Deep End-to-end Causal Inference

Tomas Geffner[† 1 *] **Javier Antoran**[† 2 *] **Adam Foster**[3 *] **Wenbo Gong**[3] **Chao Ma**[3] **Emre Kiciman**[3]
Amit Sharma[3] **Angus Lamb**[† 4] **Martin Kukla**[3] **Nick Pawlowski**[3] **Miltiadis Allamanis**[3] **Cheng Zhang**[3]

Abstract

Causal inference is essential for data-driven decision making across domains such as business engagement, medical treatment or policy making. However, research on causal discovery and inference has evolved separately, and the combination of the two domains is not trivial. In this work, we develop Deep End-to-end Causal Inference (DECI), a single flow-based method that takes in observational data and can perform both causal discovery and inference, including conditional average treatment effect (CATE) estimation. We provide a theoretical guarantee that DECI can recover the ground truth causal graph under mild assumptions. In addition, our method can handle heterogeneous, real-world, mixed-type data with missing values, allowing for both continuous and discrete treatment decisions. Moreover, the design principle of our method can generalize beyond DECI, providing a general End-to-end Causal Inference (ECI) recipe, which enables different ECI frameworks to be built using existing methods. Our results show the superior performance of DECI when compared to relevant baselines for both causal discovery and (C)ATE estimation in over a thousand experiments on both synthetic datasets and other causal machine learning benchmark datasets.

1. Introduction

Causal-aware decision making is pivotal in many fields such as economics (Zhang & Chan, 2006; Battocchi et al., 2021) and healthcare (Tu et al., 2019; Bica et al., 2019; Huang, 2021). For example, in healthcare, caregivers may wish to understand the effectiveness of different treatments given only historical data when randomized control trials cannot be carried out. In this case, we must estimate causal effects from historical data alone, with no or partial knowledge about the causal relationships between variables. This is the *end-to-end causal inference* problem. We must go from observational data to an estimate of causal quantities, such as average treatment effect (ATE) and conditional average treatment effect (CATE), with no, or incomplete, *a priori* knowledge of the causal graph.

Existing methods for estimating causal quantities from data, which we refer to as *causal inference methods*, require complete *a priori* knowledge of the causal graph. On the other hand, existing *causal graph discovery methods* often return a set of plausible graphs under a set of assumptions on the statistical properties of the data (Spirtes & Glymour, 1991), as shown in Figure 1. As causal inference and causal discovery have been developed separately, their assumptions are often not aligned, making the task of answering causal queries in an end-to-end manner highly non-trivial. Moreover, this becomes even more challenging in real-world scenarios, where the observational data may have missing values and mixed data types.

In this work, we tackle the question of end-to-end causal inference (ECI). We aim to provide a framework that can help practitioners optimize their actions using only observational data as input. Our contributions are:

- *A deep learning-based end-to-end causal inference framework named DECI, which can be used in general real-world scenarios.* DECI is a flow-based method that uses variational inference and extends previous work in functional causal discovery (Zheng et al., 2020; Lachapelle et al., 2019). It can model complex nonlinear relationships between variables and noise distributions, and can efficiently handle missing values and mixed data types. By taking a probabilistic approach, DECI learns an approximate posterior distribution over causal graphs. Additionally, we show how the functions learnt by DECI can later be used for simulation-based estimation of ATE and CATE. Theoretically, we show that with infinite data and under certain assumptions, DECI recovers the true underlying structure and the data generation process.
- *A general recipe for end-to-end causal inference that allows arbitrary combinations of causal discovery and*

[*]Equal contribution. † contributed during internship or residency in Microsoft Research [1]University of Massachusetts Amherst [2]University of Cambridge [3]Microsoft Research [4]G-Research. Correspondence to: Cheng Zhang <cheng.zhang@microsoft.com>.

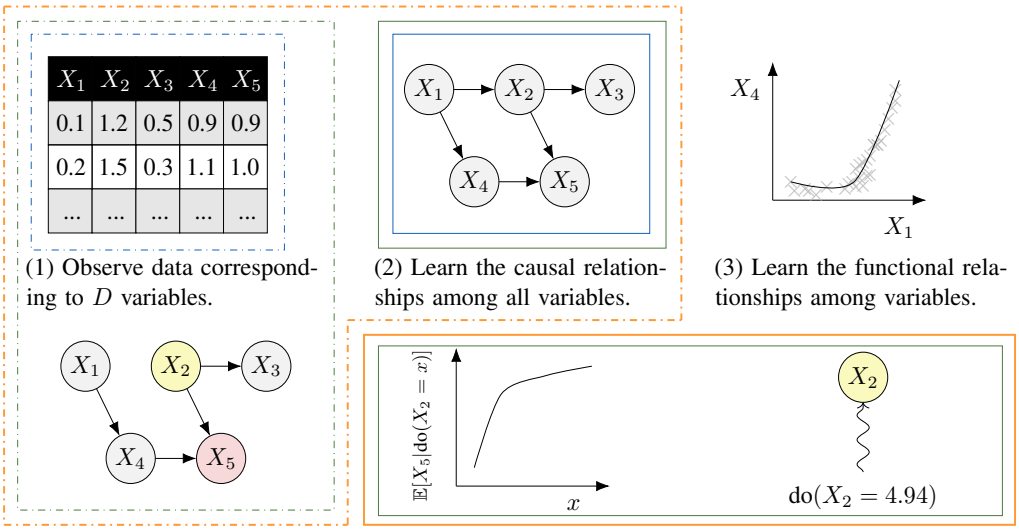


Figure 1. An overview of our end-to-end causal inference pipeline compared to traditional causal discovery and causal inference. The dashed line boxes show the inputs and the solid line boxes show the outputs. In causal discovery, a user provides observational data (1) as input. The output is mainly the causal relationship (2) which typically are DAGs or partial DAGs. In causal inference, the user needs to provide both the data (1) and the causal graph (2) as input and provide a causal question by specifying treatment and effect (4), a model is learned and outputs the causal quantities (5) which helps decision making (6). In this work, we aim to answer causal questions end-to-end. DECI allows the user to provide the observational data only and specify any causal questions and output both the discovered causal relationship (2) and the causal quantities (5) that helps decision making (6).

causal inference methods. We provide a unified view for deep learning-based causal discovery methods, showing that that DECI generalizes a large number of them, including linear and nonlinear variants of *Notears* (Zheng et al., 2018a; 2020) and *Grandag* (Lachapelle et al., 2019), among others (Ng et al., 2019; 2020). More importantly, we show that simulation-based causal inference can be applied beyond DECI and its deep learning-based framework. This allows us to not only combine DECI with other discovery and inference methods but also to combine arbitrary existing causal discovery and inference methods.

- *Insights into end-to-end inference performance with more than 1000 experiments.* We evaluated DECI, along with a large range of causal discovery methods, and many potential combinations of discovery and inference algorithms in systematic experiments. We show that, with our general framework, a user can perform ECI efficiently. With DECI in particular, we obtain very competitive causal discovery and inference performance, outperforming strong baselines in both domains.

2. Related Work and Preliminaries

Related Work. Our work relates to both causal discovery and causal inference research. Approaches for causal discovery from observational data can be classified into three broad groups: constraint-based, score-based, and functional causal models (Glymour et al., 2019). Recently, Zheng et al. (2018a) framed the causal discovery problem as a continuous optimisation problem in the linear setting; this has been extended to work with flexible nonlinear function approximators, such as neural networks (Zheng et al., 2020; Lachapelle et al., 2019; Ng et al., 2019). Despite good performance, the applicability of these methods to real-world scenarios remains limited, as they cannot handle mixed data types nor datasets with missing values. These functional causal models aside, the functional relationships between variables (as shown in Figure 1(3)) are not typically learned by causal discovery algorithms, making subsequent inference of causal quantities non-trivial. Our framework addresses all these points regarding real-world applicability and extends this to the ECI setting.

On the other hand, causal inference methods assume that either the graph structure is provided (Pearl, 2009) or relevant structural assumptions are provided without the graph (Imbens & Rubin, 2015). In both cases, causal inference can be decomposed into two steps: identification and estimation. Identification focuses on converting the causal estimand (e.g. $P(Y|\text{do}(X = x), W)$) into an estimand that can be estimated using the observed data distribution (e.g. $P(Y|X, W)$). Common examples of identification methods include the back-door criterion, front-door criterion (Pearl, 2009), and instrumental variables (Angrist et al., 1996). Causal estimation computes the identified estimand using statistical methods, such as simple conditioning, inverse propensity weighting (Li et al., 2018), or match-

ing (Stuart, 2010; Rosenbaum & Rubin, 1983). Machine learning-based estimators for CATE have also been proposed (Chernozhukov et al., 2018; Wager & Athey, 2018). Recently, there have been efforts to weaken the graph requirement (Jung et al., 2021), so instead of a single directed acyclic graph (DAG), partially directed acyclic graphs (PAGs) and completed partially directed acyclic graphs (CPDAGs) can be considered as well, but this only considers one type of graph at a time, and is not generalized to a distribution over graphs. Our work proposes an end-to-end inference solution where the causal quantities can be estimated without a user-specified graph. In addition, our ECI framework can handle distributions over graphs.

Structural Equation Models. Let $\mathbf{x} = (x_1, \dots, x_D)$ be a collection of random variables. Structural equation models (SEM) (Pearl, 2009) are often used to model causal relationships between the individual variables x_i. Given a DAG G on nodes $\{1, \dots, D\}$, $\mathbf{x}$ can be described by $x_i = F_i\left(\mathbf{x}_{\text{pa}(i;G)}, \varepsilon_i\right)$, where ε_i is a noise random variable that is independent of all other variables in the model, $\text{pa}(i; G)$ is the set of parents of node i in G, and F_i specifies how variable x_i depends on its parents and the noise ε_i. In this paper, we focus on additive noise SEMs, i.e.

$$F_i\left(\mathbf{x}_{\text{pa}(i;G)}, \varepsilon_i\right) = f_i\left(\mathbf{x}_{\text{pa}(i;G)}\right) + \varepsilon_i. \tag{1}$$

For simplicity, we can write this in the vector form

$$\mathbf{x} = f_G(\mathbf{x}) + \boldsymbol{\varepsilon}. \tag{2}$$

Average Treatment Effects. To make a decision about an action, ATE and CATE estimates are often required. Assume that $\mathbf{x}_T$ (with $T \subset \{1, \dots, D\}$) are the treatment variables; the interventional distribution is denoted $p(\mathbf{x} \mid \text{do}(\mathbf{x}_T = \mathbf{a}))$. The ATE (Pearl, 2009) of treatment $\mathbf{a}$ and reference $\mathbf{b}$ of $\mathbf{x}_T$ on targets $\mathbf{x}_Y$ is given by

$$\begin{aligned}\text{ATE}(\mathbf{a}, \mathbf{b}) =& \mathbb{E}_{p(\mathbf{x}_Y|\text{do}(\mathbf{x}_T=\mathbf{a}))}[\mathbf{x}_Y] \\ &- \mathbb{E}_{p(\mathbf{x}_Y|\text{do}(\mathbf{x}_T=\mathbf{b}))}[\mathbf{x}_Y],\end{aligned} \tag{3}$$

and the CATE of $\mathbf{x}_T$ on $\mathbf{x}_Y$ conditional on $\mathbf{x}_C = \mathbf{c}$ is

$$\begin{aligned}\text{CATE}(\mathbf{a}, \mathbf{b}|\mathbf{c}) =& \mathbb{E}_{p(\mathbf{x}_Y|\text{do}(\mathbf{x}_T=\mathbf{a}),\mathbf{x}_C=\mathbf{c})}[\mathbf{x}_Y] \\ &- \mathbb{E}_{p(\mathbf{x}_Y|\text{do}(\mathbf{x}_T=\mathbf{b}),\mathbf{x}_C=\mathbf{c})}[\mathbf{x}_Y].\end{aligned} \tag{4}$$

3. DECI: Deep End-to-end Causal Inference

We now introduce DECI, an end-to-end deep learning-based causal inference framework. DECI provides both a method to learn a distribution over causal graphs from observational data, and a method to estimate causal quantities.

We assume that the data are generated from an underlying nonlinear additive noise model. For causal discovery, DECI learns complex non-linear relationships between variables jointly with a distribution over causal graphs. DECI is a flow-based model, extending the work of Khemakhem et al. (2021) to automatically discover causal relationships among multiple variables. Additionally, DECI fills the gaps in previous discovery approaches by allowing non-Gaussian exogenous noise, mixed type data (continuous and discrete), and partially observed data. For causal inference, we show how the generative model learnt by DECI can be used to simulate samples from intervened distributions. This represents a novel general framework for performing causal inference via simulations, instead of estimating these quantities based on identified rules from the graphs.

3.1. DECI and Causal Discovery

DECI takes a Bayesian approach to causal discovery (Heckerman et al., 1999). We model the causal graph G jointly with the observations $\mathbf{x}^1, \dots, \mathbf{x}^N$ as

$$p_\theta(\mathbf{x}^1, \dots, \mathbf{x}^N, G) = p(G) \prod_n p_\theta(\mathbf{x}^n|G). \tag{5}$$

After learning, the posterior distribution, $p_\theta(G|x^1, \dots, x^N)$ characterizes our beliefs about the causal structure. We now detail each of the model components in eq. (5).

Prior over Graphs. The graph prior $p(G)$ should characterize the graph as a DAG. We implement this by leveraging the continuous DAG penalty from Zheng et al. (2018a),

$$h(G) = \text{tr}\left(e^{G \odot G}\right) - D, \tag{6}$$

which is non-negative and zero only if G is a DAG. We then implement the prior as

$$p(G) \propto \exp\left(-\lambda_s \|G - W_0\|_F^2 - \rho\, h(G)^2 - \alpha\, h(G)\right), \tag{7}$$

where we weight the DAG penalty by α and ρ, which are gradually increased to favour DAGs during training. Additionally, this distribution admits the use of prior knowledge about the causal structure in the form of a weighted adjacency matrix $W_0 \in [0, 1]^{D \times D}$, with zero entries encouraging sparser graphs. The scalar λ_s then regulates the strength of enforcement of these prior beliefs.

Likelihood of Structural Equation Model. For the likelihood term $p_\theta(\mathbf{x}^n|G)$ we follow Khemakhem et al. (2021) and set it to an autoregressive flow with base distribution p_z (Gaussian with learnable variances, we consider other noise models in Section 3.4) and transformation $z = g_G(\mathbf{x}; \theta) = \mathbf{x} - f_G(\mathbf{x}; \theta)$, which is guaranteed to be invertible if G is a DAG. Then, using the change of variable formula for random variables we get

$$p_\theta(\mathbf{x}^n|G) = p_z\left(g_G(\mathbf{x}^n; \theta)\right), \tag{8}$$

where we have omitted the change of volume term for $g_G(\mathbf{x}; \theta)$ because the Jacobian-determinant is always equal

to one for DAGs G (Mooij et al., 2011). We formalize this in Lemma 1, proved in Appendix D.

Lemma 1. *Let G represent a binary adjacency matrix, $f_G : \mathbb{R}^d \to \mathbb{R}^d$ a function whose i-th output only depends on the parents of x_i, and $J_G(x)$ the Jacobian of $g_G(x) = x - f_G(x)$. If G is a DAG, then $|\det J_G(x)| = 1$.*

The choice for $f_G : \mathbb{R}^d \to \mathbb{R}^d$ must satisfy the adjacency relations specified by G. That is, if there is no edge $j \to i$ in G, then the function $f_i(x)$—the i-th component of the output of $f_G(x)$— must satisfy $\partial f_i(x)/\partial x_j = 0$. Inspired by Graph Neural Networks (Hamilton, 2020), we propose a flexible parameterization that satisfies this by setting

$$f_i(x) = h_i\left(\sum_{j=1}^{d} G_{j,i}\ \ell_j(x_j)\right), \tag{9}$$

where $G_{j,i} \in \{0, 1\}$ indicates the presence of the edge $j \to i$, and ℓ_i and h_i ($i = 1, \dots, d$) are MLPs. A naïve implementation would require training $2D$ neural networks. To avoid this, we parameterize these functions as $h_i(\cdot) = h(\mathbf{u}_i, \cdot)$ and $\ell_i(\cdot) = \ell(\mathbf{u}_i, \cdot)$, where $\mathbf{u}_i \in \mathbb{R}^D$ is a trainable embedding. This reduces the number of MLPs to just two.

Optimization and Inference Details. There are two challenges to work with the model described above. First, the true posterior over G is intractable. Second, maximum likelihood cannot be used to fit the model parameters, due to the presence of the latent variable G. We overcome these challenges using variational inference (Jordan et al., 1999; Blei et al., 2017; Zhang et al., 2018). We define a variational distribution $q_\phi(G)$ to approximate the intractable posterior $p_\theta(G|\mathbf{x}^1, \dots, \mathbf{x}^N)$, and use it to build the evidence lower bound (ELBO), a lower bound on the marginal likelihood that can be used as a surrogate objective. This is given by

$$\begin{aligned}\text{ELBO}(\theta, \phi) &= \mathbb{E}_{q_\phi(G)}\left[\log p(G) \prod_n p_\theta(\mathbf{x}^n|G)\right] + H(q_\phi) \\ &\leq \log p_\theta(\mathbf{x}^1, \dots, \mathbf{x}^N),\end{aligned} \tag{10}$$

where $H(q_\phi)$ represents the entropy of the distribution q_ϕ and $p_\theta(\mathbf{x}^n|G)$ takes the form of eq. (8). (Details for the derivation in Appendix B.2.) Specifically, we define $q_\phi(G)$ as the product of independent Bernoulli distributions, one for the existence and one for the orientation of each possible edge in G. Then, the SEM parameters θ and variational parameters ϕ are trained by maximizing the ELBO, using the Gumbel-softmax trick (Maddison et al., 2016; Jang et al., 2016) to get stochastic estimates of the ELBO gradient. We describe the full optimization procedure in Appendix B.1.

3.2. Theoretical Considerations for DECI

We now answer the question: when maximizing the ELBO in eq. (10) with infinite data, does DECI recover both the ground truth data generating process $p(\boldsymbol{x}; G^0)$ and the ground truth graph G^0? Here, we show that DECI recovers the truth under mild assumptions. We achieve this by showing (i) DECI's maximum likelihood estimate (MLE) of (θ, G) recover $p(\boldsymbol{x}; G^0)$; (ii) solutions from maximizing the ELBO are closely related to MLE. Specifically, we show that DECI induces the same joint likelihood as the ground truth and the posterior $q_\phi(G)$ is a delta function $\delta(G = G^0)$ concentrated on the ground truth G^0. This is formalized in the following theorem; the proof is in Appendix A.

Theorem 1 (DECI recovers true distribution)**.** *Under assumptions 1-5 (Appendix A), the solution $(\theta', q'_\phi(G))$ from maximizing the ELBO (eq. (10)) satisfies $q'_\phi(G) = \delta(G = G')$ where G' is a unique graph. In particular, $G' = G^0$ and $p_{\theta'}(\mathbf{x}; G') = p(\mathbf{x}; G^0)$.*

3.3. Estimating Causal Quantities

We now show how the generative model learnt by DECI can be used to evaluate expectations under interventional distributions, and thus estimate ATE and CATE. As explained above, DECI returns $q_\phi(G)$ an approximation of the posterior over graphs given data. Then, interventional distributions can be obtained by marginalizing over graphs

$$\mathbb{E}_{q_\phi(G)}\left[p\left(\mathbf{x}_Y|\text{do}(\mathbf{x}_T{=}\mathbf{a}), G\right)\right]. \tag{11}$$

Similarly, treatment effects can be obtained as

$$\mathbb{E}_{q_\phi(G)}[\text{ATE}(\mathbf{a}, \mathbf{b}|G)], \quad \mathbb{E}_{q_\phi(G)}[\text{CATE}(\mathbf{a}, \mathbf{b}|\mathbf{c}, G)]. \tag{12}$$

Our framework can be seen as a probabilistic relaxation of traditional causal quantity estimators. When we are certain about the causal graph, i.e. $q_\phi(G) = \delta(G - G_i)$, it matches traditional causal inference. Next, we discuss exactly how DECI estimates ATE and the more challenging CATE.

Estimating ATE. After training, we can use the model learnt by DECI to simulate new samples $\mathbf{x}$ by first sampling a graph $G \sim q_\phi(G)$ and a set of exogenous noise variables $\mathbf{z} \sim p_z$, and using them to generate a sample using DECI's learned structural equation model, following the topological order defined by G. ATE estimation requires samples from interventional distributions $p\left(\mathbf{x}_{\setminus T}|\text{do}(\mathbf{x}_T = \mathbf{b}), G\right)$, which can be sampled from by noting that

$$p\left(\mathbf{x}_{\setminus T}|\text{do}(\mathbf{x}_T = \mathbf{b}), G\right) = p\left(\mathbf{x}_{\setminus T}|\mathbf{x}_T = \mathbf{b}, G_{\text{do}(\mathbf{x}_T)}\right),$$

where $G_{\text{do}(\mathbf{x}_T)}$ is the "mutilated" graph obtained by removing incoming edges to $\mathbf{x}_T$. Then, samples from this distribution can be obtained by following the sampling procedure explained above, but fixing the values $\mathbf{x}_T = \mathbf{b}$ and using $G_{\text{do}(\mathbf{x}_T)}$ instead of G. We then use these samples to obtain a Monte Carlo estimate of the expectations required for ATE computation, defined in eq. (3).

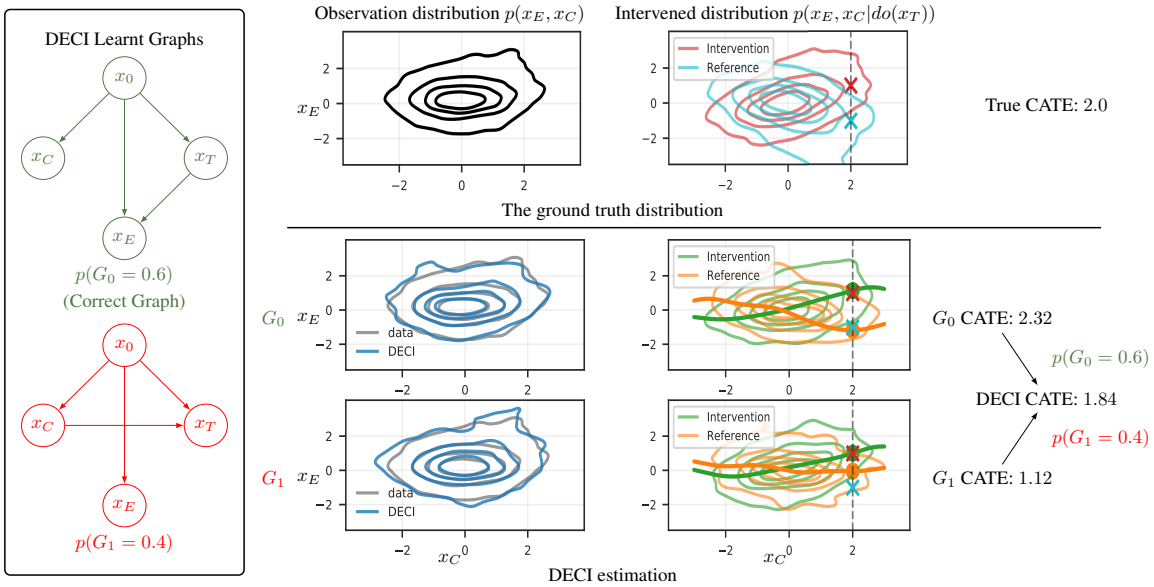


Figure 2. Illustration of DECI CATE estimation on CSuite symprod-simpson dataset. On the left, DECI's posterior has two modes with $p(G) = 0.6$ for the correct graph and $p(G) = 0.4$ for an alternative possibility with some incorrect edges. On the right side, we display the joint distribution of conditioning and effect variables in the observational setting and under interventions on $\mathbf{x}_T$. DECI captures the observational density well. The interventional distributions are shown on the right with their conditional means $\mathbf{x}_C = \mathbf{c}$ marked with crosses. DECI predicts conditional expectations by fitting functions from $\mathbf{x}_C$ to $\mathbf{x}_E$ and evaluating them at $\mathbf{c}$. DECI outputs CATE by marginalizing the result over possible graphs.

Estimating CATE. Let the index set $Y = X \setminus (T \cup C)$ denote all variables that we do not intervene or condition on. Conditional densities

$$p_\theta(\mathbf{x}_Y \mid \text{do}(\mathbf{x}_T{=}\mathbf{b}), \mathbf{x}_C{=}\mathbf{c}, G) = \frac{p_\theta(\mathbf{x}_Y, \mathbf{x}_C{=}\mathbf{c}|\mathbf{x}_T{=}\mathbf{b}, G_{\text{do}(\mathbf{x}_T)})}{p_\theta(\mathbf{x}_C{=}\mathbf{c}|\mathbf{x}_T{=}\mathbf{b}, G_{\text{do}(\mathbf{x}_T)})} \tag{13}$$

are intractable due to the intractability of the marginal distribution $p_\theta(\mathbf{x}_C{=}\mathbf{c}|\mathbf{x}_T{=}\mathbf{b}, G_{\text{do}(\mathbf{x}_T)})$. Instead, we propose training a model to estimate the intractable distribution from eq. (13). We do this by drawing samples from the joint distribution $p_\theta(\mathbf{x}_Y, \mathbf{x}_C|\mathbf{x}_T{=}\mathbf{b}, G_{\text{do}(\mathbf{x}_T)})$, and using them as training points to fit a surrogate regression model h_G[1] to predict the relationship between $\mathbf{x}_C$ and $\mathbf{x}_Y$, by minimizing

$$\mathbb{E}_{p_\theta(\mathbf{x}_Y, \mathbf{x}_C|\mathbf{x}_T=\mathbf{b}, G_{\text{do}(\mathbf{x}_T)})} \left[\|\mathbf{x}_Y - h_G(\mathbf{x}_C)\|^2\right].$$

Then, the optimal h_G is known to be the conditional mean of $\mathbf{x}_Y$. We set h_G to be a basis-function linear model with random Fourier basis functions (Yu et al., 2016). As illustrated in Figure 2, we train two separate surrogate models for our intervention $\mathbf{x}_T = \mathbf{a}$ and reference $\mathbf{x}_T = \mathbf{b}$. We estimate CATE as the difference between these two surrogate model outputs when evaluated at $\mathbf{x}_C = \mathbf{c}$, and average over posterior graphs

$$\mathbb{E}_{q_\phi(G)} \left[h_{G_{\text{do}(\mathbf{x}_T=\mathbf{a})}}(\mathbf{x}_C = \mathbf{c}) - h_{G_{\text{do}(\mathbf{x}_T=\mathbf{b})}}(\mathbf{x}_C = \mathbf{c}))\right].$$

[1]The subscript G allows us to differentiate the models learnt for different graph samples from $q_\phi(G)$.

Thus, we show that DECI, a flow-based deep learning framework, performs causal discovery and inference end-to-end.

3.4. DECI for Real-world Heterogeneous Data

We extend DECI to handle flexible noise distributions beyond Gaussians, mixed-type data and data with missing values, which are often seen in real-world applications.

Exogenous Noise Model p_z. We propose a flexible model using normalizing flows (Rezende & Mohamed, 2015) for each noise variable z_i, which yields

$$p(\mathbf{z}_i) = \mathcal{N}\left(\psi_i^{-1}(\mathbf{z}_i)|\mathbf{0}, I\right) \left|\frac{\partial \psi_i^{-1}(\mathbf{z}_i)}{\partial \mathbf{z}_i}\right|. \tag{14}$$

In practice, we choose the learnable bijection ψ_i to be a rational quadratic spline (Durkan et al., 2019). Since our SEM formulation requires independent noise variables, each noise component uses independent one-dimensional transformations ψ_i, avoiding couplings across dimensions. This is significantly more flexible than the Gaussian noise model used by previous methods (Zheng et al., 2018a; 2020; Lachapelle et al., 2019; Ng et al., 2019; 2020).

Handling Mixed-type Data. For DECI to support discrete-valued variables, we remove the additive noise model and directly parametrise parent-conditional class probabilities

$$p_\theta^{\text{discrete}}\left(x_i|\mathbf{x}_{\text{pa}(i;G)}; G\right) = P_i\left(\mathbf{x}_{\text{pa}(i;G)}; \theta\right)(x_i) \tag{15}$$

where $P_i\left(\mathbf{x}_{\text{pa}(i;G)}; \theta\right)$ is a normalised probability vector over the number of classes of x_i, obtained by applying the

softmax operator to $f_i(\mathbf{x}_{\text{pa}(i;G)})$. Note that, for discrete variables, the output of f_i is a vector of length equal to the number of classes for variable i.

Handling Missing Data. We propose an extension of DECI to partially observed data.[2] We use $\mathbf{x}_o^n$ to denote the observed components of $\mathbf{x}^n$, $\mathbf{x}_u^n$ to denote the unobserved components, and $p_\theta(\mathbf{x}_o^n, \mathbf{x}_u^n)$ to denote the probability of the sample obtained by combining $\mathbf{x}_o^n$ and $\mathbf{x}_u^n$. We propose to use a distribution to approximate the true posterior over the missing variables,

$$q_{\phi,\psi}\left(G, \mathbf{x}_u^1, \dots, \mathbf{x}_u^N | \mathbf{x}_o^1, \dots, \mathbf{x}_o^N\right) = q_\phi(G) \prod_n q_{\psi_n}(\mathbf{x}_u^n | \mathbf{x}_o^n),$$

and to optimize the resulting ELBO

$$\text{ELBO}(\theta, \phi, \psi) = H(q_\phi) + \sum_n H(q_{\psi_n}) + \mathbb{E}_{q_{\phi,\psi}}\left[\log p(G) \prod_n p_\theta(\mathbf{x}_o^n, \mathbf{x}_u^n | G)\right]. \quad (16)$$

We parameterize the Gaussian imputation distribution $q_{\psi_n}(\mathbf{x}_u^n | \mathbf{x}_o^n)$ using an amortization network (Kingma & Welling, 2013), which receives as input the observed set $\mathbf{x}_o^n$ and outputs the mean and variance of the variational imputation distribution.

3.5. DECI as a General ECI Framework

Here, we present a unified view, showing that DECI generalizes a large number of causal discovery frameworks. In addition, we show that simulation-based Bayesian inference methods for causal inference can be used in a general ECI pipeline, bridging existing causal discovery and inference methods.

Unified View. Most causal discovery methods based on continuous optimization can be framed from a probabilistic perspective as fitting a flow. This connection uses the concept of a weighted adjacency matrix $W(\theta) \in \mathbb{R}^{d\times d}$ linked to a function $f(x;\theta) : \mathbb{R}^d \to \mathbb{R}^d$. Such adjacency matrices can be constructed efficiently for a wide range of parameterizations for f (Zheng et al., 2020).

Lemma 2. *Let $f(\mathbf{x};\theta) : \mathbb{R}^d \to \mathbb{R}^d$ be a θ-parameterized function with weighted adjacency matrix $W(\theta) \in \mathbb{R}^{d\times d}$. Given a dataset $\{\mathbf{x}^1, \dots, \mathbf{x}^N\}$, fitting a flow with the transformation $\mathbf{z} = \mathbf{x} - f(\mathbf{x};\theta)$, base distribution p_z and a hard acyclicity constraint on $W(\theta)$ is equivalent to solving*

$$\max_\theta \sum_{i=1}^n \log p_z(\mathbf{x}^i - f(\mathbf{x}^i;\theta)) \text{ s.t. } h(W(\theta)) = 0, \quad (17)$$

where $h(\cdot)$ is the DAG regularizer from eq. (6).

[2] We assume that values are missing (completely) at random, which is a common assumption (Rubin, 1976; Stekhoven & Bühlmann, 2012; Ma et al., 2018; Strobl et al., 2018).

Lemma 2 is the building block in connecting flow-based methods and continuous optimization-based causal discovery, as the objectives used by many such methods (Zheng et al., 2018a; 2020) can be recovered from eq. (17) with different choices for p_z and f. See Appendix C for details.

General ECI Framework. The cornerstones of the DECI framework are our probabilistic treatment of the DAG G, and our use of a simulation based method to estimate causal inference quantities. These general principles can be applied beyond DECI. Many causal discovery methods work by constraining the space of graphs (see Appendix B.5) and outputting a set of DAGs compatible with the data. We can interpret this output as a distribution, and use eq. (11) and eq. (12) to compute causal quantities by marginalizing over this graph distribution. Additionally, given each graph sampled from this distribution, the quantities inside the expectations can be estimated using any existing causal inference method, such as Double ML (Chernozhukov et al., 2018), DoWhy (Sharma et al., 2021), etc. Thus, the DECI framework can be used as a general ECI framework. Our implementation includes, not only our DECI framework, but also a large range of end-to-end causal inference methods combining different existing approaches for causal discovery and causal inference.

4. Experiments

We evaluate DECI on both causal discovery and causal inference tasks.

4.1. Causal Discovery Evaluation

Datasets. We consider synthetic, pseudo-real, and real data. For the synthetic data we follow the approach from Lachapelle et al. (2019) and Zheng et al. (2020). We sample a DAG following two different random graph models, **Erdős-Rényi (ER)** and **scale-free (SF)**, and simulate each graph $x_i = f_i(\mathbf{x}_{\text{pa}(i;G)}) + z_i$, where f_i is a nonlinear function (randomly sampled spline). We consider two noise distributions, a standard Gaussian and a more complex one obtained by transforming samples from a standard Gaussian with a randomly sampled MLP. We consider two dimensionalities $d \in \{16, 64\}$ with number of edges $e \in \{d, 4d\}$. The resulting datasets are identified as $\mathbf{ER}(d, e)$ and $\mathbf{SF}(d, e)$. All datasets have $n = 5000$ training samples.

For the pseudo-real dataset we consider the **SynTReN** generator (Van den Bulcke et al., 2006), which creates synthetic transcriptional regulatory networks and produces simulated gene expression data that approximates experimental data. We use the datasets generated by Lachapelle et al. (2019) (dimension 20), and take $n = 400$ for training. Finally, for the real dataset, we use the protein measurements in human cells from Sachs et al. (2005). We use a training set with

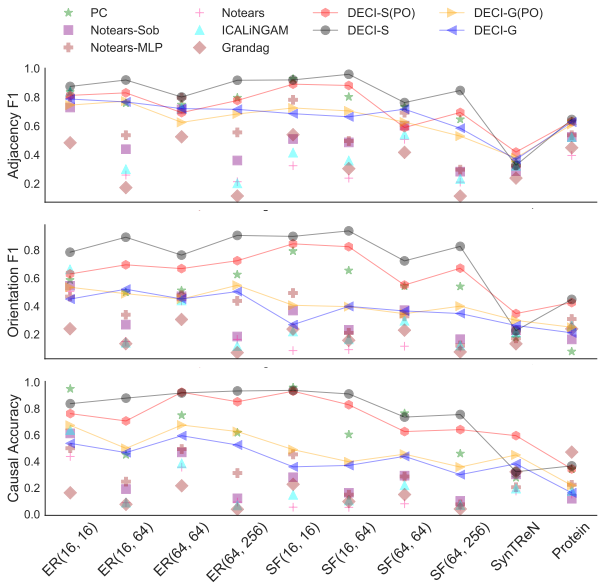


Figure 3. **DECI achieves better results than the baselines in all metrics shown.** The "(PO)" corresponds to running DECI with 30% of the training data missing. For readability, we highlight the DECI results by connecting them with soft lines. The figure shows mean results across five different random seeds.

$n = 800$ observational samples of dimension $d = 11$.

Baselines. We run DECI using two models for exogenous noise: a Gaussian with learnable variance (identified as DECI-G) and a spline flow (DECI-S). We compare against *PC* (Kalisch & Bühlman, 2007), (linear) *Notears* (Zheng et al., 2018a), the nonlinear variants *Notears-MLP* and *Notears-Sob* (Zheng et al., 2020), *Grandag* (Lachapelle et al., 2019), and *ICALiNGAM* (Shimizu et al., 2006). When a CPDAG is the output, e.g., from PC, we treat all possible DAGs under the CPDAG as having the same probability. All baselines are obtained from `gcastle` (Zhang et al., 2021).

Causality Metrics. We report F1 scores for adjacency and orientation (Glymour et al., 2019; Tu et al., 2019) and causal accuracy (Claassen & Heskes, 2012). For DECI, we report the expected values of these metrics estimated using samples from the approximated posterior over the causal graph.

Figure 3 shows the results for the data generated with non-Gaussian noise. We observe that DECI achieves the best results across all metrics. Additionally, using the flexible spline model for the exogenous noise (DECI-S) yields better results than the Gaussian model (DECI-G). This is expected, as the noise used to generate the data is non-Gaussian. For Gaussian noise (see Figure 6), both DECI-S and DECI-G perform similarly. Moreover, when data are partially observed *(PO)*, the strong performance of DECI remains, showing that DECI can handle missing data efficiently.

4.2. End-to-end Causal Inference

We now empirically validate the proposed end-to-end frameworks with different method combinations, including our specific instantiation of the framework: DECI. Our experiments cover handcrafted settings designed to test specific hypotheses, large-scale systematic evaluation and standard causal inference benchmarks. Due to the large number of experimental settings under consideration, this section only provides an overview of the results. For further results and details of the experimental set-up, see Appendices B and E.

Datasets. For causal inference evaluation, we generate samples from random interventions with at most 3 edges between the intervention and the effect variables from the **ER** and **SF** synthetic graphs described in Section 4.1. This allows us to estimate ATE. To obtain a more detailed view of different methods' behaviour, we hand-craft a suite of synthetic causal data generating models, which we name **CSuite**. Different CSuite datasets target different aspects of the end-to-end procedure, such as identifiability of the causal graph, distribution of the exogenous noise or size of the optimal adjustment set. We employ HMC to draw conditional samples from CSuite models, allowing us to evaluate both ATE and CATE. CSuite covers both continuous and mixed type data. See Appendix F.1 for details about CSuite. Finally, we include two causal inference benchmark datasets for ATE evaluation: **Twins** (twin birth datasets in the US) (Almond et al., 2005) and **IHDP** (Infant Health and Development Program data) (Hill, 2011). Due to the semi-synthetic nature of IHDP and Twins, only a subset of the ground truth causal graph is known. Thus, our 'true graph' results are given access to a part of the true graph only. For details of the generation and pre-processing of these datasets, see Appendix F.

Baselines. We evaluate a number of methods within the end-to-end causal inference framework, by combining causal discovery methods from {DECI-Gaussian, DECI-Spline, PC, and True graph}, and causal inference methods from {DECI-Gaussian, DECI-Spline, DoWhy-Linear, DoWhy-Non-linear}, which gives a total of 14 valid combinations. Among them, DECI-Gaussian and DECI-Spline are our proposed methods; true graph directly uses the ground truth causal graph (except for Twins and IHDP, see above) as the input to inference methods to provide performance upper bounds; DoWhy-Linear and DoWhy-Nonlinear (Sharma et al., 2021) implement linear adjustment and Double Machine Learning (DML) (Chernozhukov et al., 2018) methods for backdoor adjustment.

4.2.1. Results Overview

Table 1 shows the average *rank* of different methods across our full range of datasets, providing an overview of our results. In addition, we present ATE/CATE results on a

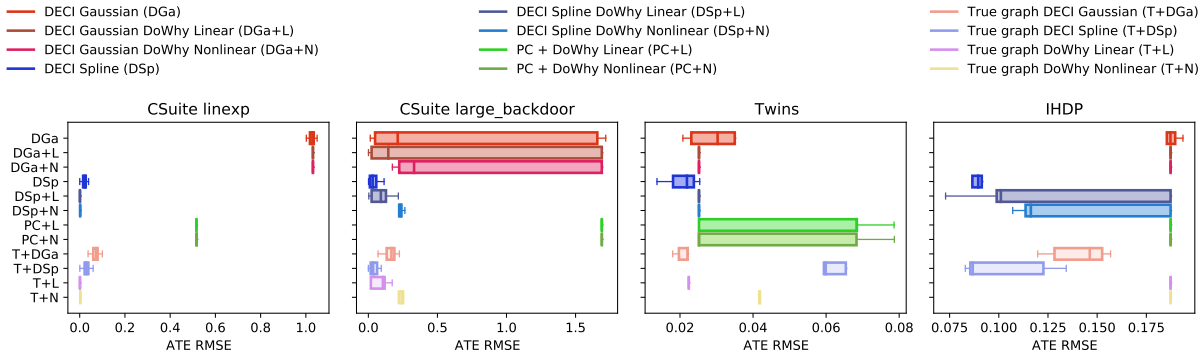


Figure 4. End-to-end ATE results on a range of datasets with different ECI combinations.

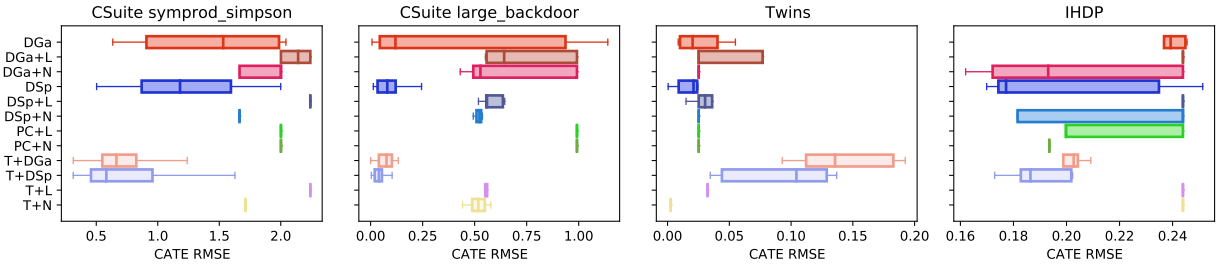


Figure 5. End-to-end CATE results on a range of datasets with different ECI combinations. Colours and acronyms as in Figure 4.

Method	Mean rank
DECI Gaussian (DGa)	$\mathbf{6.26 \pm 0.60}$
DECI Gaussian DoWhy Linear (DGa+L)	8.37 ± 0.50
DECI Gaussian DoWhy Nonlinear (DGa+N)	8.52 ± 0.51
DECI Spline (DSp)	$\mathbf{6.04 \pm 0.68}$
DECI Spline DoWhy Linear (DSp+L)	7.78 ± 0.60
DECI Spline DoWhy Nonlinear (DSp+N)	$\mathbf{6.63 \pm 0.66}$
PC + DoWhy Linear (PC+L)	8.87 ± 0.41
PC + DoWhy Nonlinear (PC+N)	7.54 ± 0.45
True graph DECI Gaussian (T+DGa)	$\mathbf{3.74 \pm 0.47}$
True graph DECI Spline (T+DSp)	$\mathbf{4.19 \pm 0.56}$
True graph DoWhy Linear (T+L)	4.87 ± 0.58
True graph DoWhy Nonlinear (T+N)	5.20 ± 0.71

Table 1. Method rank on different (CSuite, Twins, IHDP and ER/SF) datasets, ranking by median ATE RMSE. We present mean ± 1 s.e. of the rank over 27 datasets. Supporting data in Table 2. Bold indicates the possible top methods, accounting for error bars. We treat methods with access to the true graph separately.

number of datasets in Figures 4 and 5. Comprehensive results can be found in Appendix E. Across all experiments we see that DECI enables end-to-end causal inference with competitive performance. DECI particularly performs well compared to other methods when its ability to handle nonlinear functional relationship and non-Gaussian noise distributions is central to causal discovery *or* causal inference on a given dataset. Other ECI combinations with our framework also achieve strong performance when suitable discovery and inference methods are combined, but have weak performance if either step's assumption is violated. This makes DECI particularly attractive given its high degree of flexibility, which may make it applicable in real-world scenarios.

ATE (Figure 4). DECI-Spline generally provides ATE estimates competitive with baselines that use the well-established backdoor adjustment methods. For larger **CSuite** datasets where the number of possible adjustment sets is large (**CSuite large_backdoor**), DECI-Spline outperforms both linear and non-linear DML method. In general, we find that PC struggles with non-linear functional relationships and heavy-tailed noise distributions present in CSuite. On semi-synthetic datasets (**IHDP** and **Twins**), DECI-Spline outperforms most baselines w.r.t. median ATE RMSE. We also see that 'true graph' methods do not always perform better here, which matches our expectation as these graphs are known to be incomplete. DECI-Spline outperforms its true graph counterpart on IHDP, showing successful discovery has a knock-on effect on ATE.

CATE (Figure 5). On **CSuite**, DECI outperforms DML when estimating CATE, with and without access to the true graph. DECI's advantage is most apparent when the true graph is unavailable, where it seems to be robust to mistakes in causal discovery. We hypothesise that this is due to DECI learning a causal graph and functional relationships among variables simultaneously. When the inferred graph is wrong, functional relationships are learnt so that DECI still obtains a good fit (Figure 2). We see a similar message with **IHDP** and **Twins** results: DECI-Spline outperforms most combinations, except for True graph DECI-Spline on **IHDP** and True graph DoWhy on **Twins**.

5. Conclusion

In this work, we take a holistic view of end-to-end causal inference and propose a novel deep learning-based framework, DECI, designed towards aiding real-world action decision making. Furthermore, we provide a general pipeline for end-to-end causal inference utilizing existing causal discovery and inference methods. We hope that our work bridges the causal discovery and inference communities. In the future, we want to further improve DECI to handle even more challenging real-world ECI problems, e.g. data not missing at random, and the existence of hidden confounders.

Acknowledgements

We would like to thank Vasilis Syrgkanis for insightful discussions regarding causal inference methods and EconML usage; we thank Yordan Zaykov for engineering support; we thank Biwei Huang and Ruibo Tu for feedback that improved this manuscript; we thank Maria Defante, Karen Fassio, Steve Thomas and Dan Truax for insightful discussions on real-world needs which inspired the whole project.

References

Almond, D., Chay, K. Y., and Lee, D. S. The costs of low birth weight. *The Quarterly Journal of Economics*, 120 (3):1031–1083, 2005.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434): 444–455, 1996.

Battocchi, K., Dillon, E., Hei, M., Lewis, G., Oprescu, M., and Syrgkanis, V. Estimating the long-term effects of novel treatments. *arXiv preprint arXiv:2103.08390*, 2021.

Bica, I., Alaa, A. M., Jordon, J., and van der Schaar, M. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2019.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Bühlmann, P., Peters, J., and Ernest, J. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of machine learning research*, 3 (Nov):507–554, 2002.

Chickering, D. M. and Meek, C. Selective greedy equivalence search: Finding optimal bayesian networks using a polynomial number of score evaluations. *arXiv preprint arXiv:1506.02113*, 2015.

Chickering, M. Statistically efficient greedy equivalence search. In *Conference on Uncertainty in Artificial Intelligence*, pp. 241–249. PMLR, 2020.

Claassen, T. and Heskes, T. A bayesian approach to constraint based causal inference. *arXiv preprint arXiv:1210.4866*, 2012.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. Neural spline flows. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/7ac71d433f282034e088473244df8c02-Paper.pdf.

Glymour, C., Zhang, K., and Spirtes, P. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Hamilton, W. L. Graph representation learning. *Synthesis Lectures on Artifical Intelligence and Machine Learning*, 14(3):1–159, 2020.

Heckerman, D., Meek, C., and Cooper, G. A bayesian approach to causal discovery. *Computation, causation, and discovery*, 19:141–166, 1999.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds.), *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia,*

Canada, December 8-11, 2008, pp. 689–696. Curran Associates, Inc., 2008a. URL https://proceedings.neurips.cc/paper/2008/hash/f7664060cc52bc6f3d620bcedc94a4b6-Abstract.html.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., Schölkopf, B., et al. Nonlinear causal discovery with additive noise models. In *NIPS*, volume 21, pp. 689–696. Citeseer, 2008b.

Huang, B. Diagnosis of autism spectrum disorder by causal influence strength learned from resting-state fmri data. In *Neural Engineering Techniques for Autism Spectrum Disorder*, pp. 237–267. Elsevier, 2021.

Huang, B., Zhang, K., Lin, Y., Schölkopf, B., and Glymour, C. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1551–1560, 2018a.

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. Neural autoregressive flows. In *International Conference on Machine Learning*, pp. 2078–2087. PMLR, 2018b.

Hyvärinen, A. and Smith, S. M. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research*, 14(Jan):111–152, 2013.

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

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Jung, Y., Tian, J., and Bareinboim, E. Estimating identifiable causal effects on markov equivalence class through double machine learning. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5168–5179. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jung21b.html.

Kaiser, M. and Sipos, M. Unsuitability of NOTEARS for causal graph discovery. *CoRR*, abs/2104.05441, 2021. URL https://arxiv.org/abs/2104.05441.

Kalisch, M. and Bühlman, P. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

Khemakhem, I., Monti, R., Leech, R., and Hyvarinen, A. Causal autoregressive flows. In *International Conference on Artificial Intelligence and Statistics*, pp. 3520–3528. PMLR, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29:4743–4751, 2016.

Lachapelle, S., Brouillard, P., Deleu, T., and Lacoste-Julien, S. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019.

Li, F., Morgan, K. L., and Zaslavsky, A. M. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

Loh, P.-L. and Bühlmann, P. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. *arXiv preprint arXiv:1705.08821*, 2017.

Ma, C., Tschiatschek, S., Palla, K., Hernández-Lobato, J. M., Nowozin, S., and Zhang, C. Eddi: Efficient dynamic discovery of high-value information with partial vae. *arXiv preprint arXiv:1809.11142*, 2018.

Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Mooij, J. M., Janzing, D., Heskes, T., and Schölkopf, B. On causal discovery with cyclic additive noise model. 2011.

Nemirovsky, A. Optimization ii. numerical methods for nonlinear continuous optimization. 1999.

Ng, I., Zhu, S., Chen, Z., and Fang, Z. A graph autoencoder approach to causal structure learning. *arXiv preprint arXiv:1911.07420*, 2019.

Ng, I., Ghassami, A., and Zhang, K. On the role of sparsity and dag constraints for learning linear dags. *arXiv preprint arXiv:2006.10201*, 2020.

Pearl, J. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

Peters, J. and Bühlmann, P. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Peters, J., Janzing, D., and Schölkopf, B. Identifying cause and effect on discrete data using additive noise models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 597–604. JMLR Workshop and Conference Proceedings, 2010.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. 2014.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Reisach, A. G., Seiler, C., and Weichwald, S. Beware of the simulated dag! varsortability in additive noise models. *CoRR*, abs/2102.13647, 2021. URL https://arxiv.org/abs/2102.13647.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Rubin, D. B. Inference and missing data. *Biometrika*, 63(3): 581–592, 1976.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

Sharma, A., Syrgkanis, V., Zhang, C., and Kıcıman, E. Dowhy: Addressing challenges in expressing and validating causal assumptions. *arXiv preprint arXiv:2108.13518*, 2021.

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

Spirtes, P. and Glymour, C. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

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

Spirtes, P. L. Directed cyclic graphical representations of feedback models. *arXiv preprint arXiv:1302.4982*, 2013.

Stekhoven, D. J. and Bühlmann, P. Missforest—nonparametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Strobl, E. V., Visweswaran, S., and Spirtes, P. L. Fast causal inference with non-random missingness by test-wise deletion. *International journal of data science and analytics*, 6(1):47–62, 2018.

Stuart, E. A. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1): 1, 2010.

Tu, R., Zhang, C., Ackermann, P., Mohan, K., Kjellström, H., and Zhang, K. Causal discovery in the presence of missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1762–1770. PMLR, 2019.

Van den Bulcke, T., Van Leemput, K., Naudts, B., van Remortel, P., Ma, H., Verschoren, A., De Moor, B., and Marchal, K. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics*, 7(1):1–12, 2006.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Yu, F. X. X., Suresh, A. T., Choromanski, K. M., Holtmann-Rice, D. N., and Kumar, S. Orthogonal random features. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/53adaf494dc89ef7196d73636eb2451b-Paper.pdf.

Zhang, C., Bütepage, J., Kjellstrom, H., and Mandt, S. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

ZHANG, K. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 647. AUAI Press, 2009.

Zhang, K. and Chan, L.-W. Extensions of ica for causality discovery in the hong kong stock market. In *International Conference on Neural Information Processing*, pp. 400–409. Springer, 2006.

Zhang, K. and Hyvarinen, A. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

Zhang, K., Wang, Z., Zhang, J., and Schölkopf, B. On estimation of functional causal models: general results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–22, 2015.

Zhang, K., Zhu, S., Kalander, M., Ng, I., Ye, J., Chen, Z., and Pan, L. gcastle: A python toolbox for causal discovery. *arXiv preprint arXiv:2111.15155*, 2021.

Zheng, X., Aragam, B., Ravikumar, P., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning. *arXiv preprint arXiv:1803.01422*, 2018a.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018b. URL https://proceedings.neurips.cc/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf.

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425. PMLR, 2020.

A. Theoretical Considerations for DECI

DECI training can be categorized as a score-based causal discovery approach, which aims to find the model parameters θ and mean-field posterior $q_\phi(G)$ by maximizing the ELBO (eq. (10)). A key statistical property of DECI is whether it is capable of recovering the ground truth data generating distribution and true graph G^0 when DECI is **correctly specified** with infinite data. In the following, we will show that DECI is indeed capable of this under mild assumptions. The main idea is to first show that the maximum likelihood estimate (MLE) recovers the ground truth due to the correctly specified model. Then, we prove that optimal solutions from maximizing the ELBO are closely related to the MLE under mild assumptions.

A.1. Notation and Assumptions

First, let's define the notation and assumptions required for the proof. We denote a random variable $\mathbf{x} \in \mathbb{R}^D$ with a ground truth data generating distribution $p(\mathbf{x}; G^0)$, where G^0 is a binary adjacency matrix representing the true DAG. DECI uses the additive noise model (ANM), defining the structural assignment $x_j = f(\mathbf{x}_{\text{pa}(j;G)}; \theta) + z_j$, where $\text{pa}(j; G)$ are the parents of node j specified by the adjacency matrix G and z_j are mutually independent noise variables with a joint distribution $p_\theta(z_1, \dots, z_D)$. The mean-field variational distribution $q_\phi(G)$ is a product of independent Bernoulli distribution, and $p(G)$ is the soft prior over the graph defined by eq. (7).

Assumption 1 (Minimality)**.** *For a distribution $p_\theta(\mathbf{x}; G)$ generated by DECI with graph G and parameter θ, we assume the minimality condition holds (Spirtes, 2013). Namely, the distribution $p_\theta(\mathbf{x}; G)$ does not satisfy the local Markov condition to any sub-graph of G.*

To satisfy this assumption in practice, one can leverage Proposition 17 from Peters et al. (2014), stating that the minimality condition can be satisfied if the the model is a continuous additive noise model (ANM) and its structural assignments are not a constant w.r.t. any of its arguments. In practice, one can always add edge pruning step to remove such spurious edges (Lachapelle et al., 2019; Bühlmann et al., 2014).

Assumption 2 (DECI Structural Identifiability)**.** *We assume that the model DECI satisfies the structural identifiability. Namely, for a distribution $p_\theta(\mathbf{x}; G)$, the graph G is said to be structural identifiable from $p_\theta(\mathbf{x}; G)$ if there exists no other distribution $p_{\theta'}(\mathbf{x}; G')$ such that $G \neq G'$ and $p_\theta(\mathbf{x}; G) = p_{\theta'}(\mathbf{x}; G')$.*

For general SEM, this assumption does not hold. In fact, one can always search for functions resulting in the independence of cause and mechanisms in both directions (Peters et al., 2017; Zhang et al., 2015). However, by correctly restricting the function family and the form of structural assignments, one can obtain structural identifiability (Shimizu et al., 2006; Hoyer et al., 2008a; Zhang & Hyvarinen, 2012; Peters & Bühlmann, 2014; Peters et al., 2014; 2010). From the formulation of DECI, it is a special case of the non-linear ANM (Hoyer et al., 2008a; Peters et al., 2014). Together with the minimality condition and some mild assumptions, Theorem 20 from Peters et al. (2010) proved that it is structural identifiable.

Assumption 3 (Correctly Specified Model)**.** *We assume the DECI model is correctly specified. Namely, there exists a parameter θ^* such that $p_{\theta^*}(\mathbf{x}; G^0) = p(\mathbf{x}; G^0)$.*

In practice, this assumption is hard to check in general. However, we can leverage the universal approximation capacity of neural networks (Hornik et al., 1989), meaning that it can approximate continuous functions arbitrarily well. This flexibility gives us a higher chance that this assumption indeed holds.

Assumption 4 (Causal Sufficiency)**.** *We assume DECI and the ground truth are causally sufficient. Namely, there are no latent confounders in the model.*

Assumption 5 (Regularity of log likelihood)**.** *We assume for all parameters θ and possible graphs G, the following holds:*

$$\mathbb{E}_{p(\mathbf{x};G^0)} \left[|\log p_\theta(\mathbf{x}; G)| \right] < \infty.$$

A.2. MLE Recovers Ground Truth

Maximum likelihood has often been used as the score function for causal discovery. For example, *Careful* (Khemakhem et al., 2021) adopts the likelihood ratio test (Hyvärinen & Smith, 2013) in the bivariate case, which is equivalent to selecting the causal directions with the maximized likelihood. However, they did not explicitly show that the resulting model recovers the ground truth for the multivariate case. In addition, Zhang et al. (2015) proved that maximizing likelihood for bivariate causal discovery is equivalent to minimizing the dependence between the cause and the noise variable. With the correctly

specified, structural identifiable model, the resulting noise and cause are independent through maximizing the likelihood, indicating the graph is indeed causal. However, it is non-trivial to generalize it to the multivariate case like DECI. In the following, we will show that under a correctly specified model and with maximum likelihood training with infinite data, DECI can recover the unique ground truth graph $G^* = G^0$ and the true data generating distribution $p_{\theta^*}(\mathbf{x}; G^*) = p(\mathbf{x}; G^0)$, where (θ^*, G^*) are MLE solutions.

Proposition 1. *Assuming assumptions 1-5 hold, we denote (θ^*, G^*) as the MLE solution with infinite training data. Then, we have*

$$p_{\theta^*}(\mathbf{x}; G^*) = p(\mathbf{x}; G^0)$$

In particular, we have $G^* = G^0$.

Proof. The key idea is to show that with arbitrary (θ, G), we have the following:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(\mathbf{x}_i; G) \leq \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \log p(\mathbf{x}_i; G^0)$$

By law of large numbers, we have

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(\mathbf{x}_i; G) = \mathbb{E}_{p(\mathbf{x};G^0)} \left[\log p_\theta(\mathbf{x}; G)\right]$$

Then, we can show

$$\begin{aligned}
&\mathbb{E}_{p(\mathbf{x};G^0)} \left[\log p_\theta(\mathbf{x}; G)\right] - \mathbb{E}_{p(\mathbf{x};G^0)} \left[\log p(\mathbf{x}; G^0)\right] \\
=&\mathbb{E}_{p(\mathbf{x};G^0)} \left[\log \frac{p_\theta(\mathbf{x}; G)}{p(\mathbf{x}; G^0)}\right] \\
\leq&\mathbb{E}_{p(\mathbf{x};G^0)} \left[\frac{p_\theta(\mathbf{x}; G)}{p(\mathbf{x}; G^0)} - 1\right] = \int p_\theta(\mathbf{x}; G) d\mathbf{x} - 1 = 0
\end{aligned}$$

where the inequality is due to $\log t \leq t - 1$. With assumption 3-4, we know there are no latent confounders and the model is correctly specified. Then, the above equality holds when (θ^*, G^*) induces the same join likelihood $p(\mathbf{x}; G^0)$. Since the model is structural identifiable, we have $G^* = G^0$. □

A.3. DECI Recovers the Ground Truth

To show that DECI can indeed recover the ground truth by maximizing the ELBO, we first introduce an important lemma showing the KL regularizer $\mathrm{KL}[q_\phi(G) \| p(G)]$ is negligible in the infinite data limit.

Lemma 3. *Assume an variational distribution $q_\phi(G)$ over a space of graphs $\mathcal{G}_\phi$, where each graph $G \in \mathcal{G}_\phi$ has a non-zero associated weight $w_\phi(G)$. With the soft prior $p(G)$ defined as eq. (7) and bounded λ, ρ, α, we have*

$$\lim_{N\to\infty} \frac{1}{N} \mathrm{KL}[q_\phi(G) \| p(G)] = 0. \tag{18}$$

Proof. First, we write down the definition of KL divergence

$$KL[q_\phi(G) \| p(G)] = \sum_{G \in \mathcal{G}_\phi} w_\phi(G) \left[\log w_\phi(G) + \lambda \|G\|_F^2 + \rho h(G)^2 + \alpha h(G) + \log Z\right]$$

where Z is the normalizing constant for the soft prior. From the definition and assumptions, it is trivial to know that $\log w_\phi(G)$, $\lambda \|G\|_F^2$ are bounded for all $G \in \mathcal{G}_\phi$. In the following, we show that $h(G)$ and $\log Z$ are also bounded.

From the definition of the DAG penalty, we have $h(G) = tr(\exp(G \odot G)) - D$. The matrix exponential is defined as

$$\begin{aligned}\mathrm{tr}(\exp(G \odot G)) &= \sum_{k=0}^{\infty} \frac{1}{k!}\mathrm{tr}((G \odot G)^k) \\ &= \sum_{k=0}^{\infty} \frac{1}{k!}\mathrm{tr}((G)^k) \\ &= \sum_{k=0}^{D} \frac{1}{k!} tr((G)^k)\end{aligned}$$

where the second equality is due to the fact that G is a binary adjacency matrix. From Zheng et al. (2018b), we know that $\mathrm{tr}(G^k)$ counts for the number of closed loops with length k. Since the graph has finite number of nodes, the longest possible closed loop is D, resulting in the third equality.

Thus, it is obvious that for any k, the number of closed loops with length k must be finite. Hence, it is trivial that $h(G) < \infty$. Therefore, with bounded λ, ρ, α, the un-normalized soft prior

$$|\exp(-\lambda\|G\|_F^2 - \rho h(G)^2 - \alpha h(G))| < \infty$$

Thus, the normalizing constant Z must be finite since there are only finite number of possible graphs.

Therefore, these must exists a constant $M_{\phi,G}$ such that $\log w_\phi(G) + \lambda\|G\|_F^2 + \rho h(G)^2 + \alpha h(G) + \log Z < M_{\phi,G}$. Hence, we have

$$0 \leq \mathrm{KL}[q_\phi(G)\|p(G)] < \sum_{G\in\mathcal{G}_\phi} w_\phi(G) M_{G,\phi} \leq \sqrt{\sum_{G\in\mathcal{G}_\phi} w_\phi^2(G)}\sqrt{\sum_{G\in\mathcal{G}_\phi} M_{G,\phi}^2} < \infty.$$

Thus, we have

$$\lim_{N\to\infty} \frac{1}{N}\mathrm{KL}[q_\phi(G)\|p(G)] = 0$$

where the third inequality is obtained by using Cauchy-Schwarz inequality. □

Now, we can prove that DECI can recover the ground truth. Recalling Theorem 1,

Theorem 1 (DECI recovers the true distribution)**.** *Assuming assumptions 1-5 are satisfied, the solution $(\theta', q'_\phi(G))$ from maximizing ELBO (eq. (10)) satisfies $q'_\phi(G) = \delta(G = G')$ where G' is a unique graph. In particular, we have $G' = G^0$ and $p_{\theta'}(\mathbf{x}; G') = p(\mathbf{x}; G^0)$.*

Proof. In terms of optimization, it is equivalent to re-write the ELBO (eq. (10)) into

$$\frac{1}{N}\mathbb{E}_{q_\phi}\left[\log p_\theta(\mathbf{x}_1, \ldots, \mathbf{x}_N)\right] - \frac{1}{N} KL\left[q_\phi(G)\|p(G)\right]$$

Now, under the infinite data limit and the definition of q_ϕ, we have

$$\begin{aligned}&\lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{q_\phi}\left[\log p_\theta(\mathbf{x}_1, \ldots, \mathbf{x}_N)\right] - \frac{1}{N} KL\left[q_\phi(G)\|p(G)\right] \\ =& \lim_{N\to\infty} \frac{1}{N} \sum_{G\in\mathcal{G}_\phi} w_\phi(G) \log p_\theta(\mathbf{x}_1, \ldots, \mathbf{x}_N|G) - \frac{1}{N} KL[q_\phi(G)\|p(G)] \\ =& \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \sum_{G\in\mathcal{G}_\phi} w_\phi(G) \log p_\theta(\mathbf{x}_i|G) \\ =& \int p(\mathbf{x}; G^0) \sum_{G\in\mathcal{G}_\phi} w_\phi(G) \log p_\theta(\mathbf{x}|G) d\mathbf{x},\end{aligned}$$

where the second and third equalities are from Lemma 3 and the law of large numbers, respectively. Let (θ^*, G^*) be the solutions from MLE (Proposition 1). Then, since $\sum_{G\in\mathcal{G}_\phi} w_\phi(G) = 1$, $w_\phi(G) > 0$, we have

$$\sum_{G\in\mathcal{G}_\phi} w_\phi(G)\mathbb{E}_{p(\mathbf{x};G^0)}\left[\log p_\theta(\mathbf{x}|G)\right] \leq \mathbb{E}_{p(\mathbf{x};G^0)}\left[\log p_{\theta^*}(\mathbf{x};G^*)\right]$$

with the equality holding when every graph $G \in \mathcal{G}_\phi$ and associated parameter θ_G satisfies

$$\mathbb{E}_{p(\mathbf{x};G^0)}\left[\log p_{\theta_G}(\mathbf{x}|G)\right] = \mathbb{E}_{p(\mathbf{x};G^0)}\left[\log p_{\theta^*}(\mathbf{x}|G^*)\right]. \tag{19}$$

From proposition 1, under correctly specified model, we have

$$\mathbb{E}_{p(\mathbf{x};G^0)}\left[\log p_{\theta^*}(\mathbf{x}|G^*)\right] = \mathbb{E}_{p(\mathbf{x};G^0)}\left[\log p(\mathbf{x};G^0)\right]$$

Thus, for a $G' \in \mathcal{G}_\phi$ and associated parameter θ', the condition in eq. (19) becomes

$$\begin{aligned}
&\mathbb{E}_{p(\mathbf{x};G^0)}\left[\log p_{\theta'}(\mathbf{x}|G')\right] = \mathbb{E}_{p(\mathbf{x};G^0)}\left[\log p(\mathbf{x}|G^0)\right] \\
\Longrightarrow &\mathbb{E}_{p(\mathbf{x};G^0)}\left[\log \frac{p_{\theta'}(\mathbf{x};G')}{p(\mathbf{x};G^0)}\right] = 0 \\
\Longrightarrow &KL[p(\mathbf{x};G^0)\|p_{\theta'}(\mathbf{x};G')] = 0,
\end{aligned}$$

which implies $p_{\theta'}(\mathbf{x};G') = p(\mathbf{x};G^0)$. Since DECI is structural identifiable, this means $G' = G^0$ and it is unique. Thus, the graph space $\mathcal{G}_\phi$ only contains one graph G', and $q'_\phi(G) = \delta(G = G')$. □

One should note that we do not explicitly restrict the noise distribution, indicating it still holds with the spline noise (Section 3.4). However, the above theorem implicitly assumes that DECI is a special case of ANM for structural indentifiability and that the data has no missing values. Thus, it is not applicable for DECI with the mixed-type and missing value extensions. We leave a more general theoretical guarantee to future work.

B. Additional Details for DECI

B.1. Optimization Details

As mentioned in the main text, we gradually increase the values of ρ and α as optimization proceeds, so that non-DAGs are heavily penalized. Inspired by *Notears*, we do this with a method that resembles the updates used by the augmented Lagrangian procedure for optimization (Nemirovsky, 1999). The optimization process interleaves two steps: (i) Optimize the objective for fixed values of ρ and α for a certain number of steps; and (ii) Update the values of the penalty parameters ρ and α. The whole optimization process involves running the sequence (i)-(ii) until convergence, or until the maximum allowed number of optimization steps is reached.

Step (i). Optimizing the objective for some fixed values of ρ and α using Adam (Kingma & Ba, 2014). We optimize the objective for a maximum of 6000 steps or until convergence, whichever happens first (we assume convergence if the loss does not improve during 1500 steps. If so, we move to step (ii)). We use Adam, initialized with a step-size of 0.01. During training, we reduce the step-size by a factor of 10 if the training loss does not improve for 500 steps. We do this a maximum of two times. If we reach the condition a third time, we do not decrease the step-size and assume optimization converged, and move to step (ii).

Iterating (i)-(ii). We initialize $\rho = 1$ and $\alpha = 0$. At the beginning of step (i) we measure the dag-penalty $P_1 = \mathbb{E}_{q_\phi(G)}h(G)$. Then, we run step (i) as explained above. At the beginning of step (ii) we measure the dag-penalty again, $P_2 = \mathbb{E}_{q_\phi(G)}h(G)$. If $P_2 < 0.65\,P_1$, we leave ρ unchanged and update $\alpha \leftarrow \alpha + \rho\,P_2$. Otherwise, if $P_2 \geq 0.65\,P_1$, we leave α unchanged and update $\rho \leftarrow 10\,\rho$. We repeat the sequence (i)-(ii) for a maximum of 25 steps or until convergence (measured as α or ρ reaching some max value), whichever happens first.

Other. We use $\lambda_s = 5$.

B.2. ELBO Derivation

The goal of maximum likelihood involves maximizing the likelihood of the observed variables. For DECI (with fully observed datasets) this corresponds to the log-marginal likelihood

$$\log p_\theta(x^1, \dots, x^N) = \log \sum_A p(G) \prod_n p_\theta(x^n|G). \tag{20}$$

Computing the marginalization from the equation above is intractable, even for moderately low dimensions, since the number of terms in the sum grows exponentially with the size of G (which grows quadratically with the problem's dimensionality).

Variational inference proposes to use a distribution $q_\phi(G)$ to build the ELBO, a lower bound of the objective from eq. (20), as follows:

$$\begin{aligned}
\log p_\theta(x^1, \dots, x^N) &= \log \sum_G p(G) \prod_n p_\theta(x^n|G) && (21)\\
&= \log \sum_G q_\phi(G) \frac{p(G)\prod_n p_\theta(x^n|G)}{q_\phi(G)} && (22)\\
&= \log \mathbb{E}_{q_\phi(G)} \left[\frac{p(G)\prod_n p_\theta(x^n|G)}{q_\phi(G)}\right] && (23)\\
&\geq \mathbb{E}_{q_\phi(G)} \left[\log \frac{p(G)\prod_n p_\theta(x^n|G)}{q_\phi(G)}\right] && \text{(Jensen's inequality)} \quad (24)\\
&= \mathbb{E}_{q_\phi(G)} \left[\log p(G) \prod_n p_\theta(x^n|G)\right] + H(q_\phi) && (25)\\
&= \text{ELBO}(\phi, \theta), && (26)
\end{aligned}$$

where we used that $H(q_\phi) = -\mathbb{E}_{q_\phi(G)} \log q_\phi(G)$ is the entropy of the distribution q_ϕ. Interestingly, the distribution q_ϕ that maximizes the ELBO is exactly the one that minimizes the KL-divergence between the approximation and the true posterior, $\text{KL}(q_\phi(G) \| p_\theta(G|x^1 \dots, x^N))$ (see, e.g. Blei et al. (2017)). This is why q_ϕ can be used as a posterior approximation.

B.3. Intervened Density Estimation with DECI

For a given graph, the density of some observation vector $\mathbf{a}$ is computed by evaluating the base distribution density after inverting the SEM

$$p_\theta(\mathbf{x} = \mathbf{a}) = \prod_i p(\mathbf{z}_i = (\mathbf{a}_i - f_i(\mathbf{a}_{\text{pa}(i;G^m)}))) \tag{27}$$

noting that the transformation Jacobian is the identity. We then marginalise the graphs using Monte Carlo:

$$p_\theta(\mathbf{x} = \mathbf{a}|G^k) = \frac{1}{M} \sum_m^M p_\theta(\mathbf{x} = \mathbf{a}|G^m); \quad G^m \sim q_\phi(G). \tag{28}$$

In the rest of this section we derive methods that allow using DECI to estimate causal quantities.

Under $G_{\text{do}(\mathbf{x}_T)}$, $i \in T$ correspond to parent nodes and we have the following factorisation: $p(\mathbf{x}|G_{\text{do}(\mathbf{x}_T)}) = p(\mathbf{x}_{\setminus T}|G_{\text{do}(\mathbf{x}_T)}) \prod_{i \in T} p(\mathbf{x}_i)$. We can then evaluate the interventional density of an observation $\mathbf{x}_{\setminus T} = \mathbf{a}$ with DECI as

$$\begin{aligned}
&p_\theta(\mathbf{x}_{\setminus T} = \mathbf{a}|\mathbf{x}_T = \mathbf{b}, G^m_{\text{do}(\mathbf{x}_T)})\\
&\quad = \frac{p_\theta(\mathbf{x}_{\setminus T} = \mathbf{a}|\mathbf{x}_T = \mathbf{b}, G^m_{\text{do}(\mathbf{x}_T)}) p_\theta(\mathbf{x}_T = \mathbf{b}|G^m_{\text{do}(\mathbf{x}_T)})}{p_\theta(\mathbf{x}_T = \mathbf{b}|G^m_{\text{do}(\mathbf{x}_T)})}\\
&\quad = \prod_{j \in \setminus T} p(\mathbf{z}_i = (\mathbf{a}_i - f_i(\mathbf{a}_{\text{pa}(i;G^m_{\text{do}(\mathbf{x}_T)})}))) \qquad (29)
\end{aligned}$$

We can then marginalise the graph using Monte Carlo as in eq. (28).

Importance Sampling The above approach might become computationally expensive if marginalising over a large number of graphs. Instead, we can note that the marginal of interest can be written as

$$p_\theta(\mathbf{x}_C{=}\mathbf{c}|\mathbf{x}_T{=}\mathbf{b}, G_{\text{do}(\mathbf{x}_T)}) = \mathbb{E}_{p_\theta(\mathbf{x}_{\setminus C}|\mathbf{x}_T=\mathbf{b}, G_{\text{do}(\mathbf{x}_T)})}[p_\theta(\mathbf{x}_C{=}\mathbf{c}|\mathbf{x}_{\setminus C}, \mathbf{x}_T{=}\mathbf{b}, G_{\text{do}(\mathbf{x}_T)})],$$

and estimated using Monte Carlo techniques. However, these could exhibit large variance when the density $p_\theta(\mathbf{x}_C{=}\mathbf{c}|\mathbf{x}_T{=}\mathbf{b}, G_{\text{do}(\mathbf{x}_T)})$ is low or when the dimensionality of $\mathbf{x}_{\setminus C}$ is large. We can make our samples more effective by using DECI's imputation network to as a proposal distribution

$$p_\theta(\mathbf{x}_C{=}\mathbf{c}|\mathbf{x}_T{=}\mathbf{b}, G_{\text{do}(\mathbf{x}_T)}) = \mathbb{E}_{q_\psi(\mathbf{x}_{\setminus C}|\mathbf{x}_C=\mathbf{c}, \mathbf{x}_T=\mathbf{b})}\left[\frac{p_\theta\left(\mathbf{x}_C{=}\mathbf{c}, \mathbf{x}_{\setminus C}|\mathbf{x}_T{=}\mathbf{b}, G_{\text{do}(\mathbf{x}_T)}\right)}{q_\psi(\mathbf{x}_{\setminus C}|\mathbf{x}_C=\mathbf{c}, \mathbf{x}_T=\mathbf{b})}\right]. \tag{30}$$

Analogously, we evaluate expectations with respect to the conditional interventional distribution for CATE estimation as:

$$\begin{aligned}&\frac{1}{M}\sum_m \mathbf{x}_Y^m \frac{p_\theta(\mathbf{x}_Y{=}\mathbf{x}_Y^m, \mathbf{x}_C{=}\mathbf{c}|\mathbf{x}_T{=}\mathbf{b}, G^m_{\text{do}(\mathbf{x}_T)})}{p_\theta(\mathbf{x}_Y{=}\mathbf{x}_Y^m, \mathbf{x}_C{=}\mathbf{x}_C^m|\mathbf{x}_T{=}\mathbf{b}, G^m_{\text{do}(\mathbf{x}_T)})};\\ &(\mathbf{x}_Y^m, \mathbf{x}_C^m) \sim p_\theta(\mathbf{x}_Y, \mathbf{x}_C|\mathbf{x}_T{=}\mathbf{b}, G^m_{\text{do}(\mathbf{x}_T)}); G^m_{\text{do}(\mathbf{x}_T)} \sim q(G_{\text{do}(\mathbf{x}_T)})\end{aligned} \tag{31}$$

By drawing a single sample from each graph, we are able to explore graph space more quickly.

B.4. Relationship with Khemakhem et al. (2021)

Khemakhem et al. (2021) introduced *Carefl*, a method that uses autoregressive flows (Kingma et al., 2016; Huang et al., 2018b) to learn causal-aware models, using the variables' causal ordering to define the autoregressive transformations. The method's main benefit is its ability to model complex nonlinear relationships between variables. However, *Carefl* alone is insufficient for causal discovery, as it requires the causal graph structure as an input. The authors propose a two-step approach. First, run a traditional constraint-based method (e.g., PC) to find the graph's skeleton and orient as many edges as possible, and second, fit several flow models to determine the orientation of the remaining edges. The drawbacks of this approach include the dependence on an external causal discovery methods (which will inherently limit *Carefl*'s performance to that of the method used), and the cost of fitting multiple flow models to orient the edges that are left unoriented after the first step. Our method extends Khemakhem et al. (2021) to learn the causal graph among multiple variables and perform end-to-end causal inference.

B.5. Discussion on Causal Discovery Methods

The posterior over graphs is computed as

$$p(G|\mathbf{X}) = \frac{p(\mathbf{X}|G)p(G)}{\sum_G p(\mathbf{X}|G)p(G)}. \tag{32}$$

In this equation, the likelihood measures the degree of compatibility of a certain DAG architecture with the observed data. For score-based discovery methods (Chickering, 2002; Chickering & Meek, 2015; Chickering, 2020; Huang et al., 2018a) we take the score to be $\log p(\mathbf{X}|G)$. For functional discovery methods (Shimizu et al., 2006; Hoyer et al., 2008b; ZHANG, 2009) we use the exogenous variable log-density. Constraint-based methods (Spirtes & Glymour, 1991; Spirtes et al., 2000) can also be cast in this light by assuming a uniform distribution over all graphs in their outputted equivalence class $\mathcal{G}$: $\log p(\mathbf{X}|G) = -\log|\mathcal{G}|, \forall\, G \in \mathcal{G}$. To what degree these methods succeed at constraining the space of possible graphs will depend on how well their respective assumptions are met and the amount of data available (Hoyer et al., 2008a).

C. Unified View of Causal Discovery Methods

This section introduces a simple analysis showing that, similarly to DECI, most causal discovery methods based on continuous optimization can be framed from a probabilistic perspective as fitting a flow. The benefits of this unified perspective are twofold. First, it allows a simple comparison between methods, shedding light on the different assumptions used by each one, their benefits and drawbacks. Second, it simplifies the development of new tools to improve these methods, since any improvements to one of them can be easily mapped to the others by framing them in this unified framework (e.g. our extensions to handle missing values and flexible noise distributions can be easily integrated with *Notears*).

The connection between causal discovery methods based on continuous optimization and flow-based models uses the concept of a weighted adjacency matrix $W(\theta) \in \mathbb{R}^{d\times d}$ linked to a function $f(\mathbf{x};\theta) : \mathbb{R}^d \to \mathbb{R}^d$. Loosely speaking, these matrices can be seen as characterizing how likely is each output of $f(\mathbf{x};\theta)$ to depend on each component of the input $\mathbf{x}$. For instance, $W(\theta)_{j,i} = 0$ indicates that $f_i(\mathbf{x};\theta)$ is completely independent of x_j. Such adjacency matrices can be constructed efficiently for a wide range of parameterizations for f, such as multi layer perceptrons and weighted combinations of nonlinear functions. We refer the reader to Zheng et al. (2020) for details.

Lemma 2. *Let $f(\mathbf{x};\theta) : \mathbb{R}^d \to \mathbb{R}^d$ be a θ-parameterized function with weighted adjacency matrix $W(\theta) \in \mathbb{R}^{d\times d}$. Given a dataset $\{\mathbf{x}^1, \dots, \mathbf{x}^N\}$, fitting a flow with the transformation $\mathbf{z} = \mathbf{x} - f(\mathbf{x};\theta)$, base distribution p_z and a hard acyclicity constraint on $W(\theta)$ is equivalent to solving*

$$\max_{\theta} \sum_{n=1}^{N} \log p_z(\mathbf{x}^n - f(\mathbf{x}^n;\theta)) \quad \text{s.t.} \quad h(W(\theta)) = 0, \tag{33}$$

where $h(\cdot)$ is the algebraic characterization of DAGs from eq. (6).

Proof. The acyclicity constraint is enforced by constraining the optimization domain to $\Theta = \{\theta : h(W(\theta)) = 0\}$. Then, the maximum likelihood objective can be written as

$$\sum_n \log p_\theta(\mathbf{x}^n) = \sum_n \log p_z(\mathbf{x}^n - f(\mathbf{x}^n;\theta)) + \log \left| \det \frac{\mathrm{d}(\mathbf{x}^n - f(\mathbf{x}^n;\theta))}{\mathrm{d}\mathbf{x}^n} \right| \tag{34}$$

$$= \sum_n \log p_z(\mathbf{x}^n - f(\mathbf{x}^n;\theta)) \qquad \text{(Lemma 1)}, \tag{35}$$

where the first equality we use the change of variable formula, valid because the transformation $\mathbf{z} = g(\mathbf{x};\theta) = \mathbf{x} - f(\mathbf{x};\theta)$ is invertible for any $\theta \in \Theta$. □

Lemma 2 is the main building block in the formulation of continuous optimization-based causal discovery methods from a probabilistic perspective as fitting flow models. This is simply because the objective used by each of these methods can be exactly recovered from eq. (33) with specific choices for $f(\mathbf{x};\theta)$ and p_z.

***Notears* (Zheng et al., 2018a)** uses a standard Gaussian for p_z and a linear transformation for $f(\mathbf{x},\theta)$. (This is similar to DECI-Gaussian, although DECI permits fully nonlinear functions.)

***Notears-MLP* (Zheng et al., 2020)** uses a standard Gaussian for p_z and d independent multi-layer perceptrons, one for each component of $f(\mathbf{x},\theta)$.

***Notears-Sob* (Zheng et al., 2020)** uses a standard Gaussian for p_z and a weighted linear combination of nonlinear basis functions.

***GAE* (Ng et al., 2019)** uses a standard Gaussian for p_z and a GNN for $f(\mathbf{x},\theta)$.

***Grandag* (Lachapelle et al., 2019)** uses a factorized Gaussian with mean zero and learnable scales for p_z and d multi layer perceptrons, one for each component of $f(\mathbf{x},\theta)$.

***Golem* (Ng et al., 2020).** This is a linear method whose original formulation was already in a probabilistic perspective, using a linear transformation for $f(\mathbf{x};\theta)$.

In summary, recently proposed causal discovery methods based on continuous optimization can be formulated from a probabilistic perspective as fitting a flow with different constraints, transformations, and base distributions. This unified formulation sheds light on the assumptions done by each method (e.g. a Gaussian noise assumption, either implicitly as in *Notears* or explicitly as in *Grandag*) and, more importantly, simplifies the development of new tools to improve them. For instance, the ideas proposed to deal with partially-observed datasets and non-Gaussian noise are readily applicable to any of the causal discovery methods mentioned in this section, addressing some of their limitations (Loh & Bühlmann, 2014; Kaiser & Sipos, 2021; Reisach et al., 2021).

D. Proof of Lemma 1

We split the proof in several simple steps.

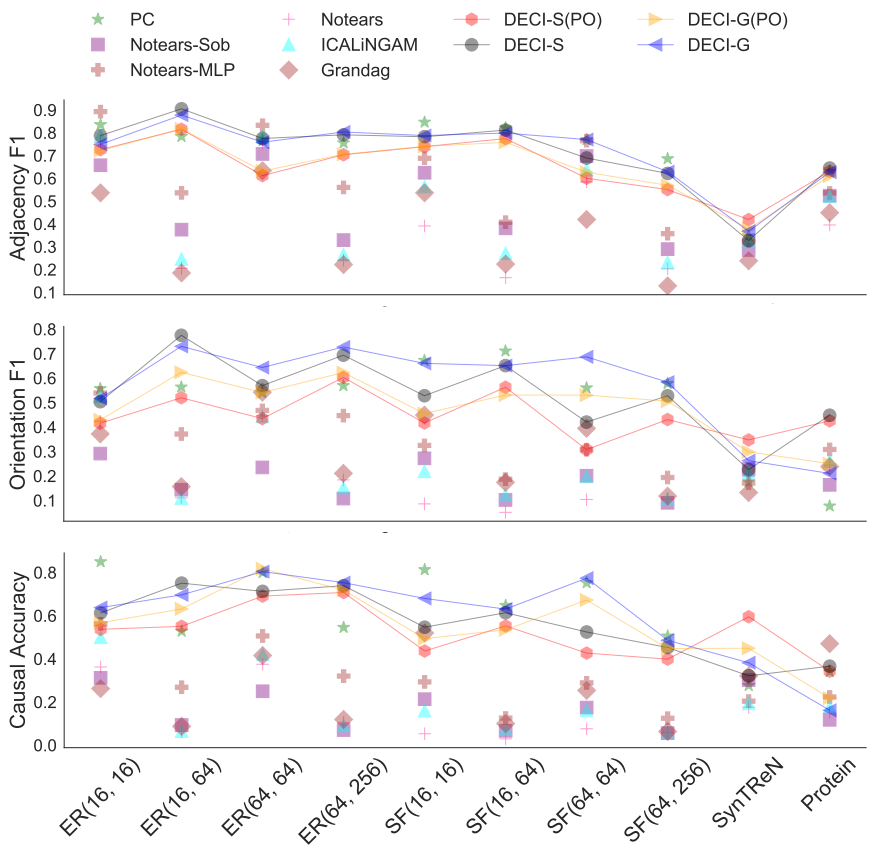


Figure 6. **DECI achieves better results than the baselines in all metrics shown.** The plots show the results for causal discovery for synthetic data generated using Gaussian noise. The legend "DECI-G" and "DECI-S" correspond to DECI using a Gaussian and spline noise model. Additionally, the "(PO)" corresponds to running DECI with 30% of the training data missing completely at random. For readability, we highlight the DECI results by connecting them with soft lines. The figure shows mean results across five different random seeds.

1. $g_G(x) = x - f_G(x)$ has Jacobian-determinant $\det(I - J_G(x))$, where $J_G(x)$ is the Jacobian of $f_G(x)$.
2. $J_G(x)$ has non-zero entries exactly in the positions where $G^\top$ is non-zero. Therefore, it retains the DAG structure.
3. Matrices with a DAG structure are nilpotent (i.e., all eigenvalues are zero). Thus, $J_G(x)$ can be factorized as $J_G(x) = QUQ^*$, where Q is unitary and U is *strictly* upper triangular (Schur factorization).
4. Finally, $\det(I - J_G(x)) = \det(I - QUQ^*) = \det(I - U) = 1$.

E. Additional Results

E.1. Causal Discovery Results under Gaussian Exogenous Noise

Figure 3 in the main text shows causal discovery results for the case where synthetic data was generated using non-Gaussian noise. In that case it was observed that using DECI together with a flexible noise model performed better than DECI with a Gaussian noise model. Figure 6 shows results for synthetic data generated using Gaussian noise. As expected, in this case using a Gaussian noise model is beneficial. (Though DECI with a spline noise mode still outperforms baselines.)

E.2. Detailed CSuite Results

Comprehensive results on CSuite ATE and CATE performance are shown in figures 7 and 8. We find DECI to perform consistently well in our 2 node datasets. It learns a uniform posterior over graphs in the non-identifiable setting, it fits

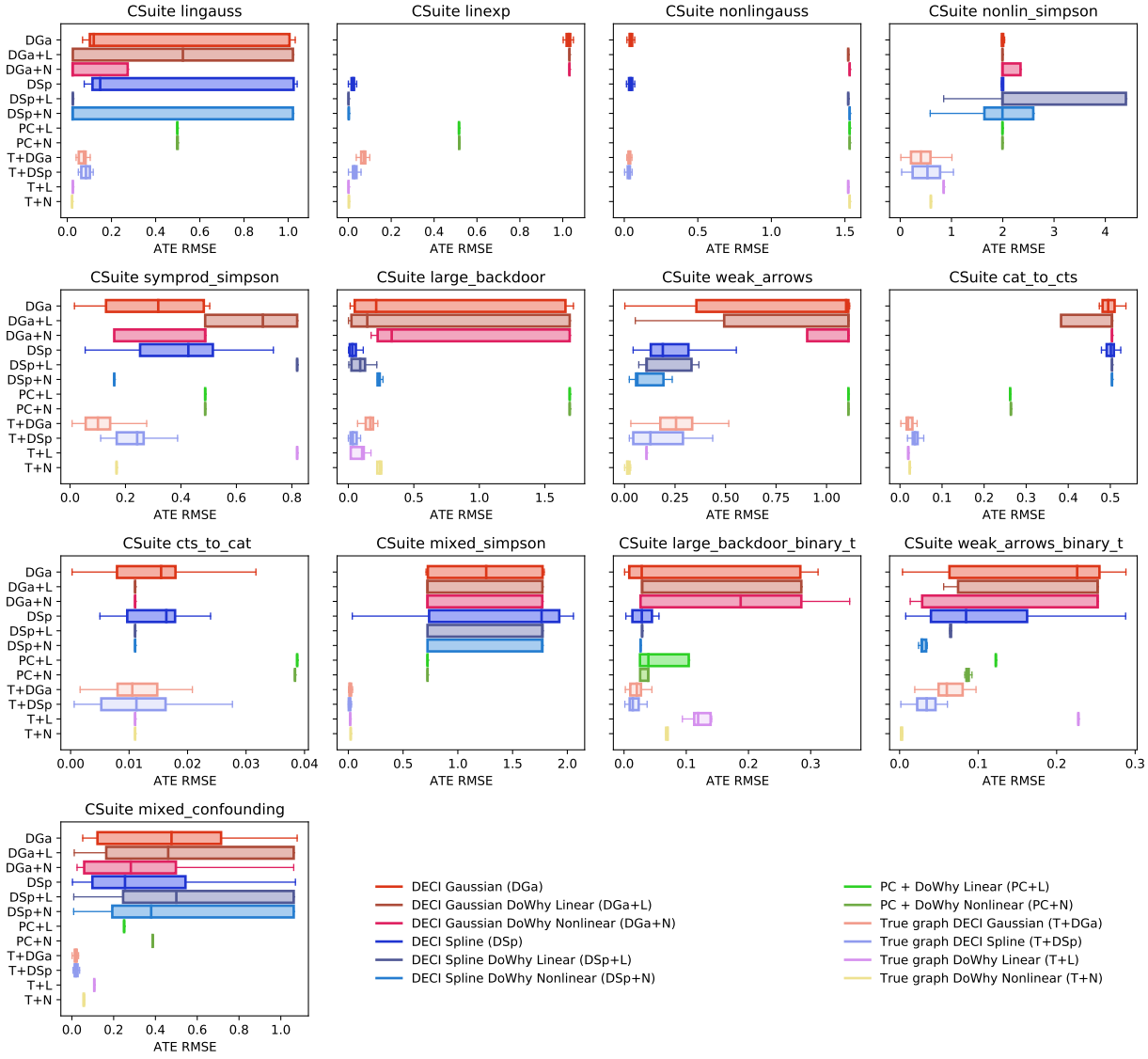


Figure 7. End-to-end ATE results on CSuite.

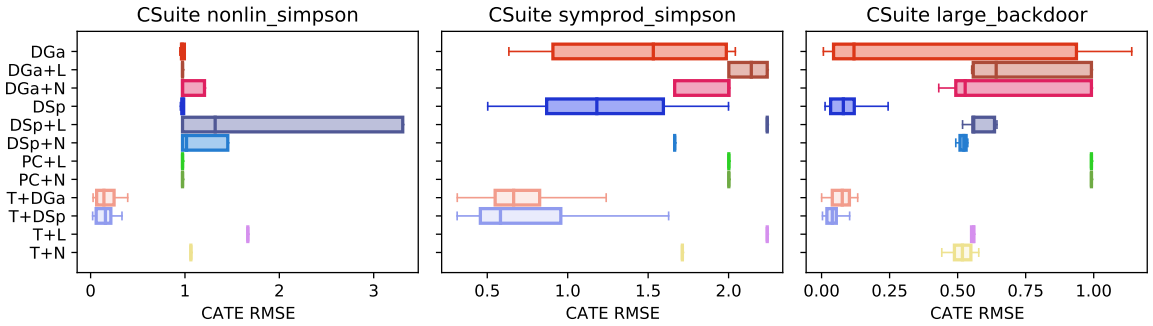


Figure 8. End-to-end CATE results on CSuite. Colours and acronyms as in Figure 7.

Method Dataset	DGa	DGa+L	DGa+N	DSp	DSp+L	DSp+N	PC+L	PC+N	T+L	T+N	T+DGa	T+DSp
ER(16, 16) - G	1.280	1.141	1.129	1.648	1.374	1.393	1.491	1.475	0.945	0.936	1.083	1.332
ER(16, 16) - S	1.726	1.776	1.780	1.829	1.776	1.755	1.869	1.790	1.755	1.793	1.643	1.660
ER(16, 64) - G	1.699	1.422	1.334	1.501	1.442	1.202	1.335	1.440	1.310	1.369	1.460	1.644
ER(16, 64) - S	2.311	2.465	2.600	2.174	2.452	2.421	2.584	2.276	2.742	2.641	2.510	2.428
ER(64, 64) - G	1.208	1.420	1.190	1.287	1.450	1.397	1.325	1.273	1.284	1.250	1.158	1.124
ER(64, 64) - S	2.246	1.626	1.626	1.892	2.325	2.292	1.526	2.446	2.442	2.441	2.490	2.481
SF(16, 16) - G	1.156	1.030	1.409	1.699	2.052	1.343	1.574	1.233	1.131	1.078	1.127	1.375
SF(16, 16) - S	2.870	1.805	2.284	2.431	2.363	1.805	3.008	2.520	2.518	2.502	2.424	2.477
SF(16, 64) - G	1.702	-	-	1.539	-	1.635	1.464	1.551	1.510	1.463	1.559	1.486
SF(16, 64) - S	3.594	-	-	4.139	-	-	3.877	4.145	4.162	4.106	4.161	3.861
SF(64, 64) - G	1.049	0.998	1.134	1.010	1.035	1.309	0.972	1.343	1.006	-	1.087	1.288
SF(64, 64) - S	2.754	3.239	3.239	3.242	3.239	3.239	3.227	2.591	2.815	-	2.883	3.034
csuite_cat_to_cts	0.495	0.504	0.504	0.501	0.504	0.504	0.262	0.264	0.020	0.023	0.019	0.036
csuite_cts_to_cat	0.015	0.011	0.011	0.016	0.011	0.011	0.039	0.038	0.011	0.011	0.011	0.011
csuite_large_backdoor	0.213	0.144	0.331	0.031	0.091	0.232	1.690	1.690	0.105	0.241	0.167	0.035
csuite_large_backdoor_bt	0.028	0.029	0.187	0.029	0.029	0.027	0.039	0.039	0.119	0.070	0.021	0.014
csuite_linexp	1.029	1.031	1.031	0.022	0.001	0.002	0.516	0.517	0.001	0.003	0.073	0.028
csuite_lingauss	0.120	0.523	0.024	0.149	0.025	0.024	0.498	0.498	0.025	0.022	0.076	0.085
csuite_mixed_confounding	0.477	0.461	0.282	0.254	0.500	0.380	0.250	0.387	0.107	0.057	0.019	0.018
csuite_mixed_simpson	1.259	0.723	0.723	1.765	1.772	1.771	0.723	0.723	0.017	0.022	0.014	0.013
csuite_nonlin_simpson	1.995	1.994	1.994	1.997	1.994	1.994	1.994	1.994	0.848	0.597	0.404	0.531
csuite_nonlingauss	0.042	1.522	1.532	0.043	1.522	1.532	1.532	1.532	1.522	1.532	0.034	0.034
csuite_symprod_simpson	0.318	0.695	0.487	0.427	0.819	0.160	0.487	0.487	0.819	0.168	0.101	0.242
csuite_weak_arrows	1.097	1.108	1.108	0.189	0.110	0.064	1.108	1.108	0.109	0.015	0.255	0.128
csuite_weak_arrows_bt	0.226	0.252	0.252	0.085	0.065	0.029	0.123	0.086	0.228	0.003	0.060	0.034
IHDP	0.187	0.187	0.187	0.090	0.101	0.116	0.187	0.187	0.187	0.187	0.146	0.087
Twins	0.030	0.025	0.025	0.022	0.025	0.025	0.068	0.025	0.022	0.042	0.022	0.060

Table 2. Median ATE RMSE data underling our rank table. The median is taken across multiple seeds, with the number of seeds shown in Table 3. Missing values indicate that the method exceeded the computational budget—this typically occurred for larger graphs.

non-linear functions well and it is robust to heavy tailed noise when employing the spline noise model. We find linear and non-linear DML inference to also perform acceptably, with the exception of the heavy tailed noise case, where the methods overfit to outliers and thus estimate ATE poorly.

On the larger (4 and 12 node) datasets, when the true graph is available, DECI provides ATE estimates competitive with the well-established non-linear DML method. Notably, DECI outperforms backdoor adjustment methods when the number of possible adjustment sets is large. Choosing the optimal adjustment set is an np-hard problem and the most common approach is to simply choose the largest one. This leads to doWhy suffering from variance. DECI's simulation-based approach avoids having to choose an adjustment set. On the other hand, for densely connected graphs where the strength of the connection between nodes is low, DECI struggles to capture the funcitonal relationships in the data and DML is most competitive. For CATE estimation DECI provides superior performance in all datasets and is able to completely solve all tasks but one.

Despite their small size, the non-linear nature of our datasets together with their heavy tailed noise make the discovery problem very challenging. We find that the PC algorithm provides very poor results or fails to find any causal DAGs compatible with the data when working with these datasets. We find both DECI to provide more acceptable performance with the DECI-spline variant producing more reliable results. When the true graph is not available, causal inference performance deteriorates sharply as a consequence of imperfect causal discovery. However, our findings in terms of relative performance among inference methods stay the same.

E.3. Discussion of Continuous CSuite Results

1. **lingauss**: When the true graph is available, all our causal inference methods are able to solve this problem. However, when the graph needs to be identified from the data, causal discovery accuracy is around 50%. DECI discovery converges to a posterior with half of its mass on the right distribution resulting in DECI inference methods showing the lowest error.

2. **linexp**: The non-Gaussian noise causes difficulties for DECI-Gaussian, which identifies the wrong orientation in a majority of cases. As a result, inference algorithm yield poor results. Surprisingly, the PC algorithm is also unable to identify the causal graph, leading to overall poor inference performance. With the spline noise model, DECI

Method Dataset	DGa	DGa+L	DGa+N	DSp	DSp+L	DSp+N	PC+L	PC+N	T+L	T+N	T+DGa	T+DSp
ER(16, 16) - G	5	5	5	5	5	5	5	5	5	5	5	5
ER(16, 16) - S	5	5	5	5	5	5	5	5	5	5	5	5
ER(16, 64) - G	5	5	5	5	5	5	5	5	5	5	5	5
ER(16, 64) - S	5	5	4	5	3	5	5	5	5	5	5	5
ER(64, 64) - G	5	1	1	5	1	1	4	5	5	5	5	5
ER(64, 64) - S	5	1	1	5	2	2	1	5	5	5	5	5
SF(16, 16) - G	5	2	3	5	4	5	4	5	5	5	5	5
SF(16, 16) - S	5	1	2	5	2	1	3	5	5	5	5	5
SF(16, 64) - G	5	0	0	5	0	1	4	5	5	5	5	5
SF(16, 64) - S	5	0	0	5	0	0	4	5	5	5	5	5
SF(64, 64) - G	5	4	3	5	3	2	3	4	5	0	5	5
SF(64, 64) - S	5	5	5	5	5	5	3	4	5	0	5	5
csuite_cat_to_cts	20	20	20	20	20	20	20	20	20	20	20	20
csuite_cts_to_cat	20	20	20	20	20	20	20	20	20	20	20	20
csuite_large_backdoor	20	20	20	20	20	20	20	20	20	20	20	20
csuite_large_backdoor_bt	20	20	20	20	20	20	20	20	20	20	20	20
csuite_linexp	20	20	20	20	20	20	20	20	20	20	20	20
csuite_lingauss	20	20	20	20	20	20	20	20	20	20	20	20
csuite_mixed_confounding	20	20	20	20	20	20	20	20	20	20	20	20
csuite_mixed_simpson	20	20	20	20	20	20	20	20	20	20	20	20
csuite_nonlin_simpson	20	20	20	20	20	20	20	20	20	20	20	20
csuite_nonlingauss	20	20	20	20	20	20	20	20	20	20	20	20
csuite_symprod_simpson	20	20	20	20	20	20	20	20	20	20	20	20
csuite_weak_arrows	20	20	20	20	20	20	20	20	20	20	20	20
csuite_weak_arrows_bt	20	20	20	20	20	20	20	20	20	20	20	20
IHDP	5	5	5	5	5	5	5	5	5	5	5	5
Twins	5	5	5	5	5	5	5	5	5	5	5	5

Table 3. Number of seeds run when computing values in Table 2. For ER/SF graphs, where fewer than 5 seeds were used, this indicates that some runs exceeded the computational budget.

successfully identifies the causal graph, allowing for all inference algorithms to solve the problem.

3. **nonlingauss**: The non-linear relationship between variables leads all DECI discovery runs to successfully recover the edge direction for this dataset while PC consistently identifies the wrong edge direction. As expected, linear ATE estimation performs poorly on this task. However, we find DoWhy non-linear to not fare much better, likely this is because DML still assumes a linear relationship between treatment and target. DECI solves the task successfully.

4. **nonlin_simpson**: Even when the true graph is available, none of our inference methods are able to recover the true ATE on this more difficult task. We observe non-linear methods (DECI and DoWhy-nonlinear) to perform similarly to each other and more strongly than the simple linear adjustment. For the CATE task, the true value is close to 0. This is correctly identified by both DECI-Gaussian and DECI-Spline. Interestingly, we find both linear and non-linear DoWhy variants to overestimate the causal effect when using the backdoor criterion. We attribute this to DECI solving a lower dimensional problem when estimating CATE. While DECI simply regresses the conditioning variable onto the effect variable. The backdoor adjustment employed by DoWhy requires regression from the joint space of conditioning variables and confounders onto the effect variables. The latter procedure involves estimating the relative strength of confounders and conditioning variables, which is a more challenging task.

   This dataset provides a challenging causal discovery task. DECI identifies the correct edges with probability 0.9. It capacity to recover the edge orientation is slightly worse 0.65. This imperfect causal discovery leads to poor inference for all methods. (potentially because they get Simpson's paradox the wrong way around).

5. **symprod_simpson**: Even with access to the true graph, no inference method is able to solve this problem. However, we find non-linear methods to clearly outperform linear adjustment for both CATE and ATE estimation. Among non-linear methods, performance is similar for ATE estimation, with DECI-Gaussian performing slightly better than DoWhy-nonlinear and DECI-Spline slightly worse. However, when estimating CATE, DECI inference present an error twice as low as nonlinear DoWhy. Again, we attribute this to the backdoor adjustment employed by DoWhy being a more challenging inference task than the 1d regression on simulated data employed by DECI.

   In terms of causal discovery, results are similar to nonlin-simpson with PC failing completely and DECI obtaining an adjacency score of 0.92 and orientation of 0.7. The imperfect graph knowledge hurts causal inference. Again we see

the non-linear backdoor adjustment to perform similarly to DECI for ATE estimation while DECI shows decisively stronger performance when estimating CATE. As expected, the linear adjustment method fares poorly in this strongly non-linear setting.

6. **weak_arrows**: When the true causal DAG is available we find that both DoWhy methods solve this ATE problem while both DECI methods predict slightly suboptimal ATE values.

   In terms of causal discovery, DECI clearly outperforms PC with the spline noise model again proving more reliable and leading to better ATE estimates. Although no methods are able to solve the task, we find that non-linear DoWhy with the DECI-spline graphs performs best. We hypothesize that the amortised function structure employed by DECI suffers in very densely connected graphs with weak edges, like is the case here.

7. **large_backdoor**: With access to the true graph, DECI methods outperform both Dowhy variants for both ATE and CATE estimation. DECI-spline performs best and is able to solve both problems. When faced with many confounders, adjustment procedures suffer from large variance. As a result, despite the non-linearity of the functional relationships at play, the simpler linear backdoor adjustment outperforms the non-linear DML approach. On the other hand, DECI's simulation based approach is not disadvantaged in this setting.

   Following the trend of the previous datasets, PC performs poorly in terms of causal discovery, biasing downstream inference methods which perform poorly in terms of ATE and CATE estimation. DECI discovery is more reliable, an effect most noticeable when using the spline noise models. With the DECI-Spline posterior over graphs, both DECI-spline and linear DoWhy are able to solve the ATE problem and DECI-spline is the only method capable of solving the CATE task. For both tasks and noise models DECI outperforms non-linear DoWhy, again showing its invariance to the size of potential adjustment set.

E.4. Synthetic Graph Experiments

We test the performance of DECI on ATE estimation with random graphs as described in section 4.1. For each graph, we randomly generate interventional data for up to five random interventions. We chose the effect variable as the last variable in the causal order that has not yet been used for data generation. For each effect variable we chose the intervention by randomly traversing the graph up to three edges away from the effect variable.

Table 2 shows the performance of the ATE estimation of DECI and all baselines on the synthetic graph data. We only show results for methods that have a runtime of less than one day. Figure 9 shows the runtimes for the different methods. DECI has consistently the lowest runtime and scales best to larger graphs. While the runtime of DECI stays approximately constant for various graphs, the runtime of the ATE estimation baselines increases with more complex graphs. In general, the methods using the true graph outperform the methods that also perform causal discovery. Further, no method strongly outperforms all other methods with DECI being a strong competitor to the already established DML methods. Lastly, we can see that DECI is capable of performing causal discovery, data imputation and ATE estimation in an end-to-end fashion without degrading performance.

E.5. Learning in Non-identifiable Settings with the Help of Graph Priors

We investigate the utility of prior knowledge over causal graphs for causal discovery and end2end inference in non-identifiable and difficult to identify settings. Specifically, we generate 2 datasets composed of 2000 training examples each. The first is composed of only linear relationships between variables and Gaussian additive noise, making the causal graph non-identifiable. The second dataset also uses linear functions but has a mix of exponential and Tanh-Gaussian noise. Although identifiable, discovery in this latter setting is challenging.

We compare DECI inference with access to the true graph to end2end DECI inference. In the latter case we consider a PC prior, which has as its mean the CP-DAG provided by PC. We consider different prior strengths, from 0 to 1, which represent how much prior mass is placed at the mean graph. The rest is distributed uniformly across graphs. We do the same with the true-graph, yielding what we refer to as the "informed prior".

In the non-identifiable case, we find both DECI (prior strength 0) and PC discovery to provide incorrect graphs. Interestingly, providing the PC CPDAG as a prior for DECI can yield large gains in terms of causal discovery due to a variance reduction effect. These gains do not translate to better ATE estimation, where performance is not improved over the uninformative prior. Providing knowledge of the true graph does help causal inference, with a more confident prior yielding better results.

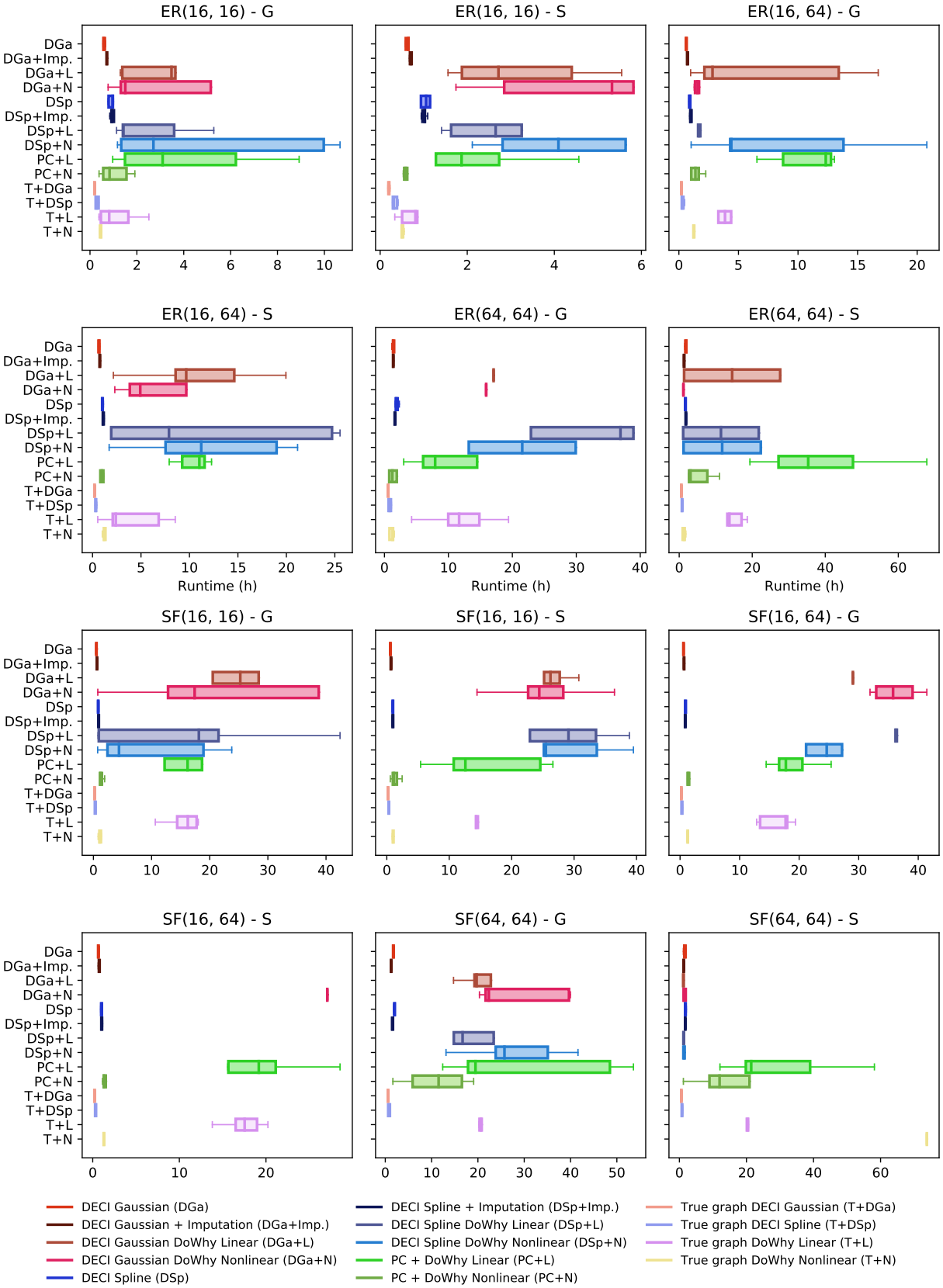


Figure 9. Runtime of End-to-end ATE estimation methods on synthetic graphs.

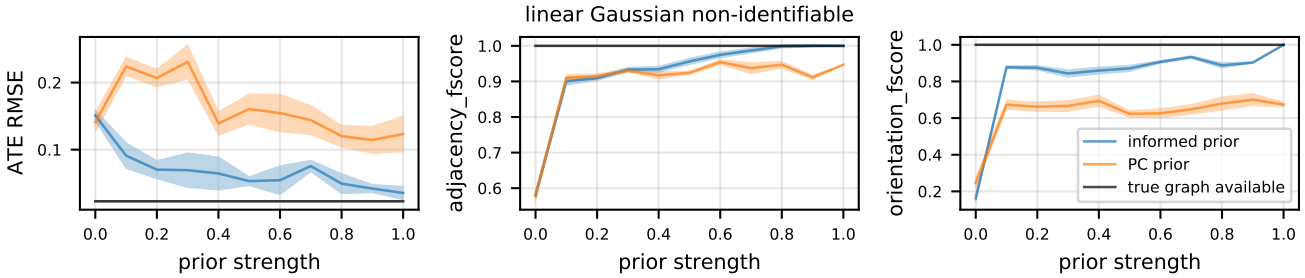


Figure 10. Causal discovery and inference results obtained on a 9 node linear Gaussian dataset, where the graph is non-identifiable without prior knowledge. We perform DECI inference with the true graph and DECI end2end inference with different priors. Informed prior refers to using a smoothed version of the true graph as the prior. PC prior refers to using the CP-DAG outputted by PC as the prior mean W_0. The prior strength indicates how much prior mass is placed on the mean prior graph W_0 and how much is spread across all other DAGs.

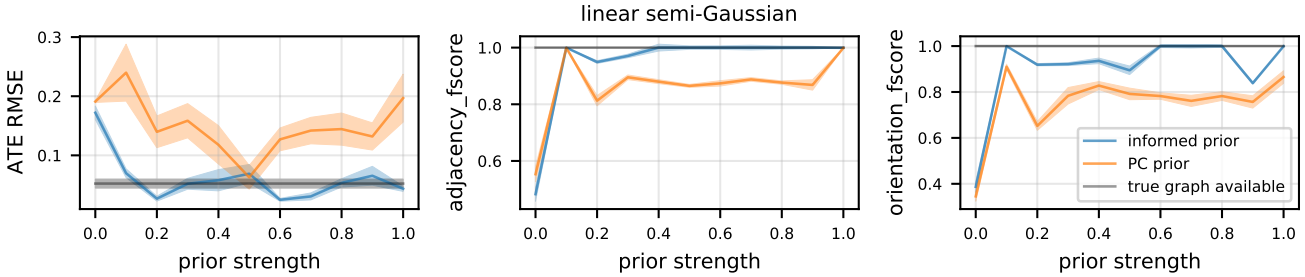


Figure 11. Causal discovery and inference results obtained on a 9 node linear non-Gaussian dataset, difficult to identify without prior knowledge. We perform DECI inference with the true graph and DECI end2end inference with different priors. Informed prior refers to using a smoothed version of the true graph as the prior. PC prior refers to using the CP-DAG outputted by PC as the prior mean W_0. The prior strength indicates how much prior mass is placed on the mean prior graph W_0 and how much is spread across all other DAGs.

In the difficult identifiable case, the PC prior does provide gains to DECI. We find the optimal prior strength to be 0.5: a balanced combination of PC and DECI discovery is most reliable, while using exclusively one of the two algorithms yields worse results. In this identifiable setting informed DECI discovery is able to obtain perfect ATE estimation performance with a prior strength as low as 0.2.

F. Datasets Details

Our two benchmark datasets are constructed following similar procedures described in Louizos et al. (2017).

IHDP (Hill, 2011). This dataset contains measurements of both infants (birth weight, head circumference, etc.) and their mother (smoked cigarettes, drank alcohol, took drugs, etc) during real-life data collected in a randomized experiment. The main task is to estimate the effect of home visits by specialists on future cognitive test scores of infants. The outcomes of treatments are simulated artificially as in (Hill, 2011); hence the outcomes of both treatments (home visits or not) on each subject are known. Note that for each subject, our models are only exposed to only one of the treatments; the outcomes of the other potential/counterfactual outcomes are hidden from the mode, and are only used for the purpose of ATE/CATE evaluation. To make the task more challenging, additional confoundings are manually introduced by removing a subset (non-white mothers) of the treated children population. In this way we can construct the IHDP dataset of 747 individuals with 6 continuous covariates and 19 binary covariates. We use 10 replicates of different simulations based on setting B (log-linear response surfaces) of (Hill, 2011), which can downloaded from `https://github.com/AMLab-Amsterdam/CEVAE`. We use a 70%/30% train-test split ratio. Before training our models, all continuous covariates are normalized.

TWINS (Almond et al., 2005). This dataset consists of twin births in the US between 1989-1991. Only twins which with the same sex born weighing less than 2kg are considered. The treatment is defined as being born as the heavier one in each twins pair, and the outcome is defined as the mortality of each twins in their first year of life. Therefore, by definition for each pair of twins, we can observe the outcomes of both treatments (the lighter twin and heavier twin). However, during training, only one of the treatment is visible to our models, and the other potential outcome is unknown to the model and are

only used for evaluation. The raw dataset is downloaded from https://github.com/AMLab-Amsterdam/CEVAE. Following Louizos et al. (2017), we also introduce artificial confounding using the categorical GESTAT10 variable. This is done by assigning treatments (factuals) using the conditional probability $t_i|\mathbf{x}_i, z_i = \text{Bern}(\sigma(w_0^T\mathbf{x}_i + w_h(z_i/10 - 0.1)))$, where t_i is the treatment assignment for subject i, z_i is the corresponding GESTAT10 covariate, $\mathbf{x}_i$ denotes the other remaining covariates. Both w_0 and w_h are randomly generated as $w_0 \sim \mathcal{N}(0, 0.1 * I)$, $w_h \sim \mathcal{N}(5, 0.1)$. All continuous covariates are normalized.

Ground Truth ATE and CATE Estimation for TWINS and IHDP. In both benchmark datasets, since the held-out hypothetical outcomes of counterfactual treatments are already known, the the ground truth ATE can be naively estimated by averaging the difference between the factual and counterfactual outcomes across the entire dataset. The CATE estimation is a bit tricky, since both datasets contains covariates collected from real-world experiments, in which the underlying ground truth causal graph structure is unknown. As a result, exact CATE estimation is generally impossible for continuous conditioning sets. Therefore, when evaluating the CATE estimation performance on **TWINS** and **IHDP**, we focus only on discrete variables (binary and categorical) as conditioning set. This allows unbiased estimation of ground truth CATE by simply averaging the treatment effects on subgroups of subjects in the dataset, that have the corresponding discrete value in the conditioning set. We consider only single conditioning variable at a time, and estimate the corresponding CATE for evaluation.

F.1. CSuite

We develop Causal Suite (CSuite), a number of small to medium (2–12 nodes) synthetic datasets generated from hand-crafted Bayesian networks with the intention of testing different capabilities of causal discovery and inference methods. All datasets take the form of additive noise models.

Each dataset comes with a training set of 2000 samples, and between 1 and 2 intervention test sets. Each intervention test set has a treatment variable, treatment value, reference treatment value and effect variable. We estimate the ground truth ATE by drawing 2000 samples from the treated and reference intervened distributions. For the datasets used to evaluate CATE, we generate samples from *conditional* intervened distributions by using Hamiltonian Monte Carlo. We employ a burn-in of 10k steps and a thinning factor of 5 to generate 2000 conditional samples, which we then use to compute our ground truth CATE estimate. We note that because all ground truth causal quantities are estimated from samples, there is a lower bound on the expected error that can be obtained by our methods. When methods obtain an error equal or lower we say that they have solved the task.

lingauss A two node graph (Figure 12a) with a linear relationship and Gaussian noise. We have $X_1 \sim N(0, 1)$ and $X_2 = \frac{1}{2}X_1 + \frac{\sqrt{3}}{2}Z_2$ where $Z_2 \sim N(0, 1)$ is independent of X_1. The observational distribution is symmetrical in $X_1 \leftrightarrow X_2$. The graph is not identifiable. The best achievable performance on this dataset is obtained when there is a uniform distribution over edge direction.

linexp A two node graph (Figure 12a) with a linear functional relationship, but with exponentially distributed additive noise. We have $X_1 \sim N(0, 1)$ and $X_2 = \frac{1}{2}X_1 + \frac{\sqrt{3}}{2}(Z_2 - 1)$ where $Z_2 \sim \text{Exp}(1)$ is independent of X_1. By using non-Gaussian noise, the graph becomes identifiable. However, the inference problem will be more challenging for methods sensitive to outliers, such as those that assume Gaussian noise.

nonlingauss A two node graph (Figure 12a) with a nonlinear relationship and Gaussian additive noise. We have $X_1 \sim N(0, 1)$ and $X_2 = \sqrt{6}\exp(-X_1^2) + \alpha Z_2$ where $Z_2 \sim N(0, 1)$ is independent of X_1 and $\alpha^2 = 1 - 6\left(\frac{1}{\sqrt{5}} - \frac{1}{3}\right)$. Note $\text{Var}(X_2) = 1$ and $\text{Cov}(X_1, X_2) = 0$. By having a linear correlation of zero between X_1 and X_2, this dataset creates a potential failure mode for causal inference methods that assume linearity.

nonlin_simpson A synthetic Simpson's paradox, using the graph Figure 12b: if the confounding factor X_3 is not adjusted for, the relationship between the treatment X_1 and effect X_2 reverses. The variable X_4 correlates strongly with the effect, but must not be used for adjustment. Choosing an incorrect adjustment set when estimating $\mathbb{E}[X_2|\text{do}(X_1)]$ leads to a significantly incorrect ATE estimate. All variables are continuous, with nonlinear structural equations and non-Gaussian additive noise.

symprod_simpson Another Simpson's paradox using the graph Figure 12c. This dataset is similar to nonlin_simpson with 2 key differences: 1) the effect variable is the result of a product between the confounding variable and the treatment variable. This makes drawing causal inferences require non-linear function estimation. Additionally, the ATE is close to 0. The conditioning variable for the CATE task is a descendant of the confounding variable. This dataset probes for methods' capacity to reduce their uncertainty about a confounding variables based on values of its child variables.

large_backdoor A nine node graph, as shown in Figure 12d. This dataset is constructed so that there are many possible choices of backdoor adjustment set. While both minimal and maximal adjustment sets can result in a correct solution, the a minimal adjustment set results in a much lower-dimensional adjustment problem and thus will result in lower variance solutions. The conditioning node for the CATE task is a child of the root variable. Thus the CATE task probes for methods' capacity to infer the value of an observed confounder from one of its children. All variables are continuous, with nonlinear structural equations and non-Gaussian additive noise.

weak_arrows A nine node graph, as shown in Figure 12e. Unlike the previous dataset, when the true graph is known, a large adjustment set must be used. The causal discovery challenge revolves around finding all arrows, which are scaled to be relatively weak, but which have significant predictive power for X_9 in aggregate. This dataset tests methods' capacity to identify the full adjustment set and adjust for a large number of variables simultaneously.

cat_to_cts A two node (Figure 12a) graph with categorical X_1 and continuous X_2 with an additive noise model. We have $X_1 \sim \text{Cat}\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right)$ takes values in $\{0, 1, 2\}$ and $X_2 = X_1 + \frac{8}{5}(s(Z_2) - 1)$ where $s(x) = \log(\exp(x) + 1)$ is the softplus function, and $Z_2 \sim N(0, 1)$ is independent of X_1.

cts_to_cat A two node (Figure 12a) graph with continuous X_1 and categorical X_2. We take $X_1 \sim U(-\sqrt{3}, \sqrt{3})$ and X_2 categorical on $\{0, 1, 2\}$ with the following conditional probabilities

$$p(X_2|X_1 = x_1) = \begin{cases} \left(\frac{6}{13}, \frac{6}{13}, \frac{1}{13}\right) & \text{if } x_1 < -\frac{\sqrt{3}}{3} \\ \left(\frac{1}{8}, \frac{3}{4}, \frac{1}{8}\right) & \text{if } -\frac{\sqrt{3}}{3} \leq x_1 < \frac{\sqrt{3}}{3} \\ \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) & \text{if } x_1 > \frac{\sqrt{3}}{3} \end{cases} \tag{36}$$

In this problem, we treat X_2 as the treatment and X_1 as the target, giving a theoretical ATE of zero.

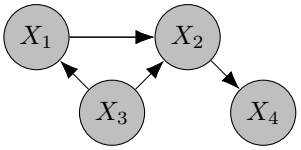
mixed_simpson Similar to the nonlin_simpson dataset, using the graph of Figure 12b, but with X_3 categorical on three categories, and X_1 binary.

large_backdoor_binary_t Similar to the large_backdoor dataset, using the graph of Figure 12d, but with X_8 binary.

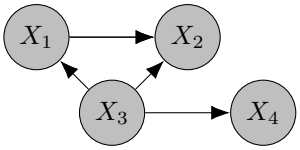
weak_arrows_binary_t Similar to the weak_arrows dataset, using the graph of Figure 12e, but with X_8 binary.

mixed_confounding A large, mixed type dataset with 12 variables, as shown in Figure 12f. In this dataset, X_1, X_5 are binary, X_3, X_6, X_8 are categorical on three categories, and other variables are continuous. We utilise nonlinear structural equations and non-Gaussian additive noise.

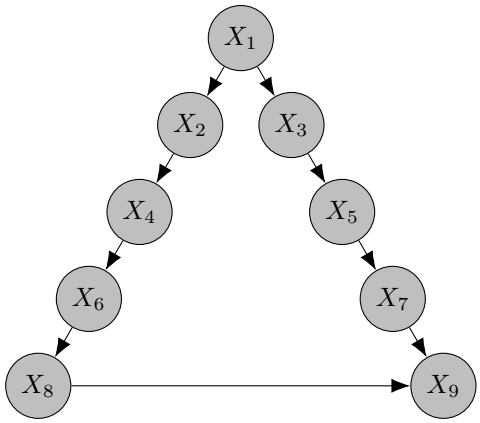
(a) Two node



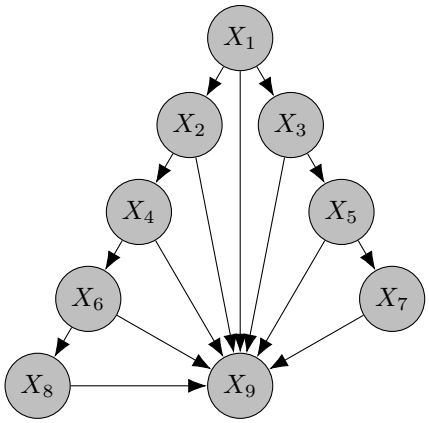
(b) Nonlin. Simpson



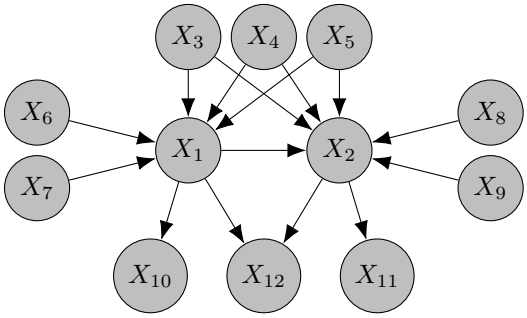
(c) Symprod. Simpson



(d) Large backdoor. Treatment X_8, outcome X_9, condition X_7.



(e) Weak arrows. Treatment X_8, outcome X_9, condition X_7.



(f) Mixed confounding

Figure 12. CSuite graphs. Unless otherwise stated, we take X_1 as the treatment, X_2 as the outcome, and for CATE we take X_3 as the conditioning variable.