Deep Multi-Modal Structural Equations For Causal Effect Estimation With Unstructured Proxies

Anonymous Author(s)
Affiliation
Address
email

Abstract

Estimating the effect of an intervention while accounting for confounding variables is a key task in causal inference. Oftentimes, the confounders are unobserved, but we have access to large amounts of unstructured data (images, text) that contain valuable proxy signal about the missing confounders. This paper demonstrates that leveraging unstructured data that is often left unused by existing algorithms improves the accuracy of causal effect estimation. Specifically, we introduce deep multi-modal structural equations, a generative model in which confounders are latent variables and unstructured data are proxy variables. This model supports multiple multi-modal proxies (images, text) as well as missing data. We empirically demonstrate on tasks in genomics and healthcare that our approach corrects for confounding using unstructured inputs, potentially enabling the use of large amounts of data that were previously not used in causal inference.

1 Introduction

An important goal of causal inference is to understand from observational data the causal effect of performing an intervention—e.g., the effect of a behavioral choice on an individual's health [33]. As an initial motivating example for this work, consider the problem of determining the effect of smoking on an individual's risk of heart disease.

This problem is complicated by the presence of confounders: e.g., it is possible that individuals who smoke have a higher likelihood to be sedentary, which is a lifestyle choice that also negatively impacts their heart disease risk. If the individual's lifestyle is available to us as a well-defined feature, we may adjust for this factor while computing treatment effects [33]. However, confounders are often not observed and not available in the form of features, making accurate causal inference challenging.

Oftentimes, datasets in domains such as medicine or genomics come with large amounts of *unstructured data*—e.g., medical images, clinical notes, wearable sensor measurements [7]. This data contains strong proxy signal about unobserved confounding factors—e.g., wearable sensor measurements can help reveal individuals who are sedentary. However, existing causal inference methods usually do not leverage this "dark data" for causal effect estimation [18].

The goal of this paper is to develop methods in causal inference that improve the estimation of causal effects in the presence of unobserved confounders by leveraging additional sources of unstructured multi-modal data, such as images and text. For example, given time series from patients' wearables, our methods may disentangle the effects of being sedentary from the effects of smoking by using clusters in the sensor measurements (which would correspond to groups of active and sedentary individuals) as a proxy for the patients' lifestyles and without requiring explicit lifestyle features.

Concretely, our paper formalizes the task of estimating causal effects using rich, unstructured multi-modal proxy variables (e.g. images, text, time series) and introduces deep multi-modal structural

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

equations, a generative model in which confounders are latent variables. This model supports multiple multi-modal proxy variables and can perform causal effect estimation with missing data by leveraging approximate variational inference and learning algorithms [6] developed for this problem.

We evaluate our methods on an important real-world causal inference problem—estimating the effects of genetic mutations in genome-wide association studies (GWASs)—as well as on benchmarks derived from popular causal inference datasets. On the GWAS task, we demonstrate that our algorithms enable incorporating new types of environmental data (e.g., historical weather time series) which enables discovering causal genetic factors in plants more accurately than using modern state-of-the-art GWAS analysis methods. Our experiments demonstrate the potential of using large amounts of data that were previously left unused by existing methods to improve causal effect estimation.

Contributions In summary, this paper makes three contributions: (1) we define the task of estimating causal effects using rich, unstructured multi-modal proxy variables; (2) we introduce deep multi-modal structural equations, a generative model tailored to this problem, and we describe associated variational learning and inference algorithms; (3) we demonstrate on an important real-world problem (GWAS) that unstructured data can improve causal effect estimation, enabling the use of large amounts of "dark data" that was previously not used in causal inference.

2 Background

Notation Formally, we are given an observational dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)}, t^{(i)})\}_{i=1}^n$ consisting of n individuals, each characterized by features $x^{(i)} \in \mathcal{X} \subseteq \mathbb{R}^d$, a binary treatment $t^{(i)} \in \{0, 1\}$, and a scalar outcome $y^{(i)} \in \mathbb{R}$. We assume binary treatments and scalar outcomes, although our method naturally extends beyond this setting.

We also use $z^{(i)} \in \mathbb{R}^p$ to model latent confounding factors that influence both the treatment and the outcome [26]. We assume that $(x^{(i)}, y^{(i)}, t^{(i)}, z^{(i)})$ represent i.i.d. realizations of random variables $X, Y, T, Z \sim \mathbb{P}$ from a data distribution $\mathbb{P}$ that decomposes as $\mathbb{P}(X|Z)\mathbb{P}(Y|Z,T)\mathbb{P}(T|Z)\mathbb{P}(Z)$ (thus Z is indeed a confounder). We are interested in recovering the true effect of $T = t$ in terms of its individual and average treatment effects:

$$Y[x,t] = \mathbb{E}[Y|X = x, \text{do}(T = t)] \qquad \text{ITE}(x) = Y[x,1] - Y[x,0] \qquad \text{ATE} = \mathbb{E}[\text{ITE}(X)], \tag{1}$$

where $\text{do}(\cdot)$ denotes an intervention [35].

Structural Equations One of the main approaches to causal inference are structural equation models of the form

$$X = f_1(Z, \varepsilon_1) \qquad T = f_2(Z, \varepsilon_2) \qquad Y = f_3(Z, T, \varepsilon_3), \tag{2}$$

where $Z \sim p(Z)$ is drawn from a prior and the ε_i are noise variables drawn independently from their distributions [10]. Structural equations define a *generative model* of the data. When this model encodes the true dependency structure of $\mathbb{P}$, we can estimate the true effect of an intervention by clamping $T = t$ and drawing samples.

Deep Structural Equations Equations 2 can be parameterized with deep neural networks, which yields deep structural equation models. Expressive neural networks may better fit the data and recover the true distribution, which improves causal inference; they also often scale well to large datasets. Such models have been used for GWAS analysis [46] and to correct for proxy variables [26].

3 Causal Effect Estimation With Unstructured Proxy Variables

Oftentimes, datasets in domains such as medicine or genomics come with large amounts of unstructured data (medical images, clinical notes), which contains strong proxy signal about unobserved confounding factors. Our paper seeks to develop methods that leverage unstructured data within causal inference. We start by formalizing this task as causal effect estimation with unstructured proxy variables; these proxies may come from multiple diverse modalities (images, text).

3.1 Task Definition

Formally, consider a causal inference dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)}, t^{(i)})\}_{i=1}^n$ in which $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)})$ is a vector of m distinct input modalities $x_j^{(i)} \in \mathcal{X}_j$ (e.g., images, text, time series, etc.). In other words, $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_m$, where each $\mathcal{X}_j$ corresponds to a space of images, time series, or other unstructured modalities. Here, $t^{(i)} \in T$ (binary or continuous) is the treatment and $y^{(i)} \in Y$ is the output. Some modalities may also be missing at training or inference time.

We assume that the $x_j^{(i)}$ are i.i.d. realizations of a random variable X_j with $X = (X_1, \ldots, X_m)$. We are interested in recovering the true effect of $T = t$ in terms of the individual and average treatment effects. We are specifically interested in estimating the individual treatment effect (ITE) from arbitrary subsets of modalities $\mathcal{M} \subseteq \{1, 2, ..., m\}$, indicating that certain inputs may be missing at test time.

$$Y[x, t, \mathcal{M}] = \mathbb{E}[Y|\text{do}(T = t), X_j = x_j \text{ for } j \text{ in } \mathcal{M}] \quad \text{ITE}(x, \mathcal{M}) = Y[x, t = 1, \mathcal{M}] - Y[x, t = 0, \mathcal{M}] \tag{3}$$

3.2 Motivating Applications

To help make this setup more concrete, we define two motivating applications.

Healthcare Consider the task of determining the effect of smoking on heart disease from an observational dataset of patients. The observational study may contain additional unstructured data about individuals, e.g., clinician notes, medical images, wearable sensor data, etc. This data may hold information about hidden confounders: for example, raw wearable sensor data can be clustered to uncover sedentary and active indivudals, revealing a latent confounding factor, sedentary lifestyle.

Genomics Consider the problem of estimating the effects of genetic variants via a genome-wide association study (GWAS). Modern GWAS datasets in plants or humans feature large amounts of unstructured inputs [7]: clinical notes, medical records, meteorological time series. Historical weather data (e.g., precipitation, wind strength, etc.) can reveal distinct climatic regions that affect phenotypes in a plant dataset and whose confounding effects should be corrected in a GWAS [51].

4 Deep Structural Equations for Causal Effect Estimation

Next, we derive models and inference algorithms for the task of causal effect estimation with unstructured proxy variables. Our approach uses deep structural equations to extract confounding signal from the multi-modal proxies $x_j^{(i)}$. We use neural networks because they naturally handle unstructured modalities via specialized architectures (e.g., convolutions for images) that can learn learn high-level representations over raw unstructured inputs (e.g., pixels).

Parameterizing structural models with neural networks also presents challenges: they induce complex latent variable models that require the development of efficient approximate inference algorithms [10, 6]. We present an instantiation of Equations 2 that admits such efficient algorithms.

4.1 Deep Multi-Modal Structural Equations

We start by introducing deep multi-modal structural equations (DMSEs), a generative model for estimating causal effects using unstructured proxy variables in which confounders are latent variables and unstructured data are proxy variables. We define a DMSE model as follows:

$$Z \sim \mathcal{N}(0_p, I_p) \quad X_j \sim \mathcal{P}_{X_j}(\theta_{X_j}(Z)) \;\; \forall j \quad T \sim \text{Ber}(\pi_T(Z)) \quad Y \sim \mathcal{P}_Y(\theta_Y(Z, T))), \tag{4}$$

where $\mathcal{P}_{X_j}, \mathcal{P}_Y$ are probability distributions with a tractable density and the μ, σ, π, θ are functions parameterized by neural networks that output the parameters of their respective probability distribution as a function of ancestor variables in the causal graph (Figure 1). The $\mathcal{P}_{X_j}, \mathcal{P}_Y$ are typically Gaussian when X_j, Y are continuous and Bernoulli when they are discrete, but other choices are possible.

These equations define a density $p(X, Y, T, Z) = p(Z)p(T|Z)p(Y|Z, T) \prod_{j=1}^m p(X_j|Z)$. Note that we make the natural modeling assumption that each auxiliary modality is independent of the others

conditioned on Z. Figure 1 shows the dependencies. These modeling choices will shortly enable us to apply stochastic variational inference algorithms.

Note that the models for μ_{X_j}, σ_{X_j} are domain-specific—certain modalities will benefit from specialized architectures, e.g., a convolutional model for images and a recurrent one for time series. See Appendix B for details on recommended architectures for all the variables μ, σ, π, θ.

4.2 Approximate Inference and Learning Algorithms

In the full multi-modal setting, the X_j are conditionally independent given Z (Figure 1), which enables us to apply efficient algorithms inspired by Wu & Goodman [52]. These algorithms offer the following improvements: (1) we may perform learning, inference, and causal estimation with missing modalities X_j; (2) the process for performing causal inference is simpler than in previous work [26] and does not require training auxiliary inference networks.

The DMSE model induces a tractable joint density $p(X, Y, T, Z)$, which allows us to fit its parameters using stochastic variational inference by optimizing the evidence lower bound (ELBO):

$$\text{ELBO}_X = \sum_{i=1}^{n} \mathbb{E}_q \big[\sum_{j=1}^{m} \log p(x_j^{(i)}|z) + \log p(y^{(i)}, t^{(i)}, z) - \log q(z|x^{(i)}, y^{(i)}, t^{(i)}) \big], \tag{5}$$

where $p(y^{(i)}, t^{(i)}, z) = p(y^{(i)}|t^{(i)}, z)p(t^{(i)}|z)p(z)$ and $q(z|x^{(i)}, y^{(i)}, t^{(i)})$ is the approximate variational posterior. We assume a total of m modalities.

Structured Multi-Modal Variational Inference We may use the independence structure of p (Figure 1) to derive a more efficient structured form for q than the one prescribed by classical variational inference. First, observe that because the true posterior factorizes as $p(z|x, t, y) \propto (p(z|t, y) \prod_{j=1}^{m} p(z|x_j)) / \prod_{j=1}^{m-1} p(z)$, the optimal approximate posterior q must also factorize as $q(z|x, t, y) \propto (q(z|t, y) \prod_{j=1}^{m} q(z|x_j)) / \prod_{j=1}^{m-1} p(z)$. This decomposition implies that we can maintain the optimal structure of q by training modality-specific inference networks $\tilde{q}(z|t, y)$ and $\tilde{q}(z|x_j)$ such that $q(z|x_j) = \tilde{q}(z|x_j)p(z)$ and $q(z|t, y) = \tilde{q}(z|t, y)p(z)$ and by defining a joint posterior as

$$q(z|x, y, t) \propto p(z)\tilde{q}(z|y, t) \prod_{j=1}^{m} \tilde{q}(z|x_j). \tag{6}$$

This network can be seen as a product of experts (PoE) [52].

Computing the density q is in general not possible. However, because $p(z)$, $q(z|t, y)$ and $q(z|x_j)$ are Gaussians, we may use the fact that a product of Gaussians with means μ_i and covariances V_i is $\mu = (\sum \mu_i T_i)/(\sum T_i)$ and $V = (\sum_i T_i)^{-1}$, where $T_i = 1/V_i$. Thus, computing $q(z|x, t, y)$ for any subset of modalities is possible without having to train an inference network for each subset of modalities separately.

A Multi-Modal ELBO Training this architecture with ELBO_X will not necessarily yield good single-modality inference networks, while training with each modality separately will prevent the network from learning how the modalities are related to each other. Hence, we train our model with a sub-sampled ELBO objective [52] that is computed on the full set of modalities, each individual modality, and a few subsets of modalities together. For this, we randomly pick at each gradient step s non-empty subsets $\{\mathcal{M}_k\}_{k=1}^{s}$ of the set of modalities $\mathcal{M}_k \subseteq \{1, 2, ..., m\}$. The final objective is $\text{ELBO}_X + \sum_{j=1}^{m} \text{ELBO}_{\{j\}} + \sum_{k=1}^{s} \text{ELBO}_{\mathcal{M}_k}$, where

$$\text{ELBO}_{\mathcal{M}} = \sum_{i=1}^{n} \mathbb{E}_q \left[\sum_{j \in \mathcal{M}} \log p(x_j^{(i)}, z) + \log p(y^{(i)}, t^{(i)}, z) - \log(p(z)\tilde{q}(z|y^{(i)}, t^{(i)}) \prod_{j \in \mathcal{M}} \tilde{q}(z|x_j^{(i)})) \right]$$

Our model doesn't train extra auxiliary inference networks, unlike that of Louizos et al. [26].

4.3 Deep Gaussian Structural Equations

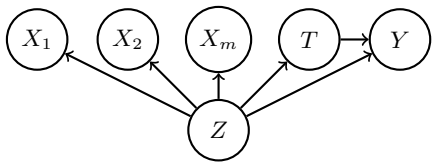

Figure 1: Causal graph for the deep multimodal structural equations model

DMSE can be simplified in settings in which there is only one type of auxiliary data x (i.e., $m = 1$). This yields the same model as Equations (8) except that we have only one equation $X \sim \mathcal{P}_X(\theta_X(Z))$ for the proxy variables. This simplified model, which we refer to as deep gaussian structural equations (DGSEs), has a tractable joint density $p(X, Y, T, Z) = p(Z)p(X|Z)p(T|Z)p(Y|Z, T)$, where the latent Z is Gaussian.

The DGSE model can also be fit using stochastic variational inference by optimizing the ELBO objective $\sum_{i=1}^{n} \mathbb{E}_q \left[\log p(x^{(i)}, y^{(i)}, t^{(i)}, z) - \log q(z|x^{(i)}, y^{(i)}, t^{(i)})\right]$, where $q(z|x, y, t)$ is an approximate variational posterior. We optimize the above objective using gradient descent, applying the reparameterization trick to estimate the gradient. We compute the counterfactual $Y[x, t]$ using auxiliary inference networks as in earlier work [26]. See Appendix B for the full derivation.

4.4 Properties of Deep Multi-Modal Structural Equations

Recovering Causal Effects The DMSE and DGSE models determine the true causal effect when their causal graph is correct and they recover the true data distribution. The following argument is analogous to that made in most previous methods on causal deep generative models [26, 53, 32].

Theorem 1. *The DMSE and DGSE models recover the true ITE*$(x, \mathcal{M})$ *for any subset* $\mathcal{M} \subseteq \{1, 2, ..., m\}$ *of observed modalities whenever they represent the true data distribution* $p(x, y, t, z)$.

Proof: We establish the theorem for DMSEs; the proof for DGSEs is analogous with $m = 1$. Let $x_{\mathcal{M}} = \{x_j \mid j \in \mathcal{M}\}$ be the data from the observed subset of modalities. We need to show that $p(y|x_{\mathcal{M}}, \text{do}(t = t'))$ is identifiable for any t'. Observe that

$$p(y|x_{\mathcal{M}}, \text{do}(t = t')) = \int_z p(y|z, x_{\mathcal{M}}, do(t = t'))p(z|x_{\mathcal{M}}, do(t = t'))dz = \int_z p(y|z, x_{\mathcal{M}}, t')p(z|x_{\mathcal{M}})dz,$$

where the second equality follows from the rule of do-calculus (applying backdoor adjustment). Since our proof holds for any t' and all elements on the right-hand side are identifiable, the claim follows. ■

Note that in practice our assumption may not hold (e.g., neural network optimization is non-convex and may fail), but there is evidence of both real-world failure modes [42] as well as success stories where deep latent variable models provide useful causal estimates [53, 32, 27, 45]. See our Discussion section for additional details.

Computing Causal Effects Given a subset of modalities $\mathcal{M}$, we can compute the ATE & ITE as $\mathbb{E}(\text{ITE}(x, \mathcal{M}))$, where $\text{ITE}(x, \mathcal{M}) = Y[x, t = 1, \mathcal{M}] - Y[x, t = 0, \mathcal{M}]$ and

$$p(y|x, \text{do}(t = t')) = \int_z p(y, t = t'|z)p(z|x)dz \approx \int_z \big(p(y|t = t', z)p(z) \prod_{j=1}^{m} \tilde{q}(z|x_j)\big)dz, \qquad (7)$$

where we use our variational posterior formulation from Equation 6 to approximate the true posterior $p(z|x)$.

5 Experimental Results

5.1 Synthetic Demonstration Dataset

We start with a demonstration that provides intuition for why proxy variables are important, and how unstructured proxies can serve in place for featurized (structured) ones. The following small-scale synthetic setup (Louizos et al. [26]) involves a data distribution $\mathbb{P}$ over binary variables y, t, z, x:

$$\mathbb{P}(z = 1) = \mathbb{P}(z = 0) = 0.5 \qquad \mathbb{P}(x = 1|z = 1) = \mathbb{P}(x = 0|Z = 0) = \rho_x = 0.5$$
$$y = t \oplus z \qquad \mathbb{P}(t = 1|z = 1) = \mathbb{P}(t = 0|z = 0) = \rho_t = 0.3$$

where $0 < \rho_x, \rho_t < 1$ are parameters. Crucially, we introduce an unstructured proxy variable $\mathbf{X}$ that represents an "image version" of x. The variable $\mathbf{X}$ will be a random MNIST image of a zero or one, depending on whether $x = 0$ or $x = 1$. Formally, $\mathbf{X}$ is distributed as follows:

$$\mathbb{P}(\mathbf{X}|x = 1) \text{ is unif. over MNIST images of '1'} \quad \mathbb{P}(\mathbf{X}|x = 0) \text{ is unif. over MNIST images of '0'}$$

Table 1: Treatment effect estimation on the synthetic demonstration dataset.

	Setting	ε_{ATE} (Train)	ε_{ATE} (Test)
Causal	Binary	0.003	0.013
	Image	0.011	0.012
Non-Causal	Binary	0.002	0.018
	Image	0.008	0.022

First, this is a setup that requires us to model proxies (see also Appendix A): treating $\mathbf{X}$ as a confounder as using a model of $\mathbb{P}(y \mid \mathbf{X}, t)$ recovers the true ATE only when $\rho_t = \rho_x = 0.5$ (i.e., when $\mathbf{X}$ is perfectly informative of z), otherwise it fails.

We also show that structural equations solve this task. We fit DGSE models to 10000 training data points of $\{x, y, t\}$ (the BINARY setting) as well as of $\{\mathbf{X}, y, t\}$ (the IMAGE setting). We note the Average Treatment Effect (ATE) on 600 test datapoints sampled independently from $\mathbb{P}$ and we report results in Table 1. DGSE performs as well as models trained on the featurized proxy variables x.

5.2 Benchmark Datasets for Causal Effect Estimation

IHDP The Infant Health and Development Project (IHDP) is a popular benchmark for causal inference algorithms [13] that contains the outcomes of comprehensive early interventions for premature, low birth weight infants. We create a benchmark for multi-modal causal inference based on IHDP in which we can replace certain features with their "unstructured version". We choose 9 of the 25 features available in IHDP in order to magnify their relative importance and accurately measure the effects of their removal. Please refer to Appendix C and D for detailed setup.

STAR The Student-Teacher Achievement Ratio (STAR) experiment [1] studied the effect of class size on the performance of students. We consider small class size as treatment; the outcome is the sum of the reading and math scores of a student. We 'derandomize' this dataset by removing 80% of the data corresponding to white students in the treated population. Similarly to IHDP, we select 8 attributes for the multi-modal experiment. Further details can be found in the Appendix C and D.

Table 2: Multimodal Experiments on IHDP Dataset: With deep structural equations, replacing baby's gender with corresponding image embedding (8 attrs + image) shows some increase in ATE error as compared to IHDP-Mini setting (9 attrs) but is better than dropping this modality altogether (8 attrs).

Model	ε_{ATE} (Train+Val)	ε_{ATE} error(Test)
Deep Str Eqns		
9 attrs	0.259 (0.037)	0.487 (0.078)
8 attrs	0.392 (0.141)	0.620 (0.158)
8 attrs + image	0.372 (0.107)	0.575 (0.130)
CFRNet		
9 attrs	0.433 (0.063)	0.549 (0.090)
8 attrs	0.412 (0.062)	0.608 (0.107)
8 attrs + image	0.501 (0.076)	0.617 (0.114)
OLS		
9 attrs	0.424 (0.061)	0.584 (0.100)
8 attrs	0.429 (0.066)	0.593 (0.103)
8 attrs + image	0.428 (0.064)	0.590 (0.101)

Adding Unstructured Modalities We create a benchmark for multi-modal causal inference derived from IHDP and STAR in which we replace features with unstructured inputs that contain the same information as their featurized versions. On IHDP, we replace the attribute 'baby's gender' with the CLIP embedding [22] of an image of a child between ages 3 to 8 years, drawn from the UTK dataset [58]. On STAR, we replace the attributes corresponding to the student's ethnicity and gender by selecting an image of a child with the same ethnicity and gender from the UTK dataset.

We train and evaluate models on datasets where the image 'replaces' the attribute. (e.g., 8 ATTRS + IMAGE). We also consider two other settings for comparison: a) the original attribute is included (e.g., 9 ATTRS) and b) the attribute is dropped from the reduced set of input features (8 ATTRS).

Results As seen in Table 2, the degradation in ATE error from replacing the baby's gender by a photograph is lower as compared to removing the attribute entirely. This shows our models leverage signal found in the unstructured image modality with the help of deep neural networks.

We compare these results with a simple Ordinary Least Squares model (OLS) baseline as described by Shalit et al. [44] to predict treatment effect. OLS shows a similar behavior when replacing baby's gender with corresponding image, however ATE errors are generally worse as compared to ATE error produced by DGSE. We also compare this with Counterfactual Regression Network (CFRNet) [16] baseline. However CFRNet did not show benefits of using image modality unlike our approach.

In Table 6, replacing gender and ethnicity attributes on STAR with the corresponding image improves ATE errors as compared to dropping these two attributes entirely. This shows that we can use an image to extract multiple attributes while doing causal inference. The CFRNet baseline shows a similar behavior, but the difference between average ATE errors across different setups is small.

5.3 Genome-Wide Association Studies

We evaluate our methods on an important real-world causal inference problem—genome-wide association study analysis (GWASs). A GWAS is a large observational study that seeks to determine the causal effects of genetic markers (or genotypes) on specific traits (known as phenotypes).

Background and Notation As motivation, consider the problem of linking a plant's genetic variants $t \in \{0,1\}^d$ with *nutritional yield*, which we model via a variable $y \in \mathbb{R}$. Our goal is to determine if each variant t_j is *causal* for yield, meaning that it influences biological mechanisms which affect this phenotype [8]. We also want to leverage large amounts of unstructured data x (e.g., health records, physiological data) that are often available in modern datasets [7, 23].

Table 3: Comparison of standard VAE methods with linear baselines. The ℓ_1 column refers to $\|\widehat{\gamma} - \gamma^\star\|_1$ where $\widehat{\gamma}$ is the vector of estimated causal effects, and $\gamma^\star$ is the vector of ground truth causal effects. Precision and recall are defined in Appendix F. Standard error of the mean (sem) is computed over 10 seeds.

Model	$\ell_1 (\downarrow)$ mean (sem)	Precision ($\uparrow$) mean (sem)	Recall ($\uparrow$) mean (sem)
Optimal	0.22 (0.04)	0.97 (0.03)	1.0 (0.00)
DSE (2 modalities)	0.30 (0.06)	0.93 (0.04)	1.0 (0.00)
LMM	0.44 (0.06)	0.85 (0.08)	1.0 (0.00)
DSE (1 modality)	0.60 (0.09)	0.78 (0.08)	1.0 (0.00)
PCA (1 component)	0.93 (0.17)	0.58 (0.09)	1.0 (0.00)
FA (1 component)	1.08 (0.17)	0.62 (0.08)	1.0 (0.00)
PCA (2 components)	1.38 (0.24)	0.44 (0.09)	0.9 (0.07)
FA (2 components)	1.44 (0.30)	0.55 (0.09)	1.0 (0.00)
PCA (3 components)	1.66 (0.23)	0.37 (0.08)	0.8 (0.08)
FA (3 components)	1.89 (0.45)	0.44 (0.08)	0.9 (0.07)

A key challenge in finding causal variants is ancestry-based confounding [2, 49]. Suppose that we are doing a GWAS of plants from Countries A and B; plants in Country A get more rainfall, and thus grow faster and are more nutritious. A simple linear model of y and t will find that any variant that is characteristic of plants in Country A (e.g., bigger leaves to capture rain) is causal for nutritional yield.

Methods and Baselines Most existing GWAS analysis methods for estimating the effect of a variant t_j rely on latent variable models: (1) they treat all remaining variants x as proxies and obtain z via *a linear projection* (e.g., PCA [36, 37] or LMM [56, 24]) of x into a lower dimensional space where genomes from Countries A and B tend to form distinct clusters (because plants from the same country breed and are similar); (2) we assume a *linear model* $\beta^\top t$ of y and add z into it, which effectively adds the country as a feature (z reveals the cluster for each country); this allows the model to *regress out* the effects of ancestry and assign the correct effect to variants t (one at a time).

Baselines. Our main baselines are Principal Component Analysis (PCA) and Linear Mixed Model (LMM), as described above and implemented via the popular LIMIX library [25]. We also compare against Factor Analysis (FA), a standard linear technique for deriving latent variables, Uniform Manifold Approximation and Projection (UMAP), a manifold learning technique for dimensionality reduction [28] and a linear model with no correction for confounding.

DSE Models. We compare against deep structural equation models that leverage one or more sets of proxy variables coming from the following unstructured modalities: genomic sequences, weather time

series, simulated physiological time series. We fit DSE models via a stagewise strategy analogous to how classical GWAS models are fit: (1) we fit the component $p(z_k|x_k)$ for each proxy x_k; (2) we fit a *linear model* of y given t and the z to estimate causal effects. Thus, our $p(x_k|z)$ components are deep, while $p(y|x, z)$ is shallow (following standard assumptions on epistasis in GWAS).

5.3.1 Simulated Human GWAS

We used the 1000 Human Genomes [3] dataset to generate a simulated multi-modal GWAS dataset, following the biologically-inspired "Spatial" strategy studied in [46]. In addition to genotypes, we generated random physiological time series by sampling Fourier series conditioned on the confounders. Please see Appendix F for details.

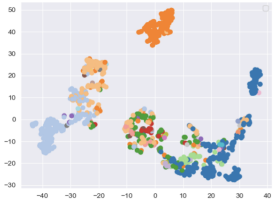

Figure 2: Latent z extracted by DSEs. Plants from different countries form clusters (in color).

Table 3 shows our results. Amongst the non-oracle baselines, multi-modal DSE has the smallest error in estimating causal effect, as well as the highest precision and recall at identifying causal SNPs. The uni-modal DSE, while worse than LMM, still beats PCA and FA. Note that the LMM model is not compatible with multiple unstructured proxies. In general, we see that precision starts to deteriorate faster than recall, suggesting that false positives are more likely from the weaker linear deconfounding methods such as PCA/FA. Our results here further support that additional sources of unstructured multi-modal data can improve GWAS.

5.3.2 Real-World Plant GWAS

We also tested our methods on real-world plant GWAS dataset from the 1001 Genomes Project for *Arabidopsis Thaliana* plants [51]. The challenge here is that we don't know true causal effects; therefore, we define a phenotype y for which the true causal effect is zero—specifically, we use the GDP of the country where each plant grows. Because the genomes of plants from the same country are similar, there exist spurious correlation caused by latent subpopulation groups z. Our goal is to detect and correct for these confounding effects. In addition to genomic data, we gather historical weather time series from the location of each plant (see Appendix E) and use both modalities to correct for confounding.

Table 4: Multimodal Experiments on IHDP and STAR Datasets comparing DMSE and DGSE methods. When the image modality is missing, DMSE still produces reasonable ATE errors as it can compute the latent representation on a subset of modalities better as compared to DMSE

Dataset	Model	ε_{ATE} (train+val)	ε_{ATE} error(test)
IHDP	DMSE	0.433 (0.057)	0.627 (0.094)
	DGSE	0.794 (0.308)	1.080 (0.337)
STAR	DMSE	32.575 (1.634)	33.743 (1.890)
	DGSE	59.102 (4.734)	60.152 (4.788)

Table 5: Correcting for confounding on a plant GWAS dataset. DSEs can discover complex, non-linear clusters over genomic and weather data to identify the latent confounding variable better.

Model	Input	$R^2(\downarrow)$
Deep Str Eqns (Ours)	Weather+SNPs	0.049 (0.022)
PCA	Weather+SNPs	0.097 (0.019)
UMAP	Weather+SNPs	0.412 (0.026)
Deep Str Eqns (Ours)	Weather	0.555 (0.027)
PCA	Weather	0.545 (0.017)
UMAP	Weather	0.653 (0.013)
Deep Str Eqns (Ours)	SNPs	0.068 (0.029)
PCA	SNPs	0.130 (0.020)
UMAP	SNPs	0.406 (0.020)
LMM	SNPs	0.804 (0.009)
Linear Model	-	0.665 (0.012)

We evaluate whether each model learned to correctly account for latent confounding effects by measuring the predictive power of $\beta^\top t$ for y, where β are the causal effects and t is the vector of variants. Specifically, we compare the R^2 correlation between $\beta^\top t$ and y—here, *lower is better*, since a model that has learned the causal effects will not correctly predict the phenotype. In Table 5, we see the effect of extracting confounding variables using DSE as opposed to using the standard PCA technique. We can see that the R^2 values produced using DGSE for extracting confounders are closer to 0 as compared to using PCA. This experiment shows that neural network architectures are effective in deal-

ing with unstructured genomic and weather data while correcting for confounding. Please refer to Appendix E for details.

5.4 Multimodal Experiments With Missing Modalities

We demonstrate the ability of the DMSE model to handle missing data. Our IHDP and STAR benchmarks involve two modalities: images (e.g., baby's gender in IHDP) and tabular data (e.g., the remaining features). We compare DMSE and DGSE models on these datasets when some of the modalities may be missing.

For DMSE, we define two different modalities X_1 and X_2 for the tabular and image modalities respectively. For DGSE, we concatenate the image embedding to the tabular modality while training the network. We evaluate ATE while randomly dropping 50% of the images. DMSE utilizes its product-of-experts inference network to approximate the posterior distribution when modalities are missing. DGSE cannot do this, and we resort to feeding it a vector of zeros when an image is missing. Table 4 shows that DMSE produces improved ATE estimates as compared to the DGSE modeling modalities.

6 Related Work

Multi-Modal Causal Inference While previous work analyzed unstructured interventions t consisting of natural language [39–41] (e.g., determining the effect of a polite vs. a rude response) as well as unstructured y [5, 32] (e.g., MRI images), our work explores unstructured x, not y, t. Veitch et al. developed models that correct for confounders from a single unstructured confounder x derived from text [30, 48] or a graph [47]. These approaches rely on a propensity scoring framework—they train a discriminative model of $p(t|x)$—hence do not support proxy variables or missing data, and require pre-trained text embeddings. Our method (i) works across all modalities (beyond text or graphs), (ii) supports arbitrary numbers of proxies, (iii) supports missing data by virtue of being generative.

Deep Latent Variable Models Representation learning in causal inference has been studied by Johansson et al. [16, 15, 17] and Schölkopf et al. [43]. Deep latent variable models find applications throughout causal inference [26, 27, 53, 32, 57, 50, 19]. Pawlowski et al. [32] study unstructured outcomes y (MRI scans), but do not support proxies. Louizos et al. [26] use variational auto-encoders to estimate confounders from proxies; we introduce a more structured model that handles multiple proxies that can be missing, and obviates the need for auxiliary modules. Tran & Blei [46] propose implicit deep structural equations for GWAS; ours are explicit and thus easier to train.

7 Discussion and Conclusion

Identifiability Approaches to causal effect estimation that rely on deep learning [26, 44, 32, 45, 27, 11, 57, 55] can never guarantee the recovery of causal effects: neural network is non-convex and has no guarantees itself. Other failure modes of deep latent variable models (DLVMs) include not having a sufficiently expressive model, not having enough data to learn the model and failures of approximate inference. That said, there is evidence of both real-world failures [42] and successes of DLVMs. [53, 32, 27].The DLVM approach is appealing to us because: (i) it naturally handles proxies as observed variables conditioned on a latent and (ii) neural networks are effectively necessary to process unstructured data (alternatives are worse in practice). Rissanen & Marttinen [42] empirically identify multiple failure modes of DLVMs; however we also identify successes (e.g., in GWAS), and ultimately the validation of DLVM methods needs to be empirical in specific domains [45, 55, 47, 11, 19].

Missing Data We make the common assumption that data is missing at random (MAR). This poses challenges if, for example, patients missing outcomes are ones that are more likely to be sick. When two modalities and their missingness are correlated, their x_j, x_k nodes in Figure 1 could be merged, somewhat addessing the issue. We leave the full exploration of non-MAR models to future work.

In conclusion, we proposed an approach based on deep structural equations to utilize useful data present in unstructured form to perform causal inference. We showed that deep structural equations can estimate causal effects with unstructured proxies across a real-world genomics task and on popular benchmarks, including with missing data.

References

[1] Achilles, C., Bain, H. P., Bellott, F., Boyd-Zaharias, J., Finn, J., Folger, J., Johnston, J., and Word, E. Tennessee's Student Teacher Achievement Ratio (STAR) project, 2008. URL `https://doi.org/10.7910/DVN/SIWH9F`.

[2] Astle, W. and Balding, D. J. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471, 2009.

[3] Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., Mardis, E. R., Marth, G. T., McVean, G. A., Nickerson, D. A., Schmidt, J. P., Sherry, S. T., Wang, J., Wilson, R. K., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., Zhu, Y., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Gupta, N., Gharani, N., Toji, L. H., Gerry, N. P., Resch, A. M., Barker, J., Clarke, L., Gil, L., Hunt, S. E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W. M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R. E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Sudbrak, R., Albrecht, M. W., Amstislavskiy, V. S., Borodina, T. A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.-L., Fulton, L., Fulton, R., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O'Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T. M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Davies, C. J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Campbell, C. L., Kong, Y., Marcketta, A., Yu, F., Antunes, L., Bainbridge, M., Sabo, A., Huang, Z., Coin, L. J. M., Fang, L., Li, Q., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Garrison, E. P., Kural, D., Lee, W.-P., Fung Leong, W., Stromberg, M., Ward, A. N., Wu, J., Zhang, M., Daly, M. J., DePristo, M. A., Handsaker, R. E., Banks, E., Bhatia, G., del Angel, G., Genovese, G., Li, H., Kashin, S., McCarroll, S. A., Nemesh, J. C., Poplin, R. E., Yoon, S. C., Lihm, J., Makarov, V., Gottipati, S., Keinan, A., Rodriguez-Flores, J. L., Rausch, T., Fritz, M. H., Stütz, A. M., Beal, K., Datta, A., Herrero, J., Ritchie, G. R. S., Zerbino, D., Sabeti, P. C., Shlyakhter, I., Schaffner, S. F., Vitti, J., Cooper, D. N., Ball, E. V., Stenson, P. D., Barnes, B., Bauer, M., Keira Cheetham, R., Cox, A., Eberle, M., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E. E., Batzer, M. A., Konkel, M. K., Walker, J. A., MacArthur, D. G., Lek, M., Herwig, R., Ding, L., Koboldt, D. C., Larson, D., Ye, K., Gravel, S., 1000 Genomes Project Consortium, T., authors, C., committee, S., group, P., of Medicine, B. C., BGI-Shenzhen, of MIT, B. I., Harvard, for Medical Research, C. I., European Molecular Biology Laboratory, E. B. I., Illumina, for Molecular Genetics, M. P. I., at Washington University, M. G. I., of Health, U. N. I., of Oxford, U., Institute, W. T. S., group, A., Affymetrix, of Medicine, A. E. C., University, B., College, B., Laboratory, C. S. H., University, C., Laboratory, E. M. B., University, H., Database, H. G. M., at Mount Sinai, I. S. o. M., University, L. S., Hospital, M. G., University, M., and National Eye Institute, N. I. H. A global reference for human genetic variation. *Nature*, 526(7571):68–74, Oct 2015. ISSN 1476-4687. doi: 10.1038/nature15393. URL `https://doi.org/10.1038/nature15393`.

[4] Balding, D. J. and Nichols, R. A. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96(1): 3–12, 1995.

[5] Batmanghelich, N. K., Dalca, A., Quon, G., Sabuncu, M., and Golland, P. Probabilistic modeling of imaging, genetics and diagnosis. *IEEE transactions on medical imaging*, 35(7):1765–1779, 2016.

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

[7] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[8] Crouch, D. J. and Bodmer, W. F. Polygenic inheritance, gwas, polygenic risk scores, and the search for functional variants. *Proceedings of the National Academy of Sciences*, 117(32): 18924–18933, 2020.

[9] Dorie, V. Non-parametrics for causal inference, 2021. URL `https://github.com/vdorie/npci`.

[10] Duncan, O. D. *Introduction to structural equation models*. Elsevier, 2014.

[11] Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. Counterfactual prediction with deep instrumental variables networks, 2016. URL `https://arxiv.org/abs/1612.09596`.

[12] Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=Sy2fzU9gl`.

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

[14] Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[15] Johansson, F. D., Kallus, N., Shalit, U., and Sontag, D. Learning weighted representations for generalization across designs, 2018.

[16] Johansson, F. D., Shalit, U., and Sontag, D. Learning representations for counterfactual inference, 2018.

[17] Johansson, F. D., Sontag, D., and Ranganath, R. Support and invertibility in domain-invariant representations, 2019.

[18] Kallus, N. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *International Conference on Machine Learning*, pp. 5067–5077. PMLR, 2020.

[19] Kim, H., Shin, S., Jang, J., Song, K., Joo, W., Kang, W., and Moon, I.-C. Counterfactual fairness with disentangled causal effect variational autoencoder, 2020. URL `https://arxiv.org/abs/2011.11878`.

[20] Kingma, D. P. and Welling, M. Auto-encoding variational bayes, 2013. URL `https://arxiv.org/abs/1312.6114`.

[21] LeCun, Y., Bengio, Y., et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[22] Li, M., Xu, R., Wang, S., Zhou, L., Lin, X., Zhu, C., Zeng, M., Ji, H., and Chang, S.-F. Clip-event: Connecting text and images with event structures. *arXiv preprint arXiv:2201.05078*, 2022.

[23] Li, R., Chen, Y., Ritchie, M. D., and Moore, J. H. Electronic health records and polygenic risk scores for predicting disease risk. *Nature Reviews Genetics*, 21(8):493–502, 2020.

[24] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833–835, 2011.

[25] Lippert, C., Casale, F. P., Rakitsch, B., and Stegle, O. Limix: genetic analysis of multiple traits. *BioRxiv*, 2014.

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

[27] Mayer, I., Josse, J., Raimundo, F., and Vert, J.-P. Missdeepcausal: Causal inference from incomplete data using deep latent variable models, 2020. URL https://arxiv.org/abs/2002.10837.

[28] McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction, 2018. URL https://arxiv.org/abs/1802.03426.

[29] Menne, M., Durre, I., Vose, R., Gleason, B., and Houston, T. An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29:897–910, 07 2012. doi: 10.1175/JTECH-D-11-00103.1.

[30] Naoki, E., Christian, J. F., Justin, G., Margaret, E. R., and Brandon, M. S. How to make causal inferences using texts, 2018.

[31] Nussbaumer, H. J. Fast fourier transform and convolution algorithms, 1981.

[32] Pawlowski, N., Coelho de Castro, D., and Glocker, B. Deep structural causal models for tractable counterfactual inference. *Advances in Neural Information Processing Systems*, 33: 857–869, 2020.

[33] Pearl, J. *Causality*. Cambridge university press, 2009.

[34] Pearl, J. An introduction to causal inference., 2010.

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

[36] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[37] Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature reviews genetics*, 11(7):459–463, 2010.

[38] Pritchard, J. K., Stephens, M., and Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[39] Pryzant, R., Chung, Y., and Jurafsky, D. Predicting sales from the language of product descriptions. In *eCOM@ SIGIR*, 2017.

[40] Pryzant, R., Basu, S., and Sone, K. Interpretable neural architectures for attributing an ad's performance to its writing style. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 125–135, 2018.

[41] Pryzant, R., Card, D., Jurafsky, D., Veitch, V., and Sridhar, D. Causal effects of linguistic properties. *arXiv preprint arXiv:2010.12919*, 2020.

[42] Rissanen, S. and Marttinen, P. A critical look at the identifiability of causal effects with deep latent variable models, 02 2021.

[43] Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Towards causal representation learning. *CoRR*, abs/2102.11107, 2021. URL https://arxiv.org/abs/2102.11107.

[44] Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms, 2017.

[45] Shi, C., Blei, D. M., and Veitch, V. Adapting neural networks for the estimation of treatment effects, 2019. URL https://arxiv.org/abs/1906.02120.

[46] Tran, D. and Blei, D. M. Implicit causal models for genome-wide association studies. *arXiv preprint arXiv:1710.10742*, 2017.

[47] Veitch, V., Wang, Y., and Blei, D. M. Using embeddings to correct for unobserved confounding in networks, 2019.

[48] Veitch, V., Sridhar, D., and Blei, D. M. Adapting text embeddings for causal inference, 2020.

[49] Vilhjálmsson, B. J. and Nordborg, M. The nature of confounding in genome-wide association studies. *Nature Reviews Genetics*, 14(1):1–2, 2013.

[50] Vowels, M. J., Camgoz, N. C., and Bowden, R. Targeted vae: Variational and targeted learning for causal inference, 2020. URL `https://arxiv.org/abs/2009.13472`.

[51] Weigel, D. and Mott, R. The 1001 genomes project for arabidopsis thaliana. *Genome biology*, 10(5):1–5, 2009.

[52] Wu, M. and Goodman, N. Multimodal generative models for scalable weakly-supervised learning. *arXiv preprint arXiv:1802.05335*, 2018.

[53] Wu, P. and Fukumizu, K. Towards principled causal effect estimation by deep identifiable models, 2021. URL `https://arxiv.org/abs/2109.15062`.

[54] Xia, K., Lee, K.-Z., Bengio, Y., and Bareinboim, E. The causal-neural connection: Expressiveness, learnability, and inference, 2021. URL `https://arxiv.org/abs/2107.00793`.

[55] Yoon, J., Jordon, J., and van der Schaar, M. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *ICLR*, 2018.

[56] Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203–208, 2006.

[57] Zhang, W., Liu, L., and Li, J. Treatment effect estimation with disentangled latent factors, 2020. URL `https://arxiv.org/abs/2001.10652`.

[58] Zhang, Z., Song, Y., and Qi, H. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5810–5818, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Will be released later
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Provided in Appendix
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Details of the toy experiment.

1. Data domains. In this experiment, there are 5 variables: a latent binary variable Z, a binary variable X^*, an MNIST image $\mathbf{X}$, and two other binary variables y, t, that is:

$$
\begin{aligned}
&Z_i, X_i^* \in \{0,1\} \quad \text{: binary values} \\
&\mathbf{X}_i \in \mathbb{R}^{28*28} \quad \text{: an MNIST image of 0 or 1} \\
&y_i, t_i \in \{0,1\} \quad \text{: binary values} \\
&i \in \{1, 2, \cdots, N\}, N = 3000 \quad \text{: datapoint index}
\end{aligned}
$$

2. Data distribution.
The previously mentioned variables are sampled from the following distribution to form our synthetic dateset:

$$
\begin{aligned}
&P[Z=0] = P[Z=1] = 0.5 \\
&P[t=0|Z=0] = P[t=1|Z=1] = \rho_t = 0.2 \\
&P[X^*=0|Z=0] = P[X^*=1|Z=1] = \rho_x = 0.1 \\
&\mathbf{X} = \begin{cases} \text{a binarized image of 0 when } X^* = 0 \\ \text{a binarized image of 1 when } X^* = 1 \end{cases} \\
&y = t \oplus Z
\end{aligned}
$$

3. ATE computation.
As mentioned before, the definition of the Average Treatment Effect (ATE) is as follows:

$$
\begin{aligned}
&ATE = \mathbb{E}[ITE(x)] \\
= \ &\mathbb{E}\Big[\mathbb{E}[y \mid \mathbf{X} = x, do(t=1)] - \mathbb{E}[y \mid \mathbf{X} = x, do(t=0)]\Big]
\end{aligned}
$$

Now we consider the first term within the outer expectation:

$$
\begin{aligned}
&\mathbb{E}[\ y \mid \mathbf{X} = x,\ do(t=1)] \\
= \ &1 * P[y=1|\mathbf{X}=x, do(t=1)] + 0 \\
= \ &\sum_z P[y=1|\mathbf{X}=x, do(t=1), Z=z] \cdot \\
&P[Z=z|\mathbf{X}=x, do(t=1)] \\
= \ &\sum_z P[y=1|\mathbf{X}=x, t=1, Z=z] \cdot P[Z=z|\mathbf{X}=x] \\
= \ &P[y=1|\mathbf{X}=x, t=1, Z=0] \cdot P[Z=0|\mathbf{X}=x] + 0 \\
= \ &P[Z=0|\mathbf{X}=x]
\end{aligned}
$$

where the second last equation is due to the fact that $y = t \oplus Z$. Similarly, we can compute the second term as:

$$
\mathbb{E}[\ y \mid \mathbf{X} = x, do(t=0)] = P[Z=1|\mathbf{X}=x]
$$

meanwhile, according to the generation process of our data, we know that

$$
\begin{aligned}
P[Z=z|\mathbf{X}=x] = \ &\frac{P[Z=z] \cdot P[\mathbf{X}=x|Z=z]}{P[\mathbf{X}=x]} \\
= \ &\frac{P[Z=z] \cdot P[\mathbf{X}=x|Z=z]}{\sum_z P[Z=z] \cdot P[\mathbf{X}=x|Z=z]} \\
= \ &\frac{0.5 \cdot P[\mathbf{X}=x|Z=z]}{0.5 \cdot [\rho_x + (1-\rho_x)]} \\
= \ &P[\mathbf{X}=x|Z=z] \\
= \ &P[X^*=x^*|Z=z]
\end{aligned}
$$

where x is an image of x^*. Plugging int he previous results, we can compute the individual treatment effect (ITE):

$$\begin{aligned}
& ITE(x) \\
= \; & \mathbb{E}[y|\mathbf{X} = x, do(t = 1)] \\
& - \mathbb{E}[y|\mathbf{X} = x, do(t = 0)] \\
= \; & P[Z = 0|\mathbf{X} = x] - P[Z = 1|\mathbf{X} = x] \\
= \; & P[\mathbf{X} = x|Z = 0] - P[\mathbf{X} = x|Z = 1] \\
= \; & P[X^* = x^*|Z = 1] - P[X^* = x^*|Z = 0] \\
= \; & \begin{cases} 2\rho_x - 1, \text{ if } x = \text{image of } 0 \\ 1 - 2\rho_x, \text{ if } x = \text{image of } 1 \end{cases}
\end{aligned}$$

Notice that,

$$\begin{aligned}
P[X^* = x^*] = \; & \sum_z P[Z = z] \cdot P[X^* = x^*|Z = z] \\
= \; & 0.5 \cdot [\rho_x + (1 - \rho_x)] \\
= \; & 0.5
\end{aligned}$$

therefore we can plug in the previous equation and get the final result of ATE:

$$\begin{aligned}
ATE = \; & \mathbb{E}[ITE(x)] \\
= \; & \sum_x P[\mathbf{X} = x] \cdot ITE(x) \\
= \; & \sum_{x^*} P[X^* = x^*] \cdot ITE(\text{image of } x^*) \\
= \; & 0
\end{aligned}$$

B Neural Architecture of Deep Structural Equations and Approximate Inference Networks

Causality in Machine Learning Pearl [34] describes structural equations to mathematically represent causal relations between variables, extending the idea from linear models to non-parametric models and graphs. Xia et al. [54] provide theoretical analysis of inductive biases needed to learn causal effects using neural models. Neural networks have been used to learn counterfactual model and propensity score model under no unobserved confounding setup [16, 55, 45].

Our Architecture In this section, we will introduce the details of the DGSE and DMSE architecture we used. $\mathbf{X}_i$ denotes an input datapoint, i.e. the feature vector (possibly containing multiple modalities), t_i is the treatment assignment, y_i denotes the corresponding outcome and $\mathbf{Z}_i$ is the latent hidden confounder. Within DGSE and DMSE, the latent variable is modelled to be Gaussian distributed. For DGSE, we can model each of the corresponding factors (similar to Louizos et al. [26]) as:

$$\begin{aligned}
p[\mathbf{Z}_i] &= \prod_{j=1}^{D_z} \mathcal{N}(Z_{ij} \mid 0, 1) \\
p[t_i \mid \mathbf{Z}_i] &= \mathsf{Bern}(\sigma(\mathrm{NN}_1(\mathbf{Z}_i))) \\
p[\mathbf{X}_i \mid \mathbf{Z}_i] &= \prod_{j=1}^{D_x} p[X_{ij} \mid \mathbf{Z}_i]
\end{aligned}$$

where $\sigma(\cdot)$ is the sigmoid function, Bern is the Bernoulli distribution, D_x, D_z are the dimensions of $\mathbf{X}$ and $\mathbf{Z}$ respectively, $p[X_{ij} \mid \mathbf{Z}_i]$ is an appropriate probability distribution for the covariate j. If the treatment variable is not binary, we can modify the distribution appropriately to work with that. Within DMSE, it is possible to further factorize the distribution $p(X_i|Z_i)$ into product of distributions over component modalities owing to the conditional independence.

If the outcome y is discrete, we parameterize its probability distribution as a Bernoulli distribution:

$$\begin{aligned} p[y_i \mid t_i, \mathbf{Z}_i] = & \ \mathsf{Bern}\left(\pi = \hat{\pi}_i\right) \\ \hat{\pi}_i = & \ \sigma\left(\mathrm{NN}_2(\mathbf{Z}_i, t_i)\right) \end{aligned}$$

and if it is continuous, we parameterize its distribution as a Gaussian with a fixed variance $\hat{v}$, defined as:

$$\begin{aligned} p[y_i \mid t_i, \mathbf{Z}_i] = & \ \mathcal{N}\left(\mu = \hat{\mu}_i, \sigma^2 = \hat{v}\right) \\ \hat{\mu}_i = & \ \mathrm{NN}_2(\mathbf{Z}_i, t_i) \end{aligned}$$

Here each of the $\mathrm{NN}_i(\cdot)$ is a neural network.

The posterior distribution for DGSE is approximated as

$$q[\mathbf{Z}_i \mid \mathbf{X_i}, t_i, y_i] = \prod_{j=1}^{D_z} q[Z_{ij} \mid \mathbf{X_i}, t_i, y_i] = \prod_{j=1}^{D_z} \mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$$

where

$$\mu_{ij}, \sigma_{ij}^2 = \mathrm{NN}_4(\mathbf{X}_i, y_i, t_i).$$

For DMSE, the posterior distribution is computed differently using Product-of-Experts (PoE) [52] formulation, due to which it can handle missing modalities during training and inference gracefully.

The objective of DGSE model is the variational lower bound defined as:

$$\begin{aligned} \mathcal{L} = \sum_{i=1}^{N} \mathbb{E}_{q[\mathbf{Z}_i \mid \mathbf{X}_i, t_i, y_i]} \Big[& \log p[\mathbf{Z}_i] + \log p[\mathbf{X}_i, t_i \mid \mathbf{Z}_i] \\ & + \log p[y_i \mid \mathbf{Z}_i, t_i] - \log q[\mathbf{Z}_i \mid \mathbf{X}_i, t_i, y_i] \Big] \end{aligned}$$

The DMSE, on the other hand, requires a sub-sampled training objective to ensure that the modality specific posterior networks are trained and the relationships between individual modalities is captured. For DGSE, we also define the auxiliary encoders and the extra term in the variational lower bound following Louizos et al. [26].

Auxiliary Encoder:

$$q[t_i \mid \mathbf{X}_i] = \ \mathsf{Bern}(\pi = \sigma(\mathrm{NN}_5(t_i)))$$

For discrete y_i:

$$\begin{aligned} q[y_i \mid t_i, \mathbf{X}_i] = & \ \mathsf{Bern}\left(\pi = \hat{\pi}_i\right) \\ \hat{\pi}_i = & \ \sigma(\mathrm{NN}_6(\mathbf{X}_i, t_i)) \end{aligned}$$

For continuous y_i:

$$\begin{aligned} p[y_i \mid t_i, \mathbf{X}_i] = & \ \mathcal{N}\left(\mu = \bar{\mu}_i, \sigma^2 = \hat{v}\right) \\ \hat{\mu}_i = & \ \mathrm{NN}_6(\mathbf{X}_i, t_i). \end{aligned}$$

This introduces the following extra term in the Variational lower bound:

$$\mathcal{L}' = \sum_{i=1}^{N} \log q[t_i \mid \mathbf{X}_i] + \log q[y_i \mid \mathbf{X}_i, t_i] \Big]$$

DMSE does not involve these extra terms within its ELBO objective.

Compared with Louizos et al. [26], DGSE can be extended to be applied over different types of architecture for the posterior distribution $q[\mathbf{Z}_i \mid \mathbf{X_i}, t_i, y_i]$. When $\mathbf{X}$ is an image (e.g. medical scans, patient photos), we can use a convolutional neural network, which has been shown to effective in extracting information in image files [21]. In our experiments with image modality, we used pretrained CLIP embeddings [22] in the first layer to extract relevant features from the images. To avoid the overwhelming difference between the image and two binary variables t, y, we also apply dimension reduction techniques such as Principle Component Analysis to the embeddings of the image before feeding it into the network that is shared with t, y. When $\mathbf{X}$ is time-series data, (e.g. text, recording), we can change the architecture to recurrent neural networks such as Long Short Term Memory [14]. More generally, we can choose modality specific architectures and make appropriate design choices to perform learning and inference over unstructured modalities as inputs.

C Setups used for IHDP and STAR Dataset experiments

C.1 IHDP Experiments

The data corresponding to non-white mothers in the treated set of children is removed so that causal effect of the intervention cannot be estimated directly. The column corresponding to mother's race is removed so that this confounder cannot be obtained directly from the input. We consider 100 replicates of this dataset, where the output is simulated according to setting 'A' of NPCI package [9]. The true treatment effect is known as the simulation provides expected output values for both values of binary treatment variable. We train a DGSE model on each replicate with a 63/27/10 ratio of training, validation and test dataset size. We set the latent dimension to be 20 units and the number of hidden layers to be 2. The hidden layers have size of 20 units.

The IHDP-Full Setting There are 25 input features in this experimental setting. We report the absolute error in ATE produced by DGSE and OLS for this setting in Table 7.

The IHDP-Mini Setting Here, we choose 9 features from the 25 input features so that removal of the feature 'baby's gender' produces statistically significant treatment effect. We used mutual information and F-statistics between each of the original 25 features and the target variable y to assess the importance of each feature in the initial 100 replicates of IHDP. While making sure that the absolute ATE errors don't deviate too much from the corresponding errors produced by IHDP-Full setting, we experimented with several combinations of the high ranking features to select the following 9 features in the IHDP-Mini setting.

1. Feature 6: 'sex of baby'
2. Feature 0: 'birth-weight'
3. Feature 1: 'b.head'
4. Feature 2: 'preterm'
5. Feature 3: 'birth.o'
6. Feature 8: 'mom married?'
7. Feature 9: 'mom's education lower than high school?'
8. Feature 12: 'Smoked cig during pregnancy?'
9. Feature 20: 'harlem'

Table 7 shows a comparison of absolute ATE errors between the IHDP-Full setting and IHDP-Mini setting. Table 2 shows the comparison of our approach with CFRNet [16] and Ordinary Least Squares (OLS) approach. OLS takes the concatenation of the covariates and treatment variable value as input to produce output. We see that the CFRNet baseline does not utilize the image effectively while OLS shows small difference between the setting where baby's gender was removed (8 attrs) and the setting where baby's gender was retained (9 attrs) setting. Hence the replacement of image (8 attrs+image) produces a small average improvement as compared to the setting where baby's gender was dropped (8 attrs). We also note that we have dropped 6 replicates from the 100 IHDP replicates under consideration. These 6 replicates showed a large degradation in ATE estimates by adding baby's gender (9 attrs) as compared to removing it (8 attrs).

C.2 STAR Experiments

We 'derandomize' this dataset by removing 80% of the data corresponding to white students in the treated population. The dataset [1] has 15 input attributes. The true treatment effect can be estimated directly since the original dataset corresponds to a randomized controlled trial. Hence, it is possible to compute ATE error as the absolute difference between true ATE estimate and the ATE as predicted by the model. Similar to IHDP, we choose the following subset of attributes for the STAR experiment

1. Feature 2: 'Student grade'
2. Feature 3: 'Student class-type'
3. Feature 4: 'Highest degree obtained by teacher'

4. Feature 5: 'Career ladder position of teacher'
5. Feature 6: 'Number of years of experience of teacher'
6. Feature 7: 'Teacher's race'
7. Feature 10: 'Student's gender'
8. Feature 11: 'Student's ethnicity'

Table 6: Multimodal Experiments on STAR Dataset: Removing student gender and ethnicity (6 attrs) shows increased ATE errors when compared with retaining these attributes (8 attrs), signaling that these two attributes are important for predicting treatment effect. Replacing these with image of a child shows no degradation in ATE estimation.

Setting	ε_{ATE} (Train+Val)	ε_{ATE} error(Test)
Deep Str Eqns		
8 attrs	36.479 (1.770)	34.039 (2.336)
6 attrs	43.682 (1.520)	40.651 (2.425)
6 attrs + image	35.581 (1.723)	33.654 (2.476)
CFRNet		
8 attrs	61.835 (1.025)	25.436 (2.332)
6 attrs	62.055 (1.001)	25.649 (2.339)
6 attrs + image	61.350 (1.109)	25.219 (2.313)

In Table 6, we repeat the experiment 100 times and report average ATE errors along with standard error. We removed 8 repetitions in the experiment where DGSE or CFRNet showed lack of convergence as evidenced by very high validation loss on any setting of input attributes.

Table 7: Treatment effects on IHDP Dataset. Using a reduced set of features in the IHDP Mini setting produces comparable absolute ATE errors as the degradation is small. Numbers in round braces indicate standard deviations. Since this is a simulated dataset, we can directly compute the treatment effect using the simulated factual and counterfactual outputs. ATE error is the absolute difference between true ATE and predicted ATE.

MODELS	INPUT	ε_{ATE} (TRAIN+VAL)	ε_{ATE} ERROR(TEST)
DGSE	FULL	0.289 (0.027)	0.358 (0.041)
OLS	FULL	0.535 (0.089)	0.718 (0.132)
DGSE	MINI	0.404 (0.107)	0.720 (0.159)

D Evaluating Quality of Pre-Trained Embeddings

We demonstrate that our pre-trained embeddings contain useful signal by building a neural network model that predicts the gender, age and ethnicity from the CLIP embedding of the corresponding image.

We build a simple neural network model that takes the CLIP embedding of an image as input and predicts the age of the person in that image. We use 5-dimensional PCA embeddings of 500 randomly chosen images of people aged 10-45 years (corresponding to the age-group in the IHDP experiment). We have an independent test dataset corresponding to 100 images chosen in a similar way. We see that the R^2 value for the age prediction on test dataset is 0.45. If we increase the size of PCA embedding to 50, this R^2 value increases to 0.58. Thus, it is possible to extract the age information from randomly chosen images using a simple neural network.

In the above setting, we also studied the classification accuracy of separate neural networks that predict gender and ethnicity from the CLIP embeddings. We saw that gender was predicted with 94% accuracy and ethnicity was predicted with 58% accuracy using a 5-dimensional PCA embedding.

After increasing the size of PCA embedding to 50 dimensions, the gender prediction accuracy increased to 95% and ethnicity prediction accuracy increased to 77%. This further supports our idea of replacing the attribute corresponding to 's baby's gender or student ethnicity/gender with an appropriate image.

E Plant GWAS

Setup We apply our deep structural equations framework for correcting the effects of confounding. We fit a DGSE model in two stages: (1) first, we only fit the model of $p(T|Z)$ using the DGSE ELBO objective; (2) then we fit $p(Y|Z,T)$ with a fixed Z produced by the auxiliary model $q(Z|T)$. We found this two-step procedure to produce best results. The subsampled SNPs corresponding to each genome are taken as input X. The encoder and decoder use a single hidden layer of 256 units while a 10-dimensional latent variable Z is used. This network is optimized using ClippedAdam with learning rate of 0.01, further reduced exponentially over 20 training epochs. The confounding variable for each genome can now be computed as latent representation produced by the VAE. To measure the success of confounding correction, we compute the R^2 values between the true GDP of the region and the GDP output as predicted via our model and the baselines. If we have corrected for confounding, then we should get low R^2 values.

Historical Weather Data We used historical weather data collected from Menne et al. [29]. We use per day precipitation data from year 2000 collected by weather station closest in distance to the latitude/longitude coordinates of the location from which the SNPs of plant were collected. For the locations where weather data was missing, we replaced those entries with zeros.

F Simulated GWAS Experiments

We demonstrate that representations trained via standard deep latent variable models (i.e. variational autoencoders [20]) are good representations of confounding variables in a simulated GWAS setting, with a time-series proxy.

F.1 Data generating process

To simulate the confounders, SNPs (genotypes), and the outcomes (phenotypes), we follow the "Spatial" simulation setup from Appendix D.1 & D.2 of [46]. Namely, we randomly generate a low-rank factorization of the allele frequency logits $F = \sigma^{-1}(\Gamma S)$, where σ is sigmoid, as is common in the literature [4, 38] – here, and in "Spatial" of [46], the S matrix is interpreted as geographic spatial positions of the individuals. For the m-th SNP of the n-th individual, we generate the SNP $X_{nm} \sim \text{Bin}(3, \sigma(F_{nm}))$. In this simulation, we considered $M = 10$ SNPs with $N = 10000$ individuals.

Individuals are clustered into $K = 3$ groups based on their locations, and the individual's cluster is the *unobserved* confounder. Then, the outcomes are calculated from *both* the SNPs, where only $c = 2$ SNPs have a non-zero causal effect, plus a confounding term that is a function of the cluster, and some i.i.d. Gaussian noise. One small deviation from [46] is that our noise is i.i.d. Gaussian, rather than having the variance depend on the confounder as well.

We further augmented the dataset to include a time-series proxy that can help identify the unobserved spatial position of individuals. For each cluster, we came up with some Fourier coefficients,

$$\text{Cluster 0: } a_0^{(0)} = 0.0, a_1^{(0)} = -1.0, a_2^{(0)} = 1.0, b_1^{(0)} = -1.0, b_2^{(0)} = 1.0$$
$$\text{Cluster 1: } a_0^{(1)} = 1.0, a_1^{(1)} = -5.0, a_2^{(1)} = 2.0, b_1^{(1)} = -5.0, b_2^{(1)} = 2.0$$
$$\text{Cluster 2: } a_0^{(2)} = -1.0, a_1^{(2)} = -2.0, a_2^{(2)} = 5.0, b_1^{(2)} = -2.0, b_2^{(2)} = 5.0.$$

For each individual in the k-th cluster, we sampled at $N_{samples} = 50$ uniformly spaced times, across $N_{periods} = 2$ periods of length $T = 5$. That is, our time series proxy consists of points

$\{x_i\}_{i \in [N_{samples}]}$,

$$x_i = \frac{1}{2} a_0^{(k)} + \sum_{\ell=1}^{2} a_\ell^{(k)} \cos\left(\frac{2\pi t_i(\ell - 1)}{T}\right) + \sum_{\ell=1}^{2} b_\ell^{(k)} \sin\left(\frac{2\pi t_i(\ell - 1)}{T}\right),$$

$$t_i = \frac{T N_{periods} i}{N_{samples}}.$$

Please see Fig. 3 for a visualization of the time-series generated for each cluster.

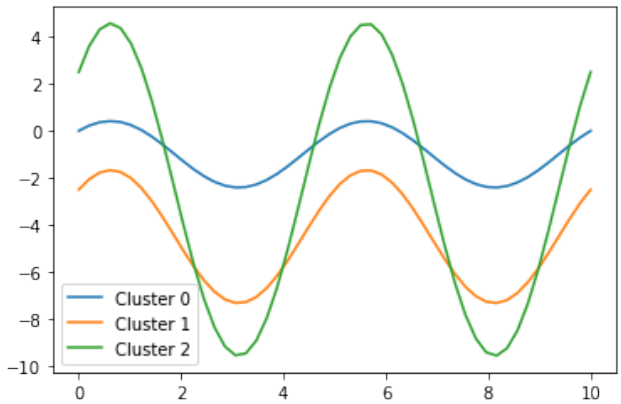

Figure 3: Visualization of time-series proxy.

F.2 VAE Modeling and evaluation setup

To handle the SNP data, we trained an MLP-based β-variational autoencoder [12] with ReLU activations, which output parameters of a Gaussian latent distribution. The MLPs consisted of 3 hidden layers, each with 64 units, and our latent dimension z was 2. To handle time series data (data of another modality), we trained a β-VAE that instead encoded and decoded with 1D convolutions, as 1D convolutions have been shown to strongly approximate fast Fourier transforms [31]. There were 3 hidden convolutional layers, each of which had 32 output channels, with a kernel of length 3 and a stride of 1. Following the convolution was an MLP with 1 hidden layer with 64 units. ReLU activations were used throughout the architecture.

In our experiments, we swept over $\beta \in \{0.1, 0.2, 0.5, 1.0, 1.2, 1.5\}$. Our general hypothesis was that lower β values would make our VAE perform better, as unlike in the standard VAE evaluation scheme, we sample from the posterior to generate latents for individuals, and not the prior, so the KL divergence term in the ELBO doesn't matter as much.

When evaluating on a set of SNPs and time series data corresponding to a set of individuals, we generate latents by passing the data of each respective type into the appropriate encoder and sampling from the resulting distributions. The latent z for each individual is the concatenation of the SNP latent z_{snp} and the time series latent z_{ts}.

Once we have the latent confounder z for an individual, we calculate the causal effect of the mth SNP by performing the linear regression,

$$y = \beta_0 + \sum_i \beta_i z_i + \gamma_m x_m + \epsilon_i,$$

where ϵ_i are i.i.d. standard normal. We solved this using the closed form solution for least-squares linear regression, given by

$$\widehat{\gamma} = (A^T A)^{-1} A^T y,$$

where $A_i = [z_i, 1, x_m]$ is the ith row in the matrix A. Then $\widehat{\gamma}_m$ is our estimated causal effect for the mth SNP. We let $\widehat{\gamma}$ be a vector of length M to denote our estimates for each SNP's causal effect.

Comparison to CEVAE [26] [26] proposed the causal effect VAE (CEVAE) to estimate causal effect with unobserved confounders, but their approach is specialized to the *uni-modal, binary treatment* setting. The CEVAE uses two separate encoders and decoders to infer the outcomes of y given both values of the treatment t. It does this by first learning the treatment distribution $p_\theta(t \mid x)$, followed by the outcome distribution $p_\theta(y \mid t, x)$ for $t \in \{0, 1\}$ and then the latent distribution $p(z \mid x, t, y)$ given a sampled y from the appropriate treatment distribution. It is possible to extend CEVAE to K different treatments, but this would require training K different encoders and decoders to learn $p_\theta(y \mid t = k, x), \forall k \in [K]$, etc. This linear scaling is not ideal, especially for GWAS, where for a SNPs vector of length M, there are 3^M possible SNP combinations; if we assume that each SNP is independent, each SNP can be interpreted as a ternary treatment, but CEVAE would nonetheless require fitting $\mathcal{O}(M)$ encoders, which is prohibitively expensive since M is typically large.

For the unimodal setting, we propose to simply encode the SNPs vector using a vanilla VAE, and term our approach "SNP VAE". The main benefit of a SNP VAE over a causal effect VAE, then, is that our VAE is scalable to treatments over a variety of different modalities and sizes. Our proposed method does not encode and decode the treatment t for every individual x—it merely encodes from x, assuming we have all the necessary information about the individual to infer possible confounders for them.

For the multi-modal setting, we handle the data of different modalities (image, text, time series) using appropriate deep learning methods (for images, we will use convolutional neural networks, for text or time series data, 1D convolutions or recurrent neural networks/transformers). We note that while the number of SNPs can be very large, the number of different data modalities seen in GWAS is generally much smaller, allowing our model to scale appropriately without a very expensive computational burden.

F.3 Results

We tested our VAE model and multi-modal VAE model against several linear latent model baselines. Two baselines, Principle Components Analysis (PCA) and Factor Analysis (FA), also generate a latent z, which we then use our linear regression approach to calculate $\widehat{\boldsymbol{\gamma}}$. We also ran the LMM implementation by Limix. We also plot two oracle baselines. The first is when $\widehat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^\star$, labeled "truth". The second is when z is the true confounder used when generating the data, according to "Spatial".

Let $\boldsymbol{\gamma}^\star$ denote the ground truth causal effect vector. We report $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^\star\|$ for ℓ_1 and ℓ_2 norms. We also report true/false positive/negatives, which we define as follows. For an individual, let $\tau = \min_{i:\gamma_i \neq 0} |\gamma_i|/2$. Then, SNP m is a

1. True Positive (tp): if $\widehat{\gamma}_m \geq \tau$ and $\gamma^\star_m \geq \tau$ or $\widehat{\gamma}_m \leq -\tau$ and $\gamma^\star_m \leq -\tau$.
2. True Negative (tn): if $|\widehat{\gamma}_m| < \tau$ and $|\gamma^\star_m| < \tau$.
3. False Negative (fn): if $|\widehat{\gamma}_m| < \tau$ but $|\gamma^\star_m| > \tau$.
4. False Positive (fp): if none of the above hold. Concretely, there are two cases. First, if $|\widehat{\gamma}_m| > \tau$ but $|\gamma^\star_m| < \tau$. Second, if the direction is wrong, i.e. $\widehat{\gamma}_m > \tau$ but $\gamma^\star_m < -\tau$, or $\widehat{\gamma}_m < -\tau$ but $\gamma^\star_m > \tau$.

Finally, recall that precision is $tp/(tp + fp)$ and recall is $tp/(tp + fn)$, and higher is better for both. In the following table, we show the mean and standard error of the mean (sem) over 10 seeds at evaluation time. Here, we use a VAE to instantiate DSEs. The unimodal VAE only encodes the SNPs vector into a 2 dimensional embedding. The multimodal VAE also encodes the time-series vector into a 2 dimensional embedding, and we concatenate both 2-dimensional embeddings into a 4 dimensional embedding to use in our linear regression. PCA_k denotes PCA with k components, and similarly for FA. Amongst the non-oracle baselines, multi-modal VAE has the smallest errors and highest precision and recall, and importantly beats all linear baselines including LMM. Unimodal VAE, while worse than LMM, still beats PCA and FA. In general, we see that precision starts to deteriorate faster than recall, suggesting that false positives are more likely from the weaker linear deconfounding techniques such as PCA/FA. It's also interesting that when PCA/FA used more components, we see worse performance across the board. This illustrates the importance of picking the right latent dimension, which is an open question. If the latent dimension is too small then we may not capture all

the confounders. If the latent dimension is too large, the latent encoding may also end up capturing spurious correlations.

Model	$\ell_1(\downarrow)$ mean (sem)	$\ell_2(\downarrow)$ mean (sem)	tp ($\uparrow$) mean (sem)	fp ($\downarrow$) mean (sem)	tn ($\uparrow$) mean (sem)	fn ($\downarrow$) mean (sem)
Optimal	0.22 (0.04)	0.09 (0.02)	2.0 (0.00)	0.1 (0.10)	7.9 (0.10)	0.0 (0.00)
DSE (2 modalities)	0.30 (0.06)	0.14 (0.03)	2.0 (0.00)	0.2 (0.13)	7.8 (0.13)	0.0 (0.00)
LMM	0.44 (0.06)	0.17 (0.02)	2.0 (0.00)	0.7 (0.42)	7.3 (0.42)	0.0 (0.00)
DSE (1 modality)	0.60 (0.09)	0.26 (0.05)	2.0 (0.00)	1.1 (0.59)	6.9 (0.59)	0.0 (0.00)
PCA (1 component)	0.93 (0.17)	0.40 (0.07)	2.0 (0.00)	2.3 (0.72)	5.7 (0.72)	0.0 (0.00)
FA (1 component)	1.08 (0.17)	0.59 (0.13)	2.0 (0.00)	2.0 (0.71)	6.0 (0.71)	0.0 (0.00)
PCA (2 components)	1.38 (0.24)	0.56 (0.09)	1.8 (0.13)	3.4 (0.83)	4.6 (0.83)	0.2 (0.13)
FA (2 components)	1.44 (0.30)	0.71 (0.19)	2.0 (0.00)	2.6 (0.75)	5.4 (0.75)	0.0 (0.00)
PCA (3 components)	1.66 (0.23)	0.68 (0.09)	1.6 (0.16)	3.8 (0.80)	4.2 (0.80)	0.4 (0.16)
FA (3 components)	1.89 (0.45)	0.95 (0.25)	1.8 (0.13)	3.0 (0.56)	5.0 (0.56)	0.2 (0.13)

Model	precision ($\uparrow$) mean (sem)	recall ($\uparrow$) mean (sem)
Optimal	0.97 (0.03)	1.0 (0.00)
DSE (2 modalities)	0.93 (0.04)	1.0 (0.00)
LMM	0.85 (0.08)	1.0 (0.00)
DSE (1 modality)	0.78 (0.08)	1.0 (0.00)
PCA (1 component)	0.58 (0.09)	1.0 (0.00)
FA (1 component)	0.62 (0.08)	1.0 (0.00)
PCA (2 components)	0.44 (0.09)	0.9 (0.07)
FA (2 components)	0.55 (0.09)	1.0 (0.00)
PCA (3 components)	0.37 (0.08)	0.8 (0.08)
FA (3 components)	0.44 (0.08)	0.9 (0.07)

G Alternative Causal Graph Structures

This section derives estimators of ATE and ITE under assumptions on the causal graph that are different from the ones described in Figure 1. The key takeaway of this section is that ATE and ITE can be computed under arbitrary causal graphs using minor modifications of our algorithms, which we describe in this section.

G.1 Alternative Causal Graph Structures

We start by formally defining the class of causal graphs that will be studied in this section. Formally, we identify two sets of modification to Figure 1. We will derive small modifications to our algorithms to allow the new types of causal graphs implied by these modifications.

Observed confounders We consider an expanded set of graphs over a space of variables (X, Y, T, Z, V), where V represents observed (non-proxy) confounders. We look at causal graphs implied by structural equations of the form:

$$Z \sim \mathcal{P}_Z \quad V \sim \mathcal{P}_V \quad X_j \sim \mathcal{P}_{X_j}(\theta_{X_j}(Z))) \;\; \forall j \quad T \sim \mathrm{Ber}(\pi_T(Z, V)) \quad Y \sim \mathcal{P}_Y(\theta_Y(Z, V, T))), \tag{8}$$

where again we note that V is variable that is assumed to be **always observed** (just like $y^{(i)}$ and $t^{(i)}$). All the other technical terms are defined as in the main body of the paper. The above equations result in a causal graph with edges between V and Y, T and define a distribution $p(x, y, t, z, v)$.

Dependent proxies Another possible set of modifications to Figure 1 is the presence of edges between proxies X_i, X_j, which can be denoted as

$$X_j \sim \mathcal{P}_{X_j}(\theta_{X_j}(Z, \mathrm{pa}(X_j)))) \;\; \forall j$$

where $\mathrm{pa}(X_j))$ denotes the set of parents of X_j among the other unstructured proxy variables X_i.

G.2 A General Estimator Class

The following Theorem shows that we can estimate ATE and ITE when the data distribution follows the structure in Figure 1, plus the two types of modifications outlined above (observed confounders and dependent proxies).

Theorem 2. *The true ITE$(x, \mathcal{M})$ for any subset $\mathcal{M} \subseteq \{1, 2, ..., m\}$ of observed modalities is identifiable when the true data distribution $p(x, y, t, z, v)$ has the causal graph structure of a DMSE model (Figure 1) in addition to having observed confounders v and possibly dependent confounders.*

Proof: Let $x_{\mathcal{M}} = \{x_j \mid j \in \mathcal{M}\}$ be the data from the observed subset of modalities. Let v be the observed proxy variable. We need to show that $p(y|x_{\mathcal{M}}, v, \text{do}(t = t'))$ is identifiable for any t'. Observe that

$$\begin{aligned} p(y|x_{\mathcal{M}}, v, \text{do}(t = t')) &= \int_z p(y|z, x_{\mathcal{M}}, v, do(t = t'))p(z|x_{\mathcal{M}}, v, do(t = t'))dz \\ &= \int_z p(y|z, x_{\mathcal{M}}, v, t')p(z|x_{\mathcal{M}})dz, \end{aligned}$$

where the second equality follows from the rule of do-calculus (applying backdoor adjustment). Since our proof holds for any t' and all elements on the right-hand side are identifiable, the claim follows. ■

G.3 Observed Confounders

Next, we derive an extension of the DMSE model to the setting in which we have observed confounders V. We refer to this modified model as **DMSE-V**.

As earlier, the **DMSE-V** model induces a tractable joint density $p(X, Y, T, Z, V)$, which allows us to fit its parameters using stochastic variational inference by optimizing the evidence lower bound (ELBO) on the marginal log-likelihood $p(y^{(i)}, x^{(i)}, t^{(i)}, v^{(i)})$ defined over an expanded dataset $\{y^{(i)}, x^{(i)}, t^{(i)}, v^{(i)}\}_{i=1}^n$:

$$\text{ELBO}_X = \sum_{i=1}^{n} \mathbb{E}_q \big[\sum_{j=1}^{m} \log p(x_j^{(i)}|z) + \log p(y^{(i)}, t^{(i)}, v^{(i)}, z) - \log q(z|x^{(i)}, y^{(i)}, t^{(i)}, v^{(i)}) \big], \quad (9)$$

where $p(y^{(i)}, t^{(i)}, v^{(i)}, z) = p(y^{(i)}|t^{(i)}, v^{(i)}, z)p(t^{(i)}|z, v^{(i)})p(z)p(v^{(i)})$ and $q(z|x^{(i)}, y^{(i)}, t^{(i)}, v^{(i)})$ is the approximate variational posterior. Note that since $v^{(i)}$ *is always observed* (just like $y^{(i)}$ and $t^{(i)}$), the $p(v^{(i)})$ term can be ignored.

In practice, this reduces to the vanilla DMSE model with the following modifications:

- The $\log p(y^{(i)}|t^{(i)}, v^{(i)}, z)$ term becomes additionally conditioned on $v^{(i)}$.
- The $\log p(t^{(i)}|v^{(i)}, z)$ term becomes additionally conditioned on $v^{(i)}$.
- The approximate posterior $q(z|x^{(i)}, y^{(i)}, t^{(i)}, v^{(i)})$ becomes additionally conditioned on $v^{(i)}$.

Crucially, the specialized inference algorithms derived for the DMSE model remain unchanged. Since V is always observed, learning a **DMSE-V** model is equivalent to learning a model $p(y, x, t, z|v)$, which has the same structure as a DMSE model. In particular, all the X_i are conditionally independent given Z. Hence, the same learning and inference algorithms apply.

G.4 Causal Links Among Proxies

Another type of causal graph that we consider is one in which proxies X_i are connect by causal edges. First, we note that when proxies are unstructured, such causal edges are expected to be rare, i.e., we do not expect the pixels of an image X_i to have a direct influence on other variables.

When X_i takes on a structured form and directly influences other proxies, our strategy is to "collapse" any sets of variables X_j that have edges among them, until we have the conditionally independent structure in Figure 1. In the extreme case, we might need to collapse all proxies X_i into a single proxy X that consists of their concatenation.

The result is a model that has the same structure as DMSE, and that can be learned using the same set of inference and learning algorithms. The only drawback is increased computational efficiency.

G.5 Additional Synthetic Data Experiment Details

We generate synthetic data to simulate a process with modified causal links as covered in subsection G.3 and G.4. Specifically, we simulate four different datasets with generative process detailed in the following section. We generate $\{X_i, t_i, y_i\}_{i=0}^{i=m}$ where $m = 10000$. We use train/val/test split of 63/27/10. Here, $\oplus$ stands for logical XOR.

1. **Case of Original Causal Graph (Dataset A)**
   Variables: $X1$ is unstructured input, $X2$ is structured input, T is treatment, Y is output, Z is confounder
   Edges in graph: $Z \to X1, Z \to X2, \{Z\} \to T, \{T, Z\} \to Y$
   Thus we added no extra edges.
   Generative process:
   (a) $P(z = 1) = 0.5$
   (b) $P(x1' = 1|z = 1) = P(x1' = 0|z = 0) = 0.1$ ($x1'$ is an intermediate variable)
   (c) $P(x2 = 1|z = 1) = P(x2 = 0|z = 0) = 0.2$
   (d) $P(t = 1|z = 1) = P(t = 0|z = 0) = 0.2$
   (e) $y = (z \oplus t)$
   (f) $P(x1|x1' = 1)$ is unif. over MNIST images of '1', $P(x1|x1' = 0)$ is unif. over MNIST images of '0'
2. **Case of the Observed Confounder (Dataset B)**
   Variables: $X1$ is unstructured input, $X2$ is structured input, T is treatment, Y is output, Z is confounder
   Edges in graph: $Z \to X1, \{X2, Z\} \to T, \{X2, T, Z\} \to Y$
   Thus we added extra edges $X2 \to T$ and $X2 \to Y$.
   Generative process:
   (a) $P(z = 1) = 0.5, P(x2 = 1) = 0.5$
   (b) $P(x1' = 1|z = 1) = P(x1' = 0|z = 0) = 0.1$ ($x1'$ is an intermediate variable)
   (c) $P(t = 1|z = 1, x2 = 1) = P(t = 0|z = 0, x2 = 1) = 0.2$;
   $P(t = 1|z = 1, x2 = 0) = P(t = 0|z = 0, x2 = 0) = 0.9$
   (d) $y = x2$ AND $(z \oplus t)$
   (e) $P(x1|x1' = 1)$ is unif. over MNIST images of '1', $P(x1|x1' = 0)$ is unif. over MNIST images of '0'
3. **Case of the Observed Confounder (Dataset C)**
   Variables: $X1$ is unstructured input, $X2$ is structured input, T is treatment, Y is output, confounder $Z = (Z1, Z2)$
   Edges in graph: $Z \to X1, Z \to X1, Z- > X2, \{X2, Z\} \to T, \{X2, T, Z\} \to Y$
   Thus we added extra edges $Z \to X2, X2 \to T$ and $X2 \to Y$.
   Generative process:
   (a) $P(z1 = 1) = 0.5, P(z2 = 1) = 0.5$
   (b) $P(x1' = 1|z1 = 1) = P(x1' = 0|z1 = 0) = 0.1$ ($x1'$ is an intermediate variable)
   (c) $P(x2 = 1|z2 = 1) = P(x2 = 0|z2 = 0) = 0.9$
   (d) $z = z_1 \oplus z_2$
   (e) $P(t = 1|z = 1, x2 = 1) = P(t = 0|z = 0, x2 = 1) = 0.9$; $P(t = 1|z = 1, x2 = 0) = P(t = 0|z = 0, x2 = 0) = 0.1$
   (f) $y = x2 \oplus t \oplus (z1 \oplus z2)$
   (g) $P(x1|x1' = 1)$ is unif. over MNIST images of '1', $P(x1|x1' = 0)$ is unif. over MNIST images of '0'

4. **Case of Causal Links Among Proxies (Dataset D)**
   Variables: $X1$ is structured input, $X2$ is structured input, $X3$ is unstructured input, T is treatment, Y is output, confounder $Z = (Z1, Z2, Z3)$
   Edges in graph:$Z \to X1, Z \to X2, Z \to X3, X1 \to X2, \{Z\} \to T, \{T, Z\} \to Y$
   Thus we added extra edge: $X1 \to X2$
   Data generation:
   (a) $P(z1 = 1) = P(z2 = 1) = P(z3 = 1) = 0.5$
   (b) $P(x1 = 1|z1 = 1) = P(x1 = 0|z1 = 0) = 0.1$
   (c) $P(x2 = 1|z2 = 1, x1 = 0) = P(x2 = 0|z2 = 0, x1 = 0) = 0.8$
   $P(x2 = 1|z2 = 1, x1 = 1) = P(x2 = 0|z2 = 0, x1 = 1) = 0.2$
   (d) $P(x3' = 1|z3 = 1) = P(x3' = 0|z3 = 0) = 0.3$ ($x3'$ is an intermediate variable)
   (e) $z = z_1 \oplus z_2 \oplus z_3$
   (f) $P(t = 1|z = 1) = P(t = 0|z = 0) = 0.2$
   (g) $y = t \oplus z$
   (h) $P(x3|x3' = 1)$ is unif. over MNIST images of '1', $P(x3|x3' = 0)$ is unif. over MNIST images of '0'
5. **Case of Increasing Number of Proxies (Dataset E)**
   Variables: $\{X_1, X_2, ..X_m\}$ are unstructured inputs, T is treatment, Y is output, confounder $Z = (Z_1, Z_2, .., Z_m)$. m is number of modalities Edges in graph: $Z_i \to X_i, Z \to T, \{T, Z\} \to Y$
   (a) $P(z_i = 1) = \frac{i}{m}$
   (b) $P(x'_i = 1|z_i = 1) = P(x'_i = 0|z_i = 0) = 1$ (x'_i are intermediate variables)
   (c) $z = \oplus_{i=1}^{m} z_i$ (xor over all z_i)
   (d) $P(t = 1|z = 1) = P(t = 0|z = 0) = 0.25$
   (e) $P(y = 1|t = 1) = \text{sigmoid}(3z + 2), P(y = 1|t = 0) = \text{sigmoid}(3z - 2)$
   (f) $P(x_i|x'_i = 1)$ is unif. over MNIST images of '1', $P(x_i|x'_i = 0)$ is unif. over MNIST images of '0'

H Additional Details on Mathematical Proofs

H.1 Evidence Lower Bound for Deep Structural Equations

We present a proof of the evidence lower bound (ELBO), which was presented in the main text in Section 4.2.

The left-hand-side of the ELBO can be written as

$$\sum_{i=1}^{n} \mathbb{E}_q \left[\sum_{j=1}^{m} \log p(x_i^j|z) + \log p(y_i|z) + \log p(y_i, t_i, z) - \log q(z|x_i, t_i, y_i) \right]. \tag{10}$$

Noting, due to the causal graph structure we assume, that

$$p(y_i, t_i, z) = \log p(y_i|t_i, z) + \log p(t_i|z) + \log p(z). \tag{11}$$

Thus, we can rewrite (10) as

$$\sum_{i=1}^{n} \mathbb{E}_q \left[\sum_{j=1}^{m} \log p(x_i^j|z) + \log p(y_i|z) + \log p(y_i|t_i, z) + \log p(t_i|z) + \log p(z) - \log q(z|x_i, t_i, y_i) \right] \tag{12}$$

Note that the first four terms in the expectation over the posterior distribution q make up a reconstruction loss of the original data $\boldsymbol{x}, y, t$ in terms of the latent variable z, while the last two terms form the negative KL divergence $-D_{KL}\left(q(z|x_i, t_i, y_i) \| p(z)\right)$ between the posterior q and the prior $p(z)$. Thus, we can write our ELBO as

$$\sum_{i=1}^{n} \mathcal{L}_{\text{reconstruction}}(x_i, y_i, t_i, z) - D_{KL}\left(q(z|x_i, t_i, y_i) \| p(z)\right), \tag{13}$$

which is exactly our variational objective. We can sum or take the average over the number of datapoints n to form our empirical objective.

H.2 Derivation of posterior

Here we show why the true posterior factorizes as

$$p(z \mid x, t, y) \propto (p(z \mid t, y) \prod_{j=1}^{m} p(z \mid x_j)) / \prod_{j=1}^{m-1} p(z)$$

.

$$\begin{aligned}
p(z \mid x, t, y) &= p(x, t, y, z)/p(x, t, y) \\
&= \left(\prod_{j=1}^{m} p(x_j \mid z) \right) p(t, y \mid z) p(z)/p(x, t, y) \qquad \text{(by cond. indep. of causal graph)} \\
&= \left(\prod_{j=1}^{m} p(z \mid x_j) p(x_j)/p(z) \right) \cdot (p(z \mid t, y) p(t, y) p(z)/p(x, t, y)) \\
&\propto \left(\prod_{j=1}^{m} p(z \mid x_j)/p(z) \right) \cdot (p(z \mid t, y) p(z)) \,. \qquad \text{(removing terms independent of } z\text{)}
\end{aligned}$$