Deep End-to-end Causal Inference

Tomas Geffner[† 1 *] **Javier Antorán**[† 2 *] **Adam Foster**[3 *] **Wenbo Gong**[3] **Chao Ma**[3]
Emre Kıcıman[3] **Amit Sharma**[3] **Angus Lamb**[† 4] **Martin Kukla**[† 5] **Agrin Hilmkil**[3]
Joel Jennings[3] **Nick Pawlowski**[3] **Miltiadis Allamanis**[† 6] **Cheng Zhang**[3]
[1] University of Massachusetts Amherst [2] University of Cambridge
[3] Microsoft Research [4] G-Research [5] TikTok [6] Google Research
cheng.zhang@microsoft.com

Abstract

Causal inference is essential for data-driven decision making across domains such as business engagement, medical treatment and policy making. However, research on causal discovery has evolved separately from causal inference, preventing straightforward combination of methods from both fields. In this work, we develop Deep End-to-end Causal Inference (DECI), a non-linear additive noise model with neural network functional relationships 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 asymptotically recover the ground truth causal graph and treatment effects when correctly specified. Our results show the competitive 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 causal machine learning benchmarks.

1 Introduction

Causal-aware decision making is pivotal in many fields such as economics [72, 4] and healthcare [5, 22, 66]. For example, in healthcare, caregivers may wish to understand the effectiveness of different treatments given only historical data. They aspire to estimate treatment effects from observational data, with incomplete or no knowledge of the causal relationships between variables. This is the *end-to-end causal inference* problem, where we discover the causal graph and estimate treatment effects together using causal assumptions and observational data.

It is well known that any causal conclusion drawn from observational data requires assumptions that are not testable in the observational environment [49]. Existing methods for estimating causal quantities from data, which we refer to as *causal inference methods*, commonly assume complete *a priori* knowledge of the causal graph. This is rarely available in real-world applications, especially when many variables are involved. On the other hand, existing *causal graph discovery methods*, i.e. those that seek to infer the causal graph from observational data, require assumptions about statistical properties of the data, which often require less human input [60]. These methods often return a large set of plausible graphs. This incompatibility of assumptions and inputs/outputs makes the task of answering causal queries in an *end-to-end manner* non-trivial.

Continuous non-linear additive noise models are a promising class of models to enable end-to-end causal inference. Indeed, if the data generating process is assumed to be in the class of non-linear additive noise structural equation models with no unobserved confounders (with a few extra technical assumptions), then both the causal graph and causal treatment effects are identifiable from observational data.

[*]Equal contribution. † Contributed during internship or residency in Microsoft Research.

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).

To take advantage of this, we propose Deep End-to-end Causal Inference (DECI), a highly flexible non-linear additive noise model which combines neural network structural equations with spline flows for the exogenous additive noise variables. We train DECI on observational data using variational inference on the causal graph, extending and generalizing recent work such as *Notears* [74, 76] and others [36, 45, 46]. The fitted DECI model can be used for simulation-based estimation of (C)ATE without additional training. We prove that, when correctly specified, variational inference with DECI asymptotically recovers the true causal graph and data generating process. DECI therefore allows practitioners to estimate causal quantities using only observational data as input.

Motivated by real-world problems, we extend DECI to handle for mixed type (continuous and categorical) variables and missing value imputation, although we sacrifice identifiability in such cases.

We systematically evaluate DECI, along with a range of combinations of existing discovery and inference algorithms in over 1000 experiments. DECI performs very competitively with baselines from both the causal discovery and inference domains.

2 Preliminaries

Structural Equation Models (SEM) describe causal relationships between variables [49]. Let $\mathbf{x} = (x_1, \ldots, x_d)$ be a collection of random variables. Given a causal graph, a DAG G on nodes $\{1, \ldots, D\}$, we can define an observational distribution $p(\mathbf{x})$ via the *structural equations* $x_i = F_i\left(\mathbf{x}_{\text{pa}(i;G)}, z_i\right)$ where z_i are independent exogenous noise variables, $\text{pa}(i; G)$ are the parents of node i in G, and F_i is the ith functional relationship. The SEM also describes how the distribution changes under *do-interventions* $p(\mathbf{x} \mid \text{do}(\mathbf{x}_T = \mathbf{a}))$ in which variables $\mathbf{x}_T$ are set to fixed values $\mathbf{a}$; here $T \subseteq \{1, \ldots, d\}$ are treatment indices. The do-intervention is used to define (conditional) average treatment effects ((C)ATE) that describe the impact of treatments [49]

$$\text{ATE}(\mathbf{a}, \mathbf{b}) = \mathbb{E}_{p(\mathbf{x}_Y|\text{do}(\mathbf{x}_T=\mathbf{a}))}[\mathbf{x}_Y] - \mathbb{E}_{p(\mathbf{x}_Y|\text{do}(\mathbf{x}_T=\mathbf{b}))}[\mathbf{x}_Y], \quad \text{and} \tag{1}$$

$$\text{CATE}(\mathbf{a}, \mathbf{b}|\mathbf{c}) = \mathbb{E}_{p(\mathbf{x}_Y|\text{do}(\mathbf{x}_T=\mathbf{a}),\mathbf{x}_C=\mathbf{c})}[\mathbf{x}_Y] - \mathbb{E}_{p(\mathbf{x}_Y|\text{do}(\mathbf{x}_T=\mathbf{b}),\mathbf{x}_C=\mathbf{c})}[\mathbf{x}_Y]. \tag{2}$$

Additive noise SEMs, which are the focus on this paper, have structural equations of the form

$$x_i = F_i\left(\mathbf{x}_{\text{pa}(i;G)}, z_i\right) = f_i\left(\mathbf{x}_{\text{pa}(i;G)}\right) + z_i \quad \text{or} \quad \mathbf{x} = f_G(\mathbf{x}) + \mathbf{z} \quad \text{in vector form.} \tag{3}$$

Importantly, with minor technical assumptions to exclude e.g. linear Gaussian SEMs, continuous additive noise models are structural identifiable from observational data [51]. This makes them a promising starting point to develop models for end-to-end causal inference.

3 DECI: Deep End-to-end Causal Inference

We propose a non-linear additive noise model with neural network functional relationships

$$f_i\left(\mathbf{x}_{\text{pa}(i;G)}\right) = \zeta_i\left(\sum_{j=1}^{d} G_{j,i}\ \ell_j(x_j)\right), \tag{4}$$

where $G_{j,i} \in \{0, 1\}$ indicates the presence of the edge $j \to i$. The functions ℓ_i and ζ_i ($i = 1, \ldots, d$) are defined using a weight sharing scheme $\zeta_i(\cdot) = \zeta(\mathbf{u}_i, \cdot)$ and $\ell_i(\cdot) = \ell(\mathbf{u}_i, \cdot)$, where ζ and ℓ are multi-layer perceptrons (MLPs). The trainable parameters of the SEM functions are the parameters of the networks ζ, ℓ and the node embeddings $(\mathbf{u}_i)_{i=1}^{d}$.

Exogenous Noise Model $p_{\mathbf{z}}$. We consider two possible models for the distribution of $\mathbf{z}$: 1) a simple Gaussian $p_{z_i}(\cdot) = \mathcal{N}\left(\cdot|0, \sigma_i^2\right)$ with trainable variances σ_i^2; 2) a flow [53]

$$p_{z_i}(z_i) = \mathcal{N}\left(\kappa_i^{-1}(z_i)|0, 1\right)\left|\frac{\partial \kappa_i^{-1}(z_i)}{\partial z_i}\right|. \tag{5}$$

We choose the learnable bijections κ_i to be a rational quadratic splines [14], parametrized independently across dimensions. Spline flows are significantly more flexible than the Gaussian distributions employed in previous work [36, 45, 46, 74, 76].

Unified View of functional Causal Discovery. DECI employs neural nets for f_i and flexible, potentially non-Gaussian, distributions for $p_{\mathbf{z}}$, making it the most flexible member of a larger family of SEMs for functional causal discovery [36, 45, 46, 74, 76]. We expand on this in Appendix D.

3.1 DECI and Bayesian Causal Discovery

DECI takes a Bayesian approach to causal discovery [17]. We therefore consider a joint distribution over both the causal graph G and the the observations $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$ given by

$$p_\theta(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}, G) = p(G) \prod_n p_\theta(\mathbf{x}^{(n)}|G) = p(G) \prod_n \prod_{i=1}^{d} p_{z_i}\left(x_i^{(n)} - f_i\left(\mathbf{x}^{(n)}_{\text{pa}(i;G)}\right)\right) \tag{6}$$

where the final form reflects our additive noise model structure and θ denotes the trainable parameters of the SEM functions and of the exogeneous noise model. We aim to fit a point estimate θ' for θ by maximizing a lower bound on the marginal likelihood of observational data $p_\theta(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)})$. The posterior $p_{\theta'}(G|\mathbf{x}^1, \ldots, \mathbf{x}^N)$ then characterizes our beliefs about the causal graph.

Prior over Graphs. We propose a graph prior $p(G)$ that encapsulates the knowledge that G must be a DAG, and that encourages graph sparsity,

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

where $h(G) = \operatorname{tr}\left(e^{G \odot G}\right)$ is the DAG penalty [74] and is zero if and only if G is a DAG; α and ρ control the weights of this penalty. These are gradually increased during training following an augmented Lagrangian scheme, ensuring only DAGs remain at convergence. Graph sparseness is controlled by penalising the norm $\lambda_s \|G\|_F$, with λ_s a scalar. In some circumstances, we have additional domain knowledge about the graph that does not go so far as completely specifying G. In such cases, we consider a prior matrix $W_0 \in [0,1]^{D \times D}$ that need not be binary, nor a DAG, and propose an informed prior

$$p(G) \propto \exp\left(-\lambda_s \|G - W_0\|_F^2 - \rho\, h(G)^2 - \alpha\, h(G)\right); \tag{8}$$

see Sec. G.5 for full details.

Optimization and Inference. Unfortunately, the graph posterior of the DECI model is intractable. To overcome this issue, we adopt variational inference [6] with $q_\phi(G)$ to approximate the intractable posterior $p_\theta(G|\mathbf{x}^1, \ldots, \mathbf{x}^N)$, and use it to build a lower bound on the marginal likelihood of the observational data (derivation in Appendix C.3), given by

$$\text{ELBO}(\theta, \phi) = \mathbb{E}_{q_\phi(G)}\left[\log p(G) \prod_n p_\theta(\mathbf{x}^{(n)}|G)\right] + H[q_\phi] \leq \log p_\theta(\mathbf{x}^1, \ldots, \mathbf{x}^N), \tag{9}$$

where $H[q_\phi]$ represents the entropy of the distribution q_ϕ. We choose $q_\phi(G)$ to be the product of independent Bernoulli distributions for each potential directed edge in G. We parametrize edge existence and orientation separately, using the ENCO parametrization [38]. Parameters θ and ϕ are trained by maximizing the ELBO. The Gumbel-softmax trick [28, 43] is used to estimate the gradients with respect to ϕ. Appendix C.1 details the full optimization procedure.

Theoretical considerations. One key statistical guarantee of DECI is to show that optimizing the variational objective eq. (9) leads to the recovery of the ground truth data generation mechanism. To prove this when DECI is correctly specified, i.e. when there exists a θ^*, G^0 corresponding to the true data generating process and there are no unobserved confounding variables, we need two additional technical assumptions: (1) f is non-invertible, 3rd-order differentiable and is not constant w.r.t. any of its inputs; (2) proper noise densities with bounded likelihood. These assumptions exclude, e.g. non-identifiable linear Gaussian SEMs. See Theorem 1 in Appendix B for the full statement and proof.

3.2 Estimating Treatment Effects

We use our trained DECI model consisting of the posterior $q_\phi(G)$ and a point estimate for θ to estimate (C)ATE by marginalizing over the graph posterior

$$\mathbb{E}_{q_\phi(G)}[\text{ATE}(\mathbf{a}, \mathbf{b}|G)] \quad \text{and} \quad \mathbb{E}_{q_\phi(G)}[\text{CATE}(\mathbf{a}, \mathbf{b}|\mathbf{c}, G)]. \tag{10}$$

This can be seen as a probabilistic relaxation of traditional causal quantity estimators [68].

Estimating ATE. We further propose a Monte Carlo approach to estimating $\text{ATE}(\mathbf{a}, \mathbf{b}|G)$ for a given graph. Specifically, we draw a sample from $p\left(\mathbf{x}_{\setminus T}|\text{do}(\mathbf{x}_T = \mathbf{b}), G\right)$ by (1) sampling exogenous noise variables $\mathbf{z} \sim p_{\mathbf{z}}$; (2) 'mutilating' the graph by removing incoming edges to $\mathbf{x}_T$ and setting $\mathbf{x}_T = \mathbf{b}$; (3) simulating the DECI SEM functions using eq. (3) and eq. (4) in the topological order defined by $G_{\text{do}(\mathbf{x}_T)}$ to give $\mathbf{x}_{\backslash T}(\mathbf{z}, G, \mathbf{b})$. Our sample-based ATE estimate is therefore

$$\text{ATE}(\mathbf{a}, \mathbf{b}) \approx \frac{1}{N_G} \sum_{i=1}^{N_G} \frac{1}{N_z} \sum_{j=1}^{N_z} [\mathbf{x}_Y(\mathbf{z}_{ij}, G_i, \mathbf{a}) - \mathbf{x}_Y(\mathbf{z}_{ij}, G_i, \mathbf{b})] \tag{11}$$

where $G_i \sim q_\phi(G), \mathbf{z}_{ij} \sim p_\mathbf{z}(\mathbf{z})$. Rather than simply taking a mean, we can also consider a density estimate of the samples $\mathbf{x}_Y(\mathbf{z}_{ij}, G_i, \mathbf{a}) - \mathbf{x}_Y(\mathbf{z}_{ij}, G_i, \mathbf{b})$ which incorporates uncertainty that arises from (1) incomplete knowledge of the causal graph, provided the posterior $q_\phi(G)$ has not collapsed to one graph, (2) differences between individuals in the population. See also Appendix C.5.

CATE estimation. We focus on CATE estimation for which the treatment $\mathbf{x}_T$ is not the cause of the conditioning set $\mathbf{x}_C$. Under this assumption, we can estimate CATE by sampling from the interventional distribution $p\left(\mathbf{x}_{\backslash T} | \text{do}(\mathbf{x}_T), G\right)$ and then estimating the conditional distribution of $\mathbf{x}_Y$ given $\mathbf{x}_C$. We describe our specific approach to estimate this conditional density in Appendix C.4.

Theoretical considerations. We previously showed that variational inference asymptotically recovers the correct SEM parameters θ^*. In Appendix B, we show further that a correctly specified DECI model satisfies causal effect identifiability and positivity, meaning that treatment effects will be correctly estimated in the large data limit.

3.3 General end-to-end causal inference (ECI) framework

The probabilistic treatment of the DAG, and the re-use of functional causal discovery generative models for simulation-based causal inference are principles that can be applied beyond DECI. Constraint-based [60] and score-based [8] discovery methods often output a set of DAGs compatible with the data, i.e. a PAG or CPDAG. It is natural to interpret these equivalence classes as uniform distributions over members of sets of graphs. We can then use eq. (10) to estimate treatment effects by marginalizing over these distributions. The quantities inside the expectations over graphs can be estimated using any existing causal inference method, such as linear regression [58], Double ML [7], etc. Our experiments explore combinations of discovery methods that return graph equivalence classes with standard causal inference methods.

Handling Heterogeneous Data. For real-world applications, DECI can be further extended to handle mixed-type (continuous and discrete) data and missing value imputation. Appendix E provides further details about these extensions.

4 Experiments

We evaluate DECI on both causal discovery and causal inference tasks. A full list of results and details of the experimental set-up are presented in Appendices F and G. Code to reproduce our experiments is available at `https://github.com/microsoft/causica`.

4.1 Causal Discovery Evaluation

Datasets. We consider synthetic, pseudo-real, and real data. For the synthetic data, we follow Lachapelle et al. [36] and Zheng et al. [76] by sampling a DAG from two different random graph models, **Erdős-Rényi (ER)** and **scale-free (SF)**. We consider number of nodes $d \in \{16, 64\}$ with number of edges $e \in \{d, 4d\}$. The resulting datasets are identified as **ER**(d, e) and **SF**(d, e).

For the pseudo-real data we consider the **SynTReN** generator [67], which creates synthetic transcriptional regulatory networks and produces simulated gene expression data that mimics experimental data. We use the datasets generated by [36] (d=20), and take n=400 for training. Finally, for the real dataset, we use the protein measurements in human cells from Sachs et al. [57].

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* [31], (linear) *Notears* [74], the nonlinear variants *Notears-MLP* and *Notears-Sob* [76], *Grandag* [36], and *ICALiNGAM* [59]. When a CPDAG is the output, e.g., from PC, we treat all possible DAGs under the CPDAG as having the same probability.

Causality Metrics. We report F1 scores for adjacency, orientation [15, 66] and causal accuracy [11]. For DECI, we report the expected values of these metrics estimated over the graph posterior.

Figure 1 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.

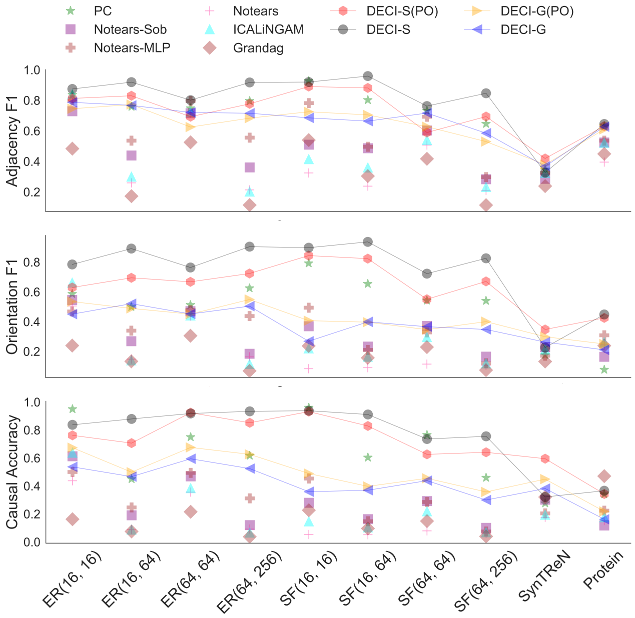


Figure 1: Causal discovery on benchmark datasets. The figure shows mean results across five random seeds.

4.2 End-to-end Causal Inference

We evaluate the *end-to-end* pipeline, taking in observational data and returning (C)ATE estimates.

Datasets. We generate ground-truth treatment effects to compare against for the **ER** and **SF** synthetic graphs that were described in Section 4.1 by applying random interventions on these synthetic SEMs, ensuring at most 3 edges between the intervention and effect variables. For more detailed analysis, we hand-craft a suite of synthetic SEMs, which we name **CSuite**. CSuite datasets are designed to elucidate particular features of the model.[2] Finally, we include two semi-synthetic causal inference benchmark datasets for ATE evaluation: **Twins** (twin birth datasets in the US) [1] and **IHDP** (Infant Health and Development Program data) [18].

Baselines. We consider different ways of *combining* discovery and inference algorithms. For DECI, we can use a trained model to immediately estimate (C)ATE. We also consider using the learned DECI graph posterior in combination with existing methods for causal inference on a known graph: DoWhy-Linear and DoWhy-Nonlinear [58] which implement linear adjustment and Double Machine Learning (DML) [7] methods for backdoor adjustment respectively. We also pair other *discovery* methods with DECI and DoWhy treatment effect estimation, namely the PC algorithm [31] as a baseline and the ground truth graph (when available) as a check. We evaluate end-to-end causal inference on all valid combinations that arise from combining *discovery* methods in {DECI-Gaussian (DGa), DECI-Spline (DSp), PC, and True graph (T)} with causal *inference* methods in {DECI-Gaussian (DGa), DECI-Spline (DSp), DoWhy-Linear (L), DoWhy-Nonlinear (N)}.

Metrics. We report RMSE between (C)ATE estimates and the ground truth.

In Table 1, we present results for six CSuite datasets. **Lin. Exp** is a two node linear SEM with exponential noise, only DECI Spline can recover the true graph; ATE estimation quality is similar for different estimator given the true graph. **Nonlin. Gauss** is a two node non-linear SEM with Gaussian noise, only DECI can fit the highly non-linear functional relationship, with equal performance between DECI-Gaussian and -Spline. **Large backdoor** is a larger non-linear SEM with non-Gaussian noise. For DECI-Spline, which performs well on discovery, the ATE estimation is best using DECI, as DoWhy takes the maximal adjustment set, which increases estimator variance on this dataset. **Weak arrows** is a larger SEM which is harder for discovery due to the presence of weak functional relationships. Here, DECI-Spline is best for discovery, but is somewhat less accurate for ATE estimation given the right graph. **Nonlin. Simpson** is an adversarially constructed dataset where 1) the true graph is theoretically identifiable, but difficult to discover in practice, 2) ATE estimation is very poor given the wrong graph (Simpson's paradox). All methods perform equally badly. **Symprod Simpson** is a similar but slightly easier dataset, for which DECI-Spline with DML does well.

[2] All CSuite datasets are publicly available at `https://github.com/microsoft/csuite`.

Table 1: Median ATE RMSEs from 20 seeds for six CSuite datasets.

	Lin. Exp	Nonlin. Gauss	Large backdoor	Weak arrows	Nonlin. Simpson	Symprod Simpson
DECI-Gaussian	1.029	**0.042**	0.213	1.097	1.995	0.318
DECI-Gaussian + DoWhy-Linear	1.031	1.522	0.144	1.108	1.994	0.695
DECI-Gaussian + DoWhy-Nonlin	1.031	1.532	0.331	1.108	1.994	0.487
DECI-Spline	0.022	**0.043**	**0.031**	0.189	1.997	0.427
DECI-Spline + DoWhy-Linear	**0.001**	1.522	0.091	0.110	1.994	0.819
DECI-Spline + DoWhy-Nonlin	0.002	1.532	0.232	**0.064**	1.994	**0.160**
PC + DoWhy-Linear	0.516	1.532	1.690	1.108	1.994	0.487
PC + DoWhy-Nonlin	0.517	1.532	1.690	1.108	1.994	0.487
True graph DECI-Gaussian	0.073	0.034	0.167	0.255	0.404	0.101
True graph DECI-Spline	0.028	0.034	0.035	0.128	0.531	0.242
True graph DoWhy-Linear	0.001	1.522	0.105	0.109	0.848	0.819
True graph DoWhy-Nonlin	0.003	1.532	0.241	0.015	0.597	0.168

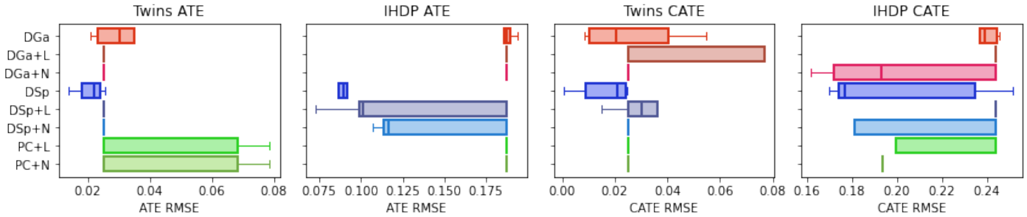


Figure 2: Box plots showing end-to-end ATE and CATE estimation error on the semi-synthetic Twins and IHDP datasets with different method combinations. Method acronyms are defined in Section 4.2. Box plots show the meadian and interquartile range over 5 seeds.

On the semi-synthetic benchmark datasets, Twins and IHDP, we evaluated both ATE and CATE estimation as shown in Figure 2. For ATE estimation, DECI-Spline is fractionally better than baselines on Twins and significantly better for IHDP. On IHDP, it appears that only DECI-Spline was successful at causal discovery, and given the right graph, DECI-Spline is the best method for computing ATE. For CATE estimation, a similar pattern emerges.

We performed similar analysis for ATE estimation on additional CSuite datasets that contain discrete variables or are not theoretically identifiable (Appendix G.2), on ER and SF datasets as large as 64 nodes (Appendix G.3), and CATE estimation on a subset of CSuite (Appendix G.2). A high-level summary of our ATE results aggregated over all 27 datasets is shown in Table 5 (page 31). We find that DECI Spline has the overall best (lowest) rank. We include additional experiments that demonstrate the utility of our domain-informed prior (Appendix G.5) and explore how DECI fares when unobserved confounders are present (Appendix G.6).

Summary Across all experiments we see that DECI enables end-to-end causal inference with competitive performance on both synthetic and more realistic data. DECI particularly performs well compared to other methods when its ability to handle nonlinear functional relationship and non-Gaussian noise distributions comes into play in causal discovery *or* causal inference.

5 Discussion, scope, and limitations

Motivated by a real-world application where our knowledge of causal relationships is incomplete, DECI combines ideas from causal discovery and inference to go directly from observations to causal predictions. This formulation requires us to adopt assumptions, namely, that the data is generated with a non-linear ANM and that there are no unobserved confounders. Empirically, we find DECI to perform well when these assumptions are satisfied, validating the viability of an end-to-end approach. Unfortunately, the non-linear ANM assumptions made by DECI are impossible to check in most real-world scenarios. Interestingly, even in our experiments where DECI's assumptions are violated , we do not find its performance to degrade severely. This encouraging result motivates us to extend our theoretical analysis to the mixed type and missing data settings in future work.

Acknowledgments and Disclosure of Funding

We 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 this project.

References

[1] Douglas Almond, Kenneth Y Chay, and David S Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.

[2] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

[3] Yashas Annadani, Jonas Rothfuss, Alexandre Lacoste, Nino Scherrer, Anirudh Goyal, Yoshua Bengio, and Stefan Bauer. Variational causal networks: Approximate bayesian inference over causal structures. *arXiv preprint arXiv:2106.07635*, 2021.

[4] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Miruna Oprescu, and Vasilis Syrgkanis. Estimating the long-term effects of novel treatments. *arXiv preprint arXiv:2103.08390*, 2021.

[5] Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2019.

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

[7] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[8] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[9] David Maxwell Chickering and Christopher Meek. Selective greedy equivalence search: Finding optimal bayesian networks using a polynomial number of score evaluations. *arXiv preprint arXiv:1506.02113*, 2015.

[10] Max Chickering. Statistically efficient greedy equivalence search. In *Conference on Uncertainty in Artificial Intelligence*, pages 241–249. PMLR, 2020.

[11] Tom Claassen and Tom Heskes. A bayesian approach to constraint based causal inference. *arXiv preprint arXiv:1210.4866*, 2012.

[12] Chris Cundy, Aditya Grover, and Stefano Ermon. Bcd nets: Scalable variational approaches for bayesian causal discovery. *Advances in Neural Information Processing Systems*, 34:7095–7110, 2021.

[13] Tristan Deleu, António Góis, Chris Emezue, Mansi Rankawat, Simon Lacoste-Julien, Stefan Bauer, and Yoshua Bengio. Bayesian structure learning with generative flow networks. *arXiv preprint arXiv:2202.13903*, 2022.

[14] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/7ac71d433f282034e088473244df8c02-Paper.pdf`.

[15] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[16] Richard Guo and Emilija Perkovic. Minimal enumeration of all possible total effects in a markov equivalence class. In *International Conference on Artificial Intelligence and Statistics*, pages 2395–2403. PMLR, 2021.

[17] David Heckerman, Christopher Meek, and Gregory Cooper. A bayesian approach to causal discovery. *Computation, causation, and discovery*, 19:141–166, 1999.

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

[19] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[20] Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors, *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*, pages 689–696. Curran Associates, Inc., 2008. URL `https://proceedings.neurips.cc/paper/2008/hash/f7664060cc52bc6f3d620bcedc94a4b6-Abstract.html`.

[21] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, Bernhard Schölkopf, et al. Nonlinear causal discovery with additive noise models. In *NIPS*, volume 21, pages 689–696. Citeseer, 2008.

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

[23] Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1551–1560, 2018.

[24] Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. In *International Conference on Machine Learning*, pages 2078–2087. PMLR, 2018.

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

[26] Guido W Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–79, 2020.

[27] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[28] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[29] Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating identifiable causal effects on markov equivalence class through double machine learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5168–5179. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/jung21b.html`.

[30] Marcus Kaiser and Maksim Sipos. Unsuitability of NOTEARS for causal graph discovery. *CoRR*, abs/2104.05441, 2021. URL `https://arxiv.org/abs/2104.05441`.

[31] Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

[32] Ilyes Khemakhem, Ricardo Monti, Robert Leech, and Aapo Hyvarinen. Causal autoregressive flows. In *International Conference on Artificial Intelligence and Statistics*, pages 3520–3528. PMLR, 2021.

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[34] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[35] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29:4743–4751, 2016.

[36] Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019.

[37] Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

[38] Phillip Lippe, Taco Cohen, and Efstratios Gavves. Efficient neural causal discovery without acyclicity constraints. *arXiv preprint arXiv:2107.10483*, 2021.

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

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

[41] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *arXiv preprint arXiv:1705.08821*, 2017.

[42] Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. *arXiv preprint arXiv:1809.11142*, 2018.

[43] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[44] AS Nemirovsky. Optimization ii. numerical methods for nonlinear continuous optimization. 1999.

[45] Ignavier Ng, Shengyu Zhu, Zhitang Chen, and Zhuangyan Fang. A graph autoencoder approach to causal structure learning. *arXiv preprint arXiv:1911.07420*, 2019.

[46] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *arXiv preprint arXiv:2006.10201*, 2020.

[47] J Pearl. Causal diagrams and the identification of causal effects. *Causality: Models, Reasoning, and Inference. Cambridge University Press: New York*, 2000.

[48] J Pearl. The logic of structure-based counterfactuals. *Causality: Models, Reasoning and Inference*, pages 201–259, 2009.

[49] Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[50] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19, 2000.

[51] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 2014.

[52] Alexander G. Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! varsortability in additive noise models. *CoRR*, abs/2102.13647, 2021. URL `https://arxiv.org/abs/2102.13647`.

[53] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

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

[55] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[56] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[57] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

[58] Amit Sharma, Vasilis Syrgkanis, Cheng Zhang, and Emre Kıcıman. Dowhy: Addressing challenges in expressing and validating causal assumptions. *arXiv preprint arXiv:2108.13518*, 2021.

[59] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

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

[61] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[62] Peter L Spirtes. Directed cyclic graphical representations of feedback models. *arXiv preprint arXiv:1302.4982*, 2013.

[63] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

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

[65] Elizabeth A Stuart. 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.

[66] Ruibo Tu, Cheng Zhang, Paul Ackermann, Karthika Mohan, Hedvig Kjellström, and Kun Zhang. Causal discovery in the presence of missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1762–1770. PMLR, 2019.

[67] Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics*, 7(1):1–12, 2006.

[68] Jussi Viinikka, Antti Hyttinen, Johan Pensar, and Mikko Koivisto. Towards scalable bayesian learning of causal dags. *Advances in Neural Information Processing Systems*, 33:6584–6594, 2020.

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

[70] Felix Xinnan X Yu, Ananda Theertha Suresh, Krzysztof M Choromanski, Daniel N Holtmann-Rice, and Sanjiv Kumar. Orthogonal random features. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/53adaf494dc89ef7196d73636eb2451b-Paper.pdf`.

[71] K ZHANG. 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.

[72] Kun Zhang and Lai-Wan Chan. Extensions of ica for causality discovery in the hong kong stock market. In *International Conference on Neural Information Processing*, pages 400–409. Springer, 2006.

[73] Kun Zhang, Zhikun Wang, Jiji Zhang, and Bernhard Schölkopf. 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.

[74] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *arXiv preprint arXiv:1803.01422*, 2018.

[75] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf`.

[76] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR, 2020.

A Related Work

Our work relates to research on both causal discovery and causal inference. Approaches for causal discovery from observational data can be classified into three groups: constraint-based, score-based, and functional causal models [15]. Recently, Zheng et al. [74] framed the directed acyclic graph (DAG) structure learning problem as a continuous optimisation task. Extensions building on this continuous optimisation framing [36, 76] employ nonlinear function approximators, like neural networks, to model the relationships among connected variables. Similarly, recent work has combined the continuous DAG learning approach with variational inference [40, 3, 12, 13]. However, these works primarily focus on linear Gaussian models [3, 12, 13]. Only Lorch et al. [40] considers non-linear functional forms but remains limited to Gaussian noise models. Our model allows both non-linear functions and a learned non-Gaussian noise model. Appendix D shows that DECI can be seen as a generalization of the *Notears* family of discovery methods. By combining the continuous optimisation approach to discovery with standard causal assumptions [45], we obtain our main theorem about causal graph learning. To the best of our knowledge, we are the first to present an identifiability proof for causal discovery using VI. We also extend functional methods to handle mixed data types and missing values, which was not previous addressed in this context of continuous causal discovery. Outside of functional causal discovery, functional relationships between variables are typically not learned by discovery algorithms [60]. Thus, distinct models, with potentially incompatible assumptions or inputs, must be relied upon for causal inference. However, when a DAG cannot be fully identified given the available data, constraint and score-based methods often return partially directed acyclic graphs (PAGs) or completed partially directed acyclic graphs (CPDAGs) [61]. Instead of returning a summary graph representing a set, DECI returns a distribution over DAGs in such situation.

Causal inference methods assume that either the graph structure is provided [49] or relevant structural assumptions are provided without the graph [27]. 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 and front-door criteria [49], and instrumental variables [2]. Causal estimation computes the identified estimand using statistical methods, such as simple conditioning, inverse propensity weighting [37], or matching [54, 65]. Machine learning-based estimators for CATE have also been proposed [7, 69]. Recent efforts to weaken structural assumption requirements [16, 29] allow for PAGs and CPDAGs. Our work takes steps in this direction, allowing inference with distributions over graphs.

A.1 Relationship with Khemakhem et al. [32]

Khemakhem et al. [32] introduced *Carefl*, a method that uses autoregressive flows [24, 35] 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, they run a traditional constraint-based method (e.g., PC) to find the graph's skeleton and orient as many edges as possible, and second, they 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. [32] to learn the causal graph among multiple variables and perform end-to-end causal inference. See Appendix D for further details on the connection between DECI and flow-based causal discovery.

B Theoretical Considerations for DECI

DECI can be categorized as a functional score-based causal discovery approach, which aims to find the model parameters θ and mean-field posterior $q_\phi(G)$ by maximizing the ELBO given in eq. (9). A key statistical question we may ask 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** and with infinite data. In the following, we will show that DECI is indeed capable of this under standard assumptions.

The main idea is to first show that the maximum likelihood estimate (MLE) recovers the ground truth graph and data generating distribution due to the correctly specified nin-linear additive noise model. Then, we prove that optimal solutions from maximizing the ELBO are closely related to the MLE in the large data limit. This is formalized in the following theorem.

Theorem 1 (DECI recovers the true data generating process)**.** *Assume DECI belongs to continuous additive noise structural equation models with (1) no latent confounders; (2) non-invertible, third order differentiable functional relations between variables that are not constant w.r.t. any of their inputs (i.e. non-trivial); (3) proper noise density and regular log likelihood, then under a correctly specified model, the solution $(\theta', q'_\phi(G))$ from maximizing the ELBO (eq. (9)) satisfies $q'_\phi(G) = \delta(G = G')$ where G' is a unique graph. In particular, $G' = G^0$ and $p_{\theta'}(\mathbf{x}; G') = p(\mathbf{x}; G^0)$.*

Our proof strategy is as follows. In Appendix B.1, we show that the assumptions of the theorem, particularly the assumption that DECI is well-specified, imply certain well-known properties, such as structural identifiability. In Appendix B.2, we prove that maximum likelihood estimation identifies the correct data generating process, and in Appendix B.3, we complete the proof by showing that maximizing the ELBO asymptotically recovers the MLE. Appendix B.4 discusses additional theoretical results for treatment effect estimation.

B.1 Notation and Assumptions

First, we define the notation and explain the implication of assumptions required for our theory. 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 causal DAG. DECI is an additive noise model (ANM), defining the structural equations $x_j = f(\mathbf{x}_{\text{pa}(j;G)}; \theta) + z_j$, where $\text{pa}(j; G)$ are the parents of node j specified by the adjacency matrix G, f are non-invertible, third-order differentiable non-trivial functions and z_j are mutually independent noise variables with a joint density $p_\theta(z_1, \ldots, z_D)$. 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).

For completeness, we first provide the definition of causal minimality, structural identifiability, and explain our model specification and regularity assumptions.

Definition 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 [62]. Namely, the distribution $p_\theta(\mathbf{x}; G)$ does not satisfy the local Markov condition with respect to any sub-graph of G.*

Definition 2 (DECI Structural Identifiability)**.** *We assume that the DECI model 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 $p_\theta(\mathbf{x}; G) = p_{\theta'}(\mathbf{x}; G')$.*

The following assumptions are explicit assumptions of Theorem 1, and are restated more formally now, along with a discussion of when such assumptions are likely to hold in practice.

Assumption 1 (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 [19], meaning that they can approximate continuous functions arbitrarily well. This flexibility gives us a higher chance that this assumption indeed holds.

Assumption 2 (Regularity of log likelihood)**.** *We assume for all parameters θ and possible graphs G, the following holds:*

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

This assumption can be verified by analysing the tail behaviour of the noise distribution of the DECI SEM.

We can now show that the our assumptions about the DECI model imply some of the commonly used causal assumptions.

Lemma 1 (Implication of DECI assumptions)**.** *Assume DECI satisfies the assumptions mentioned in Theorem 1. Then DECI satisfies (1) the* causal Markov assumption*; (2)* causal minimality*; (3)* structural identifiability.

Proof. We show how each assumptions are implied by referring to some theorems from previous works.

Causal Markov assumption. Since DECI belongs to the class of SEMs with proper probability density functions, Theorem 1.4.1 in Pearl et al. [50] proves that its induced joint probability is Markov with respect to the graph.

Causal minimality. Since DECI's function relationships are not constant w.r.t. any of their inputs, Proposition 16 in Peters et al. [51] proves that DECI satisfies causal minimality.

Structural identifiability. Since DECI satisfies causal minimality, and the functions are non-invertible and third-order differentiable, Theorem 19 and Proposition 22 in Peters et al. [51] prove that the corresponding DECI SEM is structural identifiable. □

B.2 MLE Recovers Ground Truth

The likelihood has often been used as the score function for causal discovery. For example, *Carefl* [32] adopts the likelihood ratio test [25] 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 causal graph for the multivariate case. In addition, Zhang et al. [73] 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 this to the multivariate case that we treat in 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 conditions in Theorem 1 hold, we denote (θ^*, G^*) as the MLE solution with infinite training data. Then, we have*

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

and 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 $\log t \leq t - 1$. From the assumptions, 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 from Lemma 1, we must have $G^* = G^0$. □

B.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 2. *Assume a 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*

$$\lim_{N\to\infty} \frac{1}{N}\mathrm{KL}[q_\phi(G)\|p(G)] = 0. \tag{12}$$

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. [75], 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

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

where the third inequality is obtained by using Cauchy-Schwarz inequality. Thus, we have

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

□

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

Theorem 1 (DECI recovers the true distribution)**.** *Assuming conditions in Theorem 1 are true, the solution $(\theta', q'_\phi(G))$ from maximizing ELBO (eq.* (9)*) in the infinite data limit 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. (9)) as

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

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

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

where the second and third equalities are from Lemma 2 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

$$\mathbb{E}_{p(\mathbf{x};G^0)}\left[\log p_{\theta_G}(\mathbf{x}|G)\right]=\mathbb{E}_{p(\mathbf{x};G^0)}\left[\log p_{\theta^*}(\mathbf{x}|G^*)\right]. \tag{13}$$

From Proposition 1, under a 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. (13) 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 (Appendix E). 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.4 Theoretical considerations for treatment effect estimation

Our proposed DECI model provides an end-to-end framework to connect causal discovery and inference. The assumptions introduced in Appendix B.1 are mainly introduced for the causal discovery task. In the following, we will discuss their relationships to some common assumptions required for the causal inference task. In particular, we show that no additional assumptions are needed to guarantee that DECI can correctly estimate treatment effects.

Causal effect identifiability and positivity. Under the assumed DECI SEM (non-linear additive noise model with independent noise variables), our model is Markovian, meaning that the well-known causal effect identification theorem for Markovian models naturally holds (Theorem 3.2.5 of Pearl [47]). This result does assume positivity of observational distributions.

Fortunately, the proposed form of DECI implies positivity. In particular, DECI is based on non-linear additive noise models. One of the fundamental assumption of non-linear ANMs is that the joint distribution should have a proper density function, i.e. it should be absolutely continuous w.r.t. Lebesgue measure (see Assumption 2). This means the density function is non-negative. Additionally, since the noise model for DECI is either a Gaussian or a spline transformed Gaussian, the measure of the set where the density is zero must have a measure zero. This means DECI will have a positive density almost everywhere, which in turn implies positivity.

A note on consistency assumptions. A consistency assumption is not required by our method, since these are only standard within the potential outcome formulation for causality of Rubin [55], and are rarely seen in graphical model based approach [47] to causality used in our paper. The graphical model approach works fundamentally differently than potential outcomes (PO): we start by defining a causal model, and all other counterfactual and interventional quantities are derived from this causal model. On the contrary, the PO approach treats counterfactuals (i.e. potential outcomes) as primitive random variables that are not observed. Therefore, the PO approach needs assumptions such as consistency, no interference, etc. to ensure the coherent mathematical consistency of counterfactual statements. See Imbens [26], Pearl [48], for example, for more discussions on PO approach versus graphical approach to causality.

C Additional Details for DECI

C.1 Optimization Details for Causal Discovery

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 [44]. 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 [33]. We optimize the objective for a maximum of 6000 steps or until convergence, whichever happens first (we stop early if the loss does not improve for 1500 optimization 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 has 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 100 steps or until convergence (measured as α or ρ reaching some max value which we set to 10^{13} for both), whichever happens first.

C.2 Other Hyperparameters

We use $\lambda_s = 5$ in our prior over graphs eq. (7). For ELBO Monte Carlo gradients we use the Gumbel softmax method [43] with a hard forward pass and a soft backward pass with temperature of 0.25.

The functions eq. (4) used in DECI's SEM, ζ and ℓ, are 2 hidden layer MLPs with 128 hidden units per hidden layer. These MLPs use residual connections and layer-norm at every hidden layer.

For the non-Gaussian noise model in eq. (5), the bijection κ is an 8 bin rational quadratic spline [14] with learnt parameters.

In Section 3.2, for ATE estimation we compute expectations by drawing 1000 graphs from DECI's graph posterior q_ϕ and for each graph we draw 2 samples of $\mathbf{x}_Y$ for a total of 2000 samples. For CATE estimation, we need to train a separate surrogate predictor per graph samples. We draw 10 different graph samples and 10000 $(\mathbf{x}_C, \mathbf{x}_Y)$ pair samples for each graph. We use these to train the surrogate models.

Our surrogate predictor is a basis function linear model with 3000 random Fourier features drawn such that the model approximates a Gaussian process with a radial basis function kernel of length-scale equal to 1 [70].

C.3 ELBO Derivation

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

$$\log p_\theta(x^1, \dots, x^N) = \log \sum_{G \in \mathcal{G}} p(G) \prod_n p_\theta(x^n|G). \tag{14}$$

Marginalising G in 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 data dimensionality D).

Variational inference proposes to use a distribution $q_\phi(G)$ to build the ELBO, a lower bound of the objective from eq. (14), as follows:

$$\begin{aligned}
\log p_\theta(x^1, \dots, x^N) &= \log \sum_G p(G) \prod_n p_\theta(x^n|G) && (15)\\
&= \log \sum_G q_\phi(G) \frac{p(G) \prod_n p_\theta(x^n|G)}{q_\phi(G)} && (16)\\
&= \log \mathbb{E}_{q_\phi(G)} \left[\frac{p(G) \prod_n p_\theta(x^n|G)}{q_\phi(G)} \right] && (17)\\
&\geq \mathbb{E}_{q_\phi(G)} \left[\log \frac{p(G) \prod_n p_\theta(x^n|G)}{q_\phi(G)} \right] \qquad \text{(Jensen's inequality)} && (18)\\
&= \mathbb{E}_{q_\phi(G)} \left[\log p(G) \prod_n p_\theta(x^n|G) \right] + H[q_\phi] && (19)\\
&= \mathrm{ELBO}(\phi, \theta), && (20)
\end{aligned}$$

where we denote $H[q_\phi] = -\mathbb{E}_{q_\phi(G)} \log q_\phi(G)$ for 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, $\mathrm{KL}(q_\phi(G) \| p_\theta(G|x^1 \dots, x^N))$ (see, e.g. Blei et al. [6]). This is why q_ϕ can be used as a posterior approximation.

C.4 CATE estimation

We focus on CATE estimation for which the treatment $\mathbf{x}_T$ is not the cause of the conditioning set $\mathbf{x}_C$, i.e. there is no directed path from T to C in G. Under this assumption, we can estimate CATE by sampling from the interventional distribution $p\left(\mathbf{x}_{\setminus T}|\mathrm{do}(\mathbf{x}_T), G\right)$ and then estimating the conditional distribution of $\mathbf{x}_Y$ given $\mathbf{x}_C$. Precisely, we let $Y = X \setminus (T \cup C)$ denote all variables that we do not intervene or condition on. Conditional densities

$$p_\theta(\mathbf{x}_Y \mid \mathrm{do}(\mathbf{x}_T{=}\mathbf{b}), \mathbf{x}_C{=}\mathbf{c}, G) = \frac{p_\theta(\mathbf{x}_Y, \mathbf{x}_C{=}\mathbf{c}|\mathbf{x}_T{=}\mathbf{b}, G_{\mathrm{do}(\mathbf{x}_T)})}{p_\theta(\mathbf{x}_C{=}\mathbf{c}|\mathbf{x}_T{=}\mathbf{b}, G_{\mathrm{do}(\mathbf{x}_T)})} \tag{21}$$

are not directly tractable due to the marginal distribution. However, we can always sample from the joint interventional distribution $p_\theta(\mathbf{x}_Y, \mathbf{x}_C|\mathbf{x}_T{=}\mathbf{b}, G_{\mathrm{do}(\mathbf{x}_T)})$. These samples can be used to train a surrogate regression model g_G[3] to the relations between $\mathbf{x}_C$ and $\mathbf{x}_Y$ by minimizing the square loss

$$\mathbb{E}_{p_\theta(\mathbf{x}_Y, \mathbf{x}_C|\mathbf{x}_T=\mathbf{b}, G_{\mathrm{do}(\mathbf{x}_T)})} \left[\|\mathbf{x}_Y - g_G(\mathbf{x}_C)\|^2 \right], \tag{22}$$

[3] Subscript G allows differentiating surrogate models fit on samples from different graphs drawn from q_ϕ.

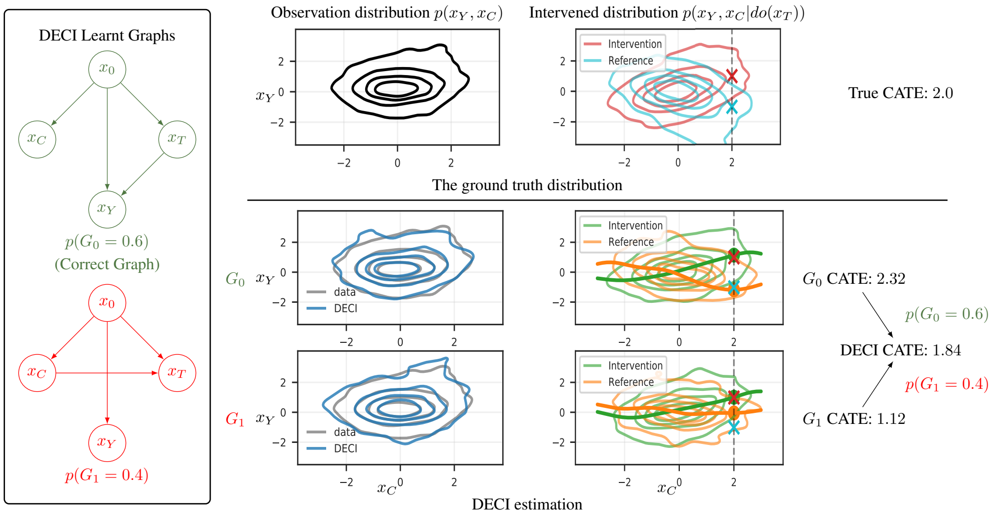


Figure 3: DECI CATE estimation on the CSuite Symprod Simpson dataset. Left: The DECI graph posterior has two modes with $p(G) = 0.6$ for the correct graph and $p(G) = 0.4$ for an alternative possibility with some incorrect edges. Middle: we display the joint distribution of conditioning and effect variables in the observational setting and under interventions on $\mathbf{x}_T$. DECI captures the observational density well. Right: interventional distributions with their conditional means $\mathbf{x}_C = \mathbf{c}$ marked with crosses. DECI predicts conditional expectations by fitting functions from $\mathbf{x}_C$ to $\mathbf{x}_Y$ and evaluating them at $\mathbf{c}$. DECI outputs CATE by marginalizing the result over possible graphs.

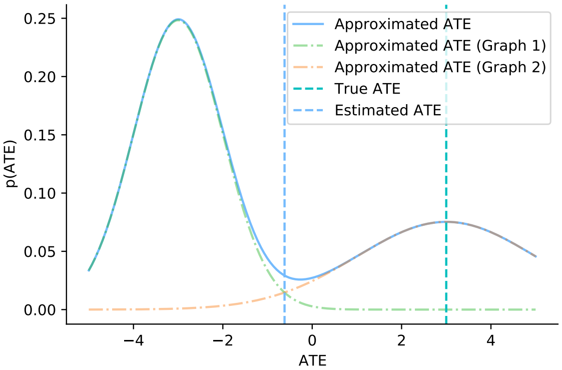


Figure 4: Marginalizing ATE estimates from graph and exogenous noise samples.

where g_G approximates the conditional mean of $\mathbf{x}_Y$. We choose g_G to be a basis-function linear model with random Fourier basis [70]. We train two separate surrogate models for intervention $\mathbf{x}_T = \mathbf{a}$ and reference $\mathbf{x}_T = \mathbf{b}$, respectively. We estimate CATE as the difference between their outputs evaluated at $\mathbf{x}_C = \mathbf{c}$, followed by marginalising over different sampled graphs, giving

$$\text{CATE}(\mathbf{a}, \mathbf{b}|\mathbf{c}) \approx \mathbb{E}_{q_\phi(G)}\left[g_{G_{\text{do}(\mathbf{x}_T=\mathbf{a})}}(\mathbf{x}_C = \mathbf{c}) - g_{G_{\text{do}(\mathbf{x}_T=\mathbf{b})}}(\mathbf{x}_C = \mathbf{c}))\right]. \tag{23}$$

Our CATE estimation procedure is illustrated in Figure 3.

C.5 Intervened Density Estimation with DECI

Apart from (C)ATE estimation, DECI may also be used to evaluate densities under intervened distributions. For a given graph, the density of some observation vector $\mathbf{a}$ is computed by evaluating the base distribution density after inverting the SEM

$$p_\theta(\mathbf{x} = \mathbf{a}|G^m) = \prod_i p(\mathbf{z}_i = (\mathbf{a}_i - f_i(\mathbf{a}_{\text{pa}(i;G^m)}))) \tag{24}$$

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

$$p_\theta(\mathbf{x} = \mathbf{a}) \approx \frac{1}{M} \sum_{m}^{M} p_\theta(\mathbf{x} = \mathbf{a} | G^m); \quad G^m \sim q_\phi(G). \tag{25}$$

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

$$\begin{aligned} & p_\theta(\mathbf{x}_{\setminus T} = \mathbf{a} | \text{do}(\mathbf{x}_T = \mathbf{b}), G^m) \\ & \quad = \frac{p_\theta(\mathbf{x}_{\setminus T} = \mathbf{a}, \mathbf{x}_T = \mathbf{b} | G^m_{\text{do}(\mathbf{x}_T)})}{p_\theta(\mathbf{x}_T = \mathbf{b} | G^m_{\text{do}(\mathbf{x}_T)})} \\ & \quad = \frac{p_\theta(\mathbf{x}_{\setminus T} = \mathbf{a} | \mathbf{x}_T = \mathbf{b}, G^m_{\text{do}(\mathbf{x}_T)}) p_\theta(\mathbf{x}_T = \mathbf{b})}{p_\theta(\mathbf{x}_T = \mathbf{b})} \\ & \quad = \prod_{j \in \setminus T} p(\mathbf{z}_i = (\mathbf{a}_i - f_i(\mathbf{a}_{\text{pa}(i; G^m_{\text{do}(\mathbf{x}_T)})}))), \end{aligned} \tag{26}$$

which amounts to evaluating the density of the exogenous noise corresponding to non-intervened variables. We can then marginalise the graph using Monte Carlo as in eq. (25). See Figure 4 for an illustration of estimating interventional densities with two graph.

C.6 Discussion on Causal Discovery Methods

When performing causal discovery, DECI returns a posterior over graphs. Most other causal discovery methods return either a single graph or an equivalence class of graphs. However, we can re-cast these methods in the probabilistic framework used by DECI by noting that a posterior over graphs takes the form

$$p(G|\mathbf{X}) = \frac{p(\mathbf{X}|G)p(G)}{\sum_G p(\mathbf{X}|G)p(G)}. \tag{27}$$

In this equation, the likelihood measures the degree of compatibility of a certain DAG architecture with the observed data. For score-based discovery methods [8, 9, 10, 23] we take the score to be $\log p(\mathbf{X}|G)$. For functional discovery methods [21, 59, 71] we use the exogenous variable log-density. Constraint-based methods [60, 61] 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 [20].

D 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. [76] for details.

Lemma 3. *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 $p_\mathbf{z}$ and a hard acyclicity constraint on $W(\theta)$ is equivalent to solving*

$$\max_\theta \sum_{n=1}^{N} \log p_\mathbf{z}(\mathbf{x}^n - f(\mathbf{x}^n;\theta)) \quad \text{s.t.} \quad h(W(\theta)) = 0, \tag{28}$$

where $h(\cdot)$ is the algebraic characterization of DAGs, $h(W) = \operatorname{tr}\left(e^{W\odot W}\right)$.

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

$$\sum_n \log p_\theta(\mathbf{x}^n) = \sum_n \log p_z(\mathbf{x}^n - f(\mathbf{x}^n;\theta)) + \log \left| \det \frac{\mathrm{d}(\mathbf{x}^n - f(\mathbf{x}^n;\theta))}{\mathrm{d}\mathbf{x}^n} \right| \tag{29}$$

$$= \sum_n \log p_z(\mathbf{x}^n - f(\mathbf{x}^n;\theta)), \tag{30}$$

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 $\theta \in \Theta$, and the second equality uses that the function $f(\mathbf{x}^n;\theta)$ has Jacobian-determinant equal to 1, due to the constraint $\Theta = \{\theta : h(W(\theta)) = 0\}$. □

Lemma 3 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. (28) with specific choices for $f(\mathbf{x};\theta)$ and $p_\mathbf{z}$.

***Notears* [74]** uses a standard isotropic Gaussian for $p_\mathbf{z}$ and a linear transformation for $f(\mathbf{x},\theta)$. (This is similar to DECI-Gaussian, although DECI permits fully nonlinear functions.)

***Notears-MLP* [76]** uses a standard isotropic Gaussian for $p_\mathbf{z}$ and D independent multi-layer perceptrons, one for each component of $f(\mathbf{x},\theta)$.

***Notears-Sob* [76]** uses a standard isotropic Gaussian for $p_\mathbf{z}$ and a weighted linear combination of nonlinear basis functions.

***GAE* [45]** uses a standard isotropic Gaussian for $p_\mathbf{z}$ and a GNN for $f(\mathbf{x},\theta)$.

***Grandag* [36]** uses a factorized Gaussian with mean zero and learnable scales for $p_\mathbf{z}$ and D multi layer perceptrons, one for each component of $f(\mathbf{x},\theta)$.

***Golem* [46].** 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 [30, 39, 52].

E DECI for Real-world Heterogeneous Data

We extend DECI to handle mixed-type (continuous and discrete) data and data with missing values, which often arise in real-world applications.

Handling Mixed-type Data. For discrete-valued variables, we remove the additive noise structure and directly parameterise parent-conditional class probabilities

$$p_\theta^{\text{discrete}}\left(x_i|\mathbf{x}_{\text{pa}(i;G)}; G\right) = P_i\left(\mathbf{x}_{\text{pa}(i;G)};\theta\right)(x_i), \tag{31}$$

where $P_i\left(\mathbf{x}_{\text{pa}(i;G)};\theta\right)$ is a normalised probability mass vector over the number of classes of x_i, obtained by applying the softmax operator to $f_i(\mathbf{x}_{\text{pa}(i;G)})$. This means that for discrete variables, the output of f_i is a vector of length equal to the number of classes for variable i. This approach gives a valid likelihood for $p_\theta(\mathbf{x}^n|G)$ which we use to train DECI. However, since the full generative model is no longer an ANM, we cannot guarantee that Theorem 1 applies in this setting.

Handling Missing Data. We propose an extension of DECI to partially observed data.[4] 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 their joint density in the observational environment is $p_\theta(\mathbf{x}_o^n, \mathbf{x}_u^n)$. We approximate the posterior $p(G, \mathbf{x}_u^n|\mathbf{x}_o^n)$ with the variational distribution,

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

which yields the following learning objective

$$\text{ELBO}(\theta, \phi, \psi) = H(q_\phi) + \sum_n H(q_\psi(\mathbf{x}_u^n|\mathbf{x}_o^n)) + \mathbb{E}_{q_{\phi,\psi}}\left[\log p(G)\prod_n p_\theta(\mathbf{x}_o^n, \mathbf{x}_u^n|G)\right]. \quad (32)$$

We parameterize the Gaussian imputation distribution $q_{\psi_n}(\mathbf{x}_u^n|\mathbf{x}_o^n)$ using an amortization network [34], whose input is $\mathbf{x}_o^n$, and output the mean and variance of the imputation distribution $q_\psi(\mathbf{x}_u^n|\mathbf{x}_o^n)$.

F Datasets Details

ER and SF graphs. For the synthetic data, we follow Lachapelle et al. [36] and Zheng et al. [76] by sampling a DAG from two different random graph models, **Erdős-Rényi (ER)** and **scale-free (SF)**, and simulating each ANM $x_i = f_i(\mathbf{x}_{\text{pa}(i;G)}) + z_i$, where f_i is a nonlinear function (randomly sampled spline). We consider two noise distributions for z_i, a standard Gaussian and a more complex one obtained by transforming samples from a standard Gaussian with an MLP with random weights. We consider number of nodes $d \in \{16, 64\}$ with number of edges $e \in \{d, 4d\}$. The resulting datasets are identified as **ER**(d, e) and **SF**(d, e). All datasets have n=5000 training samples.

Our two benchmark datasets (Twins and IHDP) are constructed following similar procedures described in Louizos et al. [41].

IHDP [18]. 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 [18]; 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 model, and are only used for the purpose of (C)ATE 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 [18], which we 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 [1]. This dataset consists of twin births in the US between 1989 and 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 twin 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 is only used for evaluation. The raw dataset is downloaded from `https://github.com/AMLab-Amsterdam/CEVAE`. Following Louizos et al. [41], 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

[4] We assume that values are missing (completely) at random, the most common setting [42, 56, 63, 64].

and w_h are randomly generated as $w_0 \sim \mathcal{N}(0, 0.1I)$, $w_h \sim \mathcal{N}(5, 0.1)$. 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 a conditioning set. This allows unbiased estimation of the 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 structural equatio models with the intention of testing different capabilities of causal discovery and inference methods. All continuous-only datasets take the form of additive noise models. Further details about CSuite, and direct access to the data is provided at `https://github.com/microsoft/csuite`.

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 5a) 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 5a) 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 5a) 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 5b: 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 5c. 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 5d. 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 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 5e. 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 5a) 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 5a) 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

$$p(X_2|X_1 = x_1) = \begin{cases} \left(\frac{6}{13}, \frac{6}{13}, \frac{1}{13}\right) & \text{if } x_1 < -\frac{\sqrt{3}}{3} \\ \left(\frac{1}{8}, \frac{3}{4}, \frac{1}{8}\right) & \text{if } -\frac{\sqrt{3}}{3} \leq x_1 < \frac{\sqrt{3}}{3} \\ \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) & \text{if } x_1 > \frac{\sqrt{3}}{3} \end{cases} \tag{33}$$

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 5b, 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 5d, but with X_8 binary.

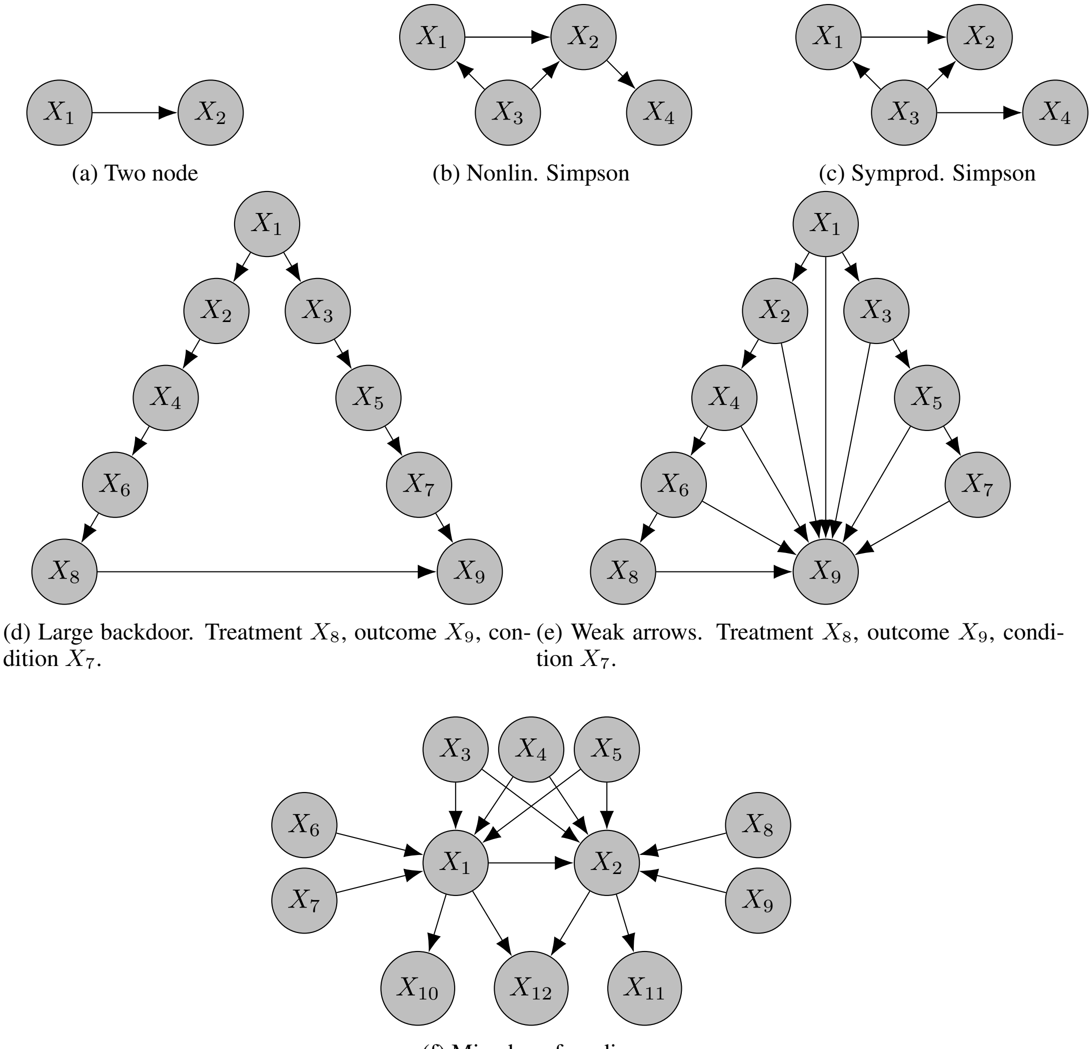
weak_arrows_binary_t Similar to the weak_arrows dataset, using the graph of Figure 5e, but with X_8 binary.

mixed_confounding A large, mixed type dataset with 12 variables, as shown in Figure 5f. 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.

G Additional Results

G.1 Causal Discovery Results under Gaussian Exogenous Noise

Figure 1 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, although DECI with a spline noise mode still performs strongly.



(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 5: 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.

G.2 CSuite Results

Comprehensive results on CSuite ATE and CATE performance are shown in Figures 7 and 8. We first provide a summary of results here and then go into per-dataset analysis in the following subsection.

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 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 high 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 funci-

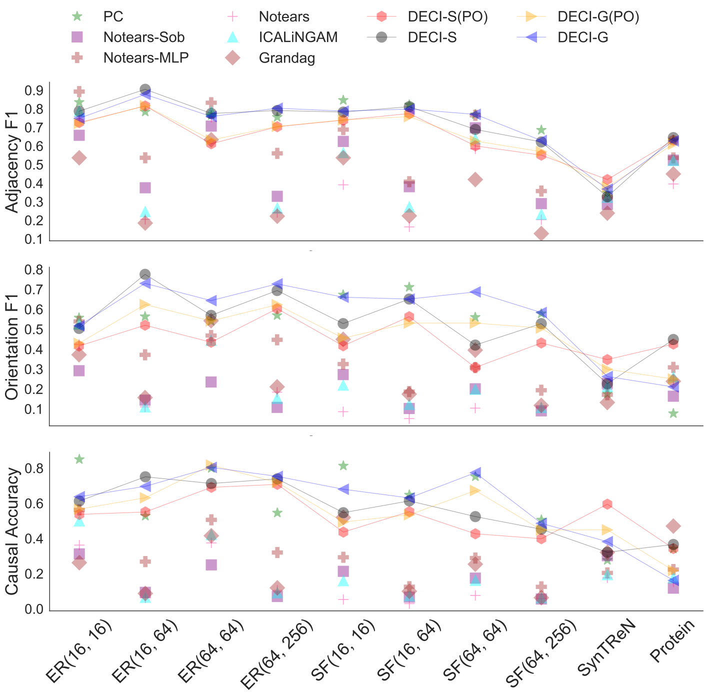


Figure 6: **DECI achieves better results than the baselines in all metrics shown.** The plots show the results for causal discovery for synthetic data generated using Gaussian noise. The legend “DECI-G" and “DECI-S" correspond to DECI using a Gaussian and spline noise model. Additionally, the “(PO)” 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.

tonal 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.

When the graph is learnt from the data, the non-linear nature of our (4 and 12 node) 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. In this learnt graph setting, 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.

G.2.1 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 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

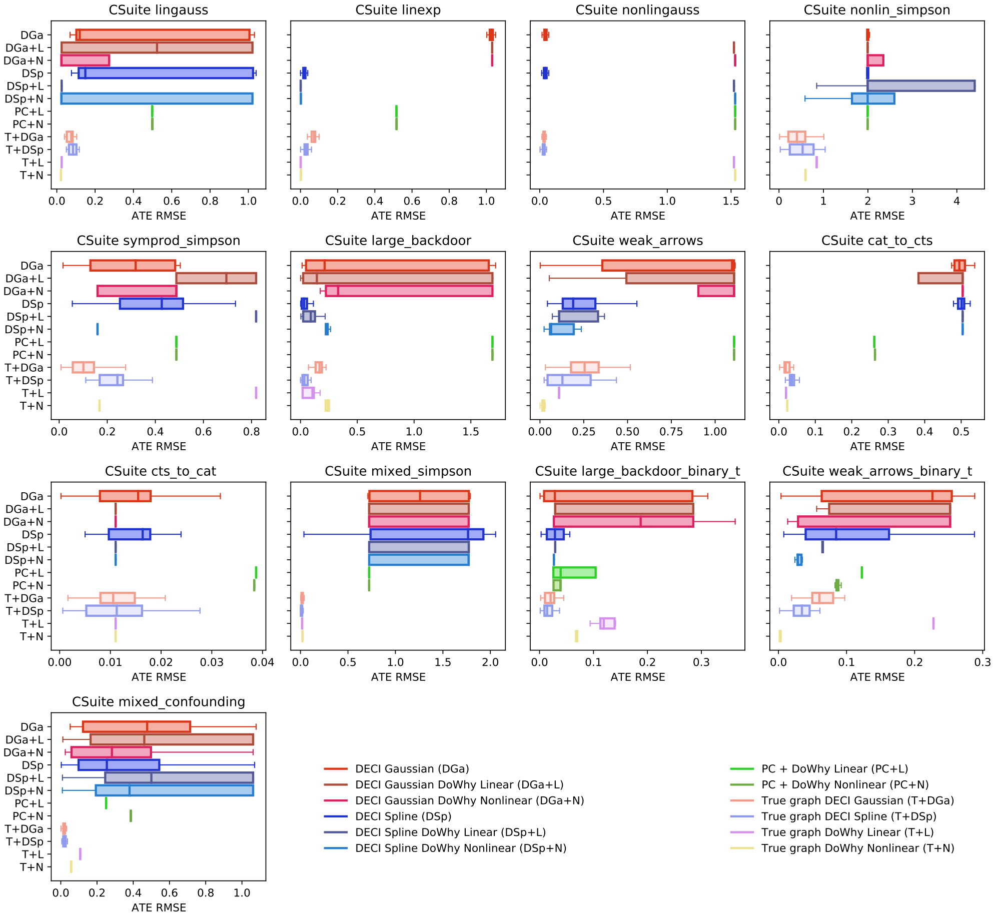


Figure 7: End-to-end ATE results on CSuite.

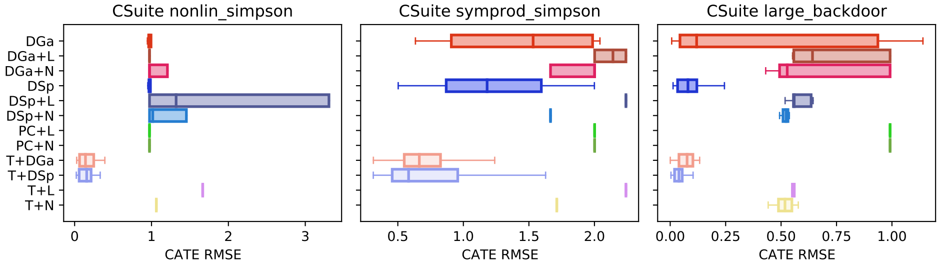


Figure 8: End-to-end CATE results on CSuite. Colours and acronyms as in Figure 7.

Method	DGa	DGa+L	DGa+N	DSp	DSp+L	DSp+N	PC+L	PC+N	T+L	T+N	T+DGa	T+DSp
Dataset												
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. Standard deviations are also shown in Table 4. Missing values indicate that the method exceeded the computational budget—this typically occurred for larger graphs.

Method	DGa	DGa+L	DGa+N	DSp	DSp+L	DSp+N	PC+L	PC+N	T+L	T+N	T+DGa	T+DSp
Dataset												
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.

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

Method	DGa	DGa+L	DGa+N	DSp	DSp+L	DSp+N	PC+L	PC+N	T+L	T+N	T+DGa	T+DSp
Dataset												
csuite_cat_to_cts	0.107	0.210	0.108	0.105	0.194	0.172	0.000	0.000	0.011	0.011	0.000	0.000
csuite_cts_to_cat	0.007	0.000	0.000	0.007	0.000	0.000	0.000	0.000	0.009	0.009	0.000	0.000
csuite_large_backdoor	0.724	0.737	0.724	0.041	0.062	0.022	0.000	0.000	0.046	0.028	0.055	0.031
csuite_large_backdoor_bt	0.134	0.125	0.124	0.081	0.077	0.056	0.033	0.006	0.011	0.010	0.037	0.004
csuite_linexp	0.013	0.000	0.000	0.014	0.000	0.000	0.000	0.000	0.015	0.015	0.000	0.000
csuite_lingauss	0.435	0.498	0.432	0.454	0.355	0.489	0.000	0.001	0.020	0.022	0.000	0.000
csuite_mixed_confounding	0.371	0.389	0.353	0.369	0.418	0.422	0.000	0.000	0.010	0.010	0.000	0.000
csuite_mixed_simpson	0.524	0.522	0.507	0.585	0.522	0.514	0.000	0.000	0.010	0.008	0.000	0.000
csuite_nonlin_simpson	1.071	0.593	0.976	1.080	1.326	1.409	0.000	0.000	0.283	0.304	0.000	0.000
csuite_nonlingauss	0.014	0.000	0.000	0.016	0.000	0.000	0.000	0.000	0.013	0.016	0.000	0.000
csuite_symprod_simpson	0.187	0.162	0.162	0.180	0.000	0.128	0.000	0.000	0.090	0.124	0.000	0.000
csuite_weak_arrows	0.433	0.401	0.426	0.256	0.118	0.075	0.000	0.000	0.160	0.131	0.000	0.008
csuite_weak_arrows_bt	0.104	0.118	0.113	0.093	0.067	0.067	0.000	0.002	0.020	0.017	0.000	0.001
IHDP	0.021	0.000	0.062	0.013	0.048	0.037	0.024	0.024	0.014	0.022	0.000	0.000
Twins	0.018	0.000	0.000	0.004	0.022	0.019	0.023	0.024	0.003	0.009	0.000	0.000

Table 4: Standard deviations for ATE RMSE results.

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.

G.2.2 Discussion of Mixed-type CSuite Results

1. **cat_to_cts**: Causal discovery proves a bottleneck for this mixed type dataset. With access the true graph, all methods to approximately as well as one another, PC appears to find the right graph more frequently than DECI.
2. **cts_to_cat**: Here, causal discovery is more successful with DECI, outperforming PC. Likely this is due to the fact that the incorrect graph $X_2 \to X_1$ does not admit an additive noise structure, and so can be discounted. With the right graph, DECI has a higher variance but similary accurate estimate of the ATE compared to DoWhy.
3. **mixed_simpson**: As with its sister dataset ‘nonlin_simpson’, causal discovery appears to be difficult here for all methods. With access to the true graph, DECI and DoWhy are equally able to estimate the treatment effects.
4. **large_backdoor_binary_t**: Here, DECI spline has a large margin over DECI Gaussian, and comes close to the true graph performance. It appears that DoWhy struggles, even with the true graph, likely due to the higher variance incurred by the maximal adjustment strategy.
5. **weak_arrows_binary_t**: Again, we see DECI spline appearing to be the most effective discovery method, but with the best ATE estimation then coming from combining it with DML.
6. **mixed_confounding**: this dataset appears to present a challenge for discovery for all methods. With access to the true graph, it appears that DECI is the most accurate ATE estimator, with linear methods in particular being somewhat off.

G.3 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.

G.4 Summary of ATE estimation results

Table 5 shows a summary of the *rank* of different methods aggregated over the following 27 datasets:

- 13 CSuite datasets: 6 are theoretically identifiable continuous non-linear ANMs, 1 is continuous linear Gaussian, 6 are mixed type
- IHDP and Twins benchmark datasets
- 6 ER graphs, 3 with Gaussian noise and 3 with spline noise
- 6 SF graphs, 3 with Gaussian noise and 3 with spline noise

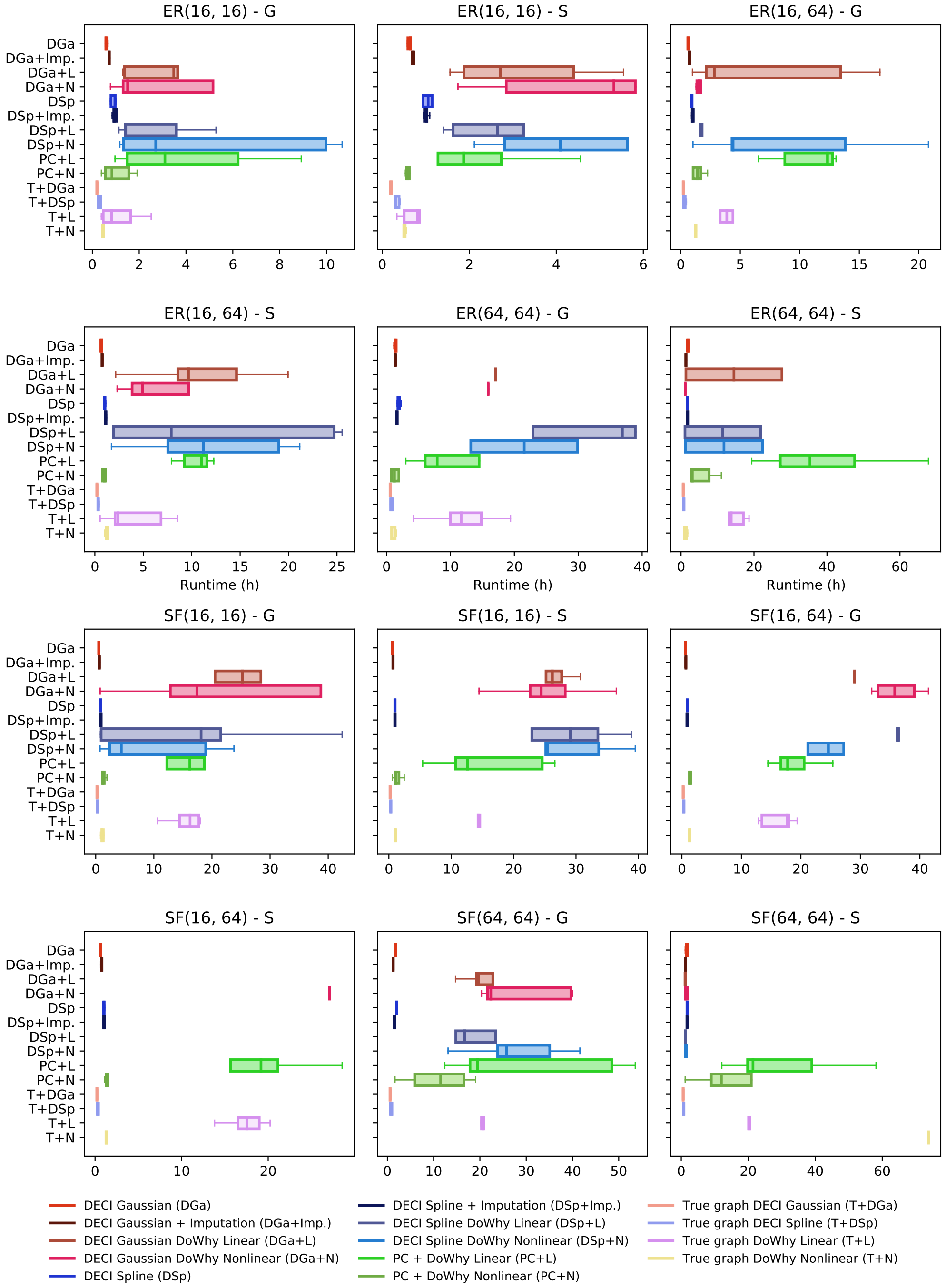


Figure 9: Runtime of End-to-end ATE estimation methods on synthetic graphs.

Table 5: 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.

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

G.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 introduce prior knowledge about graph sparseness through the weighted adjacency matrix $W_0 \in [0, 1]^{D \times D}$, with zero entries encouraging sparser graphs. The resulting informed DECI prior is

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

with the scalar λ_s regulating the strength of the prior beliefs encoded in W_0.

We compare DECI inference with access to the true graph to end-to-end DECI inference. In the latter case we consider a PC prior, which is the CP-DAG provided by PC. We consider different prior strengths, i.e. the value of the entries of W_0, between 0 and 1. We also experiment with introducing the true-graph as a prior of this form, 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.

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.

G.6 Understanding Robustness to Unobserved Confounders

To evaluate the robustness of DECI to the presence of unobserved confounders, we introduced a *new* CSuite datasets. **Fork nonlin. Gauss** has the graph shown in Figure 12a and uses nonlinear functions with Gaussian exogeneous noise. We trained DECI on the data from observed variables only (as shown in Figure 12), as well as on the full dataset treating all variables as observed. Table 6 shows the results.

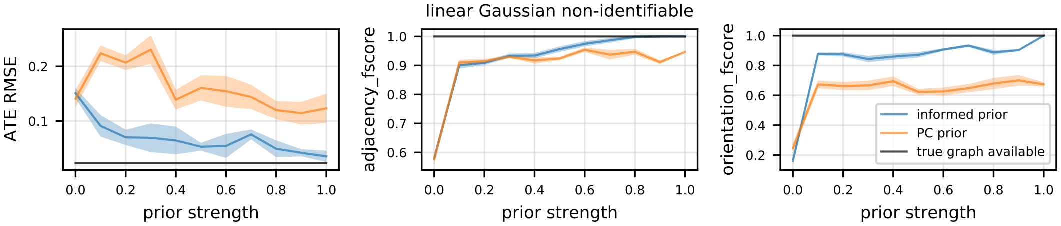


Figure 10: Causal discovery and inference results obtained on a 9 node linear Gaussian dataset, where the graph is non-identifiable without prior knowledge. We perform DECI inference with the true graph and DECI end2end inference with different priors. Informed prior refers to using a smoothed version 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.

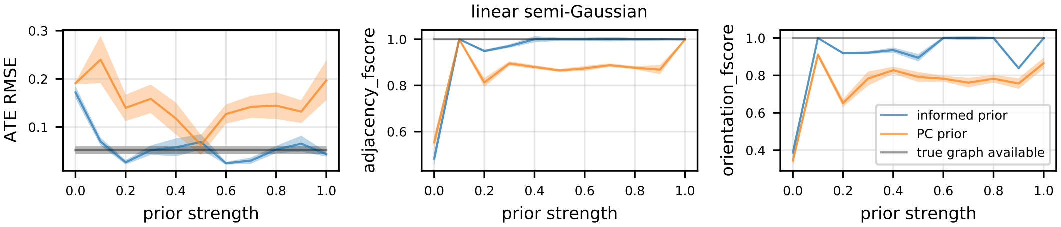
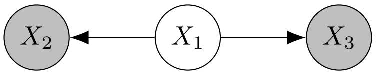


Figure 11: Causal discovery and inference results obtained on a 9 node linear non-Gaussian dataset, difficult to identify without prior knowledge. We perform DECI inference with the true graph and DECI end2end inference with different priors. Informed prior refers to using a smoothed version of the true 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.

Table 6: ATE RMSE results on CSuite Fork nonlin. Gauss dataset with unobserved confounders.

Method	ATE RMSE (mean ± s.e. from 5 seeds)
DECI Gaussian (DGa)	0.034 ± 0.007
DECI Spline (DSp)	0.036 ± 0.009
DECI Gaussian (DGa) full observability	0.019 ± 0.002
DECI Spline (DSp) full observability	0.018 ± 0.002



(a) Fork. Treatment X_2, outcome X_3.

Figure 12: CSuite graph with unobserved confounders.