On the Identifiability of Causal Abstractions

Xiusi Li
Mila, McGill University

Sékou-Oumar Kaba
Mila, McGill University

Siamak Ravanbakhsh
Mila, McGill University

Abstract

Causal representation learning (CRL) enhances machine learning models' robustness and generalizability by learning structural causal models associated with data-generating processes. We focus on a family of CRL methods that uses contrastive data pairs in the observable space, generated before and after a random, unknown intervention, to identify the latent causal model. (Brehmer et al., 2022) showed that this is indeed possible, given that all latent variables can be intervened on *individually*. However, this is a highly restrictive assumption in many systems. In this work, we instead assume interventions on *arbitrary subsets* of latent variables, which is more realistic. We introduce a theoretical framework that calculates the *degree* to which we can identify a causal model, given a set of possible interventions, up to an *abstraction* that describes the system at a higher level of granularity.

1 INTRODUCTION

Causal representation learning (CRL) (Schölkopf et al., 2021) generalizes non-linear independent component analysis (Hyvärinen and Pajunen, 1999; Hyvarinen et al., 2019) and causal discovery (Spirtes et al., 2001), aiming to extract both latent variables and their causal graph in the form of structural causal models (SCMs). The question of identifiability naturally arises since we would like to guarantee that the set of models consistent with the given observable distribution is unique up to some equivalence class. In particular, we would like to know the level of granularity at which the true latent variables and the causal graph can be recovered.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

We consider the problem of CRL assuming we have access to counterfactual data pairs $\mathbf{x}$ and $\tilde{\mathbf{x}}$ from the same observable space, before and after a random, unknown intervention. This is necessitated by the infeasibility of learning disentangled representations from unsupervised observational data alone (Hyvärinen and Pajunen, 1999; Locatello et al., 2019), and is sometimes referred to as self-supervised (Von Kügelgen et al., 2021), contrastive (Zimmermann et al., 2021), or weakly supervised learning (Shu et al., 2019; Locatello et al., 2020; Brehmer et al., 2022; Ahuja et al., 2022).

Previously, in this particular setting, it has been shown that the full causal graph can be recovered up to isomorphism (Brehmer et al., 2022), and all the latent variables can be recovered up to element-wise diffeomorphism. However, as was noted by the authors, this result relies on overly restrictive assumptions; for example, the number of nodes on the causal graph must be known in advance, and each node must be intervened upon individually with nonzero probability.

Instead, we show that when we remove these assumptions, we can still identify "coarser" versions of the causal model, known as its *abstractions*, the related works of which we will expand on in the following paragraph. This relaxation is significant as it makes the setting much more realistic; in many systems, intervening on every variable individually is infeasible.

Causal abstraction (Rubenstein et al., 2017; Beckers and Halpern, 2019; Rischel, 2020; Rischel and Weichwald, 2021; Otsuka and Saigo, 2022; Anand et al., 2023b; Massidda et al., 2023) is the study of how microscopic variables and causal mechanisms can be aggregated to macroscopic abstractions on a higher level while maintaining a notion of interventional consistency, which enables more efficient reasoning and interpretability (Geiger et al., 2021, 2024). Since this is still an emerging avenue of research, a standard unified formalism has yet to be established. However, as was highlighted previously in (Zennaro, 2022), all of the various definitions of causal abstractions, in one way or another, tend to have a *structural* map dealing with the causal graph, and a *distributional* map dealing

with the variables and stochastic causal mechanisms associated with its nodes and edges respectively.

In classic causal inference, the idea of structural abstraction can in some sense already be found in the form of Markov equivalence classes, or interventional Markov equivalence classes (Verma and Pearl, 1990; Hauser and Bühlmann, 2012; Yang et al., 2018), where in the simplest cases, directed edges on the causal graph are abstracted away to undirected ones.

In CRL, the concept of distributional abstraction is more prevalent. For example, (Von Kügelgen et al., 2021) introduces a definition for *block-identifiability*, in which the grouping of latent variables into blocks essentially serves as an abstraction of individual latent variables. More recent works (Ahuja et al., 2022; Yao et al., 2023) also examine overlapping blocks of latent variables and the identifiability of their intersections, complements, and unions.

In this paper, we investigate the identifiability of latent causal models up to their abstractions. Previous works in this setting have either focused on the problem of whether a pre-conceived latent causal model can be identified at all (Brehmer et al., 2022), or are only concerned with identifying abstractions of the latent variables without identifying abstractions of the latent causal graph (Von Kügelgen et al., 2021; Ahuja et al., 2022; Yao et al., 2023).

To the best of our knowledge, we provide the first identifiability results that give a graphical criterion for the degree of abstraction which we can identify latent causal models up to, depending on the interventional data available within the context of the weakly-supervised CRL problem, which take into account both structural and distributional properties of SCMs.

We structure the remainder of this work as follows. In Section 2.1 we introduce the weakly-supervised CRL problem setup in terms of the data generating process, which is essentially the same as the one outlined in (Brehmer et al., 2022). In Section 2.2, we proceed to give increasingly restrictive definitions of the identifiability of causal model parameters *up to equivalence*, and in Section 2.3 we move on to increasingly restrictive definitions of the identifiability of causal model parameters *up to abstraction*. In Section 3 we explain the assumptions behind our main results, before presenting the statements. We leave the detailed proofs of these results to the appendix but draw attention to some key properties of the data generating process in Section 3.3, and the proof techniques used at a high level. In Section 3.4 we provide some intuition for some of the unexpected conclusions that can be drawn from the results. Finally, in Section 4, we outline the various downstream applications of our results, as well their limitations.

2 PROBLEM FORMULATION

2.1 DATA GENERATING PROCESS

In this section, we describe the data-generating process in our problem setting, which is the functional relationship between the latent causal model parameters θ and the resultant distribution $p_\theta(\mathbf{x}, \tilde{\mathbf{x}})$ of counterfactual or contrastive pairs of observational data.

We first introduce the *structural causal model* (SCM) describing the pre-intervention latent variables. Let $\mathcal{G}$ be a directed acyclic graph, and associate each of the nodes $i \in V(\mathcal{G})$ with a vector space $\mathcal{Z}_i$, a random variable $\mathbf{z}_i$ taking values on $\mathcal{Z}_i$, as well as a conditional probability distribution $p\left(\mathbf{z}_i \mid \mathbf{z}_{Pa_\mathcal{G}(i)}\right)$. Furthermore, let each of the conditional distributions have a functional representation [1]

$$\mathbf{z}_i = f_i\left(\mathbf{z}_{Pa_\mathcal{G}(i)}, \boldsymbol{\varepsilon}_i\right), \quad \boldsymbol{\varepsilon}_i \sim p_{\boldsymbol{\varepsilon}_i} \quad \forall i \in V(\mathcal{G}), \tag{1}$$

where the distributions of the *exogenous variables* $\boldsymbol{\varepsilon}_i \in \mathcal{E}_i$ are all mutually independent, and each $f_i : \mathcal{Z}_{Pa_\mathcal{G}(i)} \times \mathcal{E}_i \to \mathcal{Z}_i$ is a deterministic function. Then we can denote $\mathcal{Z} := \bigoplus_{i \in V(\mathcal{G})} \mathcal{Z}_i$ and $\mathcal{E} := \bigoplus_{i \in V(\mathcal{G})} \mathcal{E}_i$, and define a deterministic function $\mathbf{f} : \mathcal{E} \to \mathcal{Z}$ by successively applying the causal mechanisms f_i. Therefore the distribution of the pre-intervention latents $p(\mathbf{z})$ can be parametrized by

$$\theta_{\text{SCM}} := (\mathcal{G}, \mathbf{f}, p_{\boldsymbol{\varepsilon}}). \tag{2}$$

We next describe the interventions on the latent variables. Let $\boldsymbol{\iota}$ be a random variable taking values in the power set of vertices of $\mathcal{G}$ (assumed to be finite), which tells us which latent variables are intervened upon at any one time. We denote the distribution of this discrete random variable as $P_{\boldsymbol{\iota}}$. Futhermore we write $\mathcal{I} := \text{supp}(\boldsymbol{\iota})$ and refer to its elements as *intervention targets* (i.e. the subsets of nodes which get intervened upon with nonzero probability). In the event that $\boldsymbol{\iota} = S$ for some $S \subseteq V(\mathcal{G})$, we assume that for every node i in S the causal mechanism from $Pa_\mathcal{G}(i)$ to i becomes completely severed, in what is known as a *perfect* intervention. Therefore, the post-intervention latents $\tilde{\mathbf{z}}$ conditional on $\boldsymbol{\iota} = S$ satisfy

$$\tilde{\mathbf{z}}_i = \tilde{f}_i^{(S)}(\tilde{\boldsymbol{\varepsilon}}_i) \quad \forall i \in S, \quad \tilde{\boldsymbol{\varepsilon}}_S \sim p_{\tilde{\boldsymbol{\varepsilon}}_S}, \tag{3}$$

$$\tilde{\mathbf{z}}_j = f_j\left(\tilde{\mathbf{z}}_{Pa_\mathcal{G}(j)}, \boldsymbol{\varepsilon}_j\right) \quad \forall j \in V(\mathcal{G}) \setminus S, \tag{4}$$

[1] The functional characterization of causal mechanisms, as was first introduced in (Verma and Pearl, 1990) allows us to define the effect of *interventions* on the model, and is sometimes referred to as *noise outsourcing* in contemporary literature (see e.g. Bloem-Reddy et al. (2020)).

where the new exogenous variable $\tilde{\boldsymbol{\varepsilon}}_S$ corresponding to the target set post-intervention is independent from the pre-intervention exogenous variable $\boldsymbol{\varepsilon}$, and each $\tilde{f}_i^{(S)}$ is a new function. Thus we can describe the effect of a random, unknown intervention on the structural causal model. Therefore the joint distribution of pre-intervention and post-intervention latent pairs $p(\mathbf{z}, \tilde{\mathbf{z}})$ can be parametrized by $(\theta_{\text{SCM}}, \theta_{\text{intv}})$, where

$$\theta_{\text{intv}} := (\tilde{\mathbf{f}}, p_{\tilde{\boldsymbol{\varepsilon}}}, P_{\boldsymbol{\iota}}). \tag{5}$$

Finally, let $\mathcal{X}$ be the vector space known as the observation space, and let $g : \mathcal{Z} \to \mathcal{X}$ be an invertible *mixing function* such that

$$\mathbf{x} = g(\mathbf{z}), \quad \tilde{\mathbf{x}} = g(\tilde{\mathbf{z}}). \tag{6}$$

Therefore the joint distribution of counterfactual pairs of observational data $p(\mathbf{x}, \tilde{\mathbf{x}})$ can be parametrized by

$$\theta := (\theta_{\text{SCM}}, \theta_{\text{intv}}, g)\,. \tag{7}$$

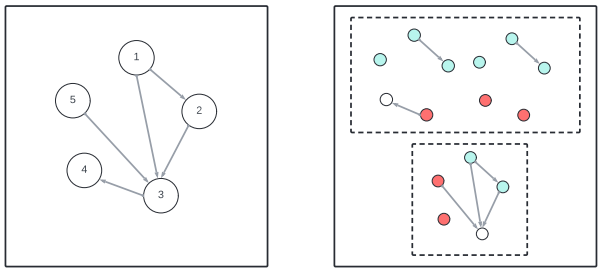

Figure 1: Data Generating Process for a Latent Causal Model. On the left we have the unintervened structural causal model generating samples of the pre-intervention latent $\mathbf{z}$, while on the right we have a mixture of intervened structural causal models generating samples of the post-intervention latent $\tilde{\mathbf{z}}$, for a collection of intervention targets $\mathcal{I} = \{\{3\}, \{3, 4\}, \{4, 5\}\}$, as denoted by the subsets of red vertices; the corresponding non-descendant sets of the intervention targets are denoted by the subsets of blue vertices.

It is important to note that the counterfactual setting outlined here differs from methods using *interventional data* (Brouillard et al., 2020; Gresele et al., 2020; Ahuja et al., 2023; von Kügelgen et al., 2024; Zhang et al., 2024) in two respects.

1. The counterfactual setting makes the more restrictive assumption of having access to the *joint* distribution of $(\mathbf{x}, \tilde{\mathbf{x}})$, whereas the interventional setting usually only assumes access to the marginal distributions of $\mathbf{x}$ and $\tilde{\mathbf{x}}$ separately.

2. The interventional setting makes the more restrictive assumption of being able to observe the *type* of intervention that occurs, even though the exact intervention target corresponding to a given type may be unknown. Concretely, in the interventional setting, we assume access to multiple marginal distributions of $\mathbf{x}^{(e)}$ indexed by an environment[2] variable e. The crucial point here is that e is *observable*, and that for any fixed e, the latent variable $\mathbf{z}^{(e)}$ is generated by a causal model with an *invariant causal structure*. In comparison, in the counterfactual setting, the intervention variable $\boldsymbol{\iota}$ is not observable, and samples of $\tilde{\mathbf{z}}$ are generated by a *mixture* of causal models with distinct graphical structures.

While CRL methods using interventional data are arguably more practical for applications such as biology (Belyaeva et al., 2021), where we have access to data generated from known experimental settings; methods using counterfactuals are better suited to cases such as temporal data from dynamical systems or offline reinforcement learning (Lippe et al., 2022; Ahuja et al., 2022; Brehmer et al., 2022), where at any given time step a *random* intervention may take place. Further justification for the assumption of availability of counterfactual data pairs will be elaborated on in Section 4.

2.2 IDENTIFIABILITY UP TO EQUIVALENCE

Identifiability in statistics refers to the ability to uniquely determine the true values of model parameters from observed data. A model is considered strongly identifiable if there is only a single set of parameter values that can generate the observed data. More generally, identifiability up to an equivalence class means that while there may be multiple sets of parameter values that can generate the same observed data, these sets are equivalent in some meaningful way.

For our purposes, we assume there exist some ground truth parameters $\theta^\star$, and that we may take unlimited samples from the latent causal model $p_{\theta^\star}(\mathbf{x}, \tilde{\mathbf{x}})$, in order to learn an estimator of $\theta^\star$. We say that $\theta^\star$ is identifiable up to some equivalence relation $\sim$ with respect to some hypothesis class of model parameters Θ, if for any $\theta \in \Theta$

$$p_\theta(\mathbf{x}, \tilde{\mathbf{x}}) = p_{\theta^\star}(\mathbf{x}, \tilde{\mathbf{x}}) \implies \theta \sim \theta^\star. \tag{8}$$

2.2.1 LATENT DISENTANGLEMENT

One instance of an equivalence relation between $\theta^\star$ and θ is defined in terms of an equivalence relation between the corresponding mixing function $g^\star$ and g.

[2]This is also sometimes referred to as a "view" Yao et al. (2023)

It is known as *disentanglement* (Bengio et al., 2013; Higgins et al., 2018), and we will draw attention to two variants of its definition.

The first variant deals with a *single latent subspace* that we want to isolate, which loosely speaking means that we want any two equivalent models to produce latent distributions that have the same marginals when restricted to this subspace.

Definition 2.1 (Latent disentanglement)**.** Given latent causal models parametrized by $\theta^\star$ and θ as defined in Section 2.1, and latent subspaces $\mathcal{W}^\star \subset \mathcal{Z}^\star$ and $\mathcal{W} \subset \mathcal{Z}$, we say that

$$\theta \sim_{\mathrm{L}} \theta^\star$$

with respect to these subspaces, if there exists a measurable function $h : \mathcal{W}^\star \to \mathcal{W}$ such that if we define the random variables $\mathbf{w}^\star$ and $\mathbf{w}$ to be the latent components of $\mathbf{z}^\star$ and $\mathbf{z}$ corresponding to the latent subspace $\mathcal{W}^\star$ and $\mathcal{W}$ respectively , then the function h satisfies

$$\mathbf{w} \stackrel{d}{=} h(\mathbf{w}^\star) \tag{9}$$

Alternatively, we can say that the *encoder* $g^{-1} : \mathcal{X} \to \mathcal{Z}$ disentangles the ground truth latent variable $\mathbf{w}^\star$ by identifying it with the variable $\mathbf{w}$ in the resultant latent representation $\mathbf{z}$.

We emphasize that the above definition is a *distributional equivalence between a single pair of latent components* in two causal models, without placing constraints on any of the other components.

The second variant of the definition of disentanglement deals with a full decomposition of the latent space into a *direct sum of latent subspaces*, such that we want any two equivalent models to produce latent distributions that have the same marginals when restricted to any of the subspaces in this decomposition.

Definition 2.2 (Full latent disentanglement)**.** Given latent causal models parametrized by $\theta^\star$ and θ as defined in Section 2.1, and latent space decompositions $\mathcal{Z}^\star = \bigoplus_{i=1}^n \mathcal{Z}_i^\star$ and $\mathcal{Z} = \bigoplus_{j=1}^n \mathcal{Z}_j$, we say that

$$\theta \sim_{\mathrm{FL}} \theta^\star$$

with respect to these decompositions, if there exist a *bijective function* $\phi : \{1, \cdots, n\} \to \{1, \cdots, n\}$ and measurable functions $h_i : \mathcal{Z}_i^\star \to \mathcal{Z}_{\phi(i)}$ for all $i \in \{1, \cdots, n\}$ such that if we define the random variables $\mathbf{z}_i^\star$ and $\mathbf{z}_j$ to be the latent components of $\mathbf{z}^\star$ and $\mathbf{z}$ corresponding to the latent subspaces $\mathcal{Z}_i^\star$ and $\mathcal{Z}_j$ respectively, then for all $i \in \{1, \cdots, n\}$ we have

$$\mathbf{z}_{\phi(i)} \stackrel{d}{=} h_i\left(\mathbf{z}_i^\star\right) \tag{10}$$

Alternatively, we can say that the encoder g^{-1} produces a fully disentangled representation $(\mathbf{z}_1, \cdots, \mathbf{z}_n)$ of the ground truth latent variables $(\mathbf{z}_1^\star, \cdots, \mathbf{z}_n^\star)$.

We emphasize that the above definition consists of *distributional equivalences between the full sets of latent components* in two causal models, up to the permutation given by ϕ, therefore it is easy to check that the following statement holds.

Lemma 2.1. *Given latent causal models parametrized by $\theta^\star$ and θ as defined in Section 2.1, suppose $\theta \sim_{FL} \theta^\star$ with respect to the decompositions $\bigoplus_{j=1}^n \mathcal{Z}_j$ and $\bigoplus_{i=1}^n \mathcal{Z}_i^\star$. Then for all $i \in \{1, \cdots, n\}$, we have $\theta \sim_L \theta^\star$ with respect to $\mathcal{Z}_{\phi(i)}$ and $\mathcal{Z}_i^\star$.*

2.2.2 STRUCTURAL CAUSAL MODEL ISOMORPHISM

Note that so far, we have only defined distributional equivalences between the variables of the latent causal model and their representations. An even stronger equivalence relation than full latent disentanglement between $\theta^\star$ and θ takes causal structure into account, and requires that in addition the causal graphs $\mathcal{G}^\star$ and $\mathcal{G}$ be *isomorphic*. This is sometimes referred to as *structural causal model isomorphism* (Fong, 2013; Brehmer et al., 2022) or the CRL identifiability class von Kügelgen et al. (2024).

Definition 2.3 (SCM isomorphism)**.** Given latent causal models parametrized by $\theta^\star$ and θ as defined in Section 2.1, note that there exist *canonical decompositions* of their latent subspaces $\mathcal{Z}^\star$ and $\mathcal{Z}$ into direct sums of subspaces $\bigoplus_{i \in V(\mathcal{G}^\star)} \mathcal{Z}_i^\star$ and $\bigoplus_{j \in V(\mathcal{G})} \mathcal{Z}_j$ respectively. We say that

$$\theta \sim_{\mathrm{SCM}} \theta^\star$$

if there exists a *graph isomorphism* $\phi : \mathcal{G}^\star \to \mathcal{G}$, and measurable functions $h_i : \mathcal{Z}_i^\star \to \mathcal{Z}_{\phi(i)}$ such that Eq(10) holds for all $i \in V(\mathcal{G}^\star)$.

Alternatively, we say that the structural causal models with parameters θ_{SCM} and $\theta_{\mathrm{SCM}}^\star$ are *isomorphic*.

Note that the above definition consists of a *structural equivalence* in the form of the graph isomorphism, as well as distributional equivalences between latent components in two causal models that are compatible with the graph isomorphism, therefore it is easy to check that the following statement holds.

Lemma 2.2. *Given latent causal models parametrized by $\theta^\star$ and θ as defined in Section 2.1, suppose $\theta \sim_{SCM} \theta^\star$. Then $\theta \sim_{FL} \theta^\star$ with respect to $\bigoplus_{j \in V(\mathcal{G})} \mathcal{Z}_j$ and $\bigoplus_{i \in V(\mathcal{G}^\star)} \mathcal{Z}_i^\star$.*

2.3 IDENTIFIABILITY UP TO ABSTRACTION

In this section, we introduce the concept of identifiability up to model abstraction, which can be thought of as the "common factor" between all models which are consistent with the observable distribution.

To do this, we need some notion of a *partial order* $\preceq$ on Θ, which compares causal models by their level of granularity or, in some sense, complexity. Broadly speaking, given any causal model $\theta^\star$, there is always another, more complex model $\theta' \succeq \theta^\star$ that produces the same observable distribution [3]. What we want to know is what all the models which are able to produce the same observable distribution as $\theta^\star$ have in common, meaning that we want to find an abstraction θ of $\theta^\star$ that is the *infimum* of all these models. Note that this principle of finding minimal causal structures consistent with the observed data, which can be thought of as a reformulation of Occam's razor, is discussed at length in Chapter 2 of Pearl (2009).

Concretely, in our problem setting, $\theta^\star$ is said to be identifiable up to the abstraction θ with respect to $\preceq$ and some hypothesis class of parameters Θ if for all $\theta' \in \Theta$

$$p_{\theta'}(\mathbf{x}, \tilde{\mathbf{x}}) = p_{\theta^\star}(\mathbf{x}, \tilde{\mathbf{x}}) \implies \theta \preceq \theta' \tag{11}$$

We will proceed to extend the definitions of equivalence relations $\sim$ on the parameter space Θ from the previous subsection to (weak) partial orders $\preceq$ on Θ.

2.3.1 LATENT ABSTRACTION

Definition 2.4 (Latent abstraction)**.** Given latent causal models parametrized by θ and θ' as defined in Section 2.1, and a latent subspace $\mathcal{W} \subset \mathcal{Z}$, together with a set of complementary latent subspaces $\mathcal{Z}'_1, \cdots, \mathcal{Z}'_k \subseteq \mathcal{Z}'$, we say that

$$\theta \preceq_{\mathrm{L}} \theta'$$

with respect to this latent subspace in $\mathcal{Z}$ and set of latent subspaces in $\mathcal{Z}'$ if there exists measurable functions $h_i : \mathcal{Z}'_i \to \mathcal{W}$ for all $i \in \{1, \cdots, k\}$ such that if we define the random variables $\mathbf{z}'_i$ to be the latent components of $\mathbf{z}'$ corresponding to the latent subspaces $\mathcal{Z}'_i$, then

$$\mathbf{w} \overset{d}{=} \sum_{i=1}^{k} h_i(\mathbf{z}'_i). \tag{12}$$

[3]For example, we can always add more "dummy variables" that increase the number of nodes on the causal graph, which, when marginalized upon do not produce any effect on the final observable distributions under intervention

Alternatively, we can say that the encoder g^{-1} produces an abstraction $\mathbf{w}$ of the latent variables $(\mathbf{z}'_1, \cdots, \mathbf{z}'_k)$.

Note that here we have defined a *single distributional equivalence* between a subset of latent components in one model, and the distribution of a single latent component in another.

Definition 2.5 (Full latent abstraction)**.** Given latent causal models parametrized by θ and θ' as defined in Section 2.1, and decompositions of the latent spaces $\mathcal{Z}'$ and $\mathcal{Z}$ into direct sums of subspaces $\mathcal{Z}'_1 \oplus \cdots \oplus \mathcal{Z}'_n$ and $\mathcal{Z}_1 \oplus \cdots \oplus \mathcal{Z}_m$ respectively, we say that

$$\theta \preceq_{\mathrm{FL}} \theta'$$

with respect to these decompositions if there exist a *surjective function* $\phi : \{1, \cdots, n\} \to \{1, \cdots, m\}$ and measurable functions $h_i : \mathcal{Z}'_i \to \mathcal{Z}_{\phi(i)}$ such that if we define the random variables $\mathbf{z}'_i$ and $\mathbf{z}_j$ to be the latent components of $\mathbf{z}'$ and $\mathbf{z}$ corresponding to the latent subspaces $\mathcal{Z}'_i$ and $\mathcal{Z}_j$ respectively, then for all $j \in \{1, \cdots, m\}$, we have

$$\mathbf{z}_j \overset{d}{=} \sum_{i \in \phi^{-1}(j)} h_i(\mathbf{z}^\star_i). \tag{13}$$

Alternatively, we can say that the encoder g^{-1} produces an abstraction $(\mathbf{z}_1, \cdots, \mathbf{z}_m)$ of the latent variables $(\mathbf{z}'_1, \cdots, \mathbf{z}'_n)$

Note that here we have defined a *full set of distributional equivalences* based on a surjection between all the latent components of two causal models, therefore it is easy to check that the following statement holds.

Lemma 2.3. *Given latent causal models parametrized by θ and θ' as defined in Section 2.1, suppose $\theta \preceq_{FL} \theta'$ with respect to the decompositions $\mathcal{Z}'_1 \oplus \cdots \oplus \mathcal{Z}'_n$ and $\mathcal{Z}_1 \oplus \cdots \oplus \mathcal{Z}_m$. Then for all $j \in \{1, \cdots, m\}$, we have $\theta \preceq_L \theta'$ with respect to the latent subspace $\mathcal{Z}_j \subseteq \mathcal{Z}$ and the set of latent subspaces $\{\mathcal{Z}'_i\}_{i \in \phi^{-1}(j)}$.*

2.3.2 STRUCTURAL CAUSAL MODEL HOMOMORPHISM

To extend the definition of a structural causal model isomorphism, we make use of a more general definition of structure preserving maps between SCMs that is explored in (Otsuka and Saigo, 2022).

Definition 2.6 (SCM homomorphism)**.** Given latent causal models parametrized by θ and θ' as defined in Section 2.1, we say that there exists a *structural causal model homomorphism* between SCMs with parameters θ'_{SCM} and θ_{SCM} if there exists a *graph homomorphism*

$\phi : \mathcal{G}' \to \mathcal{G}$, and measurable functions $h_i : \mathcal{Z}'_i \to \mathcal{Z}_{\phi(i)}$ such that Eq(13) holds for all $j \in V(\mathcal{G})$. See Fig. 2 for example.

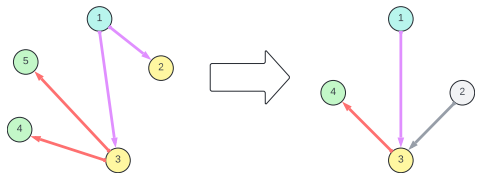

Figure 2: SCM homomorphism, as defined by a graph homomorphism ϕ that maps nodes in $\mathcal{G}'$ to nodes in $\mathcal{G}$ of the same colour (e.g. $\phi(2) = \phi(3) = 3$), as well as a set of invertible measurable functions which ensure that the latent variables represented by nodes in the image of ϕ, which in this case are $\mathbf{z}'_1$, $\mathbf{z}'_3$ and $\mathbf{z}'_4$, have equivalent distributions to their counterparts after marginalization on $\mathbf{z}'_2$.

Definition 2.7 (SCM abstraction)**.** Given latent causal models parametrized by θ and θ' as defined in Section 2.1, we say that

$$\theta \preceq_{\text{SCM}} \theta'$$

if there exists a SCM homomorphism between SCMs with parameters θ'_{SCM} and θ_{SCM}, and the associated graph homomorphism $\phi : \mathcal{G}' \to \mathcal{G}$ is surjective (i.e. an epimorphism). See Fig. 3 for example.

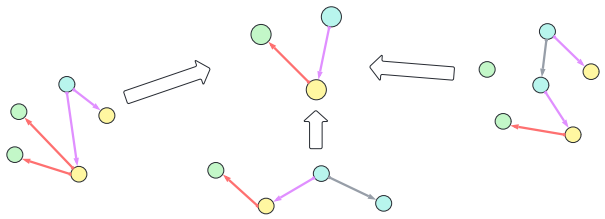

Figure 3: SCM Abstractions. The white arrows represent surjective SCM homomorphisms, which map causal models operating at higher granularity levels to lower granularity levels, such that the model at the centre is an abstraction of all the other models.

Alternatively, we can say that θ_{SCM} is an *abstraction* of the SCM parameters θ'_{SCM}.

Note that the above definition consists of a *structural map* in the form of a graph epimorphism, as well as a *full set of distributional equivalences* between all the latent components of two causal models that is compatible with the structural map, therefore it is easy to check that the following statement holds.

Lemma 2.4. *Given latent causal models parametrized by θ and θ' as defined in Section 2.1, suppose $\theta \sim_{SCM} \theta'$. Then $\theta \preceq_{FL} \theta'$ with respect to the canonical decompositions $\bigoplus_{j \in V(\mathcal{G})} \mathcal{Z}_j$ and $\bigoplus_{i \in V(\mathcal{G}')} \mathcal{Z}'_i$.*

Finally, we can check that the equivalence relations from Section 2.2 are consistent with the definitions of partial orders that give rise to the corresponding abstractions defined in this section.

Lemma 2.5. *Given latent causal models parametrized by θ and θ' as defined in Section 2.1, and any $(\sim, \preceq) \in \{(\sim_L, \preceq_L), (\sim_F L, \preceq_F L), (\sim_{SCM}, \preceq_{SCM})\}$*

$$\theta \sim \theta' \iff \theta \preceq \theta' \text{ and } \theta \succeq \theta'$$

3 IDENTIFIABILITY RESULTS

3.1 PRELIMINARIES

Before presenting our main result, we will introduce several key concepts on which it is based, as well as some notation. We will assume that the reader is familiar with the definitions of σ-algebras and partitions, which can otherwise be found in Appendix A.

Family of non-descendants Given a directed graph $\mathcal{G}$, the *non-descendants* of a subset S of its vertices shall be denoted as $\text{nd}(S)$, that is the intersection of the non-descendants of all the nodes in S. Furthermore, for a family $\mathcal{I}$ of intervention targets, we will denote the corresponding family of non-descendants as $\mathbf{nd}(\mathcal{I}) := \{\text{nd}(S) : S \in \mathcal{I}\}$

σ-algebra generated by family of sets Given a collection of subsets $\mathcal{A}$ of a set X, the *σ-algebra generated by $\mathcal{A}$*, denoted $\sigma(\mathcal{A})$, is the smallest σ-algebra on X that contains all the sets in $\mathcal{A}$. More formally, $\sigma(\mathcal{A})$ is the intersection of all σ-algebras on X that contain $\mathcal{A}$, ensuring that $\sigma(\mathcal{A})$ satisfies the properties of a σ-algebra (containing the empty set, closed under complements and countable union).

Partition generated by σ-algebra The *partition generated by a σ-algebra $\mathcal{F}$* on a set X, denoted $\mathcal{P}(\mathcal{F})$, is the collection of disjoint measurable sets in $\mathcal{F}$ that together cover X. More formally, it consists of the equivalence classes of the relation that considers two elements $x, y \in X$ to be equivalent if every set in $\mathcal{F}$ either contains both or contains neither. These equivalence classes form a partition, and each class is an element of the σ-algebra. This partition is maximal in the sense that the elements of the partition cannot be further subdivided using sets from $\mathcal{F}$.

Quotient graph generated by partition The *quotient graph* $\mathcal{G}/\mathcal{P}$ of a directed graph $\mathcal{G}$ with respect

to a partition $\mathcal{P} = \{A_1, A_2, \ldots, A_k\}$ of the vertex set $V(\mathcal{G})$ is defined such that the vertex set of $\mathcal{G}/\mathcal{P}$ is exactly $\mathcal{P}$, and for distinct blocks A_i and A_j in $\mathcal{P}$, there is a directed edge from A_i to A_j in $\mathcal{G}/\mathcal{P}$ if and only if there exists at least one directed edge from a vertex in A_i to a vertex in A_j in the original directed graph $\mathcal{G}$.

Graph condensations Note that the definition above implies that given a directed acyclic graph $\mathcal{G}$ and an arbitrary partition $\mathcal{P}$ of its vertices, the resultant quotient graph $\mathcal{G}' := \mathcal{G}/\mathcal{P}$ can contain cycles. However, it is always true that there exists another quotient graph $\mathcal{G}'_C$, known as the *condensation* of $\mathcal{G}'$, that is acyclic. This is important because it ensures that we can always obtain an abstraction of some causal model for any partition of the vertices by taking the condensation.

We construct the condensation $\mathcal{G}'_C$ by taking the quotient of $\mathcal{G}'$ with respect to the partition defined by all the *strongly connected components* of $\mathcal{G}'$, which are maximal subgraphs $\mathcal{H} \subseteq \mathcal{G}'$ such that for all vertices $u, v \in V(\mathcal{H})$, there exists a directed path from u to v and from v to u.

3.2 MAIN RESULTS

For both of the results stated in this section, we make the following assumptions on the hypothesis class of parameters Θ

1. *Faithfulness of the causal graph*: Let $\mathcal{G}$ be a perfect map (Pearl, 2009) for the distribution of $p(\mathbf{z})$, meaning that it encapsulates all the conditional independences.
2. *Absolute continuity of latent distributions*: Let $\mathcal{E}_i \cong \tilde{\mathcal{E}}_i \cong \mathcal{Z}_i$[4], let f_i, $f_i^{(S)}$ be continuously differentiable for all i and S, and let $p_{\boldsymbol{\varepsilon}}$ and $p_{\tilde{\boldsymbol{\varepsilon}}}$ be absolutely continuous.
3. *Smoothness of mixing function*: Let g be a diffeomorphism.

3.2.1 IDENTIFIABLE MODEL ABSTRACTION

Our first result shows that the parameters of a latent causal model as defined in Section 2.1 can be identified up to a SCM abstraction, depending on the *non-descendant sets* of the intervention targets.

Theorem 3.1. *Any latent causal model with parameters $\theta^\star \in \Theta$ is identifiable up to a SCM abstraction θ with causal graph*

$$\mathcal{G} = \mathcal{G}^\star/\mathcal{P}(\sigma(\mathbf{nd}(\mathcal{I}^\star))). \tag{14}$$

[4]Here we use $\cong$ to denote isomorphic vector spaces

meaning that for all $\theta' \in \Theta$

$$p_{\theta'}(\mathbf{x}, \tilde{\mathbf{x}}) = p_{\theta^\star}(\mathbf{x}, \tilde{\mathbf{x}}) \implies \theta \preceq_{SCM} \theta'. \tag{15}$$

Furthermore, we can show that the quotient graph $\mathcal{G}$ is acyclic, so we do not have to resort to taking its condensation.

Remark. Theorem 3.1 implies that in order to identify the latent variables corresponding to a subset of nodes $S \in V(\mathcal{G}^\star)$ in the SCM up to abstraction, it is not necessary to have an intervention on S directly, meaning that we require $S \in \mathcal{I}^\star$. It is sufficient to have $S \in \sigma(\mathbf{nd}(\mathcal{I}^\star))$ instead.

Example 3.1. *Let $\theta^\star$ be the latent causal model depicted in Fig. 1, so that $\mathcal{I}^\star = \{\{3\}, \{3,4\}, \{4,5\}\}$. Then the corresponding family of non-descendants is $\mathbf{nd}(\mathcal{I}^\star) = \{\{1,2,5\}, \{1,2\}\}$, as shown by the subsets of blue vertices on the panel on the right of the figure. Hence we can compute the partition $\mathcal{P}(\sigma(\mathbf{nd}(\mathcal{I}^\star))) = \{\{1,2\}, \{3,4\}, \{5\}\}$, and know that we can identify the latent causal model up to the SCM abstraction shown in Fig. 4.*

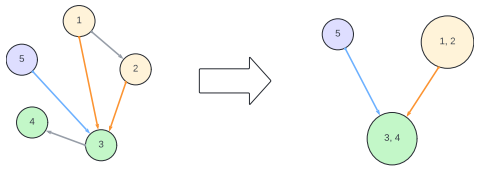

Figure 4: Identifiable SCM abstraction with graphical structure as shown on the left and $\mathcal{I}^\star = \{\{3\}, \{3,4\}, \{4,5\}\}$ as its family of intervention targets.

3.2.2 ADDITIONAL IDENTIFIABLE LATENTS

Our second result follows from the first, and uses our definition of latent disentanglement $\sim_{\mathrm{L}}$ to identify additional latents. Here we have a notion of equivalence between latent variables that correspond to the *intersections* of intervention targets with the same non-descendant sets, but no constraints on the causal mechanisms between them.

Theorem 3.2. *For any $N \in \mathbf{nd}(\mathcal{I}^\star)$ denote the intersection of all intervention targets with N as their non-descendant set as*

$$\pi(N) := \bigcap \{S \in \mathcal{I} : nd(S) = N\}.$$

Now provided that $\pi(N)$ is a singleton set $\{i\}$ and $\mathcal{Z}_i^\star \cong \mathbb{R}$, then we can identify the latent $\mathbf{z}^\star_{\pi(N)}$ up to disentanglement, meaning that for all $\theta \in \Theta$

$$p_\theta(\mathbf{x}, \tilde{\mathbf{x}}) = p_{\theta^\star}(\mathbf{x}, \tilde{\mathbf{x}}) \implies \theta \sim_L \theta^\star \tag{16}$$

with respect to $\mathcal{Z}_i^\star$ and some latent subspace in $\mathcal{Z}$.

Example 3.2. *Let $\theta^\star$ be the latent causal model depicted in Fig. 1, so that $\mathcal{I}^\star = \{\{3\}, \{3,4\}, \{4,5\}\}$ and $\mathbf{nd}(\mathcal{I}^\star) = \{\{1,2,5\}, \{1,2\}\}$ as before. Then by looking at the intersections of subsets of blue vertices in each of the boxes with dotted lines, we can see that $\pi(\{1,2\}) = \{4,5\}$ and $\pi(\{1,2,5\}) = \{3\}$. Thus we can disentangle the latent variable $\tilde{\mathbf{z}}_3^\star$, provided that it takes value on $\mathbb{R}$.*

3.3 PROOF OUTLINES

We leave the precise details of the proofs of Theorem 3.1 and Theorem 3.2 to Appendix C, and instead highlight some key properties of the data generating process that the proofs depend on. All together, these should serve as an outline of the main techniques used.

Finite mixtures Since the intervention target $\boldsymbol{\iota}$ is a discrete random variable and takes only a finite number of values, the latent distribution becomes a finite mixture of distributions with one mixture component for each value of $S \in \mathcal{I}$

$$p(\mathbf{z}, \tilde{\mathbf{z}}) = \sum_{S \in \mathcal{I}} \mathbb{P}(\boldsymbol{\iota} = S) p(\mathbf{z}, \tilde{\mathbf{z}} \mid \boldsymbol{\iota} = S). \tag{17}$$

We will show that these components can be separated up to equivalence classes of $\mathcal{I}$ where each component corresponds to an element $N \in \mathbf{nd}(\mathcal{I})$.

$$p(\mathbf{z}, \tilde{\mathbf{z}}) = \sum_{N \in \mathbf{nd}(\mathcal{I})} \mathbb{P}(\mathrm{nd}(\boldsymbol{\iota}) = N) p(\mathbf{z}, \tilde{\mathbf{z}} \mid \mathrm{nd}(\boldsymbol{\iota}) = N). \tag{18}$$

Invariance of non-descendant variables From Eq. (1) and Eq. (4), we can see that given any intervention target S, the block of latents which correspond to the non-descendant set of S (i.e. nodes in $\mathcal{G}$ which are not descendants of any member of S), is invariant across the counterfactual pair, and crucially is the "maximally" invariant block. Formally, if we let $N = \mathrm{nd}(S)$ denote the non-descendant set of S then

$$\mathbb{P}(\mathbf{z}_N \neq \tilde{\mathbf{z}}_N \mid \boldsymbol{\iota} = S) = 0, \tag{19}$$

$$\mathbb{P}(\mathbf{z}_T \neq \tilde{\mathbf{z}}_T \mid \boldsymbol{\iota} = S) > 0 \quad \forall T \not\subseteq N. \tag{20}$$

Von Kügelgen et al. (2021) made use of this property to isolate the *combined* "content" block, $\bigcap \mathbf{nd}(\mathcal{I}^\star)$, from the "style" block, which consists of the remainder of the latents, identifying the true causal model up to the abstraction $\mathbf{z}_{\text{content}} \to \mathbf{z}_{\text{style}}$. However, through a slightly more careful examination, we show that every block in $\mathcal{P}(\sigma(\mathbf{nd}(\mathcal{I}^\star)))$ can be disentangled (see Theorem 3.1). Essentially, this comes down to the fact that for each $N \in \mathbf{nd}(\mathcal{I})$, the distribution $p(\mathbf{z}, \tilde{\mathbf{z}} \mid \mathrm{nd}(\boldsymbol{\iota}) = N)$ is an absolutely continuous measure μ_N with non-zero mass on a submanifold of $\mathcal{Z} \times \mathcal{Z}$ that uniquely identifies N, since $\mathbf{z}_N \overset{d}{=} \tilde{\mathbf{z}}_N$ with respect to μ_N. Therefore we can obtain a matching of all latent blocks and their complements in the non-descendant sets of intervention targets, for any two latent causal models which produce the same observable distribution. Note that for this step, the assumption of absolute continuity of the latent distributions, together with the assumption of the smoothness of the mixing function are key.

Independence of interventional targets From Eq. (3), we can see that given any intervention target $S \in \mathcal{I}$, the block of post-intervention latents corresponding to S is statistically independent of the pre-intervention latents. i.e. $\tilde{\mathbf{z}}_S \perp\!\!\!\perp \mathbf{z} \mid \boldsymbol{\iota} = S$. Brehmer et al. (2022) made use of this property to disentangle $\mathbf{z}_S^\star$ for all $S \in \mathcal{I}$, but under the restrictive assumption that $\mathcal{I}$ consists precisely of all the *atomic* intervention targets [5], so that no distinct intervention targets share the same non-descendant set. We remove this assumption, and instead disentangle $\mathbf{z}^\star_{\pi(N)}$ (see Theorem 3.2) by making use of the fact that for all $N \in \mathbf{nd}(\mathcal{I})$

$$\tilde{\mathbf{z}}_{\pi(N)} \perp\!\!\!\perp \mathbf{z} \mid nd(\iota) = N. \tag{21}$$

Loosely, the equation above translates to the fact that with respect to the distribution of $p(\mathbf{z}, \tilde{\mathbf{z}} \mid \mathrm{nd}(\boldsymbol{\iota}) = N)$, which we managed to separate from the other mixture components of the total paired latent distribution as a result of the previous step, $\pi(N)$ corresponds to the maximal partition of $\tilde{\mathbf{z}}$ that is independent from $\mathbf{z}$. Note that the faithfulness of the causal graph is particularly important here, since we do not want to fail to take into account conditional independences which were not represented in the graph.

3.4 DISCUSSION

Theorem 3.1 implies that in order to identify the latent variables corresponding to a subset of nodes $S \in V(\mathcal{G}^\star)$ in the SCM up to abstraction, it is not necessary to have an intervention on S directly, meaning that we require $S \in \mathcal{I}^\star$, which is perhaps surprising. Instead, it is sufficient to have $S \in \sigma(\mathbf{nd}(\mathcal{I}^\star))$.

Furthermore, this result implies that all these distributional maps of identifiable latent variable blocks are *compatible with a structural map* that abstracts $\mathcal{G}$ to its quotient $\mathcal{G}'$ as defined in Eq. (14).

This is particularly relevant in cases of CRL problems where recovering the full causal graph $\mathcal{G}$ is not possible

[5] an intervention target set S is said to be atomic if $S = \{i\}$ for some node $i \in V(\mathcal{G})$

due to lack of atomic interventions. In these scenarios, we can learn the causal structure up to an abstraction $\mathcal{G}'$, meaning that while the causal relationships between certain latent variables corresponding to nodes in $\mathcal{G}$ are not clear, the causal mechanisms between aggregated subsets of these variables corresponding to nodes in $\mathcal{G}'$ can be recovered correctly.

Additionally, Theorem 3.2 tells us that we can disentangle even more of the latent variables than the ones implied by Theorem 3.1, at the cost of disregarding the causal graph.

4 DOWNSTREAM APPLICATIONS AND LIMITATIONS

In terms of downstream tasks, our method has the usual applications for causal discovery, including causal effect estimation with respect to high-level variables; although we emphasize that in our particular problem formulation, none of the causal variables are directly observable, and therefore our work differs from the settings in classic ATE estimation or those presented in (Anand et al., 2023a).

Instead, we find that our weakly-supervised CRL setup makes assumptions that are more prevalent in recent machine learning literature, in which under unknown interventions counterfactual data pairs are indeed available, sometimes at the cost of the direct observability of causal variables. For example

- In contrastive learning (Von Kügelgen et al., 2021), pairs of data samples before and after random augmentations or transformations, which can be viewed as interventions, are used in order to learn latent representations with causal dependencies
- In problems with temporal data (Lippe et al., 2022), we may consider sequential observations of the system as our counterfactual data.
- In the field of causal interpretability, specifically with respect to the method of interchange intervention training (Geiger et al., 2022), we ay generate synthetic counterfactual data pairs by activation patching of neural networks.

In most machine learning applications, access to the latent causal structure can benefit generalization to out-of-distribution data. It can also serve as a foundation for interpretable and fair ML methods.

The eventual objective in CRL is having a methodology for learning the underlying latent causal model, up to some degree of abstraction. However, this is a distinct objective from having theoretical identifiability results, which demonstrate that should one succeed in learning an estimator of the model parameters which maximize the likelihood of observable distribution, then the true causal model is identified up to abstraction.

Learning remains an important problem of its own, and we make no claim in addressing that problem. Our example in the appendix demonstrates a possible route for a small toy problem, and is by no mean a demonstration of a scalable method for identification of the abstract causal model, which we shall leave for future research.

5 CONCLUSION

We introduced a new framework for examining the identifiability of causal models up to abstraction. While previous works aiming to jointly learn the causal graph in conjunction with the latent variables have focused on fully identifying the graph up to isomorphism (Brehmer et al., 2022; von Kügelgen et al., 2024; Wendong et al., 2024), we show that with relaxed assumptions, we can still recover a quotient graph and additional latent blocks that can all be determined from the family of intervention targets, which are not necessarily atomic and do not have to include all nodes of the graph. We argue that this is meaningful because in some ways, the identifiable causal model abstraction to constitutes the "real" ground truth model given the observable distribution, by the law of parsimony.

ACKNOWLEDGMENTS

This research was supported by CIFAR AI Chairs, NSERC Discovery, and Samsung AI Labs. Mila and Compute Canada provided computational resources. The authors would like to thank Sébastien Lachapelle for insightful discussions.

References

Ahuja, K., Hartford, J. S., and Bengio, Y. (2022). Weakly supervised representation learning with sparse perturbations. *Advances in Neural Information Processing Systems*, 35:15516–15528.

Ahuja, K., Mahajan, D., Wang, Y., and Bengio, Y. (2023). Interventional causal representation learning. In *International conference on machine learning*, pages 372–407. PMLR.

Anand, T., Ribeiro, A., Tian, J., and Bareinboim, E. (2023a). Causal effect identification in cluster dags. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37:12172–12179.

Anand, T. V., Ribeiro, A. H., Tian, J., and Bareinboim, E. (2023b). Causal effect identification in cluster dags. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12172–12179.

Beckers, S. and Halpern, J. Y. (2019). Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 2678–2685.

Belyaeva, A., Squires, C., and Uhler, C. (2021). Dci: learning causal differences between gene regulatory networks. *Bioinformatics*, 37(18):3067–3069.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Bloem-Reddy, B., Whye, Y., et al. (2020). Probabilistic symmetries and invariant neural networks. *Journal of Machine Learning Research*, 21(90):1–61.

Brehmer, J., De Haan, P., Lippe, P., and Cohen, T. S. (2022). Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331.

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. (2020). Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877.

Fong, B. (2013). Causal theories: A categorical perspective on bayesian networks. *arXiv preprint arXiv:1301.6201*.

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., et al. (2024). Causal abstraction: A theoretical foundation for mechanistic interpretability. *Preprint*.

Geiger, A., Lu, H., Icard, T., and Potts, C. (2021). Causal abstractions of neural networks. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586. Curran Associates, Inc.

Geiger, A., Wu, Z., Lu, H., Rozner, J., Kreiss, E., Icard, T., Goodman, N., and Potts, C. (2022). Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning*, pages 7324–7338. PMLR.

Gresele, L., Rubenstein, P. K., Mehrjou, A., Locatello, F., and Schölkopf, B. (2020). The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. In *Uncertainty in Artificial Intelligence*, pages 217–227. PMLR.

Hauser, A. and Bühlmann, P. (2012). Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1):2409–2464.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. (2018). Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*.

Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439.

Hyvarinen, A., Sasaki, H., and Turner, R. (2019). Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR.

Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, S. (2022). Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR.

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. (2020). Weakly-supervised disentanglement without compromises. In *International conference on machine learning*, pages 6348–6359. PMLR.

Massidda, R., Geiger, A., Icard, T., and Bacciu, D. (2023). Causal abstraction with soft interventions. In *Conference on Causal Learning and Reasoning*, pages 68–87. PMLR.

Otsuka, J. and Saigo, H. (2022). On the equivalence of causal models: A category-theoretic approach. In *Conference on Causal Learning and Reasoning*, pages 634–646. PMLR.

Pearl, J. (2009). *Causality*. Cambridge university press.

Rischel, E. F. (2020). The category theory of causal models. *Master's thesis, University of Copenhagen*.

Rischel, E. F. and Weichwald, S. (2021). Compositional abstraction error and a category of causal models. In *Uncertainty in Artificial Intelligence*, pages 1013–1023. PMLR.

Rubenstein, P. K., Weichwald, S., Bongers, S., Mooij, J. M., Janzing, D., Grosse-Wentrup, M., and Schölkopf, B. (2017). Causal consistency

of structural equation models. *arXiv preprint arXiv:1707.00819*.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Shu, R., Chen, Y., Kumar, A., Ermon, S., and Poole, B. (2019). Weakly supervised disentanglement with guarantees. *arXiv preprint arXiv:1910.09772*.

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, prediction, and search*. MIT press.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. *Probabilistic and Causal Inference*.

von Kügelgen, J., Besserve, M., Wendong, L., Gresele, L., Kekić, A., Bareinboim, E., Blei, D., and Schölkopf, B. (2024). Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36.

Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. (2021). Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467.

Wendong, L., Kekić, A., von Kügelgen, J., Buchholz, S., Besserve, M., Gresele, L., and Schölkopf, B. (2024). Causal component analysis. *Advances in Neural Information Processing Systems*, 36.

Yang, K., Katcoff, A., and Uhler, C. (2018). Characterizing and learning equivalence classes of causal DAGs under interventions. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5541–5550. PMLR.

Yao, D., Xu, D., Lachapelle, S., Magliacane, S., Taslakian, P., Martius, G., von Kügelgen, J., and Locatello, F. (2023). Multi-view causal representation learning with partial observability. *arXiv preprint arXiv:2311.04056*.

Zennaro, F. M. (2022). Abstraction between structural causal models: A review of definitions and properties. *arXiv preprint arXiv:2207.08603*.

Zhang, J., Greenewald, K., Squires, C., Srivastava, A., Shanmugam, K., and Uhler, C. (2024). Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36.

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. (2021). Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

APPENDIX

A DEFINITIONS

Definition A.1 (σ-algebra)**.** A *σ-algebra* on a set X is a collection $\mathcal{F}$ of subsets of X which satisfies the following properties

1. Universality: $X \in \mathcal{F}$
2. Closure under Complements: $A \in \mathcal{F} \implies A^c \in \mathcal{F}$
3. Closure under Countable Union: $A_n \in \mathcal{F} \quad \forall n \in \mathbb{N} \implies \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$

Definition A.2 (Partition)**.** A *partition* of a set X is a collection of non-empty, pairwise disjoint subsets $\{A_i\}_{i\in I}$ of X such that:

1. $A_i \cap A_j = \emptyset$ for all $i \neq j$ (the subsets are pairwise disjoint),
2. $\bigcup_{i\in I} A_i = X$ (the union of all subsets covers X).

Definition A.3 (Group)**.** A *group* is a set G equipped with a binary operation (often denoted by $*$) that satisfies the following four axioms:

1. Closure: $\forall a, b \in G$, $a * b \in G$
2. Associativity: $\forall a, b, c \in G$, $(a * b) * c = a * (b * c)$
3. Identity: $\exists e \in G$ such that $\forall a \in G$, $a*e = e*a = a$
4. Inverse: $\forall a \in G$, $\exists a^{-1} \in G$ such that $a * a^{-1} = a^{-1} * a = e$

Definition A.4 (Group Action)**.** A *group action* of a group G on a set X is a map $\cdot : G \times X \to X$ that satisfies the following two properties:

1. Identity: For all $x \in X$, $e \cdot x = x$, where e is the identity element of G
2. Compatibility: For all $g, h \in G$ and $x \in X$, $(g * h) \cdot x = g \cdot (h \cdot x)$.

Definition A.5 (Stabilizer)**.** Let G be a group acting on a set X. For any element $x \in X$, the stabilizer of x, denoted by G_x, is the subgroup of G consisting of all elements that fix x

$$G_x := \{g \in G : g \cdot x = x\}$$

Definition A.6 (Diffeomorphism Group)**.** The diffeomorphism group on a differentiable manifold M, denoted by $\mathrm{Diff}(M)$, is the group of all smooth, invertible maps from M to itself with smooth inverses. In other words, it consists of all bijective maps $f : M \to M$ such that both f and its inverse f^{-1} are differentiable. The group operation in $\mathrm{Diff}(M)$ is composition of maps. The identity element of $\mathrm{Diff}(M)$ is the identity map $\mathrm{id}_M : M \to M$. The inverse of a diffeomorphism $f \in \mathrm{Diff}(M)$ is its inverse map f^{-1}.

Definition A.7 (Pushforward Distributions)**.** The *pushforward of a distribution* describes how a probability distribution is transformed under a given function. Let $(X, \mathcal{A}, \mu)$ be a measure space, where X is a set, $\mathcal{A}$ is a σ-algebra of measurable sets on X, and $\mathbb{P}$ is a measure for a distribution on $(X, \mathcal{A})$. Let $f : X \to Y$ be a measurable function from X to another measurable space $(Y, \mathcal{B})$. The *pushforward of the measure* $\mathbb{P}$ under f, denoted by $f_*\mathbb{P}$, is a new measure on $(Y, \mathcal{B})$ defined by:

$$(f_*\mathbb{P})(B) = \mathbb{P}(f^{-1}(B)) \quad \text{for all} \quad B \in \mathcal{B}.$$

In other words, the measure of a set $B \subset Y$ under the pushforward measure $f_*\mathbb{P}$ is the measure of its preimage $f^{-1}(B) \subset X$ under the original measure $\mathbb{P}$.

B ASIDE ON DISTRIBUTIONAL ASYMMETRY

Given random variables $\mathbf{w}, \mathbf{w}' \in M$ and $\mathbf{y} \in M'$ such that $\mathbf{w} \overset{d}{=} f(\mathbf{w}', \mathbf{y})$ for some measurable function $f : M \times M' \to M$, we are motivated by the question: does the statistical independence of $\mathbf{w}$ and $\mathbf{y}$ imply the functional independence of $\mathbf{w}$ and $\mathbf{y}$? More formally, we want to know if the following holds

$$\mathbf{w} \perp\!\!\!\perp \mathbf{y} \implies \exists \hat{f} : M \to M \text{ s.t. } f(\cdot, y) = \hat{f} \quad \forall y \in M' \tag{22}$$

At first glance, the answer might seem to be yes. But suppose that $\mathbf{w}$ has a standard Gaussian distribution on $\mathbb{R}^2$, $\mathbf{y}$ has a uniform distribution over $[0, \pi]$ and $f(w, y) := R_y(w)$, where R_y denotes a clockwise rotation about the origin in $\mathbb{R}^2$. Then it is easy to see that the distribution $p(\mathbf{w} \mid \mathbf{y} = y)$ is a 2D standard Gaussian for all y, so $\mathbf{w} \perp\!\!\!\perp \mathbf{y}$, but certainly f is not constant with respect to its second argument.

One way to guarantee that statistical independence implies function independence is by assuming that f is smooth and that $p(\mathbf{w}')$ cannot be preserved by a "smoothly varying" family of diffeomorphisms, which we will show below.

Proposition B.1. *Suppose the stabilizer of the distribution $\mathbb{P}_{\mathbf{w}'}$ is totally disconnected in the diffeomorphism group* $\mathrm{Diff}(M)$*. Then the condition in Eq.* (22) *holds.*

Proof. First, note that the diffeomorphism group $\mathrm{Diff}(M)$ can act on the space of probability distributions on M in a natural way. For any diffeomorphism $f \in \mathrm{Diff}(M)$ and any probability distribution $\mathbb{P}$ on M, the pushforward of $\mathbb{P}$ by f, denoted by $f_*\mathbb{P}$, is a new probability distribution on M. We can thus easily check that the group action $f \cdot \mathbb{P} := f_*\mathbb{P}$ is well-defined.

Next, for any metric d on M we can equip $\mathrm{Diff}(M)$ with the following metric, where for $f, g \in \mathrm{Diff}(M)$ we have $d_\infty(f, g) := \sup_{x \in M} d(f(x), g(x))$. A set $S \subseteq \mathrm{Diff}(M)$ is said to be *totally disconnected* if there is no $\epsilon > 0$ and $f \in \mathrm{Diff}(M)$ such that $B(f, \epsilon) = \{g \in \mathrm{Diff}(M) : d_\infty(f, g) < \epsilon\} \subseteq S$. Intuitively, this means there is no smoothly varying family of diffeomorphisms in S, however small.

Finally, consider the set of diffeomorphisms $S = \{f(\cdot, y) : y \in M'\} \subseteq \mathrm{Diff}(M)$. By the smoothness of f we can see that S is connected in $\mathrm{Diff}(M)$. Since the stabilizer of the distribution $\mathbb{P}_{\mathbf{w}'}$ is totally disconnected in the diffeomorphism group $\mathrm{Diff}(M)$, $\mathbf{w} \perp\!\!\!\perp \mathbf{y}$ implies that

$$f(\cdot, y)_* \mathbb{P}_{\mathbf{w}'} = \mathbb{P}_{\mathbf{w}} \quad \forall y \in M' \tag{23}$$

which means that S is contained a left coset of the stabilzer of $\mathbb{P}_{\mathbf{w}'}$, and therefore must be totally disconnected. But since S is connected this means that it is a singleton set, and hence there exists $\hat{f} \in \mathrm{Diff}(M)$ such that $f(\cdot, y) = \hat{f}$ for all $y \in M'$. □

C PROOFS

Theorem 3.1 Under the following assumptions on the hypothesis class of parameters Θ, latent causal model with parameters $\theta^\star \in \Theta$ is identifiable up to a causal model abstraction θ with directed acyclic causal graph $\mathcal{G} = \mathcal{G}^\star / \mathcal{P}(\sigma(\mathbf{nd}(\mathcal{I}^\star)))$

1. *Faithfulness of the causal graph*: Let $\mathcal{G}$ be a perfect map for the distribution of $p(\mathbf{z})$, meaning that it encapsulates all the conditional independences.
2. *Absolute continuity of latent distributions*: Let $\mathcal{E}_i \cong \tilde{\mathcal{E}}_i \cong \mathcal{Z}_i$, let f_i, $f_i^{(S)}$ be continuously differentiable for all i and S, and let $p_{\boldsymbol{\varepsilon}}$ and $p_{\tilde{\boldsymbol{\varepsilon}}}$ be absolutely continuous.
3. *Smoothness of mixing function*: Let g be a diffeomorphism.

Proof. Suppose we have $\theta' \in \Theta$ such that $p_{\theta'}(\mathbf{x}, \tilde{\mathbf{x}}) = p_{\theta^\star}(\mathbf{x}, \tilde{\mathbf{x}})$. Denote the measures representing the distributions of $p_{\theta'}(\mathbf{z}, \tilde{\mathbf{z}})$ and $p_{\theta^\star}(\mathbf{z}, \tilde{\mathbf{z}})$ as μ and ν respectively. Denote their marginals $p_{\theta'}(\mathbf{z})$ and $p_{\theta^\star}(\mathbf{z})$ as μ_0 and ν_0 respectively. Let $h := (g')^{-1} \circ g^\star$, and write $h^{\otimes 2} : (z, \tilde{z}) \mapsto (h(z), h(\tilde{z}))$. Then by definition of the data generating process

$$\mu = h_*^{\otimes 2} \nu \tag{24}$$

Finite Mixtures conditioned on Non-Descendants We can decompose these distributions into finite mixtures, so that we have

$$\mu = \sum_{M \in \mathbf{nd}(\mathcal{I}')} \alpha_M \mu_M \tag{25}$$

$$\nu = \sum_{N \in \mathbf{nd}(\mathcal{I}^\star)} \beta_N \nu_N \tag{26}$$

where μ_M and ν_N denote the distributions $p_{\theta'}(\mathbf{z}, \tilde{\mathbf{z}} \mid \mathbf{nd}(\boldsymbol{\iota}) = M)$ and $p_{\theta^\star}(\mathbf{z}, \tilde{\mathbf{z}} \mid \mathbf{nd}(\boldsymbol{\iota}) = N)$, and α_M and β_N denote the positive constants $p_{\theta'}(\mathbf{nd}(\boldsymbol{\iota}) = M)$ and $p_{\theta^\star}(\mathbf{nd}(\boldsymbol{\iota}) = N)$.

Distributions of Latent Differences By a slight abuse of notation define the following operator on both $\mathcal{Z}^\star \times \mathcal{Z}^\star$ and $\mathcal{Z}' \times \mathcal{Z}'$.

$$\Delta : (\mathbf{z}, \tilde{\mathbf{z}}) \mapsto \mathbf{z} - \tilde{\mathbf{z}} \tag{27}$$

For any $N \in \mathbf{nd}(\mathcal{I}^\star)$ note that the support of $\Delta_* \nu_N$ is restricted to the subspace $\{\mathbf{0}_N\} \times \mathcal{Z}^\star_{N^c}$, and moreover N is the maximal subset $S \subseteq V(\mathcal{G})$ such that the support of $\Delta_* \nu_N$ can be restricted to the subspace $\{\mathbf{0}_S\} \times \mathcal{Z}^\star_{S^c}$.

So for distinct $N_1, N_2 \in \mathbf{nd}(\mathcal{I}^\star)$, $N_1 \cap N_2$ is the maximal subset $S \subseteq V(\mathcal{G}^\star)$ such that the union of the supports of $\Delta_* \nu_{N_1}$ and $\Delta_* \nu_{N_2}$ can be restricted to the subspace $\{\mathbf{0}_S\} \times \mathcal{Z}^\star_{S^c}$, and $N_1 \cup N_2$ is the maximal subset $S \subseteq V(\mathcal{G}^\star)$ such that the intersection of the supports of $\Delta_* \nu_{N_1}$ and $\Delta_* \nu_{N_2}$ can be restricted to the subspace $\{\mathbf{0}_S\} \times \mathcal{Z}^\star_{S^c}$.

Similarly, for distinct $M_1, M_2 \in \mathbf{nd}(\mathcal{I}')$, $M_1 \cap M_2$ is the maximal subset $S \subseteq V(\mathcal{G}')$ such that the union of the supports of $\Delta_* \mu_{M_1}$ and $\Delta_* \mu_{M_2}$ can be restricted to the subspace $\{\mathbf{0}_S\} \times \mathcal{Z}'_{S^c}$, and $M_1 \cup M_2$ is the maximal subset $S \subseteq V(\mathcal{G}')$ such that the intersection of the supports of $\Delta_* \mu_{M_1}$ and $\Delta_* \mu_{M_2}$ can be restricted to the subspace $\{\mathbf{0}_S\} \times \mathcal{Z}'_{S^c}$.

Separating Non-Descendant Mixtures For all $N \in \mathbf{nd}(\mathcal{I}^\star)$ we can let $\varphi(N)$ be the maximal subset $M \in \mathbf{nd}(\mathcal{I}')$ such that $h(\mathrm{supp}(\nu_N)) \subseteq \{\mathbf{0}_M\} \times \mathcal{Z}'_{M^c}$, so that this defines a bijection

$$\varphi : \mathbf{nd}(\mathcal{I}^\star) \to \mathbf{nd}(\mathcal{I}') \tag{28}$$

such that for all $N \in \mathbf{nd}(\mathcal{I}^\star)$

$$\alpha_{\varphi(N)} = \beta_N \tag{29}$$
$$\mu_{\varphi(N)} = h_*^{\otimes 2}\nu_N \tag{30}$$

Disentangling Non-Descendant Sets The projection of ν onto $\Delta^{-1}(\{\mathbf{0}_N\} \times \mathcal{Z}^\star_{N^c})$ consists only of components $\nu_{N'}$ such that $N' \supseteq N$. So under ν, given $\mathbf{z}_N = \tilde{\mathbf{z}}_N$, the joint distribution of the latent variables $h(\mathbf{z})$ and $h(\tilde{\mathbf{z}})$ is represented by a corresponding mixture of components $\mu_{\varphi(N')}$ where $\varphi(N') \supseteq \varphi(N)$, so the support of each $\Delta_*\mu_{\varphi(N')}$ is restricted to $\{\mathbf{0}_{\varphi(N)}\} \times \mathcal{Z}'_{\varphi(N)^c}$. Therefore $\mathbf{z}_N = \tilde{\mathbf{z}}_N$ almost surely implies $h(\mathbf{z})_{\varphi(N)} = h(\tilde{\mathbf{z}})_{\varphi(N)}$ almost surely, meaning that $h(\mathbf{z})_{\varphi(N)}$ is conditionally independent of $\mathbf{z}_{N^c}$ given $\mathbf{z}_N$. Similarly, we can show that $\mathbf{z}_N$ is conditionally independent of $h(\mathbf{z})_{\varphi(N)^c}$ given $h(\mathbf{z})_{\varphi(N)}$. Hence there exists h_N such that

$$\mathbf{z}_{\varphi(N)} \overset{d}{=} h_N(\mathbf{z}^\star_N) \tag{31}$$

Disentangling Complements of Non-Descendants The projection of ν onto $\Delta^{-1}(\{\mathbf{0}_{N^c}\} \times \mathcal{Z}^\star_N)$ only contains components $\nu_{N'}$ if $N' \supseteq N^c$. Thus $\mathbf{z}_{N^c} = \tilde{\mathbf{z}}_{N^c}$ almost surely implies $h(\mathbf{z})_M = h(\tilde{\mathbf{z}})_M$ almost surely for all $M \in \mathbf{nd}(\mathcal{I}')$ with $\varphi^{-1}(M) \supseteq N^c$, so $h(\mathbf{z})_{\varphi(N)^c} = h(\tilde{\mathbf{z}})_{\varphi(N)^c}$ almost surely. Therefore $h(\mathbf{z})_{\varphi(N)^c}$ is conditionally independent of $\mathbf{z}_N$ given $\mathbf{z}_{N^c}$. Similarly, we can show that $\mathbf{z}_{N^c}$ is conditionally independent of $h(\mathbf{z})_{\varphi(N)}$ given $h(\mathbf{z})_{\varphi(N)^c}$. Hence there exists h_{N^c} such that

$$\mathbf{z}_{\varphi(N)^c} \overset{d}{=} h_{N^c}(\mathbf{z}^\star_{N^c}) \tag{32}$$

Disentangling Intersections If there are functions h_A and h_B identifying $\mathbf{z}^\star_A$ and $\mathbf{z}^\star_B$ then naturally their marginals would agree on $\mathcal{Z}^\star_{A\cap B}$ such that they define a new function $h_{A\cap B}$ that identifies $\mathbf{z}^\star_{A\cap B}$.

Identifying Quotient Graph Given that we have shown that we can identify all latents corresponding to complements of non-descendant sets as well as intersections, we can extend the φ such that for any $A, B \in \sigma(\mathbf{nd}(\mathcal{I}^\star))$ we have $\varphi(A^c) = \varphi(A)^c$ and $\varphi(A \cap B) = \varphi(A) \cap \varphi(B)$, and note that this is a bijection from $\sigma(\mathbf{nd}(\mathcal{I}^\star))$ to $\sigma(\mathbf{nd}(\mathcal{I}'))$. Furthermore, using $\mathcal{P}^\star$ and $\mathcal{P}'$ to denote $\mathcal{P}(\sigma(\mathbf{nd}(\mathcal{I}^\star)))$ and $\mathcal{P}(\sigma(\mathbf{nd}(\mathcal{I}')))$ respectively, we can now restrict φ to a bijection

$$\varphi : \mathcal{P}^\star \to \mathcal{P}' \tag{33}$$

This matching of partitions constitutes a graph isomorphism between the quotient graphs $\mathcal{G}^\star/\mathcal{P}^\star$ and $\mathcal{G}'/\mathcal{P}'$, which is equivalent to a graph epimorphism from $\mathcal{G}$ to $\mathcal{G}^\star/\mathcal{P}^\star$. To show this, note that any edge in the quotient graph $\mathcal{G}^\star/\mathcal{P}^\star$ from a source block A to a terminal block B is the result of causal dependency between some vertices $i \in A$ and $j \in B$, by faithfulness of the original graph. So since $\mathbf{z}_j \not\perp\!\!\!\perp \mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(j)}$, we must have $\mathbf{z}_B \not\perp\!\!\!\perp \mathbf{z}_A \mid \mathbf{z}_{\mathrm{Pa}(B)}$, which means $h_B(\mathbf{z}_B) \not\perp\!\!\!\perp h_A(\mathbf{z}_A) \mid h_{\mathrm{Pa}(B)}(\mathbf{z}_{\mathrm{Pa}(B)})$. Therefore there is an edge from $h(A)$ to $h(B)$ in $\mathcal{G}'/\mathcal{P}'$. The same of course holds for φ^{-1} and h^{-1}.

Finally, we can quickly note that $\hat{G} = \mathcal{G}^\star/\mathcal{P}^\star$ is indeed acyclic, since any cycle in $\hat{\mathcal{G}}$ implies that there exists a directed edge from the complement $V(\mathcal{G}^\star) \setminus N$ of a non-descendant set $N \in \mathbf{nd}(\mathcal{I}^\star)$ to N itself, which violates the definition of a non-descendant set.

□

Theorem C.1. *For any $N \in \mathbf{nd}(\mathcal{I}^\star)$ denote the intersection of all intervention targets with N as their non-descendant set as*

$$\pi(N) := \bigcap \{S \in \mathcal{I} : nd(S) = N\}$$

Now provided that $\pi(N)$ is a singleton set $\{i\}$ and $\mathcal{Z}^\star_i \cong \mathbb{R}$, then we can identify the latent $\mathbf{z}^\star_{\pi(N)}$ up to disentanglement, meaning that for all $\theta \in \Theta$

$$p_\theta(\mathbf{x}, \tilde{\mathbf{x}}) = p_{\theta^\star}(\mathbf{x}, \tilde{\mathbf{x}}) \implies \theta \sim_L \theta^\star \quad wrt\ \mathcal{Z}^\star_i \tag{34}$$

Proof. To show that we can additionally identify $\pi(N)$, note that for any $N \in \mathbf{nd}(\mathcal{I}^\star)$, by definition of the data generating process, there exists some deterministic function $F_{\theta^\star}$ such that

$$\tilde{\mathbf{z}}^\star_{\pi(N)^c} = F_{\theta^\star}(\mathbf{z}^\star, \tilde{\mathbf{z}}_{\pi(N)}) \tag{35}$$

Thus we can deduce

$$\tilde{\mathbf{z}}_{\pi(N)} \overset{d}{=} h(\tilde{\mathbf{z}}^\star) \tag{36}$$
$$= h(\tilde{\mathbf{z}}^\star_{\pi(N)} + F_{\theta^\star}(\mathbf{z}^\star, \tilde{\mathbf{z}}^\star_{\pi(N)}))_{\pi(N)} \tag{37}$$
$$= h(\tilde{\mathbf{z}}^\star_{\pi(N)} + F_{\theta^\star}(h^{-1}(\mathbf{z}), \tilde{\mathbf{z}}^\star_{\pi(N)}))_{\pi(N)} \tag{38}$$

Note also that the latent distribution $(\mathbf{z}, \tilde{\mathbf{z}})$ under θ satisfies $\tilde{\mathbf{z}}_{\pi(N)} \perp\!\!\!\perp \mathbf{z} \mid \mathrm{nd}(\boldsymbol{\iota}) = N$. Therefore by Proposition B.1, there exists $h_{\pi(N)}$ such that $\tilde{\mathbf{z}}_{\pi(N)} \overset{d}{=} h_{\pi(N)}(\tilde{\mathbf{z}}^\star_{\pi(N)})$, since $\tilde{\mathbf{z}}^\star_{\pi(N)})$ is a random variable with domain on $\mathbb{R}$, so the stabilizer of its distribution contains at most two elements – the identity and a reflection about a point on $\mathbb{R}$ – and thus must be totally disconnectd.

□

D EXPERIMENTS

SETUP We generate synthetic data for linear Gaussian models, where the parameter space Θ is defined as follows. For any directed acyclic graph $\mathcal{G}$ with n nodes, we let each of the nodes $i \in V(\mathcal{G})$ be equipped with the latent space $\mathbb{R}$. We let the distribution of the exogenous variables be a standard isotropic unit Gaussian $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ and furthermore define $\mathbf{z}_i = \sum_{j \in Pa_{\mathcal{G}}(i)} a_{ij}\mathbf{z}_j + \boldsymbol{\varepsilon}_i$, where the coefficients a_{ij} are sampled from a Gaussian mixture model that has two equal components with means ± 1 and standard deviation 0.25. For the intervention parameters, we sample the new exogenous variables $\tilde{\boldsymbol{\varepsilon}}_S$ from unit Gaussians again and let $\tilde{f}_i^{(S)}$ be the identity for all subsets $S \subseteq V(\mathcal{G})$ and $i \in S$. We then uniformly sample a rotation matrix $Q \in SO(n)$ with respect to the Haar measure for our mixing function g.

RESULTS We maximize the likelihood of a set of latent causal parameters θ over the parameter space Θ for linear Gaussian models as described above via gradient descent, with respect to the observed data. We validate our theory by showing that for parameters with sufficiently high likelihood, the learned encoder Q^T inverts the ground truth mixing function $Q^\star$ up to the required level of abstraction, meaning that in particular, as a result of Theorem 3.1, $Q^T Q^\star$ is approximately a block diagonal matrix where each of the blocks corresponds to an element in $\mathcal{P}(\sigma(\mathbf{nd}(\mathcal{I}^\star)))$.