VACA: Designing Variational Graph Autoencoders for Causal Queries

Pablo Sánchez-Martín,[1, 2] **Miriam Rateike,** [1,2] **Isabel Valera** [2]

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2] Department of Computer Science of Saarland University, Saarbrücken, Germany
psanchez@tue.mpg.de, mrateike@tue.mpg.de, ivalera@cs.uni-saarland.de

Abstract

In this paper, we introduce VACA, a novel class of variational graph autoencoders for causal inference in the absence of hidden confounders, when only observational data and the causal graph are available. Without making any parametric assumptions, VACA mimics the necessary properties of a *Structural Causal Model* (SCM) to provide a flexible and practical framework for approximating interventions (*do-operator*) and *abduction-action-prediction* steps. As a result, and as shown by our empirical results, VACA accurately approximates the interventional and counterfactual distributions on diverse SCMs. Finally, we apply VACA to evaluate counterfactual fairness in fair classification problems, as well as to learn fair classifiers without compromising performance.

Introduction

Graph Neural Networks (GNNs) are a powerful tool for graph representation learning and have been proven to excel in practical complex problems like neural machine translation (Bastings et al. 2017), traffic forecasting (Derrow-Pinion et al. 2021; Yu, Yin, and Zhu 2018) or drug discovery (Gilmer et al. 2017).

In this work, we investigate to which extent the inductive bias of GNNs–encoding the causal graph information–can be exploited to answer interventional and counterfactual queries. More specific, to approximate the interventional and counterfactual distributions induced by interventions on a casual model. To this end, we assume i) causal sufficiency–i.e., absence of hidden confounders; and, ii) access to observational data and the true causal graph. We stress that the causal graph can often be inferred from expert knowledge (Zheng and Kleinberg 2019) or via one of the approaches for causal discovery (Glymour, Zhang, and Spirtes 2019; Vowels, Camgoz, and Bowden 2021). With this analysis we aim to complement the concurrent line of research that theoretically studies the use of Neural Networks (NN) (Xia et al. 2021), and more recently GNNs (Zečević et al. 2021), for causal inference.

To this end, we describe the architectural design conditions that a variational graph autoencoder (VGAE)–as a density estimator that leverages a priori graph structure–must fulfill so that it can approximate causal interventions (*do-operator*) and *abduction-action-prediction* steps (Pearl 2009b). The resulting Variational Causal Graph Autoencoder, referred to as VACA, enables *approximating* the observational, interventional and counterfactual distributions induced by a causal model with unknown structural equations. We remark that parametric assumptions on the structural causal equations are in general not testable, may thus not hold in practice (Peters, Janzing, and Schölkopf 2017) and may lead to inaccurate results, if misspecified. VACA addresses this limitation by including uncertainty, i.e., a probabilistic model, in the estimation of the causal-parent relationships.

We show in extensive synthetic experiments that VACA outperforms competing methods (Karimi et al. 2020; Khemakhem et al. 2021) on complex datasets at estimating not only the mean of the interventional/counterfactual distribution (as in previous work), but also the overall distribution (measured in terms of Maximum Mean Discrepancy (Gretton et al. 2012)). Finally, we show a practical use-case in which VACA is used to assess counterfactual fairness of different classifiers trained on the real-world German Credit dataset (Dua and Graff 2017a), as well as to learn counterfactually fair classifiers without compromising performance.

Related Work

Deep generative models are enjoying increasing attention for causal queries in complex data (Moraffah et al. 2020; Parafita and Vitria 2019a). Existing approaches for causal inference focus on i) estimating the Average Treatment Effect (ATE)–a specific type of group-level causal queries–by assuming a fixed causal graph that includes a treatment variable (Kim et al. 2021; Louizos et al. 2017; Rakesh et al. 2018; Schwab, Linhardt, and Karlen 2018; Vowels, Camgoz, and Bowden 2020; Zhang, Zhang, and Li 2020); ii) discovering and intervening on the causal latent structure of the (e.g., image) data (Kim et al. 2021; Parafita and Vitria 2019a,b; Shen et al. 2020; Yang et al. 2020); or iii) addressing interventional and/or counterfactual queries by fitting a conditional model for each observed variable given its causal parents (Garrido et al. 2021; Karimi et al. 2020; Kocaoglu et al. 2018; Parafita and Vitria 2020; Pawlowski, Coelho de Castro, and Glocker 2020).

Within the scope of causality, GNNs have predominantly been used for causal discovery (Yu et al. 2019; Zhang et al.

2019) and only very recently, concurrent with us, exploited to answer interventional queries (Zečević et al. 2021).

Khemakhem et al. (2021) propose CAREFL, an autoregressive normalizing flow for both causal discovery and inference. The authors focus on (multi-dimensional) bi-variate graphs, but their approach can be extended to more general directed acyclic graphs (DAGs) using e.g., neuronal spline flows (Durkan et al. 2019). However, causal assumptions in a graph are modeled not only by the direction of edges, but also the absence of edges (Pearl 2009a). For the task of causal inference CAREFL is unable to exploit the absence of edges fully as it reduces a causal graph to its causal ordering (which may not be unique). Further, the authors only evaluate interventions in root nodes (which reduces to conditioning on the intervened-upon variable).

Karimi et al. (2020) answer interventional queries by fitting a conditional variational autoencoder (CVAE) to each conditional in the Markov factorization implied by the causal graph. As each observed variable is independently fitted, the mismatch between the true and generated distribution can cause errors that propagate to the distribution of its descendants. This can be problematic, especially for long causal paths. Pawlowski, Coelho de Castro, and Glocker (2020) propose an approach similar to Karimi et al. (2020), and additionally propose an approach based on normalizing flows to approximate the causal parent-child effect.

In contrast, VACA leverages i) GNNs to encode the causal graph information (inductive bias), ii) the GNN message passing algorithm to approximate the effect of interventions (*do-operator* (Pearl 2009b)) in the causal graph, and iii) jointly optimizes the observational distribution for all observed variables to avoid error propagation along the Markov factorization. We thoroughly evaluate the performance of VACA and compare it with related work, at approximating both interventional and counterfactual distributions induced by interventions on both root and non-root nodes in a wide variety of causal models.

Background

In this section, we first provide a brief overview on SCMs and then introduce the main building block of VACA, i.e., variational graph autoencoders.

Structural causal models

An SCM $\mathcal{M} = (p(\mathbf{U}), \tilde{\mathbf{F}})$ determines how a set of d endogenous (observed) random variables $\mathbf{X} := \{X_1, \dots X_d\}$ is generated from a set of exogenous (unobserved) random variables $\mathbf{U} := \{U_1, \dots U_d\}$ with prior distribution $p(\mathbf{U})$ via the set of *structural equations* $\tilde{\mathbf{F}} = \{X_i := \tilde{f}_i\left(\mathbf{X}_{\text{pa}(i)}, U_i\right)\}_{i=1}^{d}$. Here $\mathbf{X}_{\text{pa}(i)}$ refers to the set of variables directly causing X_i, i.e., parents of i. Similarly to (Karimi et al. 2020; Khemakhem et al. 2021; Pearl 2009a), we consider SCMs that are associated with a directed acyclic *causal graph* (although Section relaxes this assumption). We here denote the causal graph by $\mathcal{G} := (\mathcal{V}, \mathcal{E})$, where each node $i \in \mathcal{V}$ corresponds to an endogenous variable X_i. The set of directed edges $(j, i) \in \mathcal{E}$ represent the causal parent-child relationship between endogenous variables (Pearl 2009a), i.e. X_j is a parent of X_i. $\mathcal{E}$ can be represented by the adjacency matrix $\mathbf{A} \in \{0, 1\}^{d \times d}$, such that $A_{ij} = 1$ if $(j, i) \in \mathcal{E}$ and $A_{ij} = 0$, otherwise. We also define the set of neighbors, a.k.a. parents, of node i as $\text{pa}(i) = \mathcal{N}_i = \{j\}_{(j,i)\in\mathcal{E}}$ and $\text{pa}^*(i) := \text{pa}(i) \cup i$.

Given an SCM, there are two types of causal queries of general interest: interventional queries, e.g., "What would happen to the population $\mathbf{X}$, if variable X_i would be set to a fixed value α?", and counterfactual queries, e.g.,"What would have happened to a specific factual sample $\mathbf{x}^F$, had X_i been set to a value α?". In more detail, *interventional queries* aim to evaluate the effect at the population level (*rung 2*) of a specific intervention on, or equivalently manipulations of, a subset of the endogenous variables $\mathcal{I} \subseteq [d] := \{1, \dots, d\}$. Interventions on an SCM $\mathcal{M}$ are often represented with the *do-operator* $do(X_i = \alpha_i)$ (Pearl 2009b) and lead to a modified SCM $\mathcal{M}^{\mathcal{I}}$ which induces a new distribution over the set of endogenous variables $p(\mathbf{X} \mid do(X_i = \alpha_i))$, which is referred to as the *interventional distribution*. In $\mathcal{G}$ an intervention removes incoming edges to node i and sets $X_i = \alpha$ (see Figure 1b). A *counterfactual query* for a given factual instance $\mathbf{x}^F$ aims to estimate what would have happened had $\mathbf{X}_{\mathcal{I}}$ instead taken value $\boldsymbol{\alpha}$. This effect is captured by the *counterfactual distribution* $p(\mathbf{x}^{CF} \mid \mathbf{x}^F, do(X_{\mathcal{I}} = \boldsymbol{\alpha}))$, which can be computed using the *abduction-action-prediction* procedure by Pearl (2009b). Refer to Section for further details on the computation of the interventional and counterfactual distributions within our framework.

(a) $\mathcal{G}$ without intervention (b) $\mathcal{G}$ with intervention

Figure 1: *triangle* SCM $\mathcal{M} = \{p(\mathbf{U}), \tilde{\mathbf{F}}\}$, $\mathbf{U} \sim p(\mathbf{U})$ with $d = |\mathbf{X}| = 3$ endogenous variables where $X_1 := \tilde{f}_1(U_1)$, $X_2 := \tilde{f}_2(X_1, U_2)$, $X_3 := \tilde{f}_3(X_1, X_2, U_3)$ with (a) the corresponding causal graph $\mathcal{G}$ and (b) the causal graph corresponding to $\mathcal{M}^{\mathcal{I}}$ after intervention $do(X_2 = \alpha)$. Blue (red) arrows highlight the direct (indirect) causal path from X_1 to X_3 (via X_2).

Variational Graph Autoencoder and Graph Neural Networks

Variational Autoencoders (VAEs). VAEs (Kingma and Welling 2014) are powerful latent variable models based on neural networks (NNs) for jointly i) learning expressive density estimators $p(\mathbf{X}) \approx \int p_\theta(\mathbf{X} \mid \mathbf{Z}) p(\mathbf{Z}) d\mathbf{Z}$, where the likelihood function (a.k.a. *decoder*) is parameterized using a NN with parameters θ, and ii) performing approximate posterior inference over the latent variables $\mathbf{Z}$ made possible via a variational distribution (a.k.a. *encoder*) $q_\phi(\mathbf{Z} \mid \mathbf{X})$ parameterized using a NN with parameters ϕ. The parameters θ and ϕ can be learned by maximizing a lower bound on

the log-evidence (Burda, Grosse, and Salakhutdinov 2016; Nowozin 2018; Rainforth et al. 2018; Tucker et al. 2018).

Variational Graph Autoencoders (VGAEs). Kipf and Welling (2016) extend VAEs to account for prior graph structure information on the data (Yu et al. 2019). VGAEs define a (potentially multidimensional) latent variable Z_i per observed variable X_i, i.e., $\mathbf{Z} := \{Z_1, \dots, Z_d\}$. Additionally, VGAEs rely on an adjacency matrix $\mathbf{A}$, which is used by two GNNs, one for the encoder and one for the decoder, to enforce structure on the posterior approximation $q_\phi(\mathbf{Z} \mid \mathbf{X}, \mathbf{A})$ and the likelihood $p_\theta(\mathbf{X} \mid \mathbf{Z}, \mathbf{A})$. Hence, $\mathbf{A}$–given as prior–determines which variables X_i influence Z_j $\forall i, j \in [d]$.

Graph Neural Networks (GNNs). In its most general form, a GNN is a composition of message passing layers (Gilmer et al. 2017), where each layer updates the state of each node in $\mathcal{G}$. In particular, the state of node i at the output of layer l, i.e., $\mathbf{h}_i^l$, is specified as:

$$\mathbf{h}_i^l = f^u\left(\mathbf{h}_i^{l-1}, f^a\left(\{\mathbf{m}_{ij}^l\}_{j\in\mathcal{N}_i}\right); \theta_u^l\right). \tag{1}$$

First, node i receives a message $\mathbf{m}_{ij}^l = f^m(\mathbf{h}_i^{l-1}, \mathbf{h}_j^{l-1}; \theta_m^l)$ from each of its neighbors $j \in \mathcal{N}_i$. Then, these messages are aggregated via f^a. Finally, $\mathbf{h}_i^l$ is computed as a function f^u of the node's previous state $\mathbf{h}_i^{l-1}$ and the aggregated message. Note, if a GNN has N_h hidden layers, then the output for node i depends not only on its direct neighbors $\mathcal{N}_i$, but also on its neighbors up to order $N_h + 1$ (hops). For example, if $N_h = 0$ ($N_h = 1$) then the output for each node only depend on its direct neighbors, i.e., *parents* (2-hop neighbors, i.e., *grand-parents*). For a detailed description of GNNs, please refer to Appendix .

Observational, interventional and counterfactual distributions

In this section, we introduce the observational, interventional and counterfactual distributions (triggered by any intervention of the form $do(\mathbf{X}_\mathcal{I} = \boldsymbol{\alpha})$) that are induced by an SCM $\mathcal{M} = \{p(\mathbf{U}), \tilde{\mathbf{F}}\}$. Specifically, we summarize the main properties of an SCM that will allow us to propose a novel class of VGAEs, namely VACA, to compute accurate estimates of these distributions using observational data and a known causal graph. To this end, we assume the absence of hidden confounders, i.e., we assume that $p(\mathbf{U}) = \prod_{i=1}^d p(U_i)$.

Observational distribution. The SCM $\mathcal{M}$ determines the observational distribution $p(\mathbf{X})$ over the set of endogenous variables $\mathbf{X} = \{X_1, \dots X_d\}$, which satisfies causal factorization (Schölkopf 2019), i.e., $p(\mathbf{X}) = \prod_{i=1}^d p(X_i \mid \mathbf{X}_{\mathrm{pa}(i)})$.

That is, after marginalizing out the exogenous variables $\mathbf{U}$, the distribution of each endogenous variable X_i depends only on its parents, i.e., $\mathbf{X}_{\mathrm{pa}(i)}$. The *observational distribution* can alternatively be written only in terms of the exogenous variables $\mathbf{U}$ as

$$p(\mathbf{X}) = \mathbf{F}_\#[p(\mathbf{U})], \tag{2}$$

i.e., $p(\mathbf{X})$ is the pushforward of $P(\mathbf{U})$ through $\mathbf{F}$. Here $\mathbf{F} : \mathbf{U} \to \mathbf{X}$ corresponds to the set of structural equations, which directly transform the exogenous variables $\mathbf{U}$ into the endogenous variables $\mathbf{X}$. This is equivalent to $\tilde{\mathbf{F}}$, which takes as input both the exogenous variable and the parent (endogenous) variables of a target endogenous variables to compute its value.

Let us denote by $\mathrm{an}(i)$ the set of indexes of the ancestors of i, and $\mathrm{an}^*(i) := \mathrm{an}(i) \cup \{i\}$. Then, the causal factorization induced by $\mathcal{M}$ leads to the following property of $\mathbf{F}(\mathbf{U})$:

Property 1 *Each endogenous variable X_i can be expressed as a function of its exogenous variable U_i and the ones of all its causal ancestors, i.e., $\mathbf{F}(\mathbf{U})=\{X_i = f_i(\{U_j\}_{j\in\mathrm{an}^*(i)})\}$. This, together with the causal sufficiency assumption, implies that X_i is statistically independent of $U_j, \forall j \notin \mathrm{an}^*(i)$.*

Interventional distribution. As stated in Section , interventions on a set of variables $\mathcal{I}$ can be performed using the *do-operator*, which can be seen as a mapping $do(\mathbf{X}_\mathcal{I} = \boldsymbol{\alpha}) : \mathcal{M} \mapsto \mathcal{M}^\mathcal{I} = (p(\mathbf{U}), \tilde{\mathbf{F}}^\mathcal{I})$ where $\tilde{\mathbf{F}}^\mathcal{I} = \{\tilde{f}_i\}_{i\notin\mathcal{I}} \cup \{\alpha_i\}_{i\in\mathcal{I}}$. As above, we can represent the resulting set of *intervened structural equations* as $\mathbf{F}^\mathcal{I} = \{f_i\}_{i\notin\mathcal{I}} \cup \{\alpha_i\}_{i\in\mathcal{I}}$, and thus write the *interventional distribution* as:

$$p(\mathbf{X} \mid do(\mathbf{X}_\mathcal{I} = \boldsymbol{\alpha})) = \mathbf{F}^\mathcal{I}{}_\#[p(\mathbf{U})]. \tag{3}$$

Property 2 *After an intervention $do(\mathbf{X}_\mathcal{I} = \boldsymbol{\alpha})$ on $\mathcal{M}$, all the causal paths from U_j $\forall j \in \mathrm{an}^*(i)$ to X_i that include an intervened-upon variable in $\mathbf{X}_\mathcal{I}$ (i.e., the causal paths where $\mathbf{X}_\mathcal{I}$ is a mediator) are severed in $\mathbf{F}^\mathcal{I}$, while the rest of causal paths remain untouched.*

The above property is illustrated in Figure 1, where we can observe that after an intervention $do(X_2 = \alpha)$, the indirect causal path (in red) from X_1, and thus from U_1, to X_3 via X_2 is severed, while the direct path (in blue) remains.

Counterfactual distribution. Assuming the SCM $\mathcal{M} = \{p(\mathbf{U}), \tilde{\mathbf{F}}\}$ to be known, the following three steps defined by Pearl (2009a) allow to compute counterfactuals $\mathbf{x}^{CF}$:

i) *Abduction:* infer the values of the exogenous variables $\mathbf{U}$ for a factual sample $\mathbf{x}^F$, i.e., compute $p(\mathbf{U} \mid \mathbf{x}^F)$; ii) *Action:* intervene with $do(\mathbf{X}_\mathcal{I} = \boldsymbol{\alpha}) : \mathcal{M} \mapsto \mathcal{M}^\mathcal{I} = (p(\mathbf{U}), \tilde{\mathbf{F}}^\mathcal{I})$; and iii) *Prediction:* use the posterior distribution $p(\mathbf{U} \mid \mathbf{x}^F)$ and the new structural equations $\tilde{\mathbf{F}}^\mathcal{I}$ to compute $p(\mathbf{x}^{CF} \mid \mathbf{x}^F)$. The prediction step can alternatively be computed using the new set of structural equations $\mathbf{F}^\mathcal{I}$ defined in terms of the exogenous variables $\mathbf{U}$, so that we can write the *counterfactual distribution* as:

$$p(\mathbf{x}^{CF} \mid \mathbf{x}^F, do(\mathbf{X}_\mathcal{I} = \boldsymbol{\alpha})) = \mathbf{F}^\mathcal{I}{}_\#[p(\mathbf{U} \mid \mathbf{x}^F)]. \tag{4}$$

Importantly, the posterior distribution $p(\mathbf{U} \mid \mathbf{x}^F)$ satisfies:

Property 3 *In the abduction step, statistical independence implies that conditioned on the endogenous variables of the factual sample $\mathbf{x}^F$, each exogenous variable U_i is independent of the factual value x_j^F if $j \neq i$ and the variable X_j is not a parent of X_i, i.e., $j \notin \mathrm{pa}^*(i)$.*

Variational Causal Autoencoder (VACA)

In this section, we present a novel variational causal graph autoencoder (VACA) to approximate the observational (2), interventional (3) and counterfactual (4) distributions. While the underlying SCM $\mathcal{M}$ is unknown, we assume access to the true causal graph $\mathcal{G}$ and observational data $\{\mathbf{x}_n\}_{n=1}^N$, i.e., i.i.d. samples of the observational distribution induced by $\mathcal{M}$ (in the absence of hidden confounders).

Definition 0.1 *(VACA) . Given a causal graph $\mathcal{G}$ over a set of endogenous variables $\mathbf{X} = \{X_1, \dots, X_d\}$, which establishes the set of parents $\text{pa}(i)$ for each variable X_i (including the i-th node), VACA is defined by:*

- *A causal adjacency matrix $\mathbf{A}$, which is a $d \times d$ binary matrix with elements $A_{ij} = 1$ if $j \in \text{pa}^*(i)$, i.e., when $i = j$ or j is a parent of i. Otherwise, $A_{ij} = 0$.*
- *A prior distribution $p(\mathbf{Z}) = \prod_i p(Z_i)$ over the set of latent variables $\mathbf{Z} = \{Z_1, \dots, Z_d\}$.*
- *A decoder $p_\theta(\mathbf{X} \mid \mathbf{Z}, \mathbf{A})$, which is a GNN (parameterized by θ) that takes as input the set of latent variables $\mathbf{Z}$ and the causal adjacency matrix $\mathbf{A}$, and outputs the parameters of the likelihood $p_\theta(\mathbf{X} \mid \mathbf{Z}, \mathbf{A})$.*
- *An encoder $q_\phi(\mathbf{Z} \mid \mathbf{X}, \mathbf{A})$, which is a GNN (parameterized by ϕ) that takes as input the endogenous variables $\mathbf{X}$ and the causal adjacency matrix $\mathbf{A}$, and outputs the parameters of the posterior approximation $q_\phi(\mathbf{Z} \mid \mathbf{X}, \mathbf{A})$.*

Next, we discuss how to design VACA such that it is able to capture the observational, interventional, and counterfactual distribution induced by an unknown SCM. Importantly, we derive the necessary conditions on the design of both the encoder and decoder GNNs such that VACA can mimic the SCM properties introduced in Section .

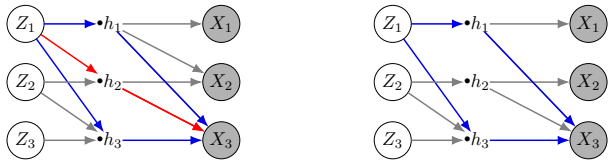
Observational distribution

VACA approximates the *observational distribution* in (2) using the generative model as

$$p(\mathbf{X}) \approx \int p_\theta(\mathbf{X} \mid \mathbf{Z}, \mathbf{A}) p(\mathbf{Z}) d\mathbf{Z}, \tag{5}$$

where $p_\theta(\mathbf{X} \mid \mathbf{Z}, \mathbf{A}) = \prod_{i=1}^d p_\theta(X_i \mid \mathbf{Z}, \mathbf{A})$. Figure 3a depicts this generative process.

Relationship between Z and U. When comparing (5) with the true observational distribution in (2), we observe that the latent variables $\mathbf{Z}$ play a similar role to the exogenous variables $\mathbf{U}$, and the decoder $p_\theta(\mathbf{X} \mid \mathbf{Z}, \mathbf{A})$ plays a similar role to the structural equations $\mathbf{F}$. We remark that $\mathbf{Z}$ do not need to correspond to the true exogenous variables (i.e., $p(\mathbf{U}) \neq p(\mathbf{Z})$), and thus, the decoder does not aim to approximate the causal structural equations. Yet, we assume that there is one independent latent variable Z_i for every observed variable X_i capturing all the information of X_i that cannot be explained by its parents. Thus, since X_i is in turn a (deterministic) function of its parents $\text{pa}(i)$ and its exogenous variable U_i, the posterior $p\left(Z_i \mid X_i, \text{pa}(i)\right)$ aims to capture the information that U_i contributes to X_i (i.e., the information of X_i not contributed by its parents). That is–similar to $p\left(U_i \mid X_i, \text{pa}(i)\right)$, the (true) posterior distribution–$p\left(Z_i \mid X_i, \text{pa}(i)\right)$ should depend only on X_i and parents $\text{pa}(i)$.



(a) Original (b) Intervened

Figure 2: VACA decoder (a) without and (b) with intervening on X_2. Message passing in the GNN correspond to direct (blue) and indirect (red) causal paths in Figure 1.

Observational noise. VACA has observational noise that is not present in the true SCM, where an observed variable is assumed to be a deterministic transformation of its exogenous variables and parents via the structural equations (SEs). As VACA does not have access to the true SEs (nor to the true distribution of the exogenous variables), the *noise* of the likelihood $p_\theta(\mathbf{X} \mid \mathbf{Z}, \mathbf{A})$ can be interpreted as an estimate of the uncertainty on the estimated observational distribution (due to the uncertainty on the true SCM).

Here, we seek to ensure that the $p(\mathbf{X})$ induced by VACA complies with causal factorization (**Property 1** in Section). To that end, the design of the decoder GNN must assure that $p_\theta(X_i \mid \mathbf{Z}, \mathbf{A}) = p_\theta(X_i \mid \mathbf{Z}_{\text{an}^*(i)})$. That is, that X_i depends only on Z_j if $j = i$ or X_j is an ancestor of X_i in the causal graph.

Proposition 1 *(Causal factorization). VACA satisfies causal factorization, $p_\theta(\mathbf{X} \mid \mathbf{Z}, \mathbf{A}) = \prod_i p_{\theta_i}(X_i \mid \mathbf{Z}_{\text{an}^*(i)})$, if and only if the number of hidden layers in the decoder is greater or equal than $\delta - 1$, with δ being the length of the longest shortest path between any two endogenous nodes.*

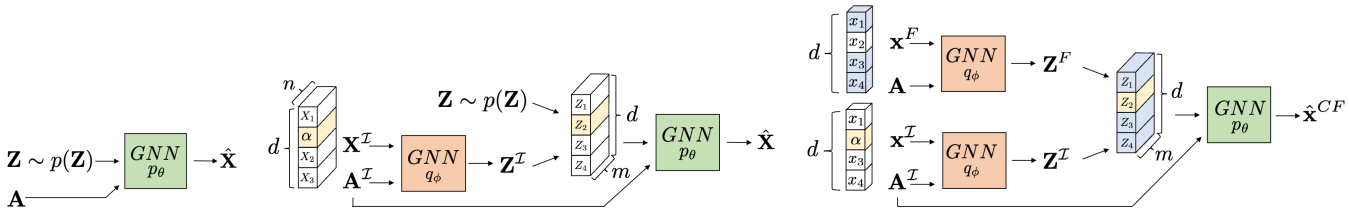
The above proposition (proved in Appendix) is based on the fact that, in a GNN with N_h hidden layers (and $N_h + 1$ layers in total), the output for the i-th node depends on its neighbors of up to $N_h + 1$ hops. As an example, consider the following *chain* causal graph: $X_1 \to X_2 \to X_3$, such that $\delta = 2$. In the decoder, the first layer yields a hidden representation for the 3-rd node $h_3^1 := f(f(Z_2), Z_3)$ that only depends on Z_2 and Z_3. Thus, we need a second layer for its output $h_3^2 := f(h_2, Z_3) = f(f(f(Z_1), Z_2), Z_3)$ to depend on Z_1 (note that X_1 is an ancestor of X_3).

Interventional distribution

VACA approximates the *interventional distribution* in (3) as (illustrated Figure 2):

$$\begin{aligned} p(\mathbf{X} \mid do(\mathbf{X}_\mathcal{I} = \boldsymbol{\alpha})) \approx \int\int & p_\theta(\mathbf{X} \mid \tilde{\mathbf{Z}}, \tilde{\mathbf{Z}}^\mathcal{I}, \mathbf{A}^\mathcal{I}) \\ & \times p(\tilde{\mathbf{Z}}) q_\phi(\tilde{\mathbf{Z}}^\mathcal{I} \mid \mathbf{A}^\mathcal{I}, \mathbf{X}_\mathcal{I}) d\tilde{\mathbf{Z}} d\tilde{\mathbf{Z}}^\mathcal{I}, \end{aligned} \tag{6}$$

where $\tilde{\mathbf{Z}}^\mathcal{I} = \{Z_i^\mathcal{I}\}_{i \in \mathcal{I}}$ is the subset of latent variables associated with the intervened-upon variables $\mathbf{X}_\mathcal{I}$, and $\tilde{\mathbf{Z}} = \{Z_i\}_{i \notin \mathcal{I}}$ denotes the subset of latent variables associated with the rest of the observed variables. Importantly, here the *do-operator* is performed on the causal adjacency matrix as $do(X_\mathcal{I} = \boldsymbol{\alpha}) : \mathbf{A} \mapsto \mathbf{A}^\mathcal{I} = \{A_{ij}\}_{\forall i \notin \mathcal{I}, j} \cup \{A_{ij} = 0\}_{\forall i \in \mathcal{I}, j}$. This ensures that X_i for $i \in \mathcal{I}$ is independent of Z_j for all $j \neq i$. Note that in order for (6) to be able to



(a) Observational (b) Interventional (c) Counterfactual

Figure 3: VACA generation of (a) observational, (b) interventional, and (c) counterfactual samples. The 'hat' in $\hat{\mathbf{X}}$ and $\hat{\mathbf{x}}^{CF}$ indicates that they are sample estimates of the true random variables.

approximate the interventional distribution in (3), an intervention on VACA should satisfy **Property 2**, i.e.:

Proposition 2 *(Causal interventions). VACA captures causal interventions if and only if the number of hidden layers in its decoder is greater than or equal to $\gamma - 1$, with γ being the length of the longest path between any two endogenous nodes in $\mathcal{G}$.*

To illustrate this, Figure 2 depicts how messages are exchanged in a one-hidden-layer decoder GNN corresponding to the causal graph $\mathcal{G}$ in Figure 1 (*triangle* with $\gamma = 2$), both (a) without and (b) with an intervention on X_2. We highlight in blue the direct messages (sent via direct causal path in $\mathcal{G}$), and in red the indirect messages (sent via indirect causal path in $\mathcal{G}$) from Z_1 to X_3. Observe that, similarly to Figure 1, in (a) there is an indirect path (via h_2) from Z_1 to X_3; while in (b) this path is severed. Hence, the hidden layer (h_1, h_2, h_3) allows to distinguish between direct and indirect paths and thus to capture interventional effects. As the condition in Proposition 2 is more restrictive than the one in Proposition 1, VACA is able to approximate the observational and interventional distributions (as empirically validated in Appendix) if:

Design condition 1 (necessary condition) *The decoder GNN of VACA has at least as many hidden layers as $\gamma - 1$, with γ being the longest directed path in the causal graph $\mathcal{G}$.*

Counterfactual distribution

VACA approximates the *counterfactual distribution* in (4) as (illustrated in Figure 3c):

$$p(\mathbf{x}^{CF} \mid do(\mathbf{X}_{\mathcal{I}} = \boldsymbol{\alpha}), \mathbf{x}^F) \approx \int \int \underbrace{p_\theta(\mathbf{X} \mid \tilde{\mathbf{Z}}^{\mathbf{F}}, \tilde{\mathbf{Z}}^{\mathcal{I}}, \mathbf{A}^{\mathcal{I}}) q_\phi(\tilde{\mathbf{Z}}^{\mathcal{I}} \mid \mathbf{x}^{\mathcal{I}}, \mathbf{A}^{\mathcal{I}})}_{action} \times \underbrace{q_\phi(\tilde{\mathbf{Z}}^{\mathbf{F}} \mid \mathbf{x}^F, \mathbf{A})}_{abduction} d\tilde{\mathbf{Z}}^{\mathcal{I}} d\tilde{\mathbf{Z}}^{\mathbf{F}}, \tag{7}$$

where $\mathbf{x}^F$ represents a sample from $\mathbf{X}$ for which we seek to compute the distribution over counterfactual $\mathbf{x}^{CF}$ and $\tilde{\mathbf{Z}}^{\mathbf{F}} = \{Z_i^F\}_{i \notin \mathcal{I}}$. Note that two different passes of the encoder are necessary: one for the *abduction* step of the factual instance $q_\phi(\tilde{\mathbf{Z}}^{\mathbf{F}} \mid \mathbf{x}^F, \mathbf{A})$; and another one for the *action* step (intervention) $q_\phi(\tilde{\mathbf{Z}}^{\mathcal{I}} \mid \mathbf{x}^{\mathcal{I}}, \mathbf{A}^{\mathcal{I}})$ with $x_i^{\mathcal{I}} = \alpha_i \; \forall i \in \mathcal{I}$ (we remark that the rest of the values in $\mathbf{x}^{\mathcal{I}}$ do not affect the overall counterfactual computation).

We then evaluate the likelihood making sure that the resulting counterfactual sample $\mathbf{x}^{CF}$ only depends on the $\tilde{\mathbf{Z}}^{\mathbf{F}}$ and $\tilde{\mathbf{Z}}^{\mathcal{I}}$. Importantly, in order for VACA to be able to approximate the counterfactual distribution, we need its abduction (and action) step(s) to comply with **Property 3**, i.e.:

Proposition 3 *(Abduction). The abduction step of an observed sample $\mathbf{x} = \{x_1, \ldots, x_d\}$ in VACA satisfies that for all i the posterior of Z_i is independent on the subset $\{x_j\}_{j \notin \mathrm{pa}^*(i)} \subseteq \mathbf{x}$, if and only if the encoder GNN has no hidden layers.*

The above result (proved in Appendix) can be shown by the message passing algorithm computed by the encoder GNN, and leads to the second design condition of VACA:

Design condition 2 (necessary condition): *The encoder GNN of VACA has no hidden layers.*

In other words, from the definition of the SCM, the posterior distribution of U_i only depends on the parents, i.e $U_i \mid X_{\mathrm{pa}(i)}$. In order for VACA to mimic this property, the GNN that parameterized the encoder contains no hidden layers: in the message passing algorithm, in the k-th iteration (layer), a node depends on its k-hop ancestors; requiring $k = 1$ in the encoder refers to a GNN without hidden layers. Note that while the above condition may look restrictive and limiting the capacity of our encoder, we may choose arbitrarily complex NNs for the message f^m and update f^u functions, as well as one or more aggregation functions f^a, e.g., sum or max, to model the encoder (Corso et al. 2020).

Practical considerations

Next, we briefly discuss practical implementation considerations to handle complex causal models, which often appear in real world applications (Dua and Graff 2017a,b) For further details on VACA implementation, refer to Appendix .

Heterogeneous causal nodes. So far, we have modeled each endogenous variable X_i as a node in the causal graph $\mathcal{G}$, and thus in the VACA GNNs. Yet, in some application domains the relationships between a subset of k_i variables may be unknown, or they may be affected by hidden confounders. In such cases, we assume that set of k_i variables to be correlated and model them as one multidimensional and potentially heterogeneous node $\mathbf{X}_i = \{X_{i1}, \ldots, X_{ik_i}\}$

that share the same latent random variable Z_i. This allows us to deal with a large variety of graphs in practice.

Heterogeneous endogenous variables. Heterogeneous causal nodes require us to model different functions for each node, i.e. nodes may now contain a mix of continuous/discrete variables. In general GNNs are parametrized such that the parameters of the message function f^m and update function f^u are shared for all the nodes and edges in the graph. However, similar to the structural equations $\mathbf{F}$, we can define a unique set of parameters θ_{mij} for each f_{ij}^m (see (1)), so that we can model a different function for every edge in the causal graph. Further, we can also assume different update functions f_i^u for each node i, by introducing different update parameters θ_{ui}. As a result, VACA fulfills the conditions to be a Neural Causal Model (NCM) Type 2 (Coll. 1 in Zečević et al. (2021)) and thereof can represent the observational, interventional, and counterfactual distributions (Thm.1 and Thm. 3 in Xia et al. (2021)).

Non-identifiability. We highlight that certain counterfactual queries are not identifiable from observational data without making assumptions on the functional relationships even under causal sufficiency (Pearl 2009a). Yet, we expect (as confirmed by our empirical validation) that sufficiently expressive GNNs will lead to accurate approximations of counterfactual queries.

Evaluation

In this section, we evaluate the potential of VACA in approximating the outcomes of causal queries and compare it to two competing methods in synthetic experiments. The synthetic setting allows us to have access the true SEs, which is necessary to evaluate interventional distributions and especially counterfactuals. We consider interventions of the form $do(x_i = \alpha_i)$ for several values of α_i on both root and non-root nodes. We compute all results over the same 10 random seeds and report mean and standard deviation. Refer to Appendix for a complete description of the experimental setup. Moreover, our code is publicly available at GitHub[1].

Datasets. We consider 6 different synthetic causal graphs that differ in the number of nodes d, diameter δ, and longest path γ. Here, we report the results for i) the *collider* ($d = 3$, $\delta = 1$, $\gamma = 1$) with linear (LIN) and non-linear (NLIN) additive noise SEs, ii) the *loan* from Karimi et al. (2020) ($d = 7, \delta = 2, \gamma = 3$), and iii) the *adult* ($d = 11, \delta = 2, \gamma = 3$) graphs. Note that the two latter ones are synthetic versions of the German Credit dataset (Dua and Graff 2017a) and the Adult datasets (Dua and Graff 2017b), respectively. See Appendix for further details on the graphs and Appendix for the results with the remaining graphs.

Metrics. We evaluate the observational distribution using the Maximum Mean Discrepancy (MMD) (Gretton et al. 2012) as distance-measure between the true and estimated distributions, i.e., the lower the MMD the better the distributions match. For the interventional distribution, we additionally report the average estimation squared error of the mean (MeanE) and of the standard deviation (StdE) over all descendants of the intervened-upon variables. For the counterfactual distribution we report the mean squared error (MSE) as well as the standard deviation of the squared error (SSE) between the true and the estimated counterfactual value. More details in Appendix .

Baselines. We compare VACA with MultiCVAE (Karimi et al. 2020) and CAREFL (Khemakhem et al. 2021) described in Section . For a fair comparison, all model hyperparameters have been cross-validated using a similar computational budget (see Appendix). In Table 1, we report for each model and SCM the best configuration according to observational MMD. We also include a time-complexity analysis in Appendix .

Results. Table 1 summarizes the results. We observe that, in general, MultiCVAE underperforms the other methods. This may be explained by the fact that MultiCVAE trains each node independently, and thus the discrepancy between the true and generated distributions in one node may be amplified in its descendants.

Comparing VACA to CAREFL, we first observe that VACA performs consistently better in terms of observational MMD, i.e., VACA is able to generate observational samples that better resemble the true ones. Second, regarding the interventional distribution, CAREFL does a good job at fitting the mean (i.e., low MeanE). However,VACA performs consistently better both at approximating the standard deviation (i.e., low StdE) and the true samples (i.e., low MMD). This can be explained by VACA i) leveraging the causal graph (contrary to CAREFL that relies on causal ordering), and ii) optimizing the log-evidence in (5) jointly (contrary to the sequential optimization of MultiCVAE). Thus, VACA approximates the distribution as a whole better, which is a desirable property for studying interventions on a population-level rather than just on average. Lastly, in the approximation of counterfactuals we observe that both CAREFL and VACA exhibit similar performance in terms of MSE and SSE. Note however that CAREFL performs exact inference while VACA is built on approximate inference and is trained on a lower bound on the log-evidence. Finding tighter bounds could boost VACA performance.

Use case: counterfactual fairness

We finally show two practical use-cases of our method: assessing counterfactual fairness and training counterfactually fair classifiers. We use the public German Credit dataset (Dua and Graff 2017a) and rely on the causal model proposed by Chiappa (2019) with the following random variables $\mathbf{X}$: sensitive feature $S = \{\textit{sex}\}$, and non-sensitive features $C = \{\textit{age}\}$, $R = \{\textit{credit amount}, \textit{repayment history}\}$ and $H = \{\textit{checking account}, \textit{savings}, \textit{housing}\}$. Then, we aim to predict the binary feature $Y = \{\textit{credit risk}\}$ from $\mathbf{X}$. See Appendix for further details.

Counterfactual fairness. Let $S \subset \mathbf{X}$ be a sensitive attribute (e.g., gender), then the counterfactual unfairness (Kusner et al. 2017) of a classifier $h : \mathbf{X} \to Y$ is measured $\forall \mathbf{x}^{CF}, \alpha' \neq \alpha, y$ as:

$$\begin{aligned} \textit{uf} = \mid & P(h(\mathbf{x}^{CF}) = y \mid do(S = \alpha), \mathbf{x}^F) \\ & - P(h(\mathbf{x}^{CF}) = y \mid do(S = \alpha'), \mathbf{x}^F) \mid \end{aligned} \tag{8}$$

[1] https://github.com/psanch21/VACA

SCM		Model	Obs. MMD	Interventional MMD	Interventional MeanE	Interventional StdE	Counterfactuals MSE	Counterfactuals SSE	Num. params
collider	LIN	MultiCVAE	30.37±8.16	44.70±12.25	13.29±4.78	46.56±2.40	87.41±3.64	65.15±2.83	553
		CAREFL	9.27±1.49	4.86±0.45	0.35±0.08	81.89±1.78	8.11±0.58	7.83±0.55	6420
		VACA	1.50±0.67	1.57±0.41	0.75±0.31	41.99±0.30	9.86±0.74	7.06±0.38	5600
	NLIN	MultiCVAE	28.03±9.12	41.60±12.62	10.49±4.12	46.48±2.43	82.32±2.61	62.05±1.87	553
		CAREFL	10.38±2.00	4.69±0.38	0.19±0.07	80.68±2.08	6.93±0.40	7.15±0.64	4308
		VACA	0.95±0.27	0.97±0.23	0.26±0.12	42.20±0.24	5.01±0.73	4.08±0.54	1805
loan	-	MultiCVAE	90.38±11.31	213.65±5.38	12.24±1.33	65.78±1.13	40.98±0.35	15.12±0.16	33717
		CAREFL	22.10±1.64	27.38±4.07	6.74±4.25	50.13±2.47	11.15±2.57	6.59±0.38	2880
		VACA	2.22±0.25	6.87±0.66	4.35±0.35	3.83±0.08	10.30±0.40	6.41±0.11	30402
adult	-	MultiCVAE	140.15±6.37	155.52±5.93	12.18±2.36	63.52±4.05	39.96±0.36	16.37±0.65	6549
		CAREFL	31.31±1.58	34.31±5.77	12.54±3.17	41.26±3.44	1.23±0.17	3.55±0.90	127420
		VACA	4.51±0.45	12.68±1.95	1.65±0.23	3.37±0.09	5.33±0.27	5.67±0.20	63432

Table 1: Performance of different methods at estimating the observational, interventional and counterfactual distribution of different complex SCMs. Values are multiplied by 100. All models have been cross-validated with a similar computational budget. The number of parameters of the best configuration is shown in the right column.

Metric	full	unaware	fair-x	fair-z
↑ *f1*	71.67	69.49	59.50	70.79 ± 5.15
↓ *uf*	14.01 ± 2.26	13.27 ± 2.28	0.14 ± 0.02	0.51 ± 0.19

Table 2: Counterfactual unfairness (*uf*) and f1-score (*f1*) of an SVM over 10 VACA seeds. Values multiplied by 100.

A classifier is counterfactually fair ($uf = 0$), if, given a factual $\mathbf{x}^F$ with sensitive attribute $S = \alpha$, had its sensitive attribute been different $S = \alpha'$, the classifier prediction would remain the same. We can use VACA to generate counterfactual estimates to *audit* the fairness level of a classifier. Following (Kusner et al. 2017), we *audit*: i) a *full* model $h_{\text{full}} : \mathbf{X} \to Y$ that takes as input the complete variable set; ii) an *unaware* model $h_{\text{unaw}} : \mathbf{X} \backslash S \to Y$ that takes as input all variables but the sensitive one; iii) and a *fair* model $h_{\text{fair-x}} : \{X_i | S \notin \text{an}^*(i)\} \to Y$ that takes as input all non-descendant variables of the sensitive attribute. Moreover, we show that we can *learn a fair classifier* $h_{\text{fair-z}} : \mathbf{Z} \backslash Z_S \to Y$, which takes as input the latent variables generated by the VACA encoder without the one of the sensitive attribute Z_S.

Fairness Auditing. Table 2 summarizes the unfairness level and f1-score for a support vector machine (SVM) classifier. See Appendix for results of a logistic regression classifier. As we do not have access to the true data generation process, we evaluate the *auditing* task by the resulting ranking of the different classifiers according to their unfairness level. Based on the counterfactual generation by VACA the *full* classifier is consistently less fair than the *unaware* and the *fair-x* classifier, respectively. This ranking is consistent with the one in (Kusner et al. 2017).

Fairness Classification Table 2 shows that for the *fair-x* classifier fairness comes at the expense of accuracy compared to the *full* classifier. On the contrary, even though VACA has been trained for representation learning without access to classification labels, *fair-z* is a fair classifier (with comparable fairness level to the *fair-x* one) while keeping the performance comparable to the unfair *full* classifier. VACA, therefore also provides a practical approach to train accurate and fair classifiers.

Conclusion, Limitations and Impact

In this work, we have proposed VACA, a variational causal autoencoder based on GNNs that: i) is specially designed to capture the properties of SCMs; ii) inherently handles heterogeneous data; and iii) provides good approximations of interventional and counterfactual distributions as a whole for SCMs of different complexities. As demonstrated by extensive synthetic experiments, VACA provides accurate results for a wide range of interventions in diverse SCMs leading to more consistent results than competing methods (Karimi et al. 2020; Khemakhem et al. 2021). Finally, we have applied VACA for counterfactually fair classification.

Practical limitations. The expressive power of VACA to model complex structural equations, e.g., in domains such as biology (Sachs et al. 2005), is limited by the GNN architectures of the encoder and the decoder. As discussed in the GNN literature (Corso et al. 2020), especially aggregation functions may limit expressiveness. We expect VACA to benefit from advances in the field. Second, long causal paths would require VACA to increase the number of layers in the decoder (see **Design condition 1**). However, the GNNs performance is known to deteriorate with depth (Gallicchio and Micheli 2020; Gu et al. 2020; Li, Han, and Wu 2018).

Social impact. Trusting counterfactuals is of great importance for decision making, e.g. in the political or medical domain. We thus encourage anyone who uses VACA (or any other ML method for causal inference) to fully understand the model assumptions and to verify (up to the possible extend) that they are fulfilled.

Future work. First, it would be important to evaluate the sensitivity of VACA to errors in the assumed causal graph, as well as to the presence of hidden confounders. We plan to extend VACA to handle more complex causal

models including, e.g., hidden confounders and non-DAG causal graphs. Second, it would be interesting to perform ablation studies on the limitations of available GNNs architectures (Wu et al. 2020) for the VACA encoder and decoder; as well as on how the performance deteriorates as we increase the length of the causal path and thus the required number of hidden layers (Li, Han, and Wu 2018). Finally, it would be intriguing to apply VACA to other causal questions such as privacy-preserving causal inference (Kusner et al. 2016) or explainable machine learning (Karimi et al. 2020).

Acknowledgments

We would like to thank Amir Hossein-Karimi, Adrián Javaloy Bornás, Jonas Kleesen and Maryam Meghdadi Esfahani for helpful feedback and discussions. Moreover, a special thanks to Diego Baptista Theuerkauf for invaluable help with formalizing proofs. Moreover, the authors would like to thank Ilyes Khemakhem for helpful insights in how to generalize their CAREFL approach to arbitrary graphs.

Pablo Sańchez Martín thanks the German Research Foundation through the Cluster of Excellence "Machine Learning – New Perspectives for Science", EXC 2064/1, project number 390727645 and Miriam Rateike thanks the German Federal Ministry of Education and Research through the AI Center Tübingen for generous funding support. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Pablo Sańchez Martín.

References

Bastings, J.; Titov, I.; Aziz, W.; Marcheggiani, D.; and Sima'an, K. 2017. Graph convolutional encoders for syntax-aware neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 3.

Burda, Y.; Grosse, R.; and Salakhutdinov, R. 2016. Importance weighted autoencoders. In *Proceedings of the International Conference onLearning Representations (ICLR)*, volume 4.

Chiappa, S. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33.

Corso, G.; Cavalleri, L.; Beaini, D.; Liò, P.; and Veličković, P. 2020. Principal neighbourhood aggregation for graph nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33.

Derrow-Pinion, A.; She, J.; Wong, D.; Lange, O.; Hester, T.; Perez, L.; Nunkesser, M.; Lee, S.; Guo, X.; Wiltshire, B.; et al. 2021. ETA Prediction with Graph Neural Networks in Google Maps. *arXiv preprint arXiv:2108.11482*.

Dua, D.; and Graff, C. 2017a. UCI Machine Learning Repository.

Dua, D.; and Graff, C. 2017b. UCI Machine Learning Repository.

Durkan, C.; Bekasov, A.; Murray, I.; and Papamakarios, G. 2019. Neural Spline Flows. In *Advances in Neural Information Processing Systems*, volume 32.

Gallicchio, C.; and Micheli, A. 2020. Fast and deep graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34.

Garrido, S.; Borysov, S.; Rich, J.; and Pereira, F. 2021. Estimating causal effects with the neural autoregressive density estimator. *Journal of Causal Inference*, 9.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 34. PMLR.

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

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *The Journal of Machine Learning Research (JMLR)*, 13.

Gu, F.; Chang, H.; Zhu, W.; Sojoudi, S.; and El Ghaoui, L. 2020. Implicit Graph Neural Networks. In *Advances in Neural Information Processing Systems*, volume 33.

Karimi, A.-H.; von Kügelgen, J.; Schölkopf, B.; and Valera, I. 2020. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In *Advances in Neural Information Processing Systems*, volume 33, 265–277.

Khemakhem, I.; Monti, R.; Leech, R.; and Hyvarinen, A. 2021. Causal autoregressive flows. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 24. PMLR.

Kim, H.; Shin, S.; Jang, J.; Song, K.; Joo, W.; Kang, W.; and Moon, I.-C. 2021. Counterfactual fairness with disentangled causal effect variational autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, volume 2.

Kipf, T. N.; and Welling, M. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

Kocaoglu, M.; Snyder, C.; Dimakis, A. G.; and Vishwanath, S. 2018. CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, volume 6.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Kusner, M. J.; Sun, Y.; Sridharan, K.; and Weinberger, K. Q. 2016. Private causal inference. In *Proceedings of the Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 19. PMLR.

Li, Q.; Han, Z.; and Wu, X.-M. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Louizos, C.; Shalit, U.; Mooij, J. M.; Sontag, D.; Zemel, R.; and Welling, M. 2017. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Moraffah, R.; Moraffah, B.; Karami, M.; Raglin, A.; and Liu, H. 2020. CAN: A causal adversarial network for learning observational and interventional distributions. *arXiv preprint arXiv:2008.11376*.

Nowozin, S. 2018. Debiasing evidence approximations: On importance-weighted autoencoders and jackknife variational inference. In *Proceedings of the International Conference on Learning Representations (ICML)*, volume 35. PMLR.

Parafita, A.; and Vitria, J. 2019a. Explaining visual models by causal attribution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 4167–4175. IEEE.

Parafita, A.; and Vitria, J. 2019b. Explaining visual models by causal attribution. In *International Conference on Computer Vision Workshop (ICCVW)*.

Parafita, A.; and Vitria, J. 2020. Causal inference with deep causal graphs. *arXiv preprint arXiv:2006.08380*.

Pawlowski, N.; Coelho de Castro, D.; and Glocker, B. 2020. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33.

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

Pearl, J. 2009b. *Causality*. Cambridge university press.

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

Rainforth, T.; Kosiorek, A.; Le, T. A.; Maddison, C.; Igl, M.; Wood, F.; and Teh, Y. W. 2018. Tighter variational bounds are not necessarily better. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 35. PMLR.

Rakesh, V.; Guo, R.; Moraffah, R.; Agarwal, N.; and Liu, H. 2018. Linked causal variational autoencoder for inferring paired spillover effects. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*. ACM.

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

Schölkopf, B. 2019. Causality for machine learning. *arXiv preprint arXiv:1911.10500*.

Schwab, P.; Linhardt, L.; and Karlen, W. 2018. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*.

Shen, X.; Liu, F.; Dong, H.; Lian, Q.; Chen, Z.; and Zhang, T. 2020. Disentangled generative causal representation learning. *arXiv preprint arXiv:2010.02637*.

Tucker, G.; Lawson, D.; Gu, S.; and Maddison, C. J. 2018. Doubly reparameterized gradient estimators for Monte Carlo objectives. In *International Conference on Learning Representations*.

Vowels, M. J.; Camgoz, N. C.; and Bowden, R. 2020. Targeted VAE: Structured inference and targeted learning for causal parameter estimation. *arXiv preprint arXiv:2009.13472*.

Vowels, M. J.; Camgoz, N. C.; and Bowden, R. 2021. D'ya like DAGs? A Survey on Structure Learning and Causal Discovery. *arXiv preprint arXiv:2103.02582*.

Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; and Philip, S. Y. 2020. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*.

Xia, K.; Lee, K.-Z.; Bengio, Y.; and Bareinboim, E. 2021. The Causal-Neural Connection: Expressiveness, Learnability, and Inference. *arXiv preprint arXiv:2107.00793*.

Yang, M.; Liu, F.; Chen, Z.; Shen, X.; Hao, J.; and Wang, J. 2020. CausalVAE: Disentangled representation learning via neural structural causal models. *arXiv preprint arXiv:2004.08697*.

Yu, B.; Yin, H.; and Zhu, Z. 2018. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, volume 27.

Yu, Y.; Chen, J.; Gao, T.; and Yu, M. 2019. DAG-GNN: DAG structure learning with graph neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 36. PMLR.

Zečević, M.; Dhami, D. S.; Veličković, P.; and Kersting, K. 2021. Relating Graph Neural Networks to Structural Causal Models. *arXiv preprint arXiv:2109.04173*.

Zhang, C.; Zhang, K.; and Li, Y. 2020. A Causal View on Robustness of Neural Networks. *Advances in Neural Information Processing Systems*, 33.

Zhang, M.; Jiang, S.; Cui, Z.; Garnett, R.; and Chen, Y. 2019. D-VAE: A variational autoencoder for directed acyclic graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Zheng, M.; and Kleinberg, S. 2019. Using domain knowledge to overcome latent variables in causal inference from time series. In *Proceedings of the Machine Learning for Healthcare Conference (MLHC)*. PMLR.