckerberg foundation.

References

Alon, U. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. CRC Press, 2019.

Bareinboim, E. and Pearl, J. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113:7345, 2016.

Bhattacharya, R., Nabi, R., and Shpitser, I. Semiparametric inference for causal effects in graphical models with hidden variables, 2020.

Bianconi, F., Baldelli, E., Ludovini, V., Crino, L., Flacco, A., and Valigi, P. Computational model of EGFR and IGF1R pathways in lung cancer: a systems biology approach for translational oncology. *Biotechnology Advances*, 30:142, 2012.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

Blei, D. M. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203, 2014.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112:859, 2017.

Cannon, W. R., Britton, S., and Alber, M. Cracking the code of metabolic regulation in biology using maximum entropy/caliber and reinforcement learning. pp. 9867. MDPI.

D'Amour, A. On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives, 2019.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. Hybrid Monte Carlo. *Physics Letters B*, 195:216, 1987.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. *Biological sequence analysis*. Cambridge University Press.

Eberhardt, F. and Scheines, R. Interventions and causal inference. *Philosophy of Science*, 74:981, 2007.

Ernst, J., Vainas, O., Harbison, C. T., Simon, I., and Bar-Joseph, Z. Reconstructing dynamic regulatory maps. pp. 74.

Evans, R. J. Graphs for margins of Bayesian networks. *Scandinavian Journal of Statistics*, 43:625, 2016.

Forré, P. and Mooij, J. M. Causal calculus in the presence of cycles, latent confounders and selection bias. 2019.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian Data Analysis*. CRC Press, 3rd edition, 2014.

Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81: 2340, 1977.

Girolami, M. and Calderhead, B. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society B*, 73:123, 2011.

Gyori, B. M., Bachman, J. A., Subramanian, K., Muhlich, J. L., Galescu, L., and Sorger, P. K. From word models to executable models of signaling networks using automated assembly. *Molecular Systems Biology*, 13, 2017.

Helske, J., Tikka, S., and Karvanen, J. Estimation of causal effects with small data in the presence of trapdoor variables, 2021.

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

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303, 2013.

Hoyt, C. T., Konotopez, A., and Ebeling, C. PyBEL: a computational framework for Biological Expression Language. *Bioinformatics (Oxford, England)*, 34(4):703–704, feb 2018. doi: 10.1093/bioinformatics/btx660.

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

Hughes, M. D., Daniels, M. J., Fischl, M. A., Kim, S., and Schooley, R. T. 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, 1998.

Jung, Y., Tian, J., and Bareinboim, E. Learning causal effects via weighted empirical risk minimization. *Advances in Neural Information Processing Systems*, 33, 2020.

Jung, Y., Tian, J., and Bareinboim, E. Estimating identifiable causal effects through double machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Keseler, I. M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martínez, C., Fulcher, C., Huerta, A. M., Kothari, A., Krummenacker, M., et al. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Research*, 41:D605, 2013.

Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martínez, C., Caspi, R., Fulcher, C., Gama-Castro, S., Kothari, A., Krummenacker, M., et al. The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Research*, 45:D543, 2017.

Kingma, D. P. and Welling, M. Stochastic gradient VB and the variational auto-encoder. In *Second International Conference on Learning Representations, ICLR*, volume 19, 2014.

Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Kondofersky, I., Fuchs, C., and Theis, F. J. Identifying latent dynamic components in biological systems. pp. 193–203.

Kuroki, M. and Pearl, J. Measurement bias and effect restoration in causal inference. *Biometrika*, 101:423, 2014.

Lattimore, F. and Rohde, D. Replacing the do-calculus with Bayes rule. *arXiv*, 2019a.

Lattimore, F. and Rohde, D. Causal inference with Bayes rule. *arXiv*, 2019b.

Lauritzen, S. L. and Spiegelhalter, D. J. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society B*, 50:157, 1988.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pp. 6446, 2017.

McNaughton, A. D., Bredeweg, E. L., Manzer, J., Zucker, J., Munoz Munoz, N., Burnet, M. C., Nakayasu, E. S., Pomraning, K. R., Merkley, E. D., Dai, Z., Chrisler, W. B., Baker, S. E., St John, P. C., and Kumar, N. Bayesian inference for integrating yarrowia lipolytica multiomics datasets with metabolic modeling.

Nabi, R., Bhattacharya, R., and Shpitser, I. Full law identification in graphical models of missing data: Completeness results, 2020.

Ostaszewski, M. and Niarakis, A. e. a. COVID19 disease map, a computational knowledge repository of virus–host interaction mechanisms. *Molecular Systems Biology*, oct 2021.

Pearl, J. Bayesian analysis in expert systems: comment: graphical models, causality and intervention. *Statistical Science*, 8:266, 1993.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82:669, 1995.

Pearl, J. *Causality*. Cambridge University Press, 2009.

Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

Pearl, J. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62: 54, 2019.

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

Richardson, T. S., Evans, R. J., Robins, J. M., and Shpitser, I. Nested markov properties for acyclic directed mixed graphs, 2017.

Robert, C. P. and Casella, G. Monte carlo statistical methods. 2004, 2004.

Schulman, L. and Srivastava, P. Stability of causal inference . UAI, 2016.

Sherman, E. S., Arbour, D., and Shpitser, I. General identification of dynamic treatment regimes under interference. *Proceedings of machine learning research*, 108: 3917–3927, aug 2020.

Shojaie, A. and Michailidis, G. Analysis of gene sets based on the underlying regulatory network. pp. 407–426.

Shpitser, I. and Pearl, J. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, pp. 1219, 2006.

Shpitser, I. and Pearl, J. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941, 2008.

Shpitser, I., Richardson, T. S., Robins, J. M., and Evans, R. Parameter and structure learning in Nested Markov Models. *arXiv*, 2012.

Shpitser, I., Evans, R. J., Richardson, T. S., and Robins, J. M. Introduction to nested markov models. *Behaviormetrika*, 41(1):3–39, Jan 2014.

Slater, T. Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discovery Today*, 19:193, 2014.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, Prediction, and Search*. MIT press, 2000.

St John, P. C., Strutz, J., Broadbelt, L. J., Tyo, K. E. J., and Bomble, Y. J. Bayesian inference of metabolic kinetics from genome-scale multiomics data. pp. e1007424.

Team, S. D. RStan: the R Interface to Stan. R package version 2.17. 3, 2018.

Tikka, S. and Karvanen, J. Identifying causal effects with the R package causaleffect. *Journal of Statistical Software*, 76(12):1–30, 2017. doi: 10.18637/jss.v076.i12.

Ulhaq, Z. S. and Soraya, G. V. Interleukin-6 as a potential biomarker of COVID-19 progression. *Medecine et Maladies Infectieuses*, 50:382, 2020.

Wainwright, M. J. and Jordan, M. I. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A., Wang, K., Wang, N. X. R., Wilhelm, C., Xie, B., Raymond, D., Weld, D. S., Etzioni, O., and Kohlmeier, S. CORD-19: The COVID-19 Open Research Dataset, 2020.

Wang, Y. and Blei, D. M. The blessings of multiple causes. *Journal of the American Statistical Association*, pp. 1, 2019.

Wilkinson, D. Package smfsb. 2018.

Zhang, C., Wu, Z., Li, J.-W., Zhao, H., and Wang, G.-Q. Cytokine release syndrome in severe COVID-19: Interleukin-6 receptor antagonist Tocilizumab may be the key to reduce mortality. *International Journal of Antimicrobial Agents*, 55:105954, 2020.

Zucker, J., Paneri, K., Mohammad-Taheri, S., Bhargava, S., Kolambkar, P., Bakker, C., Teuton, J., Hoyt, C. T., Oxford, K., Ness, R., and Vitek, O. Leveraging structured biological knowledge for counterfactual inference: A case study of viral pathogenesis. *IEEE Transactions on Big Data*, 7:25, 2021.