Neural Causal Abstractions

Kevin Xia, Elias Bareinboim

CausalAI Lab, Columbia University
{kevinmxia, eb}@cs.columbia.edu

Abstract

The ability of humans to understand the world in terms of cause and effect relationships, as well as their ability to compress information into abstract concepts, are two hallmark features of human intelligence. These two topics have been studied in tandem under the theory of causal abstractions, but it is an open problem how to best leverage abstraction theory in real-world causal inference tasks, where the true model is not known, and limited data is available in most practical settings. In this paper, we focus on a family of causal abstractions constructed by clustering variables and their domains, redefining abstractions to be amenable to individual causal distributions. We show that such abstractions can be learned in practice using Neural Causal Models, allowing us to utilize the deep learning toolkit to solve causal tasks (identification, estimation, sampling) at different levels of abstraction granularity. Finally, we show how representation learning can be used to learn abstractions, which we apply in our experiments to scale causal inferences to high dimensional settings such as with image data.

1 Introduction

Humans understand the world around them through the use of abstract notions. Biologists can study the function of the liver without understanding the interactions between its subatomic particles studied by physicists. Economists find it more practical to consider macro-level behavior through concepts like aggregate supply and demand rather than studying the purchasing behavior of individuals. At home, we choose to interpret the object in the television as a dog or a car as opposed to a collection of photons or pixels. Humans are highly capable of learning through interacting with the environment and understanding cause and effect between different concepts. Understanding causality is considered a hallmark of human intelligence and allows humans to plan a course of action, determine blame and responsibility, and generalize across environments. It follows that the ability to abstract concepts and study them causally is a key ability expected from modern intelligent systems.

AI systems are built on a foundation of generative models, which are representations of the underlying processes from which data is collected. Standard generative models simply model some joint density of a set of variables of interest, while *causal* generative models further model distributions involving causal interventions and counterfactual relations. In this paper, we study the problem of learning a causal generative model from data. One major challenge is that data is often provided in complex low level forms (e.g., pixels), while it would be more useful in applications to focus on higher level concepts (e.g., dog or car). We would therefore like to learn a more abstract causal generative model at a higher level of granularity, while guaranteeing that the queries from the coarser model match the ground truth.

To formalize this problem, we build on the semantics of a class of generative models called structural causal models (SCMs) (Pearl 2000). An SCM $\mathcal{M}^*$ describes a collection of mechanisms and distribution over unobserved factors. Each SCM induces three qualitatively different sets of distributions related to the human concepts of "seeing" (called observational), "doing" (interventional), and "imagining" (counterfactual), collectively known as the Ladder of Causation or the Pearl Causal Hierarchy (PCH) (Pearl and Mackenzie 2018; Bareinboim et al. 2022). The PCH is a containment hierarchy in which each of these distribution sets can be put into increasingly refined layers, where observational distributions go in layer 1 ($\mathcal{L}_1$), interventional in layer 2 ($\mathcal{L}_2$), and counterfactual in layer 3 ($\mathcal{L}_3$). In typical tasks of causal inference, the goal is to obtain a quantity from a higher layer when given data only from lower layers (e.g. inferring interventional quantities from observational data). Still, it is understood that this is generally impossible without additional assumptions since higher layers are underdetermined by lower layers (Bareinboim et al. 2022; Ibeling and Icard 2020).

Generative models can often be implemented in practice as neural networks. Deep learning models have achieved promising success in a variety of applications such as computer vision (Krizhevsky, Sutskever, and Hinton 2012), speech recognition (Graves and Jaitly 2014), and game playing (Mnih et al. 2013). Many of these successes are attributed to *representation learning* (Bengio, Courville, and Vincent 2013), in which the learned representation can be thought of as an abstraction of the data. Further, there has also been growing interest in the idea of incorporating causality into deep models[1]. Our work

[1]Many successful approaches have been developed to estimate causal effects from observational data under backdoor or ignorability

leverages one such model, the Neural Causal Model (NCM), which incorporates the same causal assumptions encoded in a causal diagram to identify and estimate interventional and counterfactual distributions (Xia et al. 2021; Xia, Pan, and Bareinboim 2023). Despite the soundness of this approach in theory, current NCM-based methods face challenges when applied to complex real-world settings for various reasons: (1) optimization is difficult when scaled to high dimensions, (2) unprocessed data can come in complicated forms (e.g. images, text, etc.), and (3) the causal diagram is difficult to fully specify in some high-dimensional settings. In this work, we address these challenges by studying how representation learning and causal reasoning are related to each other and by building on this understanding to develop a neural framework for causal abstraction learning.

Existing works that study causal abstractions set a solid foundation by defining various mathematical notions of abstractions (Rubenstein et al. 2017; Beckers and Halpern 2019; Beckers, Eberhardt, and Halpern 2019). Such definitions are declarative; that is, if the lower and higher level models are given, one can use the definition to decide whether the higher level model is indeed an abstraction of the lower level one. However, neither models are available in practice, and one would want to use limited lower level data to learn a higher level causal abstraction. We will expand on the current generation of causal abstractions in two ways. First, given that the true SCM is almost never available in practice, nor entirely learnable from data, we introduce a relaxed notion of abstractions that applies on the layers of the PCH. Second, we develop algorithms to systematically obtain abstractions in practice given some structural information about the data, which can then be used for downstream inferential tasks such as causal identification, estimation, and sampling.

Fig. 1 summarizes the general problem tackled by this paper. The ground truth model $\mathcal{M}_L$ (left) is defined over low level variables $\mathbf{V}_L$ (e.g., pixels), while it may be practical to work in their high level abstract counterparts $\mathbf{V}_H$ (e.g., dog or car). $\mathcal{M}_L$ induces distributions from the three layers of the PCH (i.e. $\mathcal{L}_1^*$, $\mathcal{L}_2^*$, $\mathcal{L}_3^*$), defined over $\mathbf{V}_L$. In this work, we introduce a new type of abstraction function τ that maps distributions over $\mathbf{V}_L$ to ones over $\mathbf{V}_H$ (i.e. $\tau(\mathcal{L}_1^*), \tau(\mathcal{L}_2^*), \tau(\mathcal{L}_3^*)$). Furthermore, $\mathcal{M}_L$ is unobserved, and only limited data is given (e.g., observational data from $\mathcal{L}_1^*$). The goal is to learn a high-level SCM $\widehat{M}_H$ (right) over the high-level variables $\mathbf{V}_H$ that encodes the given causal constraints ($\mathcal{G}_\mathbb{C}$ in the figure) and matches $\mathcal{M}_L$ on the available data across τ (e.g. $\widehat{\mathcal{L}}_1 = \tau(\mathcal{L}_1^*)$). Then, we investigate when and how the resulting model $\widehat{M}_H$ can be used as a surrogate, allowing one to make interventional and counterfactual inferences about the higher layers of $\mathcal{M}_L$ through the higher layers of $\widehat{M}_H$.

More specifically, our contributions are as follows: In Sec. 2, we define a new class of abstractions based on clusters of variables (intervariable) and their domains (intravariable). Building on this new class, we define a notion of abstraction consistency on the layers of the PCH. We then show how to systematically construct an abstraction consistent with all three layers of the PCH and then relate these abstractions to existing definitions. In Sec. 3, we show how to leverage NCM machinery to perform interventional (layer 2) and counterfactual (layer 3) inferences across these abstractions when the true SCM is unavailable. In Sec. 4, we introduce a variant of the NCM that learns representations of each variable and encodes causal assumptions on the representation level, allowing us to learn abstractions even in settings where the assumption of the availability of clusters is relaxed. Experiments in Sec. 5 corroborate with the theory. All appendices, including the proofs, experimental details, further discussion, and examples, can be found in the full technical report (Xia and Bareinboim 2023).

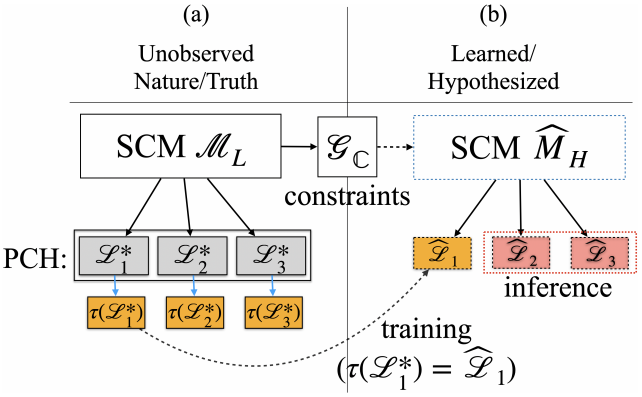


Figure 1: Overview of this paper. High-level SCM $\widehat{M}_H$ (right) is trained on available data to serve as an abstract proxy of the true, unobserved, low-level SCM $\mathcal{M}_L$ (left).

1.1 Preliminaries

We now introduce the notation and definitions used throughout the paper. We use uppercase letters (X) to denote random variables and lowercase letters (x) to denote corresponding values. Similarly, bold uppercase ($\mathbf{X}$) and lowercase ($\mathbf{x}$) letters denote sets of random variables and values respectively. We use $\mathcal{D}_X$ to denote the domain of X and $\mathcal{D}_\mathbf{X} = \mathcal{D}_{X_1} \times \cdots \times \mathcal{D}_{X_k}$ for the domain of $\mathbf{X} = \{X_1, \ldots, X_k\}$. We denote $P(\mathbf{X} = \mathbf{x})$ (often shortened to $P(\mathbf{x})$) as the probability of $\mathbf{X}$ taking the values $\mathbf{x}$ under the distribution $P(\mathbf{X})$.

We utilize the basic semantic framework of structural causal models (SCMs), as defined in (Pearl 2000, Ch. 7). An SCM $\mathcal{M}$ consists of endogenous variables $\mathbf{V}$, exogenous variables $\mathbf{U}$ with distribution $P(\mathbf{U})$, and mechanisms $\mathcal{F}$. $\mathcal{F}$ contains functions f_{V_i} (for all $V_i \in \mathbf{V}$) that map endogenous parents $\mathbf{Pa}_{V_i}$ and exogenous parents $\mathbf{U}_{V_i}$ to V_i. Each $\mathcal{M}$ induces a causal diagram $\mathcal{G}$, where every $V_i \in \mathbf{V}$ is a vertex, there is a directed arrow $(V_j \rightarrow V_i)$ for every $V_i \in \mathbf{V}$ and $V_j \in \mathbf{Pa}_{V_i}$, and there is a dashed-bidirected arrow $(V_j \dashleftarrow\dashrightarrow V_i)$ for every pair $V_i, V_j \in \mathbf{V}$ such that $\mathbf{U}_{V_i}$ and $\mathbf{U}_{V_j}$ are not independent (Markovianity is not assumed).

conditions (Shalit, Johansson, and Sontag 2017; Louizos et al. 2017; Li and Fu 2017; Johansson, Shalit, and Sontag 2016; Yao et al. 2018; Yoon, Jordon, and van der Schaar 2018; Kallus 2020; Shi, Blei, and Veitch 2019; Du et al. 2020; Guo et al. 2020), and also to answer causal queries through neural-parameterized SCMs (Kocaoglu et al. 2018; Goudet et al. 2018).

Our treatment is constrained to *recursive* SCMs, which implies acyclic causal diagrams, with finite discrete domains over endogenous variables $\mathbf{V}$.

Counterfactual quantities can be computed from SCM $\mathcal{M}$ as follows:

Definition 1 (Layer 3 Valuation)**.** An SCM $\mathcal{M}$ induces layer $\mathcal{L}_3(\mathcal{M})$, a set of distributions over $\mathbf{V}$, each with the form $P(\mathbf{Y}_*) = P(\mathbf{Y}_{1[\mathbf{x}_1]}, \mathbf{Y}_{2[\mathbf{x}_2],\ldots})$ such that

$$P^{\mathcal{M}}(\mathbf{y}_{1[\mathbf{x}_1]}, \mathbf{y}_{2[\mathbf{x}_2]}, \ldots) = \int_{\mathcal{D}_{\mathbf{U}}} \mathbf{1}\left[\mathbf{Y}_{1[\mathbf{x}_1]}(\mathbf{u}) = \mathbf{y}_1, \mathbf{Y}_{2[\mathbf{x}_2]}(\mathbf{u}) = \mathbf{y}_2, \ldots\right] dP(\mathbf{u}) \quad (1)$$

where $\mathbf{Y}_{i[\mathbf{x}_i]}(\mathbf{u})$ is evaluated under $\mathcal{F}_{\mathbf{x}_i} := \{f_{V_j} : V_j \in \mathbf{V} \setminus \mathbf{X}_i\} \cup \{f_X \leftarrow x : X \in \mathbf{X}_i\}$. $\mathcal{L}_2$ is the subset of $\mathcal{L}_3$ for which all $\mathbf{x}_i$ are equal, and $\mathcal{L}_1$ is the subset for which all $\mathbf{X}_i = \emptyset$. ■

Each $\mathbf{Y}_i$ corresponds to a set of variables in a world where the original mechanisms f_X are replaced with constants $\mathbf{x}_i$ for each $X \in \mathbf{X}_i$; this is also known as the mutilation procedure. This procedure corresponds to interventions, and we use subscripts to denote the intervening variables (e.g. $\mathbf{Y}_{\mathbf{x}}$) or subscripts with brackets when the variables are indexed (e.g. $\mathbf{Y}_{1[\mathbf{x}_1]}$). For instance, $P(y_x, y'_{x'})$ is the probability of the joint counterfactual event $Y = y$ had X been x and $Y = y'$ had X been x'.

We use the notation $\mathcal{L}_i(\mathcal{M})$ to denote the set of $\mathcal{L}_i$ distributions from $\mathcal{M}$. We use $\mathbb{Z}$ to denote a set of quantities from Layer 2 (i.e. $\mathbb{Z} = \{P(\mathbf{V}_{\mathbf{z}_k})\}_{k=1}^{\ell}$), and $\mathbb{Z}(\mathcal{M})$ denotes those same quantities induced by SCM $\mathcal{M}$ (i.e. $\mathbb{Z}(\mathcal{M}) = \{P^{\mathcal{M}}(\mathbf{V}_{\mathbf{z}_k})\}_{k=1}^{\ell}$).

This work utilizes Neural Causal Models (NCMs) for practical implementations, as follows:

Definition 2 ($\mathcal{G}$-Constrained Neural Causal Model ($\mathcal{G}$-NCM) (Xia et al. 2021, Def. 7))**.** Given a causal diagram $\mathcal{G}$, a $\mathcal{G}$-constrained Neural Causal Model ($\mathcal{G}$-NCM) $\widehat{M}(\boldsymbol{\theta})$ over $\mathbf{V}$ with parameters $\boldsymbol{\theta} = \{\theta_{V_i} : V_i \in \mathbf{V}\}$ is an SCM $\langle \widehat{\mathbf{U}}, \mathbf{V}, \widehat{\mathcal{F}}, P(\widehat{\mathbf{U}}) \rangle$ such that (1) $\widehat{\mathbf{U}} = \{\widehat{U}_{\mathbf{C}} : \mathbf{C} \in \mathbb{C}(\mathcal{G})\}$, where $\mathbb{C}(\mathcal{G})$ is the set of all maximal cliques over bidirected edges of $\mathcal{G}$; (2) $\widehat{\mathcal{F}} = \{\hat{f}_{V_i} : V_i \in \mathbf{V}\}$, where each $\hat{f}_{V_i}$ is a feedforward neural net parameterized by $\theta_{V_i} \in \boldsymbol{\theta}$ mapping $\mathbf{U}_{V_i} \cup \mathbf{Pa}_{V_i}$ to V_i for $\mathbf{U}_{V_i} = \{\widehat{U}_{\mathbf{C}} : \widehat{U}_{\mathbf{C}} \in \widehat{\mathbf{U}} \text{ s.t. } V_i \in \mathbf{C}\}$ and $\mathbf{Pa}_{V_i} = Pa_{\mathcal{G}}(V_i)$; (3) $P(\widehat{\mathbf{U}})$ is defined s.t. $\widehat{U} \sim \text{Unif}(0, 1)$ for each $\widehat{U} \in \widehat{\mathbf{U}}$. ■

In words, a $\mathcal{G}$-NCM is an SCM in which the exogenous variables $\widehat{\mathbf{U}}$ are fixed, and the mechanisms $\widehat{\mathcal{F}}$ are trainable neural nets, whose inputs are determined by the graph $\mathcal{G}$.

2 Abstractions of the Pearl Causal Hierarchy

The discussion of abstractions begins with defining causal variables. In many established causal inference tasks, it is typically assumed that there is a well-specified and known set of endogenous variables of interest $\mathbf{V}$, and nature is modeled by a collection of mechanisms that assign values to each of these variables. However, in practice, the definition of $\mathbf{V}$ may not always be clear. In particular, the variables of interest may not align with the features of the data. For example, in an economic system, perhaps data on each individual consumer is collected, but the variable of interest is an aggregate measure like gross domestic product (GDP). In image data, perhaps the pixel values are collected, but the variables of interest are related to the objects of the image, not the individual pixels.

Acknowledging that the data is not always provided in the best choice of granularity, the causal abstraction literature typically defines two sets of variables, $\mathbf{V}_L$ and $\mathbf{V}_H$, which describe the lower level and higher level settings, respectively. They are typically modeled by corresponding causal models $\mathcal{M}_L$ and $\mathcal{M}_H$, respectively.

In this section, we study on the distinction between low level variables $\mathbf{V}_L$ (e.g. pixels) and their higher level counterparts $\mathbf{V}_H$ (e.g. image) from the perspective of individual distributions of the PCH. We consider nature's underlying SCM $\mathcal{M}_L$ defined over $\mathbf{V}_L$, and the goal is to reason about the higher level variables $\mathbf{V}_H$ given data on $\mathbf{V}_L$[2]. See the full technical report (Xia and Bareinboim 2023) for detailed examples of every definition.

2.1 Constructive Abstraction Functions

The connection between $\mathbf{V}_H$ and $\mathbf{V}_L$ can be described through a mapping between their domains, $\tau : \mathcal{D}_{\mathbf{V}_L} \to \mathcal{D}_{\mathbf{V}_H}$. Here, we consider a family of abstraction functions where τ is based on clusters of the variables and values of $\mathbf{V}_L$:

Definition 3 (Inter/Intravariable Clusterings)**.** Let $\mathcal{M}$ be an SCM over variables $\mathbf{V}$.

1. A set $\mathbb{C}$ is said to be an intervariable clustering of $\mathbf{V}$ if $\mathbb{C} = \{\mathbf{C}_1, \mathbf{C}_2, \ldots \mathbf{C}_n\}$ is a partition of a subset of $\mathbf{V}$. $\mathbb{C}$ is further considered admissible w.r.t. $\mathcal{M}$ if for any $\mathbf{C}_i \in \mathbb{C}$ and any $V \in \mathbf{C}_i$, no descendent of V outside of $\mathbf{C}_i$ is an ancestor of any variable in $\mathbf{C}_i$. That is, there exists a topological ordering of the clusters of $\mathbb{C}$ relative to the functions of $\mathcal{M}$.
2. A set $\mathbb{D}$ is said to be an intravariable clustering of variables $\mathbf{V}$ w.r.t. $\mathbb{C}$ if $\mathbb{D} = \{\mathbb{D}_{\mathbf{C}_i} : \mathbf{C}_i \in \mathbb{C}\}$, where $\mathbb{D}_{\mathbf{C}_i} = \{\mathcal{D}^1_{\mathbf{C}_i}, \mathcal{D}^2_{\mathbf{C}_i}, \ldots, \mathcal{D}^{m_i}_{\mathbf{C}_i}\}$ is a partition (of size m_i) of the domains of the variables in $\mathbf{C}_i$, $\mathcal{D}_{\mathbf{C}_i}$ (recall that $\mathcal{D}_{\mathbf{C}_i}$ is the Cartesian product $\mathcal{D}_{V_1} \times \mathcal{D}_{V_2} \times \cdots \times \mathcal{D}_{V_k}$ for $\mathbf{C}_i = \{V_1, V_2, \ldots, V_k\}$, so elements of $\mathcal{D}^j_{\mathbf{C}_i}$ take the form of tuples of the value settings of $\mathbf{C}_i$). ■

In words, intervariable clusters partition the low level variables to describe each high level variable as a collection of low level variables. Intravariable clusters then describe the domains of these high level variables by partitioning the corresponding value spaces of these intervariable clusters.

Example 1. Consider a study on the effects of certain food dishes on body mass index (BMI), inspired by nutrition studies like Gamba et al. (2014). Data is collected on individuals eating at restaurants, including the restaurant (R), dish ordered (D), the amount of carbohydrates (C), fat (F), and protein (P) in the dish, and the BMI of the customer (B). That is, $\mathbf{V}_L = \{R, D, C, F, P, B\}$. One food scientist argues

[2]For concreteness, we assume that $\mathcal{M}_L$ is an SCM, but the underlying generative model can be left implicit as explained in Appendix D.1.

that any nutritional impact of the food on BMI could be abstracted based on how many calories are in each dish. One may then be tempted to cluster the variables C, F, and P together into one variable, named calories, labeled Z. This is an example of intervariable clustering.

To denote this formally, we may choose $\mathbb{C} = \{\mathbf{C}_1 = \{B\}, \mathbf{C}_2 = \{C, F, P\}, \mathbf{C}_3 = \{D\}\}$ as the intervariable clusters. In this case, B and D are placed in their own clusters, $\mathbf{C}_1$ and $\mathbf{C}_3$, respectively. C, F, and P are all clustered together into $\mathbf{C}_2$. R is not included and is abstracted away, which may be desirable if R is not relevant to the study. Collectively, $\mathbf{C}_1$, $\mathbf{C}_2$, and $\mathbf{C}_3$ form a partition of the subset of $\mathbf{V}_L$ without R. Each of the clusters of $\mathbb{C}$ will correspond to a high level variable of $\mathbf{V}_H$. In this case, for example, let Z denote the high level variable corresponding to cluster $\mathbf{C}_2$, interpreted as calories. This is shown at the top of Fig. 2 (red).

The domain of $\mathbf{C}_2$ contains every tuple of C, F, and P, but the domain of Z can be simplified. After all, the computation of calories can be specified as $Z = 4C + 9F + 4P$, which means that two sets of values, $(c_1, f_1, p_1), (c_2, f_2, p_2)$ are considered equivalent if $4c_1 + 9f_1 + 4p_1 = 4c_2 + 9f_2 + 4p_2$. This clustering of domain values is an example of intravariable clustering, shown at the bottom of Fig. 2 (blue). More formally, the intervariable clusters would be denoted $\mathbb{D} = \{\mathbb{D}_{\mathbf{C}_1}, \mathbb{D}_{\mathbf{C}_2}, \mathbb{D}_{\mathbf{C}_3}\}$, where each $\mathbb{D}_{\mathbf{C}_i}$ is a partition of $\mathcal{D}_{\mathbf{C}_i}$. In the case of $\mathbb{D}_{\mathbf{C}_2}$, we may define $\mathbb{D}_{\mathbf{C}_2} = \{\mathcal{D}^1_{\mathbf{C}_2}, \mathcal{D}^2_{\mathbf{C}_2}, \dots\}$, where each $\mathcal{D}^j_{\mathbf{C}_2}$ is a collection of tuples $(c, f, p) \in \mathcal{D}_{\mathbf{C}_2}$ corresponding to some specific value $4c + 9f + 4p$. In Fig. 2 for example, $\mathcal{D}^1_{\mathbf{C}_2} = \{(c, f, p) : 4c + 9f + 4p = 200, (c, f, p) \in \mathcal{D}_{\mathbf{C}_2}\}$. Each of the intravariable clusters correspond to a domain value of the high level variable. For example, $\mathcal{D}^1_{\mathbf{C}_2}$ corresponds to a value of $Z = 200$. ■

For the remainder of this paper, we consider settings where the intervariable clusters are admissible. Collectively, given an intervariable clustering $\mathbb{C}$ and intravariable clustering $\mathbb{D}$ of $\mathbf{V}_L$, an abstraction function τ can be defined as follows.

Definition 4 (Constructive Abstraction Function)**.** A function $\tau : \mathcal{D}_{\mathbf{V}_L} \to \mathcal{D}_{\mathbf{V}_H}$ is said to be a constructive abstraction function w.r.t. inter/intravariable clusters $\mathbb{C}$ and $\mathbb{D}$ iff

1. There exists a bijective mapping between $\mathbf{V}_H$ and $\mathbb{C}$ such that each $V_{H,i} \in \mathbf{V}_H$ corresponds to $\mathbf{C}_i \in \mathbb{C}$;
2. For each $V_{H,i} \in \mathbf{V}_H$, there exists a bijective mapping between $\mathcal{D}_{V_{H,i}}$ and $\mathbb{D}_{\mathbf{C}_i}$ such that each $v^j_{H,i} \in \mathcal{D}_{V_{H,i}}$ corresponds to $\mathcal{D}^j_{\mathbf{C}_i} \in \mathbb{D}_{\mathbf{C}_i}$; and
3. τ is composed of subfunctions $\tau_{\mathbf{C}_i}$ for each $\mathbf{C}_i \in \mathbb{C}$ such that $\mathbf{v}_H = \tau(\mathbf{v}_L) = (\tau_{\mathbf{C}_i}(\mathbf{c}_i) : \mathbf{C}_i \in \mathbb{C})$, where $\tau_{\mathbf{C}_i}(\mathbf{c}_i) = v^j_{H,i}$ if and only if $\mathbf{c}_i \in \mathcal{D}^j_{\mathbf{C}_i}$. We also apply the same notation for any $\mathbf{W}_L \subseteq \mathbf{V}_L$ such that $\mathbf{W}_L$ is a union of clusters in $\mathbb{C}$ (i.e. $\tau(\mathbf{w}_L) = (\tau_{\mathbf{C}_i}(\mathbf{c}_i) : \mathbf{C}_i \in \mathbb{C}, \mathbf{C}_i \subseteq \mathbf{W}_L)$). ■

In words, through the subfunction $\tau_{\mathbf{C}_i}$, each low level cluster $\mathbf{C}_i \in \mathbb{C}$ maps to a single high level variable $V_{H,i} \in \mathbf{V}_H$, and the value $\mathbf{c}_i \in \mathcal{D}_{\mathbf{C}_i}$ maps to a corresponding high level value $v^j_{H,i} \in \mathcal{D}_{V_{H,i}}$. Specifically, $\tau_{\mathbf{C}_i}(\mathbf{c}_i)$ maps to $v^j_{H,i}$ if $\mathbf{c}_i$ is in the intravariable cluster $\mathcal{D}^j_{\mathbf{C}_i}$. Then, the overall function τ is simply composed of the subfunctions $\tau_{\mathbf{C}_i}$. Intuitively, τ is a constructive abstraction function if it maps $\mathbf{V}_L$ to $\mathbf{V}_H$ by first grouping the variables by their corresponding intervariable cluster in $\mathbb{C}$ (red maps to yellow in Fig. 2 (top)), followed by assigning each cluster a value based on which intravariable cluster they belong in $\mathbb{D}$ (blue maps to green in Fig. 2 (bottom)). As a result, $\mathbf{V}_H$ can be interpreted such that $\mathbf{V}_H = \mathbb{C}$ and $\mathcal{D}_{V_{H,i}} = \mathbb{D}_{\mathbf{C}_i}$ for each $V_{H,i} \in \mathbf{V}_H$.

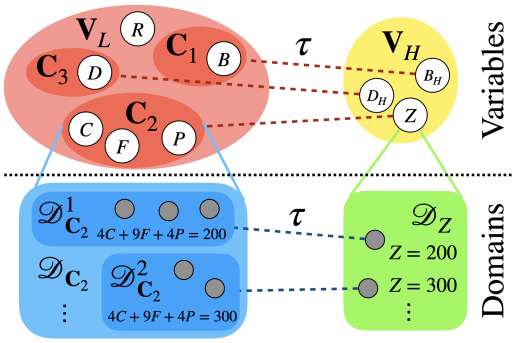


Figure 2: Example of a constructive abstraction function τ w.r.t. corresponding inter/intravariable clusters. Top (intervariable): The low-level variables, dish (D) and BMI (B), are in their own clusters while restaurant (R) is abstracted away. Carbohydrates (C), fat (F), and protein (P) are clustered together and are mapped to a single variable, calories (Z). Bottom (intravariable): The intravariable clustering for $\mathbf{C}_2 = \{C, F, P\}$ is shown. Calories Z can be computed from C, F, P using the formula $Z = 4C + 9F + 4P$. This means that the domain is partitioned such that two different values, $(c_1, f_1, p_1), (c_2, f_2, p_2)$ are in the same intravariable cluster if $4c_1 + 9f_1 + 4p_1 = 4c_2 + 9f_2 + 4p_2$.

Note that the relationship between $\mathbf{V}_L$ and $\mathbf{V}_H$ modeled by τ is not causal. Rather, the contents of $\mathbf{V}_L$ *constitute* $\mathbf{V}_H$[3]. Intuitively, two variables of $\mathbf{V}_L$ are mapped to the same intervariable cluster if they constitute the same high level variable (e.g. two pixels of the same dog), and two values are mapped to the same intravariable cluster if, from a higher level perspective, they are functionally identical (e.g. same image of the dog but rotated or cropped). In this sense, intravariable clustering can be thought of as invariances in the data.

This paper will focus on abstractions based on constructive abstraction functions τ created from intervariable and intravariable clusters. This is in contrast with the previous works on causal abstractions discussed in App. B, which leave the functional form of τ implicit. One benefit of making τ concrete is that it allows for a rigorous definition of equivalence between the distributions of a low level model and that of a high level model, as will be elaborated next.

2.2 Layer-Specific Abstractions

Ultimately, we would like to study causal properties of $\mathbf{V}_L$ through their higher level counterparts $\mathbf{V}_H$. A sensible goal is, therefore, to learn an SCM $\mathcal{M}_H$ over $\mathbf{V}_H$, which can then be queried for causal inference tasks. Still, even if $\mathbf{V}_H$ and

[3]The distinction between causal and constitutional relationships is important and is explained in detail in Appendix D.1.

$\mathbf{V}_L$ are connected through some function τ, this alone does not imply that $\mathcal{M}_H$ is an abstraction of $\mathcal{M}_L$. This is the case since the distributions over $\mathbf{V}_H$ induced by $\mathcal{M}_H$ may not have any clear connection with the distributions over $\mathbf{V}_L$.

When two SCMs are defined over the same space of variables, one can verify that they are similar if they induce the same distributions. For example, an SCM $\mathcal{M}'$ is $\mathcal{L}_2$-consistent with $\mathcal{M}$ if $\mathcal{L}_2(\mathcal{M}') = \mathcal{L}_2(\mathcal{M})$, that is, $\mathcal{M}$ and $\mathcal{M}'$ match in every interventional distribution (Bareinboim et al. 2022; Xia et al. 2021). However, when two SCMs are defined over different variable spaces, comparing their distributions is no longer well-defined. Hence, a different notion of consistency is needed to compare an SCM over $\mathbf{V}_L$ with another over $\mathbf{V}_H$ through τ.

We first note that not all low-level quantities have corresponding high-level counterparts due to the clusters. To define the low level counterfactual quantities that have high level counterparts through τ, first denote $\mathbf{Y}_{L,*}$ as a set of counterfactual variables over $\mathbf{V}_L$. That is,

$$\mathbf{Y}_{L,*} = \left(\mathbf{Y}_{L,1[\mathbf{x}_{L,1}]}, \mathbf{Y}_{L,2[\mathbf{x}_{L,2}]}, \dots\right), \tag{2}$$

where each $\mathbf{Y}_{L,i[\mathbf{x}_{L,i}]}$ corresponds to the potential outcomes of the variables $\mathbf{Y}_{L,i}$ under the intervention $\mathbf{X}_{L,i} = \mathbf{x}_{L,i}$. Each $\mathbf{Y}_{L,i}$ and $\mathbf{X}_{L,i}$ must be unions of clusters from $\mathbb{C}$ (i.e. $\mathbf{Y}_{L,i} = \bigcup_{\mathbf{C}\in\mathbb{C}'}\mathbf{C}$ for some $\mathbb{C}' \subseteq \mathbb{C}$) such that $\tau(\mathbf{Y}_{L,i})$ and $\tau(\mathbf{X}_{L,i})$ are well-defined (i.e. $\tau(\mathbf{Y}_{L,i}) = \left(\bigwedge_{\mathbf{C}\in\mathbb{C}'}\tau_{\mathbf{C}}(\mathbf{C})\right)$). For the high-level counterpart, denote

$$\mathbf{Y}_{H,*} = \tau(\mathbf{Y}_{L,*}) \tag{3}$$

$$= \left(\tau(\mathbf{Y}_{L,1[\tau(\mathbf{x}_{L,1})]}), \tau(\mathbf{Y}_{L,2[\tau(\mathbf{x}_{L,2})]}), \dots\right). \tag{4}$$

For any value $\mathbf{y}_{H,*} \in \mathcal{D}_{\mathbf{Y}_{H,*}}$, denote

$$\mathcal{D}_{\mathbf{Y}_{L,*}}(\mathbf{y}_{H,*}) = \{\mathbf{y}_{L,*} \in \mathcal{D}_{\mathbf{Y}_{L,*}} : \tau(\mathbf{y}_{L,*}) = \mathbf{y}_{H,*}\}, \tag{5}$$

that is, the set of all values $\mathbf{y}_{L,*}$ such that $\tau(\mathbf{y}_{L,*}) = \mathbf{y}_{H,*}$.

We can now define a notion of consistency relating low level counterfactual quantities to high level counterparts.

Definition 5 (Q-τ Consistency)**.** Let $\mathcal{M}_L$ and $\mathcal{M}_H$ be SCMs defined over variables $\mathbf{V}_L$ and $\mathbf{V}_H$, respectively. Let $\tau : \mathcal{D}_{\mathbf{V}_L} \rightarrow \mathcal{D}_{\mathbf{V}_H}$ be a constructive abstraction function w.r.t. clusters $\mathbb{C}$ and $\mathbb{D}$. Let

$$Q = \sum_{\mathbf{y}_{L,*}\in\mathcal{D}_{\mathbf{Y}_{L,*}}(\mathbf{y}_{H,*})} P(\mathbf{Y}_{L,*} = \mathbf{y}_{L,*}) \tag{6}$$

be a low-level Layer 3 quantity of interest (for some $\mathbf{y}_{H,*} \in \mathcal{D}_{\mathbf{Y}_{H,*}}$), as expressed in Eq. 2, and let

$$\tau(Q) = P(\mathbf{Y}_{H,*} = \mathbf{y}_{H,*}) \tag{7}$$

be its high level counterpart, as expressed in Eq. 4. We say that $\mathcal{M}_H$ is Q-τ consistent with $\mathcal{M}_L$ if

$$\sum_{\mathbf{y}_{L,*}\in\mathcal{D}_{\mathbf{Y}_{L,*}}(\mathbf{y}_{H,*})} P^{\mathcal{M}_L}(\mathbf{Y}_{L,*} = \mathbf{y}_{L,*}) = P^{\mathcal{M}_H}(\mathbf{Y}_{H,*} = \mathbf{y}_{H,*}), \tag{8}$$

that is, the value of Q induced by $\mathcal{M}_L$ is equal to the value of $\tau(Q)$ induced by $\mathcal{M}_H$[4]. Furthermore, if $\mathcal{M}_H$ is Q-τ consistent with $\mathcal{M}_L$ for all $Q \in \mathcal{L}_i(\mathcal{M}_L)$ of the form of Eq. 6, then $\mathcal{M}_H$ is said to be $\mathcal{L}_i$-τ consistent with $\mathcal{M}_L$. ■

[4]Note that the equality in Eq. 8 is consistent with the push-forward measure through τ.

Def. 5 defines the formal connection between quantities of $\mathcal{M}_L$ and $\mathcal{M}_H$. Intuitively, $\mathcal{M}_H$ can only be viewed as an abstraction of $\mathcal{M}_L$ for the quantities in which they are τ-consistent. Note that the definition naturally applies to the $\mathcal{L}_2$ case (i.e. all $\mathbf{x}_{L,i}$ are identical) and the $\mathcal{L}_1$ case (i.e. all $\mathbf{X}_{L,i} = \emptyset$). It turns out that when $\mathcal{M}_H$ is Q-τ consistent with $\mathcal{M}_L$ on all three layers of the PCH (i.e. $\mathcal{L}_3$-τ consistent), then $\mathcal{M}_H$ can be considered an abstraction of $\mathcal{M}_L$ on the SCM-level, which coincides with the definition of constructive τ-abstractions from Beckers and Halpern (2019, Def. 3.19), shown below.

Proposition 1 (Abstraction Connection)**.** *Let $\tau : \mathcal{D}_{\mathbf{V}_L} \rightarrow \mathcal{D}_{\mathbf{V}_H}$ be a constructive abstraction function (Def. 4). $\mathcal{M}_H$ is $\mathcal{L}_3$-τ consistent (Def. 5) with $\mathcal{M}_L$ if and only if there exists SCMs $\mathcal{M}'_L$ and $\mathcal{M}'_H$ s.t. $\mathcal{L}_3(\mathcal{M}'_L) = \mathcal{L}_3(\mathcal{M}_L)$, $\mathcal{L}_3(\mathcal{M}'_H) = \mathcal{L}_3(\mathcal{M}_H)$, and $\mathcal{M}'_H$ is a constructive τ-abstraction of $\mathcal{M}'_L$.* ■

All proofs are provided in Appendix A. This proposition provides the connection between the abstractions defined in this work and established definitions from previous works[5].

2.3 Algorithmic Abstraction Construction

With the abstraction function τ defined, the notion of Q-τ consistency allows for comparisons of distributions between the low level model $\mathcal{M}_L$ and the abstraction $\mathcal{M}_H$. Still, it would be desirable to be able to systematically construct $\mathcal{M}_H$ given $\mathcal{M}_L$ and τ such that $\mathcal{M}_H$ is Q-τ consistent with $\mathcal{M}_L$ for as many queries Q as possible. Moving in this direction, we first note that as a subtlety, for some cases of $\mathcal{M}_L$, there are certain choices of $\mathbb{C}$ and $\mathbb{D}$ (and corresponding τ) for which Q-τ consistency (for some queries Q) is impossible to achieve in any choice of $\mathcal{M}_H$. This phenomenon can be described formally by the following condition.

Definition 6 (Abstract Invariance Condition (AIC))**.** Let $\mathcal{M}_L = \langle \mathbf{U}_L, \mathbf{V}_L, \mathcal{F}_L, P(\mathbf{U}_L)\rangle$ be an SCM. Let $\tau : \mathcal{D}_{\mathbf{V}_L} \rightarrow \mathcal{D}_{\mathbf{V}_H}$ be a constructive abstraction function relative to $\mathbb{C}$ and $\mathbb{D}$. The SCM $\mathcal{M}_L$ is said to satisfy the abstract invariance condition (AIC) with respect to τ if, for all $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{D}_{\mathbf{V}_L}$ such that $\tau(\mathbf{v}_1) = \tau(\mathbf{v}_2)$, all $\mathbf{u} \in \mathcal{D}_{\mathbf{U}_L}$, and all $\mathbf{C}_i \in \mathbb{C}$, the following holds:

$$\begin{aligned}&\tau_{\mathbf{C}_i}\left(\left(f^L_V(\mathbf{pa}^{(1)}_V, \mathbf{u}_V) : V \in \mathbf{C}_i\right)\right)\\ &= \tau_{\mathbf{C}_i}\left(\left(f^L_V(\mathbf{pa}^{(2)}_V, \mathbf{u}_V) : V \in \mathbf{C}_i\right)\right),\end{aligned} \tag{9}$$

where $\mathbf{pa}^{(1)}_V$ and $\mathbf{pa}^{(2)}_V$ are the values corresponding to $\mathbf{v}_1$ and $\mathbf{v}_2$. Then, $\widetilde{\mathbf{pa}}_V$ is used to denote any arbitrary value s.t. $\tau(\widetilde{\mathbf{pa}}_V) = \tau(\mathbf{pa}^{(1)}_V) = \tau(\mathbf{pa}^{(2)}_V)$. ■

In words, the AIC enforces that if two low level values $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{D}_{\mathbf{V}_L}$ map to the same high level value (i.e. $\tau(\mathbf{v}_1) = \tau(\mathbf{v}_2)$), then for each cluster $\mathbf{C}_i \in \mathbb{C}$, the functions of those

[5]Note that one subtlety of this result is that it is not $\mathcal{M}_H$ that is directly a constructive τ-abstraction of $\mathcal{M}_L$, but rather their $\mathcal{L}_3$-equivalent counterparts, $\mathcal{M}'_H$ and $\mathcal{M}'_L$. Indeed, the definition of constructive τ-abstractions is stronger than $\mathcal{L}_3$-τ consistency (see proof for more details), but in tasks where we are only concerned with the layers of the PCH, this distinction is inconsequential.

clusters should map to the same value regardless of $\mathbf{U}_L$ (i.e. the outputs of $f_V^L(\mathbf{pa}_V^{(1)}, \mathbf{u}_V)$ for each $V \in \mathbf{C}_i$ should map to the same result as the outputs of $f_V^L(\mathbf{pa}_V^{(2)}, \mathbf{u}_V)$ when passed through $\tau_{\mathbf{C}_i}$). Intuitively, this implies that two values in the same intravariable cluster have the same functional effect in the higher level setting.

It turns out that the AIC describes precisely when an appropriate $\mathcal{M}_H$ exists as an abstraction of the low level model $\mathcal{M}_L$, as shown by the following result.

Proposition 2 (Abstraction Conditions)**.** *For any SCM $\mathcal{M}_L$ and constructive abstraction function τ relative to $\mathbb{C}$ and $\mathbb{D}$, there exists an SCM $\mathcal{M}_H$ over variables $\mathbf{V}_H = \tau(\mathbf{V}_L)$ such that $\mathcal{M}_H$ is $\mathcal{L}_3$-τ consistent with $\mathcal{M}_L$ if and only if there exists $\mathcal{M}'_L$ such that $\mathcal{L}_3(\mathcal{M}_L) = \mathcal{L}_3(\mathcal{M}'_L)$ and $\mathcal{M}'_L$ satisfies the abstract invariance condition with respect to τ.* ■

This critical property guarantees the existence of a high-level SCM $\mathcal{M}_H$ such that $\mathcal{L}_3$-τ consistency holds, so we will assume that the AIC holds for the rest of this work. Still, see App. D.2 for further discussion on its implications and for relaxations in cases where $\mathcal{L}_3$-τ consistency is not required.

With the notion of abstractions well-defined, we study how $\mathcal{M}_H$ can be obtained from $\mathcal{M}_L$. Interestingly, when given the admissible clusterings $\mathbb{C}$ and $\mathbb{D}$, the procedure for recovering τ and converting $\mathcal{M}_L$ to $\mathcal{M}_H$ can be done as shown in Alg. 1. Intuitively, one can obtain an abstraction $\mathcal{M}_H$ of $\mathcal{M}_L$ by first constructing the abstraction function τ using the clusterings $\mathbb{C}$ and $\mathbb{D}$ (lines 2-3), followed by designing the functions of $\mathcal{M}_H$ to wrap the original functions of $\mathcal{M}_L$ with τ (lines 4-6). This can be verified using the following result.

Proposition 3. *Let τ and $\mathcal{M}_H$ be the function and SCM obtained from running Alg. 1 on inputs $\mathcal{M}_L$, $\mathbb{C}$, and $\mathbb{D}$. Then, $\mathcal{M}_H$ is $\mathcal{L}_3$-τ consistent with $\mathcal{M}_L$.* ■

Alg. 1 can be used to systematically obtain an abstraction $\mathcal{M}_H$ of the low-level model $\mathcal{M}_L$, so long as $\mathcal{M}_L$ is provided alongside the clusters $\mathbb{C}$ and $\mathbb{D}$. Since $\mathcal{M}_L$ is almost never available in practice, the following sections show how this requirement can be relaxed.

Algorithm 1: Constructing $\mathcal{M}_H$ from $\mathcal{M}_L$.

Input : SCM $\mathcal{M}_L = \langle \mathbf{U}_L, \mathbf{V}_L, \mathcal{F}_L, P(\mathbf{U}_L) \rangle$, admissible inter/intravariable clusters $\mathbb{C}$ and $\mathbb{D}$ satisfying abstract invariance condition

Output : SCM $\mathcal{M}_H$ and $\tau : \mathcal{D}_{\mathbf{V}_H} \to \mathcal{D}_{\mathbf{V}_L}$ s.t. $\mathcal{M}_H$ is $\mathcal{L}_3$-τ consistent with $\mathcal{M}_L$

1 $\mathbf{U}_H \leftarrow \mathbf{U}_L, P(\mathbf{U}_H) \leftarrow P(\mathbf{U}_L)$
2 $\mathbf{V}_H \leftarrow \mathbb{C}, \mathcal{D}_{\mathbf{V}_H} \leftarrow \mathbb{D}$
3 $\tau \leftarrow$ `AbsFunc`$(\mathbb{C}, \mathbb{D})$ `// from Def. 4`
4 **for** $\mathbf{C}_i \in \mathbb{C}$ **do**
5 $\quad f_i^H \leftarrow \tau\left(f_V^L(\widetilde{\mathbf{pa}}_V, \mathbf{u}_V) : V \in \mathbf{C}_i\right)$
6 $\mathcal{F}_H \leftarrow \{f_i^H : \mathbf{C}_i \in \mathbb{C}\}$
7 **return** $\tau, \mathcal{M}_H = \langle \mathbf{U}_H, \mathbf{V}_H, \mathcal{F}_H, P(\mathbf{U}_H) \rangle$

3 Inferences Across Abstractions

As demonstrated by Alg. 1, converting a low level model $\mathcal{M}_L$ to a high level model $\mathcal{M}_H$ is somewhat immediate when given full observability of the underlying SCM $\mathcal{M}_L$. However, in real applications, it is rarely the case that the full specification of $\mathcal{M}_L$ is known. Typically, one will only be given partial information of $\mathcal{M}_L$ in the form of data, such as samples of the observational distribution $P(\mathbf{V}_L)$. The question we investigate in this section is: is it still possible to "learn" some $\mathcal{M}_H$ given the observed data?

We first note the impossibility result described by the Causal Hierarchy Theorem (CHT) (Bareinboim et al. 2022, Thm. 1), which states that a model trained to match another SCM on lower layers of the causal hierarchy (e.g. $\mathcal{L}_1$) will likely not match on higher layers (e.g. $\mathcal{L}_2$ or $\mathcal{L}_3$). Naturally, the same is true when it comes to inferring causal quantities across abstractions. One may be tempted to believe that $\mathcal{M}_H$ can be learned given $\mathcal{L}_1$ data from $\mathcal{M}_L$ by instantiating some expressive parametric model $\widehat{M}_H$ on $\mathbf{V}_H$, and then training $\widehat{M}_H$ on $P(\mathbf{V}_H) = P(\tau(\mathbf{V}_L))$ such that $\widehat{M}_H$ is $\mathcal{L}_1$-τ consistent with $\mathcal{M}_L$. Unfortunately, such a model $\widehat{M}_H$ will fail to generalize because even under perfect training, $\widehat{M}_H$ is not guaranteed to be $\mathcal{L}_2$-τ (or $\mathcal{L}_3$-τ) consistent with $\mathcal{M}_L$. This means that any causal quantities induced by $\widehat{M}_H$ will likely bear no relationship with causal quantities induced by $\mathcal{M}_L$. We show this in the next result.

Proposition 4 (Abstract Causal Hierarchy Theorem (Informal))**.** *Given constructive abstraction function $\tau : \mathcal{D}_{\mathbf{V}_H} \to \mathcal{D}_{\mathbf{V}_L}$, even if $\mathcal{M}_H$ is $\mathcal{L}_i$-τ consistent with $\mathcal{M}_L$, $\mathcal{M}_H$ will almost never be $\mathcal{L}_j$-τ consistent with $\mathcal{M}_L$ for $j > i$.* ■

In words, matching across abstractions on lower layers does not guarantee the same will hold for higher layers. The consequence of this result is that causal assumptions will be necessary to make progress. Given this necessity, one type of assumption prevalent throughout causal inference literature is the availability of a causal diagram (Pearl 1995), a graphical structure that qualitatively describes the functional relationships between variables. This assumption is a weaker requirement than assuming the availability of the entire SCM, since it does not require full detail of the generating mechanisms and exogenous distributions. Still, it has been shown that having the causal diagram allows certain inferences across layers, determined through the causal identification problem (Pearl 2000; Bareinboim and Pearl 2016).

In the context of abstractions however, specifying the causal diagram for the true model $\mathcal{M}_L$ requires describing the relationships between every low-level variable in $\mathbf{V}_L$. This is still unrealistic in many practical settings since there are typically too many low-level variables (e.g. 128×128 pixels in an image) to expect a description of the relationship between every pair, and many of these relationships may not even be well-defined in a causal manner. Instead, it may be more reasonable to specify a causal diagram over $\mathbf{V}_H$ (or intervariable clusters $\mathbb{C}$). When $|\mathbf{V}_H| \ll |\mathbf{V}_L|$, the amount of information required is reduced, and the causal relationships between variables may be more clear given that the higher-level variables tend to be more explainable. The causal diagram over $\mathbf{V}_H$ can be viewed as a graphical abstraction of the causal diagram over $\mathbf{V}_L$. The relationship can be formalized through the concept of cluster causal diagrams

(C-DAGs), introduced in Anand et al. (2023).

Definition 7 (Cluster Causal Diagram (C-DAG) (Anand et al. 2023, Def. 1))**.** Given a causal diagram $\mathcal{G} = \langle \mathbf{V}, \mathbf{E} \rangle$ and an admissible clustering $\mathbb{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_k\}$ of $\mathbf{V}$, construct a graph $\mathcal{G}_{\mathbb{C}} = \langle \mathbb{C}, \mathbf{E}_{\mathbb{C}} \rangle$ over $\mathbb{C}$ with a set of edges $\mathbf{E}_{\mathbb{C}}$ defined as follows:

1. A directed edge $\mathbf{C}_i \to \mathbf{C}_j$ is in $\mathbf{E}_{\mathbb{C}}$ if there exists some $V_i \in \mathbf{C}_i$ and $V_j \in \mathbf{C}_j$ such that $V_i \to V_j$ is an edge in $\mathbf{E}$.
2. A dashed bidirected edge $\mathbf{C}_i \dashleftarrow\dashrightarrow \mathbf{C}_j$ is in $\mathbf{E}_{\mathbb{C}}$ if there exists some $V_i \in \mathbf{C}_i$ and $V_j \in \mathbf{C}_j$ such that $V_i \dashleftarrow\dashrightarrow V_j$ is an edge in $\mathbf{E}$. ■

In words, the nodes of the C-DAG $\mathcal{G}_{\mathbb{C}}$ simply correspond to the clusters of $\mathbb{C}$, and edges connect clusters $\mathbf{C}_i$ and $\mathbf{C}_j$ if they connect some $V_i \in \mathbf{C}_i$ and $V_j \in \mathbf{C}_j$ in the original causal diagram $\mathcal{G}$. Interestingly, the C-DAG definition aligns with the concept of intervariable clusters, providing a way for encoding constraints in the smaller space of $\mathbf{V}_H$. Following the nutrition study in Ex. 1, Fig. 3 shows the corresponding causal diagram $\mathcal{G}$ (left) and the simpler C-DAG $\mathcal{G}_{\mathbb{C}}$ (right). With the constraints of $\mathcal{G}_{\mathbb{C}}$, we now introduce a notion of identification across abstractions to determine precisely which queries can be inferred.

Definition 8 (Abstract Identification)**.** Let $\tau : \mathcal{D}_{\mathbf{V}_H} \to \mathcal{D}_{\mathbf{V}_L}$ be a constructive abstraction function. Consider C-DAG $\mathcal{G}_{\mathbb{C}}$, and let $\mathbb{Z} = \{P(\mathbf{V}_{L[\mathbf{z}_k]})\}_{k=1}^{\ell}$ be a collection of available interventional (or observational if $\mathbf{Z}_k = \emptyset$) distributions over $\mathbf{V}_L$. Let Ω_L and Ω_H be the space of SCMs defined over $\mathbf{V}_L$ and $\mathbf{V}_H$, respectively, and let $\Omega_L(\mathcal{G}_{\mathbb{C}})$ and $\Omega_H(\mathcal{G}_{\mathbb{C}})$ be their corresponding subsets that induce C-DAG $\mathcal{G}_{\mathbb{C}}$. We say that query Q is τ-ID from $\mathcal{G}_{\mathbb{C}}$ and $\mathbb{Z}$ iff for every $\mathcal{M}_L \in \Omega_L(\mathcal{G}_{\mathbb{C}}), \mathcal{M}_H \in \Omega_H(\mathcal{G}_{\mathbb{C}})$ such that $\mathcal{M}_H$ is $\mathbb{Z}$-τ consistent with $\mathcal{M}_L$, $\mathcal{M}_H$ is also Q-τ consistent with $\mathcal{M}_L$. ■

This definition establishes a notion of identification between two different spaces of SCMs, Ω_L and Ω_H, that are connected through τ. In words, τ-identifiability implies that in every pair of SCMs $\mathcal{M}_L$ over $\mathbf{V}_L$ and $\mathcal{M}_H$ over $\mathbf{V}_H$, "matching" in graph $\mathcal{G}_{\mathbb{C}}$ and data $\mathbb{Z}$ implies a match in query Q. Since $\mathcal{M}_L$ and $\mathcal{M}_H$ are defined over different spaces of variables, the term "match" has some nuance. Specifically, "matching" in $\mathcal{G}_{\mathbb{C}}$ implies that $\mathcal{G}_{\mathbb{C}}$ is a C-DAG for $\mathcal{M}_L$ and is a causal diagram for $\mathcal{M}_H$. "Matching" in $\mathbb{Z}$ (resp. Q) implies that $\mathcal{M}_H$ is $\mathbb{Z}$-τ consistent (resp. Q-τ consistent) with $\mathcal{M}_L$. On the other hand, τ-nonidentifiability implies that there exist a pair of models $\mathcal{M}_L$ over $\mathbf{V}_L$ and $\mathcal{M}_H$ over $\mathbf{V}_H$ such that $\mathcal{M}_L$ and $\mathcal{M}_H$ match in both $\mathcal{G}_{\mathbb{C}}$ and $\mathbb{Z}$ yet still do not match in Q. This means that despite the constraints added through the C-DAG $\mathcal{G}_{\mathbb{C}}$, there are still queries that cannot be inferred across τ due to nonidentifiability. This is more acute when there is a large amount of unobserved confounding.

The definition of τ-ID provides rigorous semantics to answer whether a query can be inferred across abstractions. The next step is to establish an approach to determine τ-ID when given the available data and graph. For this purpose, one fundamental result is that the notion of τ-ID is actually equivalent to classical identification in the higher level space.

Theorem 1 (Dual Abstract ID)**.** *Q is τ-ID from $\mathcal{G}_{\mathbb{C}}$ and $\mathbb{Z}$ if and only if $\tau(Q)$ is ID from $\mathcal{G}_{\mathbb{C}}$ and $\tau(\mathbb{Z})$.* ■

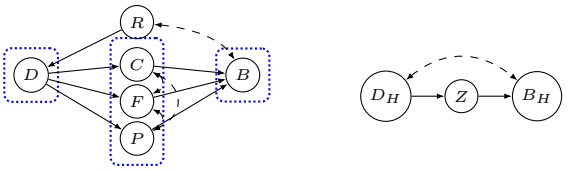


Figure 3: The causal diagram $\mathcal{G}$ over variables $\mathbf{V}_L$ for the nutrition study in Ex. 1 is on the left. Clusters $\mathbb{C} = \{D_H = \{D\}, Z = \{C, F, P\}, B_H = \{B\}\}$ are outlined in blue. The corresponding C-DAG $\mathcal{G}_{\mathbb{C}}$ is on the right.

Algorithm 2: NeuralAbstractID – Identifying and estimating queries across abstractions using NCMs.

Input : query Q, $\mathcal{L}_2$ datasets $\mathbb{Z}(\mathcal{M}_L)$, C-DAG $\mathcal{G}_{\mathbb{C}}$, and admissible inter/intravariable clusters $\mathbb{C}$ and $\mathbb{D}$ satisfying AIC

Output : $Q(\mathcal{M}_L)$ if identifiable, `FAIL` otherwise.

$\mathbf{V}_H \leftarrow \mathbb{C}, \mathcal{D}_{\mathbf{V}_H} \leftarrow \mathbb{D}$

$\tau \leftarrow \texttt{AbsFunc}(\mathbb{C}, \mathbb{D})$ `// from Def. 4`

$\widehat{M} \leftarrow \texttt{NCM}(\mathbf{V}_H, \mathcal{G}_{\mathbb{C}})$ `// from Def. 2`

$\boldsymbol{\theta}^*_{\min} \leftarrow \arg\min_{\boldsymbol{\theta}} \tau(Q)(\widehat{M}(\boldsymbol{\theta}))$ s.t. $\tau(\mathbb{Z})(\widehat{M}(\boldsymbol{\theta})) = \tau(\mathbb{Z}(\mathcal{M}_L))$

$\boldsymbol{\theta}^*_{\max} \leftarrow \arg\max_{\boldsymbol{\theta}} \tau(Q)(\widehat{M}(\boldsymbol{\theta}))$ s.t. $\tau(\mathbb{Z})(\widehat{M}(\boldsymbol{\theta})) = \tau(\mathbb{Z}(\mathcal{M}_L))$

if $\tau(Q)(\widehat{M}(\boldsymbol{\theta}^*_{\min})) \neq \tau(Q)(\widehat{M}(\boldsymbol{\theta}^*_{\max}))$ **then**

  return `FAIL`

else

  return $\tau(Q)(\widehat{M}(\boldsymbol{\theta}^*_{\min}))$ `// choose min or max arbitrarily`

This result is powerful since it implies that inferences can be made about the low level space by using existing results in the high level space. Notably, since our goal is to learn a higher level SCM $\mathcal{M}_H$ to make inferences about $\mathcal{M}_L$, we can build on the machinery of Neural Causal Models (NCMs) (Xia et al. 2021). NCMs allow one to take the graph $\mathcal{G}_{\mathbb{C}}$ as an inductive bias (a $\mathcal{G}_{\mathbb{C}}$-NCM as described in Def. 2), and they can leverage gradient methods to fit any SCM within the constrained space. Indeed, identification in NCMs can be shown to be equivalent to classical identification when considering models of the same granularity (Xia, Pan, and Bareinboim 2023, Thm. 3). When combined with Thm. 1, this implies the following result.

Corollary 1 (Abstract ID with NCMs)**.** *Q is τ-ID from $\mathcal{G}_{\mathbb{C}}$ and $\mathbb{Z}$ iff $\tau(Q)$ is Neural-ID from $\widehat{\Omega}(\mathcal{G}_{\mathbb{C}})$ and $\tau(\mathbb{Z})$. Moreover, if it is ID, then Q can be computed by computing $\tau(Q)$ by definition from any $\mathcal{G}_{\mathbb{C}}$-NCM $\widehat{M}$ that is $\tau(\mathbb{Z})$-consistent.* ■

In words, determining τ-ID is equivalent to determining neural identification (identification in the space of NCMs) on the space of $\mathbf{V}_H$. Further, to compute Q in the identifiable case, $\tau(Q)$ can be queried from any $\mathcal{G}_{\mathbb{C}}$-NCM $\widehat{M}$ that is $\tau(\mathbb{Z})$-consistent. Corol. 1 implies that we can perform causal

identification and estimation across abstractions using the NeuralID algorithm (Xia, Pan, and Bareinboim 2023, Alg. 1) on the high level space. This procedure is shown in Alg. 2. First, τ is constructed as described in Def. 4 given the clusters. Then, a $\mathcal{G}_{\mathbb{C}}$-NCM is constructed over high-level variables $\mathbf{V}_H$. Two parameterizations of the NCM are created. Both are optimized to fit the transformed data $\tau(\mathbb{Z})$, but one is optimized to maximize the transformed query $\tau(Q)$ while the other is optimized to minimize it. If both parameterizations return the same result, then it must be the true value of the query; otherwise, the query is not identifiable.

To implement this algorithm in practice, we leverage the GAN-NCM introduced in Xia, Pan, and Bareinboim (2023); see details in App. C. Alg. 2 is sound and complete for solving the abstract identification problem, as shown below.

Corollary 2 (Soundness and Completeness)**.** *Let $\mathcal{M}_L$ be the low-level SCM, $\mathbb{C}$ and $\mathbb{D}$ be inter/intravariable clusters of $\mathbf{V}_L$, $\mathcal{G}_{\mathbb{C}}$ be a C-DAG, Q be a query, and $\widehat{Q}$ be the result from running Alg. 2 with inputs $\mathbb{Z}(\mathcal{M}_L) > 0$, $\mathbb{C}$, $\mathbb{D}$, $\mathcal{G}_{\mathbb{C}}$, and Q. Then, Q is τ-ID from $\mathcal{G}_{\mathbb{C}}$ and $\mathbb{Z}$ if and only if $\widehat{Q}$ is not* `FAIL`*. Moreover, if $\widehat{Q}$ is not* `FAIL`*, then $\widehat{Q} = Q(\mathcal{M}_L)$.* ■

While Alg. 2 solves the abstract ID problem, the consequences of the results in this section are more general. Notably, if Q is indeed τ-ID (which can be verified through Alg. 2), the algorithm produces a neural model $\widehat{M}$ that serves as a proxy SCM that is Q-τ consistent with the true model $\mathcal{M}_L$. Such a model could serve as a generative model of the distribution Q, which has many uses. The samples generated from such a model could be used to estimate the query, or, in more complex settings such as with image data, it may be desirable to simply have novel generated samples consistent with the causal invariances embedded in the system.

4 Representations in Learning Abstractions

In many applications, the choice of intervariable clusters $\mathbb{C}$ is natural and can be made in tandem when deciding the assumptions of the C-DAG $\mathcal{G}_{\mathbb{C}}$[6]. However, fully specifying the intravariable clusters $\mathbb{D}$ is quite challenging when working with high-dimensional data like image data. Doing so would require an enumeration of every possible image along with some label designating each one to a cluster. In this section, we investigate the problem of learning abstractions when the intravariable clusters $\mathbb{D}$ are left unspecified.

While coarser clusters tend to be better in practice due to the dimensionality reduction, the theory in this paper can be applied for any choice of $\mathbb{D}$ so long as the AIC (Def. 6) holds. Hence, a possible constraint when learning $\mathbb{D}$ is to find a set of clusters such that the AIC is not violated. To this effect, the following result can be leveraged.

Proposition 5. *$\mathcal{M}_L$ is guaranteed to satisfy the AIC w.r.t. τ iff $\mathbb{D}_{\mathbf{C}_i} = \mathcal{D}_{\mathbf{C}_i}$ for all $\mathbf{C}_i \in \mathbb{C}$.* ■

In other words, this means that Alg. 2 can be applied in any case where $\tau_{\mathbf{C}_i}$ is a bijective mapping between $\mathcal{D}_{\mathbf{C}_i}$ and $\mathcal{D}_{V_{H,i}}$. Also implied by this result is that, without additional information, one cannot choose any coarser clustering without potentially violating the AIC[7]. While this choice of $\mathbb{D}$ does not reduce the size of the abstracted space, this means that we are not restricted to the original space of $\mathbf{V}_L$ and can choose any $\mathbf{V}_H$ with the same cardinality. In practice, this means that we can choose the option for $\mathbf{V}_H$ that is the most beneficial for our task. Leveraging this insight, we introduce the representational NCM.

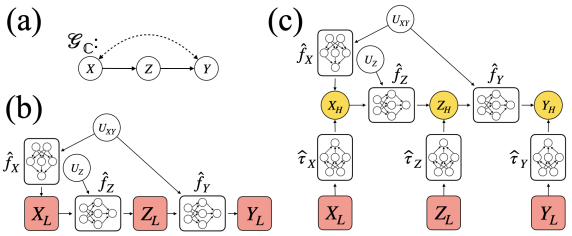


Figure 4: Example comparison between (b) the $\mathcal{G}_{\mathbb{C}}$-NCM and (c) $\mathcal{G}_{\mathbb{C}}$-RNCM, with $\mathcal{G}_{\mathbb{C}}$ shown in (a). Functions of the NCM directly output values of the lower level variables (grouped by clusters in $\mathbb{C}$), while functions of the RNCM output values of their higher level counterparts, mapped by $\widehat{\tau}$.

Definition 9 (Representational NCM (RNCM))**.** A representational NCM (RNCM) is a tuple $\langle \widehat{\tau}, \widehat{M} \rangle$, where $\widehat{\tau}(\mathbf{v}_L; \boldsymbol{\theta}_\tau)$ is a function parameterized by $\boldsymbol{\theta}_\tau$ mapping from $\mathbf{V}_L$ to $\mathbf{V}_H$, and $\widehat{M}$ is an NCM defined over $\mathbf{V}_H$. A $\mathcal{G}_{\mathbb{C}}$-constrained RNCM ($\mathcal{G}_{\mathbb{C}}$-RNCM) is an RNCM $\langle \widehat{\tau}, \widehat{M} \rangle$ such that $\widehat{\tau}$ is composed of subfunctions $\widehat{\tau}_{\mathbf{C}_i}$ for each $\mathbf{C}_i \in \mathbb{C}$ (each with its own parameters $\boldsymbol{\theta}_{\tau_{\mathbf{C}_i}}$), and $\widehat{M}$ is a $\mathcal{G}_{\mathbb{C}}$-NCM. ■

In an RNCM, the abstraction function $\widehat{\tau}$ is a trainable parameterized function, and the NCM $\widehat{M}$ is trained over the resulting space mapped by $\widehat{\tau}$. Fig. 4 shows an example illustrating the difference between the RNCM and a standard NCM. Training can be done in a two step procedure, where first $\widehat{\tau}$ is trained to map to an optimal task-specific space, and then $\widehat{M}$ can be trained on $\widehat{\tau}(\mathbf{V}_L)$ (e.g. through Alg. 2). To enforce bijectivity between $\mathcal{D}_{\mathbf{C}_i}$ and $\mathcal{D}_{V_{H,i}}$, as suggested by Prop. 5, one can train $\widehat{\tau}$ in an autoencoder-like setup (Kramer 1991; Kingma and Welling 2014) with a reconstruction loss. $\widehat{\tau}$ can be thought of as a function mapping to a representation space, making this approach amenable to the wide developments of the representation learning literature (Bengio, Courville, and Vincent 2013). We empirically demonstrate this approach below in the experiment of Sec. 5.2.

5 Experiments

In this section, we empirically evaluate the effects of utilizing abstractions in causal inference tasks. Details of data-generating models and architectures can be found in Appendix C. Implementation code is publicly available at https://github.com/CausalAILab/NeuralCausalAbstractions.

[6] Still, see App. D.1 for best practices on how to choose or learn intervariable clusters.

[7] In many cases, there may be additional information in the form of invariances (e.g. rotational invariance in image data). In such cases, this information can be leveraged to learn coarser clusters. See Appendix D.3 for more details.

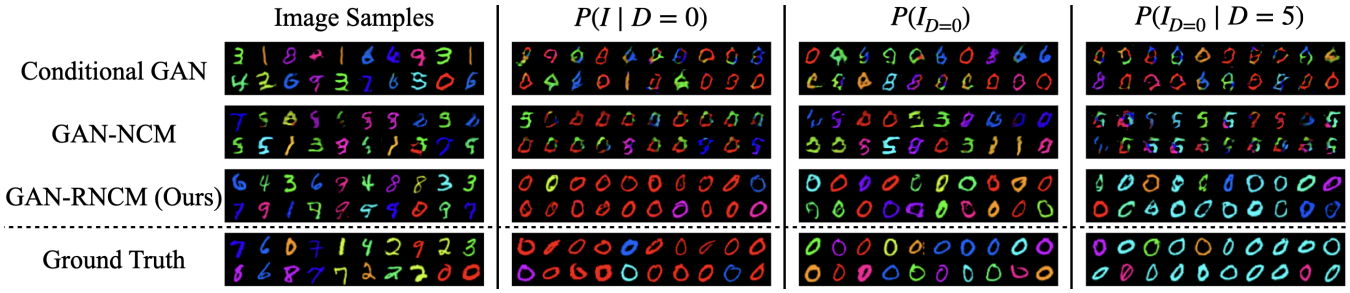


Figure 5: Colored MNIST results. Samples from various causal queries (top) are collected from competing approaches (left).

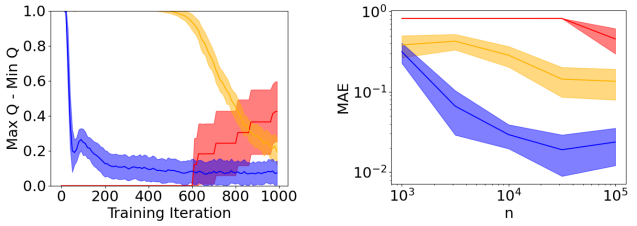
5.1 Nutritional Study

We perform the toy study on nutrition depicted in Ex. 1. Since a BMI of 25 or over is considered overweight, the goal is to identify and estimate the query $Q = P(B_{D=d} \geq 25)$ (the causal effect of diet on weight) using Alg. 2. R and D are 32-dimensional one-hot vectors, and the others are real-valued, so the query may be difficult to answer given such high-dimensional variables. Instead, it may be more effective to work in an abstract space with the proposed intervariable clusters $\mathbb{C} = \{D_H = \{D\}, Z = \{C, F, P\}, B_H = \{B\}\}$. The original graph $\mathcal{G}$ and corresponding C-DAG $\mathcal{G}_{\mathbb{C}}$ are shown in Fig. 3. We are also given intravariable clusters $\mathbb{D}$ such that all values of D_H, Z, and B_H are clustered into binary categories. Specifically, $D_H = 1$ denotes unhealthy dishes, $Z = 1$ denotes high calorie count, and $B_H = 1$ denotes an overweight BMI (≥ 25). We compare the effectiveness identifying and estimating Q with NCMs in both the original setting under $\mathbf{V}_L$, and in the abstracted setting of $\mathbf{V}_H$ computed using the constructive abstraction function τ defined on $\mathbb{C}$ and $\mathbb{D}$. The results are shown in Fig. 6. Since Q is identifiable, the gap between the max and min queries computed in Alg. 2 are expected to be as small as possible. As shown in Fig. 6a, the proposed approach converges quickly while others fail to close the gap. Fig. 6b also shows that the proposed approach can estimate Q with significantly lower error.

5.2 Colored MNIST Digits

We evaluate the RNCM in a high-dimensional image dataset of colorized MNIST (Deng 2012) digits. Each image (I) has a corresponding digit (D) and color (C) label, and their relationships are shown in the C-DAG $\mathcal{G}_{\mathbb{C}}$ in Fig. 7a. Color and digit are highly correlated (e.g. 0s are typically red, while 5s are cyan), as shown in Fig. 7b. We evaluate three approaches in the task of sampling images from causal queries. The first approach is a naïve conditional GAN that does not take causality into account. The second is a standard GAN-NCM as described in Xia, Pan, and Bareinboim (2023). The third is our approach described in Sec. 4, a representational NCM also implemented as a GAN, called GAN-RNCM.

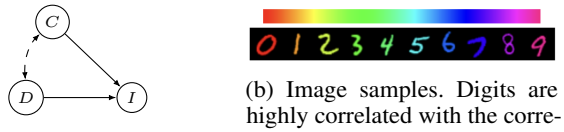
Samples of the results are shown in Fig. 5. All models are capable of producing digit images, as shown in the first column. The second column illustrates $P(I \mid D = 0)$, the images conditioned on digit $= 0$. Many red 0s are expected since most 0s are red in the dataset. The third column illustrates the interventional query $P(I_{D=0})$, the images with



(a) Gaps between max and min query across 1000 training iterations when running Alg. 2.

(b) Mean absolute error (MAE) v. dataset size (in log-log scale) for query estimation.

Figure 6: Results of the nutrition experiment. Our approach (blue) is compared with a GAN-NCM trained on raw data (red) and one trained on normalized data (yellow).



(a) $\mathcal{G}_{\mathbb{C}}$ for Colored MNIST.

(b) Image samples. Digits are highly correlated with the corresponding gradient color.

Figure 7: Colored MNIST Experimental Setup

digits forced to be 0 through intervention. As interventions ignore spurious correlations, 0s of all colors are expected. Finally, the fourth column illustrates the counterfactual query $P(I_{D=0} \mid D = 5)$, indicating what the digits would have looked like had they been 0, given that they were originally 5. Since 5s tend to be cyan, the samples are expected to be 0s that retain the cyan color of the 5s. In all cases, GAN-RNCM produces results close to the expected, while the other approaches have difficulty disentangling color from digit.

6 Conclusions

Through the notions of inter/intravariable clusters and Q-τ consistency, we introduced a new family of abstractions allowing analysis on individual PCH distributions. We proved that ID across abstractions is equivalent to classical ID (Thm. 1) and provided a sound and complete algorithm to perform such inferences (Alg. 2). We provided a relaxation of intravariable clusters leveraging representation learning through the RNCM (Def. 9). Finally, we demonstrated empirically that abstractions are vital in high-dimensional settings.

Acknowledgements

This research was supported in part by the NSF, ONR, AFOSR, DARPA, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

References

Anand, T. V.; Ribeiro, A. H.; Tian, J.; and Bareinboim, E. 2023. Causal Effect Identification in Cluster DAGs. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*. AAAI Press.

Bareinboim, E.; Correa, J. D.; Ibeling, D.; and Icard, T. 2022. On Pearl's Hierarchy and the Foundations of Causal Inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, 507–556. New York, NY, USA: Association for Computing Machinery, 1st edition.

Bareinboim, E.; and Pearl, J. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27): 7345–7352.

Beckers, S.; Eberhardt, F.; and Halpern, J. Y. 2019. Approximate Causal Abstraction. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*.

Beckers, S.; and Halpern, J. Y. 2019. Abstracting Causal Models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press. ISBN 978-1-57735-809-1.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8): 1798–1828.

Deng, L. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6): 141–142.

Du, X.; Sun, L.; Duivesteijn, W.; Nikolaev, A.; and Pechenizkiy, M. 2020. Adversarial Balancing-based Representation Learning for Causal Effect Inference with Observational Data. arXiv:1904.13335.

Gamba, R.; Schuchter, J.; Rutt, C.; and Seto, E. 2014. Measuring the Food Environment and its Effects on Obesity in the United States: A Systematic Review of Methods and Results. *Journal of Community Health*, 40(3): 464–475.

Goudet, O.; Kalainathan, D.; Caillou, P.; Guyon, I.; Lopez-Paz, D.; and Sebag, M. 2018. Learning functional causal models with generative neural networks. In *Explainable and interpretable models in computer vision and machine learning*, 39–80. Springer.

Graves, A.; and Jaitly, N. 2014. Towards End-To-End Speech Recognition with Recurrent Neural Networks. In Xing, E. P.; and Jebara, T., eds., *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, 1764–1772. Bejing, China: PMLR.

Guo, R.; Cheng, L.; Li, J.; Hahn, P. R.; and Liu, H. 2020. A Survey of Learning Causality with Data. *ACM Computing Surveys*, 53(4): 1–37.

Ibeling, D.; and Icard, T. 2020. Probabilistic reasoning across the causal hierarchy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10170–10177.

Johansson, F. D.; Shalit, U.; and Sontag, D. 2016. Learning Representations for Counterfactual Inference. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, 3020–3029. JMLR.org.

Kallus, N. 2020. DeepMatch: Balancing Deep Covariate Representations for Causal Inference Using Adversarial Training. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 5067–5077. PMLR.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In Bengio, Y.; and LeCun, Y., eds., *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Kocaoglu, M.; Snyder, C.; Dimakis, A. G.; and Vishwanath, S. 2018. CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. In *International Conference on Learning Representations*.

Kramer, M. A. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2): 233–243.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F.; Burges, C. J. C.; Bottou, L.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems*, volume 25, 1097–1105. Curran Associates, Inc.

Li, S.; and Fu, Y. 2017. Matching on Balanced Nonlinear Representations for Treatment Effects Estimation. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30, 929–939. Curran Associates, Inc.

Louizos, C.; Shalit, U.; Mooij, J.; Sontag, D.; Zemel, R.; and Welling, M. 2017. Causal Effect Inference with Deep Latent-Variable Models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6449–6459. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing Atari With Deep Reinforcement Learning. In *NIPS Deep Learning Workshop*.

Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4): 669–688.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. New York, NY, USA: Cambridge University Press, 2nd edition.

Pearl, J.; and Mackenzie, D. 2018. *The Book of Why*. New York: Basic Books.

Rubenstein, P. K.; Weichwald, S.; Bongers, S.; Mooij, J.; Janzing, D.; Grosse-Wentrup, M.; and Schölkopf, B. 2017. Causal Consistency of Structural Equation Models. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 3076–3085. International Convention Centre, Sydney, Australia: PMLR.

Shi, C.; Blei, D. M.; and Veitch, V. 2019. Adapting Neural Networks for the Estimation of Treatment Effects. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2503–2513.

Xia, K.; and Bareinboim, E. 2023. Neural Causal Abstractions. Technical Report R-101, Columbia University, Department of Computer Science, New York.

Xia, K.; Lee, K.-Z.; Bengio, Y.; and Bareinboim, E. 2021. The Causal-Neural Connection: Expressiveness, Learnability, and Inference. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 10823–10836. Curran Associates, Inc.

Xia, K.; Pan, Y.; and Bareinboim, E. 2023. Neural Causal Models for Counterfactual Identification and Estimation. In *Proceedings of the 11th International Conference on Learning Representations (ICLR-23)*.

Yao, L.; Li, S.; Li, Y.; Huai, M.; Gao, J.; and Zhang, A. 2018. Representation Learning for Treatment Effect Estimation from Observational Data. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31, 2633–2643. Curran Associates, Inc.

Yoon, J.; Jordon, J.; and van der Schaar, M. 2018. GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets. In *International Conference on Learning Representations*.