Quantifying Consistency and Information Loss for Causal Abstraction Learning

Fabio Massimo Zennaro, **Paolo Turrini** and **Theodoros Damoulas**
University of Warwick,Coventry, United Kingdom
{fabio.zennaro, p.turrini, t.damoulas}@warwick.ac.uk,

Abstract

Structural causal models provide a formalism to express causal relations between variables of interest. Models and variables can represent a system at different levels of abstraction, whereby relations may be coarsened and refined according to the need of a modeller. However, switching between different levels of abstraction requires evaluating a trade-off between the consistency and the information loss among different models. In this paper we introduce a family of interventional measures that an agent may use to evaluate such a trade-off. We consider four measures suited for different tasks, analyze their properties, and propose algorithms to evaluate and learn causal abstractions. Finally, we illustrate the flexibility of our setup by empirically showing how different measures and algorithmic choices may lead to different abstractions.

1 Introduction

In his IJCAI 2022 keynote talk, Judea Pearl argued that reasoning with causality is among the biggest challenges of modern AI. *Structural causal models* (SCM) [Pearl, 2009] were introduced to address this challenge as a rigorous graph-based formalism explicitly encoding causal relations between variables. Analyzing causes and effects, however, implicitly requires the assumption of a given *level of abstraction* (LA) at which variables are observed. The same system may indeed be modelled at different LAs depending on the resolution a modeller or a decision-making agent are considering. Choosing the appropriate scale for modelling, analyzing and controlling a system is a fundamental challenge in science and decision-making, with instances ranging from ecological multi-scale modelling [Levin, 1992] to neural population coarse-graining [Schmutz *et al.*, 2020].

Within the context of causal models, evaluating which abstraction is the "correct" one is a nontrivial challenge in itself. A few approaches have been proposed in the literature to express relationships of abstraction between SCMs [Rubenstein *et al.*, 2017; Beckers and Halpern, 2019; Rischel, 2020], with some of them offering quantitative ways to assess the degree of approximation (or error) introduced by an abstraction in terms of interventional consistency (IC) [Rischel, 2020; Rischel and Weichwald, 2021]. However, understanding which LA is the optimal one requires balancing potentially conflicting properties. For instance, while assessing among different candidate abstractions, an agent may well be concerned with information loss when probability distributions over fine-grained random variables are compressed to fit coarser variables. Although abstractions are a key part of causal reasoning, we still lack a theoretical framework specifying flexible measures of approximation and how to use them to learn optimal abstractions.

Contributions. In this paper we bridge the gap in the literature by introducing and analyzing measures of abstraction approximation that capture consistency and information loss in causal abstraction. Concretely: (i) we define a family of interventional measures of abstraction approximation (of which IC is a particular case) and analyze their properties; (ii) we introduce algorithms for evaluating and learning causal abstractions based on our properties; (iii) we illustrate how our measures are sensitive to their parameters and how they can capture different aspects of an abstraction. All in all, we provide a grounded set of measures that can be used for abstraction learning according to the specific aims at hand.

Related Literature. Abstraction is a fundamental component for general intelligence [Mitchell, 2021] and an important strategy for managing complexity and allowing artificial agents to achieve superhuman performance in challenging games [Kroer and Sandholm, 2018; Sandholm, 2015]. Coarsening of Bayesian networks or learning structures with bounded complexity in order to improve computational efficiency and representation has been studied, for instance, in [Chang and Fung, 1990; Elidan and Gould, 2008].

In causal reasoning, the problem of abstraction between SCMs was first introduced by [Rubenstein *et al.*, 2017]; the original framework was then further developed by [Beckers and Halpern, 2019; Beckers *et al.*, 2020], and it has recently found application for interpretability [Geiger *et al.*, 2021]. Our work builds on the results of [Rischel, 2020], who provides a framework for evaluating IC grounded in category theory [Spivak, 2014]. The problem of learning abstractions within this framework has been practically studied in [Zennaro *et al.*, 2023] on synthetic and real-world data. Here, we generalise the IC approach across multiple dimensions and study it both analytically and algorithmically.

Work on evaluating abstraction naturally connects to work on causal representation learning [Chalupka *et al.*, 2017] and work on defining and measuring emergence [Hoel, 2017; Eberhardt and Lee, 2022], too; these deal with causal systems at different LAs without the formalism of SCMs.

Paper Structure. In Sec. 2 we introduce mathematical preliminaries for our framework. In Sec. 3 we provide measures for abstraction approximation, and in Sec. 4 we study their properties. We rely on these properties to discuss how to learn abstractions in Sec. 5. Finally, we illustrate our contributions empirically in Sec. 6 and conclude in Sec. 7.

2 Preliminaries

We define here SCMs, interventions, and abstractions.

Definition 1 (SCM [Pearl, 2009])**.** *A structural causal model (SCM) $\mathcal{M}$ is a tuple $\langle \mathcal{X}, \mathcal{U}, \mathcal{F}, P(\mathcal{U}) \rangle$ with an underlying directed acyclic graph (DAG) $\mathcal{G}_{\mathcal{M}}$ where:*

- *$\mathcal{X}$ is a finite set of N endogenous random variables X_i; each variable X_i is associated with a finite set $\mathcal{M}[X_i] = \{x_1, ..., x_M\}$ of outcomes; sets of variables are associated with the Cartesian product of the sets.*
- *$\mathcal{U}$ is a finite set of exogenous random variables.*
- *$\mathcal{F}$ is a finite set of N measurable structural functions f_i, one for each endogenous variable X_i; a structural function $f_i : \mathcal{M}[Pa(X_i)] \times \mathcal{M}[\mathcal{U}] \to \mathcal{M}[X_i]$, where $Pa(X_i) \subseteq \mathcal{X}$ denotes parents, defines deterministically the value of the random variable X_i.*
- *$P(\mathcal{U})$ is a distribution over the exogenous variables.*

Following [Rischel, 2020], we assume we have a finite number of endogenous variables, each one with a finite domain. Our definition implies that we are working with semi-Markovian SCMs. Notice, also, that the DAG structure implies a partial ordering $X_i \prec X_j$ of the endogenous variables according to reachability. SCMs allow us to study causality via interventions:

Definition 2 (Intervention [Pearl, 2009])**.** *Given a SCM $\mathcal{M}$, a variable-value pair $(\mathbf{X}, \mathbf{x})$ such that for each X_i in the set $\mathbf{X} \subseteq \mathcal{X}$ there is a $x_i \in \mathcal{M}[X_i]$ in the set $\mathbf{x}$, an intervention $\iota : do(\mathbf{X} = \mathbf{x})$ is an operator that generates a new SCM $\mathcal{M}_\iota$ by replacing the structural functions f_i with the constants x_i.*

Let us now consider a base (low-level) model $\mathcal{M}$ and an abstracted (high-level) model $\mathcal{M}'$ and define abstraction:

Definition 3 (Abstraction [Rischel, 2020])**.** *An abstraction $\boldsymbol{\alpha}$ from SCM $\mathcal{M}$ to SCM $\mathcal{M}'$ is a tuple $\langle R, a, \alpha_i \rangle$ where:*

- *$R \subseteq \mathcal{X}$ defines a subset of relevant variables in $\mathcal{M}$;*
- *$a : R \to \mathcal{X}'$ is a surjective function mapping relevant variables R in $\mathcal{M}$ to variables in $\mathcal{M}'$;*
- *$\alpha_i : \mathcal{M}[a^{-1}(X_i')] \to \mathcal{M}'[X_i']$ is a collection of surjective functions, one for each variable in $\mathcal{M}'$, mapping the outcomes of variable(s) $a^{-1}(X_i')$ onto the outcomes of variable X_i'.*

Example 1. *Consider two laboratories, Lab A and Lab B, having defined two models of lung cancer: the model in Fig. 1a and the model in Fig. 1b, respectively. A formal abstraction $\boldsymbol{\alpha}$ from the model of Lab A to the model of Lab B may be defined by choosing $R = \{Sm, LC\}$, $a : \{Sm \mapsto Sm', LC \mapsto Hea'\}$, and setting $\alpha_{Sm'}, \alpha_{Hea'}$ to identities. A complete definition of the SCMs and the abstraction is provided in App.* **??**.

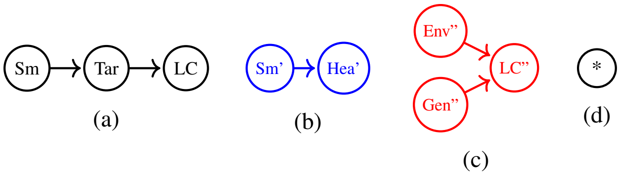


Figure 1: DAGs of four SCMs modelling causal relationships between smoking (Sm), tar deposits (Tar), genetic factors (Gen), environmental factors (Env), health index (Hea), and lung cancer (LC).

A first measure to assess quantitatively an abstraction was suggested in [Rischel, 2020] in the form of IC between results obtained on the low-level and high-level.

Definition 4 (IC error wrt an intervention [Rischel, 2020])**.** *Given an abstraction $\boldsymbol{\alpha}$ from $\mathcal{M}$ to $\mathcal{M}'$, and given two disjoint sets $\mathbf{X}', \mathbf{Y}' \in \mathcal{X}'$, we define the IC error wrt the interventional distribution $P(\mathbf{Y}'|do(\mathbf{X}'))$ by considering the following diagram:*

$$\begin{tikzcd}
\mathcal{M}[\mathrm{a}^{-1}(\mathbf{X}')] \arrow[r, "\mu_{do(a^{-1}(\mathbf{X}'))}"] \arrow[d, "\alpha_{\mathbf{X}'}"'] & \mathcal{M}[\mathrm{a}^{-1}(\mathbf{Y}')] \arrow[d, "\alpha_{\mathbf{Y}'}"'] \\
\mathcal{M}'[\mathbf{X}'] \arrow[r, "\nu_{do(\mathbf{X}')}"] & \mathcal{M}'[\mathbf{Y}']
\end{tikzcd}$$

where $\mu_{do()}$, $\nu_{do()}$ are the stochastic functions computed in the respective interventional models, and by evaluating:

$$E_{IC}(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}') = D_{JSD}(\alpha_{\mathbf{Y}'} \circ \mu_{do(a^{-1}(\mathbf{X}'))}, \nu_{do(\mathbf{X}')} \circ \alpha_{\mathbf{X}'}), \quad (1)$$

where D_{JSD} is the Jensen-Shannon distance (JSD) and $\circ$ denotes function composition.

The definition of JSD is recalled in App. **??**. This diagram evaluates the discrepancy between performing first an intervention on the low level and then abstracting, or abstracting first and then performing an intervention on the high level. Beyond a causal reading, the diagram has an algebraic and categorical reading. Algebraically, every node is associated with the set of outcomes of the given variable(s), while the stochastic functions and abstractions on the arrows can be expressed as matrices; this means that arrow composition can be efficiently computed by matrix multiplication; see App. **??** for further details. Categorically, the diagram has a rigorous meaning in the category `FinStoch` enriched in `Met` [Rischel, 2020; Fritz, 2020].

We can extend the notion of IC error from an intervention to the abstraction itself:

Definition 5 (Overall IC error [Rischel, 2020])**.** *Given an abstraction $\boldsymbol{\alpha}$ from $\mathcal{M}$ to $\mathcal{M}'$, the overall IC error is:*

$$e_{IC}(\boldsymbol{\alpha}) = \sup_{(\mathbf{X}', \mathbf{Y}') \in \mathcal{J}} E_{IC}(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}'), \quad (2)$$

where $\mathcal{J}$ is the set of all non-empty disjoint pairs $(\mathbf{X}', \mathbf{Y}') \in \mathscr{P}(\mathcal{X}') \times \mathscr{P}(\mathcal{X}')$, with $\mathscr{P}()$ being the powerset.

Example 2. *Given the abstraction in Ex. 1, Lab A can measure the overall IC error $e_{IC}(\boldsymbol{\alpha}) \approx 0.385$ wrt the set of pairs $\mathcal{J} = \{(Sm', Hea')\}$. See App.* **??** *for the exact computation.*

3 Measures of Abstraction Approximation

While IC provides a measure of interventional alignment between the low- and high-level model, this measure may not properly capture the priorities of an agent and weight candidate abstractions accordingly.

Example 3. *Suppose Lab A with its model in Fig. 1a is looking for an abstraction. Since abstraction $\boldsymbol{\alpha}$ to the model of Fig. 1b has IC error $e_{IC}(\boldsymbol{\alpha}) \approx 0.385$ as in Ex. 2, Lab A may consider an abstraction $\boldsymbol{\beta}$ to the singleton model in Fig. 1d. By definition, $e_{IC}(\boldsymbol{\beta}) = 0$. However, despite the lower error, this abstraction may be problematic given that the singleton model trivially carries no information.*

To enrich our understanding of abstraction approximation, we use the definition of IC as a template for a generalizing the notion of *error wrt an intervention* and *overall error*.

Definition 6 (Error wrt an intervention)**.** *Given an abstraction $\boldsymbol{\alpha}$ from $\mathcal{M}$ to $\mathcal{M}'$, and given two disjoint sets $\mathbf{X}', \mathbf{Y}' \in \mathcal{X}'$, we define the error wrt the interventional distribution $P(\mathbf{Y}'|do(\mathbf{X}'))$ by considering the following diagram:*

$$\begin{tikzcd}
\mathcal{M}[\mathrm{a}^{-1}(\mathbf{X}')] \arrow[r, "\mu_{do(a^{-1}(\mathbf{X}'))}"] \arrow[d, bend right, "\alpha_{\mathbf{X}'}"'] & \mathcal{M}[\mathrm{a}^{-1}(\mathbf{Y}')] \arrow[d, bend right, "\alpha_{\mathbf{Y}'}"'] \\
\mathcal{M}'[\mathbf{X}'] \arrow[r, "\nu_{do(\mathbf{X}')}"] \arrow[u, bend right, "\alpha^+_{\mathbf{X}'}"'] & \mathcal{M}'[\mathbf{Y}'] \arrow[u, bend right, "\alpha^+_{\mathbf{Y}'}"']
\end{tikzcd}$$

where $\alpha^+_{\mathbf{X}'}$ is the pseudo-inverse of $\alpha_{\mathbf{X}'}$, and by evaluating:

$$E_I(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}') = D(p, q), \tag{3}$$

where D is a distance, and p, q are two paths in the above diagram with the same start and end points (as in Tab. 1).

Definition 7 (Overall error)**.** *Given an abstraction $\boldsymbol{\alpha}$ from $\mathcal{M}$ to $\mathcal{M}'$, the overall interventional error is:*

$$e_I(\boldsymbol{\alpha}) = \underset{(\mathbf{X}', \mathbf{Y}') \in \mathcal{J}}{f} E_I(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}'), \tag{4}$$

where f is an aggregation function $f : \mathbb{R}^{|\mathcal{J}|} \to \mathbb{R}$, and $\mathcal{J}$ is an assessment set containing pairs $(\mathbf{X}', \mathbf{Y}')$.

The definitions are generic and depend on five parameters:

- D, the distance measure to assess an individual error $E_I(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}')$. We rely on JSD which guarantees abstraction compositionality for IC [Rischel, 2020] and for other interventional measures (see Sec. 4). Alternative measures guaranteeing the same property, such as p-Wasserstein distances, had already been suggested in [Rischel, 2020; Rischel and Weichwald, 2021] and could be used in place of JSD.
- $\alpha^+_{\mathbf{X}'}$, the pseudo-inverse of $\alpha_{\mathbf{X}'}$. We will adopt the standard Moore-Penrose inverse, whose definition and relevant properties are discussed in App. **??**.
- p, q, the paths to be considered on the diagram in Definition 6; possible choices are discussed in Sec. 3.1.
- f, the function aggregating the individual errors $E_I(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}')$. We adopt a supremum aggregation function, which provides a robust, worst-case scenario, evaluation of the error. Other functions may be considered in different scenarios, such as mean or weighted average. Moreover, ensembling theory [Dietterich, 2000; Rokach, 2019] may help defininig desirable properties for aggregation and analyzing correlations between errors.
- $\mathcal{J}$, the assessment set to evaluate $e_I(\boldsymbol{\alpha})$; we will discuss possible choices in detail in Sec. 3.2.

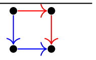
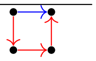
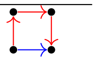
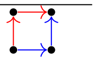
	p	q	diagram
IC	$\nu_{do()} \circ \alpha_{\mathbf{X}'}$	$\alpha_{\mathbf{Y}'} \circ \mu_{do()}$	
IIL	$\mu_{do()}$	$\alpha^+_{\mathbf{Y}'} \circ \nu_{do()} \circ \alpha_{\mathbf{X}'}$	
ISIL	$\nu_{do()}$	$\alpha_{\mathbf{Y}'} \circ \mu_{do()} \circ \alpha^+_{\mathbf{X}'}$	
ISC	$\alpha^+_{\mathbf{Y}'} \circ \nu_{do()}$	$\mu_{do()} \circ \alpha^+_{\mathbf{X}'}$	

Table 1: Interventional measures wrt different paths.

We focus on paths and assessment sets as they allow for the definition of new measures and crucially contribute to the meaning and the computational complexity of our measures.

3.1 Paths

In the diagram of Definition 6 the two horizontal arrows, $\mu_{do()}$ and $\nu_{do()}$, have a defined directionality; they capture causal mechanisms, and their inverse would represent anti-causal relationships, which are of no interest in this context. The two vertical arrows, however, may be considered in both directions: it may be desirable to move between a low-level and a high-level model in both ways. This naturally leads to the definition of four different measures, listed in Tab. 1, each one having relevance in light of specific settings and downstream tasks an agent may face. We present these measures by defining the paths p, q and illustrating their use on an example using a lung cancer model from [Guyon *et al.*, 2008].

Interventional Consistency (IC)

Discussed above, IC takes $p = \nu_{do()} \circ \alpha_{\mathbf{X}'}$ and $q = \alpha_{\mathbf{Y}'} \circ \mu_{do()}$. IC considers interventions on the low-level model and it evaluates the agreement, via abstraction, between results computed at the low-level and high-level. Low IC would be relevant when a downstream task depends on $P(\alpha_{\mathbf{Y}'}(\mathbf{Y})|do(\mathbf{X}))$, with $\alpha_{\mathbf{Y}'}$ expressing a coarsening of $P(\mathbf{Y}|do(\mathbf{X}))$, while an agent wants to rely on the higher-level distribution $P(\mathbf{Y}'|do(\alpha_{\mathbf{X}'}(\mathbf{X}))$.

Example 4. *Consider the health scenario in Fig. 2, where Lab A has developed a large lung cancer SCM (black) and*

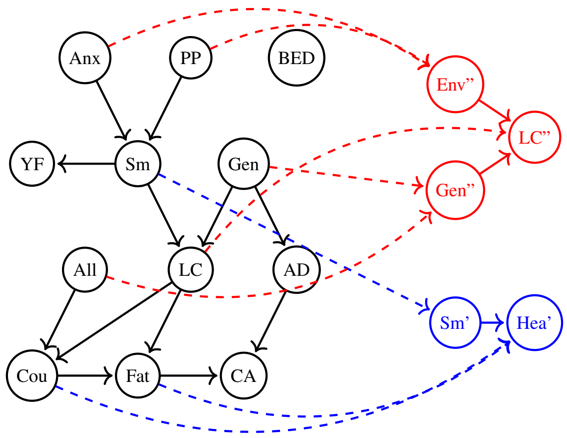


Figure 2: Base model (black), health model (blue) of Lab B from Fig. 1b with the corresponding abstraction (dashed blue line), and lung cancer model (red) of Lab C from Fig. 1c with the corresponding abstraction (dashed red line). Acronyms of variables are explained in Tab. **??**.

Lab B has produced a simpler model (blue). Lab A is performing a smoking experiment, estimating a health index from variables in its model, and feeding the result to a decision-making module to discriminate whether further exams are necessary. Lab B wants to evaluate patients in its own model and provide results to the same decision-making module, concerned for patients to be processed equivalently in both models. The original downstream task depends on $P(\alpha_{\mathbf{Y}'}(\mathbf{Y})|do(\mathbf{X}))$, where $\mathbf{Y}$ is a set of variables aggregated in a health index via $\alpha_{\mathbf{Y}'}$ and $\mathbf{X}$ is smoking; Lab B will then be concerned with evaluating IC, thus reducing the discrepancy between $P(\alpha_{\mathbf{Y}'}(\mathbf{Y})|do(\mathbf{X}))$ and $P(\mathbf{Y}'|do(\alpha_{\mathbf{X}'}(\mathbf{X}))$. The lower the IC, the more accurate the results (if the downstream decision-making module had been tuned on the low-level model) and the fairer the output (if we are concerned with the same proportions of patients being forwarded to further analysis by Lab A and Lab B).

It is worth remarking that, in this context we deal with interventional fairness, and not counterfactual fairness [Kusner *et al.*, 2017]; that is, fairness holds on the distributional level (models at different LAs produce the same interventional distributions), not the individual level (outcomes for an individual are not necessarily identical on different LAs).

Interventional Information Loss (IIL)

IIL takes $p = \mu_{do()}$ and $q = \alpha^{+}_{\mathbf{Y}'} \circ \nu_{do()} \circ \alpha_{\mathbf{X}'}$. IIL also considers interventions on the low-level model and it evaluates the information lost by working through the high-level model. Low IIL would be relevant when a downstream task depends on $P(\mathbf{Y}|do(\mathbf{X}))$ but, because of constraints, an agent can not compute this quantity directly on the low-level model but it has to rely on coarser estimations obtained through the high-level model, thus ending to use $P(\alpha^{+}_{\mathbf{Y}'}(\mathbf{Y}')|do(\alpha_{\mathbf{X}'}(\mathbf{X}))$. In algebraic terms, IIL evaluates how well a low-level mechanisms may be decomposed into two abstractions and a high-level mechanism.

Example 5. *Consider the same health scenario as in Ex. 4. Lab A is performing a smoking experiment, estimating a health index, and predicting the probability of car accidents. The downstream task can be described as depending on $P(\mathbf{Y}|do(\mathbf{X}))$, with $\mathbf{Y}$ being a set of variables for a health index and $\mathbf{X}$ smoking. Since estimating $\mathbf{Y}$ in its model is (in the relative terms of the example) computationally expensive, Lab A decides to rely on the model of Lab B. As the result will be re-used by Lab A for further computations, Lab A wants to estimate IIL, thus assessing the discrepancy between the expensive-to-compute $P(\mathbf{Y}|do(\mathbf{X}))$ and the cheaper $P(\alpha^{+}_{\mathbf{Y}'}(\mathbf{Y}')|do(\alpha_{\mathbf{X}'}(\mathbf{X}))$. This will guarantee that replacing some computations in the low-level model with higher-level model computations will produce results analogous to performing the whole computation at low-level.*

Like fairness, replaceability between the original base model and the base model with a sub-part replaced holds only in an interventional, not counterfactual, sense.

Interventional Superresolution Information Loss (ISIL)

ISIL takes $p = \nu_{do()}$ and $q = \alpha_{\mathbf{Y}'} \circ \mu_{do()} \circ \alpha^{+}_{\mathbf{X}'}$. ISIL considers interventions on the high-level and it evaluates the information mismatch by working on the low-level model. Low ISIL would be relevant when a downstream task depends on $P(\mathbf{Y}'|do(\mathbf{X}'))$ but, because of constraints, an agent is requested to compute this quantity with higher precision on the low-level model, thus ending to rely on $P(\alpha_{\mathbf{Y}'}(\mathbf{Y})|do(\alpha^{+}_{\mathbf{X}'}(\mathbf{X}')))$. In a way complementary to IIL, ISIL evaluates how well a high-level mechanisms may be factored into two abstractions and a low-level mechanism.

Example 6. *Consider the lung cancer scenario in Fig. 2, where Lab A has developed a large lung cancer SCM (black) and now Lab C has produced a simpler model (red). Lab C is performing environmental manipulation, and using the result to evaluate other high-level statistics that depend on $P(\mathbf{Y}'|do(\mathbf{X}'))$, with $\mathbf{Y}'$ being lung cancer and $\mathbf{X}'$ environment. Given the sensitivity of the evaluation, Lab C wants to match the more detailed model of Lab A by optimizing for ISIL, thus minimizing the discrepancy between its approximate $P(\mathbf{Y}'|do(\mathbf{X}'))$ and the finer-grained $P(\alpha_{\mathbf{Y}'}(\mathbf{Y})|do(\alpha^{+}_{\mathbf{X}'}(\mathbf{X}')))$.*

Notice that, because of the properties of the Moore-Penrose inverse, for an intervention $do(\mathbf{X}' = \mathbf{x}')$, $\alpha^{+}_{\mathbf{X}'}$ entails a uniform distribution of probability mass over all $do(\mathbf{X} = \mathbf{x})$ such that $a(\mathbf{X}) = \mathbf{X}'$ and $\alpha_{\mathbf{X}'}(\mathbf{x}) = \mathbf{x}'$. This uniform solution may be physically meaningless, and in [Rischel and Weichwald, 2021] an abstraction is indeed enriched with an additional explicit map between high-level interventions and low-level interventions.

Interventional Superresolution Consistency (ISC)

ISC takes $p = \alpha^{+}_{\mathbf{Y}'} \circ \nu_{do()}$ and $q = \mu_{do()} \circ \alpha^{+}_{\mathbf{X}'}$. ISC considers interventions on the high-level and it evaluates the agreement, via abstraction, between results computed at high-level and low-level. Lower ISC would be relevant when a downstream task depends on $P(\alpha^{+}_{\mathbf{Y}'}(\mathbf{Y}')|do(\mathbf{X}'))$, with $\alpha^{+}_{\mathbf{Y}'}$ expressing a refinement of $P(\mathbf{Y}'|do(\mathbf{X}'))$, and an agent is required to work with $P(\mathbf{Y}|do(\alpha^{+}_{\mathbf{X}'}(\mathbf{X}')))$.

	Original task	Abstraction task
IC	$P(\alpha_{\mathbf{Y}'}(\mathbf{Y})\|do(\mathbf{X}))$	$P(\mathbf{Y}'\|do(\alpha_{\mathbf{X}'}(\mathbf{X})))$
IIL	$P(\mathbf{Y}\|do(\mathbf{X}))$	$P(\alpha^{+}_{\mathbf{Y}'}(\mathbf{Y}')\|do(\alpha_{\mathbf{X}'}(\mathbf{X})))$
ISIL	$P(\mathbf{Y}'\|do(\mathbf{X}'))$	$P(\alpha_{\mathbf{Y}'}(\mathbf{Y})\|do(\alpha^{+}_{\mathbf{X}'}(\mathbf{X}')))$
ISC	$P(\alpha^{+}_{\mathbf{Y}'}(\mathbf{Y}')\|do(\mathbf{X}'))$	$P(\mathbf{Y}\|do(\alpha^{+}_{\mathbf{X}'}(\mathbf{X}')))$

Table 2: Relation between interventional measures of abstraction approximation and downstream tasks. *Original task* specifies on which distribution an original downstream task depends; *Abstraction task* denotes on which distribution of a higher LA the task may depend.

Example 7. *Consider the same lung cancer scenario as in Ex. 6. Lab C is performing environmental manipulation, estimating the probability of lung cancer, and using a decision-making module to recommend further treatment. As patients are undergoing a similar experiment in the more sophisticated model of Lab A, it is required that outcomes between the two models are aligned. The downstream task depends on* $P(\alpha^{+}_{\mathbf{Y}'}(\mathbf{Y}')|do(\mathbf{X}'))$*, where* $\mathbf{Y}'$ *is lung cancer refined in the low-level model via* $\alpha^{+}_{\mathbf{Y}'}$ *and* $\mathbf{X}'$ *is environment. Lab C will then evaluate ISC, which allows it to measure the discrepancy between* $P(\alpha^{+}_{\mathbf{Y}'}(\mathbf{Y}')|do(\mathbf{X}'))$ *and* $P(\mathbf{Y}|do(\alpha^{+}_{\mathbf{X}'}(\mathbf{X}')))$*.*

Tab. 2 summarizes the four interventional measures.

3.2 Assessment Set

The definition of an assessment set is crucial in the computation of an interventional measure of abstraction approximation. We discuss a few representative options, highlighting again their differences wrt potential downstream tasks.

The choice to consider non-empty disjoint pairs of sets prevents us from considering observational distributions such as $P(\mathbf{X}') = P(\mathbf{X}'|do(\emptyset))$ corresponding to the pair $(\emptyset, \mathbf{X}')$. This increases the robustness of the measure to differences in the marginal distributions of the root-nodes, providing a degree of insensitivity to root covariate shift.

Example 8. *Let Lab A and Lab B work with two SCMs having the same DAG as in Fig. 1b, and let us assume an identity abstraction between them. As long as they specify the same mechanism* $Sm' \to Hea'$*, then* $e_{IC}(\boldsymbol{\alpha}) = 0$ *independently from the marginal distributions on* Sm'*.*

However, robustness to differences in observational marginal or to anti-causal quantities is only partial.

Example 9. *Let us take the same setup as in Ex. 8, and assume Lab B wants to consider the error wrt the disjoint pair* (Hea', Sm')*. This would correspond to evaluating error wrt the interventional quantity* $P(Sm'|do(Hea'))$*. This presents two problems: (i) the stochastic matrix capturing this distribution would have an anti-causal meaning; and (ii) because of the form of the DAG,* $P(Sm'|do(Hea')) = P(Sm')$*, leading us to account for an observational quantity.*

To avoid the error being affected by anti-causal quantities, we can define a causal assessment set:

Definition 8 (Causal Assessment Set)**.** *Given an abstraction* $\boldsymbol{\alpha}$ *from* $\mathcal{M}$ *to* $\mathcal{M}'$*, let* $\mathcal{J}_c$ *be the set of all non-empty disjoint pairs* $(\mathbf{X}', \mathbf{Y}') \in \mathscr{P}(\mathcal{X}') \times \mathscr{P}(\mathcal{X}')$*, such that* $\forall Y' \in \mathbf{Y}'$*,* $\exists X' \prec Y'$ *in* $\mathcal{M}'_{do(\mathbf{X}')}$*.*

Assessment set	Downstream task dependes on...
Complete $\mathcal{J}$	any possible causal or anti-causal intervention $P(\mathbf{Y}'\|do(\mathbf{X}'))$
Causal $\mathcal{J}_c$	any possible causal intervention $P(\mathbf{Y}'\|do(\mathbf{X}'))$, potentially affected by root covariate shift
Parental $\mathcal{J}_p$	causal intervention $P(\mathbf{Y}'\|do(\mathbf{X}'))$ dependent only on causal mechanisms
Custom $\mathcal{J}_u$	user-specific set of interventions $P(\mathbf{Y}'\|do(\mathbf{X}'))$

Table 3: Relation between assessment sets and downstream tasks.

The ordering relation guarantees that every node in the outcome $\mathbf{Y}'$ is affected by $\mathbf{X}'$. An equivalent definition in terms of independence is offered in App. **??**.

Moreover, the evaluation is not, in general, robust to root covariate shift, as contributions from marginal distributions may always enter the evaluation if paths from the root nodes are not blocked.

Example 10. *Let Lab C work with an abstracted model as in Fig. 1c, and assume it wants to consider the error wrt the disjoint pair* (Env'', LC'')*. This would correspond to evaluating error wrt the interventional quantity* $P(LC''|do(Env''))$*. In the model, this corresponds to* $\sum_{Gen''} P(LC''|Env'', Gen'')P(Gen'')$*, revealing the contribution of the marginal* $P(Gen'')$*.*

This sensitivity may be desirable in certain cases, for instance when the contribution of marginals can not be suppressed or when deemed informative. However, if we were interested in assessing abstraction error robustly wrt covariate shift, that is, only wrt the actual causal mechanisms, we could consider a parental assessment set.

Definition 9 (Parental Assessment Set)**.** *Given an abstraction* $\boldsymbol{\alpha}$ *from* $\mathcal{M}$ *to* $\mathcal{M}'$*, let* $\mathcal{J}_p$ *be the set of all non-empty disjoint pairs* $(\mathbf{X}', \mathbf{Y}') \in \mathscr{P}(\mathcal{X}') \times \mathscr{P}(\mathcal{X}')$*, such that* $\mathbf{X}' = Pa(\mathbf{Y}')$*.*

Lastly, it may be worth pointing out that custom assessment sets $\mathcal{J}_u$ may always be defined by an agent, if it is interested in the abstraction only of specific sub-parts of a base model. Tab. 3 summarizes the assessment sets we considered, and the type of relevant downstream tasks of interest.

4 Properties of Abstraction Approximation

All interventional measures of abstraction approximation share common properties, relevant for learning abstractions. Complete proofs are provided in App. **??** and **??**.

4.1 Properties of Error

Let us consider two generic abstractions: $\boldsymbol{\alpha}$ from $\mathcal{M}$ to $\mathcal{M}'$, and $\boldsymbol{\beta}$ from $\mathcal{M}'$ to $\mathcal{M}''$. Thanks to JSD, a key property is (E1) *triangle inequality*, that is, $E_I(\boldsymbol{\beta\alpha}, \mathbf{X}'', \mathbf{Y}'') \leq E_I(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}') + E_I(\boldsymbol{\beta}, \mathbf{X}'', \mathbf{Y}'')$, which is grounded in horizontal compositionality (or composition of abstractions) [Rischel, 2020]. Two other forms of compositionality instead do not hold: (NE1) *vertical non-compositionality* (or stochastic function non-composition); and (NE2) *product non-compositionality*. These negative properties follow from stochastic functions being computed from different post-interventional models (see App. **??** for a more precise discussion). From (E1) we immediately derive (E2) *non-monotonicity*, stating that it is not guaranteed that $E_I(\boldsymbol{\beta\alpha}, \mathbf{X}'', \mathbf{Y}'') \geq E_I(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}')$. From the definition we can identify extrema of the error: (E3) *zero at identity*, when the abstracted model is an identity; (E4) *zero at singleton*, only for IC, when the abstracted model is a singleton.

Finally, we have properties concerning relationships and identities (see App. **??**) between interventional measures.

Proposition 1 (Relationship between measures)**.** *We have a partial ordering among the interventional measures as:* $E_{IIL}(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}') \geq E_{IC}(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}')$*,* $E_{IIL}(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}') \geq E_{ISC}(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}')$*,* $E_{IC}(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}') \geq E_{ISIL}(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}')$*,* $E_{ISC}(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}') \geq E_{ISIL}(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}')$*.*

Proof sketch. All relations can be proved by applying the property of shortness of JSD and the right inverse property of the Moore-Penrose inverse. ■

4.2 Properties of Overall Error

Properties of the error may immediately extend to the overall error according to the chosen aggregation function f. In the case of the supremum, the overall error inherits the properties of: (O1) *triangle inequality*; (O2) *non-monotonicity*; (O3) *zero at identity*; and (O4) *zero at singleton*. Also, extension of Proposition 1 hold; however, notice that, despite this extension, two interventional measures of abstraction approximation may reach their minima for different abstractions. For instance, given two abstractions $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ from $\mathcal{M}$ to $\mathcal{M}'$, such that $e_{IIL}(\boldsymbol{\alpha}) \leq e_{IIL}(\boldsymbol{\beta})$, while it holds that $e_{IIL}(\boldsymbol{\alpha}) \geq e_{IC}(\boldsymbol{\alpha})$ and $e_{IIL}(\boldsymbol{\beta}) \geq e_{IC}(\boldsymbol{\beta})$, it does not follow that $e_{IC}(\boldsymbol{\alpha}) \leq e_{IC}(\boldsymbol{\beta})$; so different abstractions between the same two SCMs $\mathcal{M}$ and $\mathcal{M}'$ may minimize different interventional measures.

A necessary condition for any error measure to be finite is:

Proposition 2 (Finiteness of the overall error)**.** $e_I(\boldsymbol{\alpha}) < \infty$ *if a is order-preserving.*

Proof sketch. It can be shown that, in absence of order-preservation, it is impossible to compose the paths required for computing any error measure. ■

This proposition holds for all measures because it is related to the directionality of the edges representing causal mechanisms. This condition implicitly asserts that an abstraction can not reverse the directionality of causation, a requirement explicit in certain abstraction frameworks [Otsuka and Saigo, 2022]. Also, the request of order-preservation has a connection to the framework of [Rubenstein *et al.*, 2017], where order-preservation is imposed on a map ω relating low-level interventions with high-level interventions. Imposing order-preservation between variables acts at a more basic level and implies order-preservation among the interventions.

5 Abstraction Evaluation and Learning

A simple algorithm for evaluating abstraction can be derived from the original specification of IC error in Definition 5 by computing the error for all pairs $(\mathbf{X}', \mathbf{Y}') \in \mathcal{J}$, as in Alg. 1. Its complexity depends on the loop in step 3: at each iteration, two matrix multiplications and a JSD computation are performed in order to evaluate $E_{IC}(\boldsymbol{\alpha}, \mathbf{X}, \mathbf{Y})$. The overall complexity then grows as $O(|\mathcal{J}|)$, which, in the base case is given by the product of two powersets, $|\mathcal{J}| \approx 2^{2|\mathcal{X}'|}$. Computational complexity can then be reduced by exploiting the structure of a SCM and shrinking the set $\mathcal{J}$. A causal assessment set relies on partial ordering to reduce the size $|\mathcal{J}_c| < |\mathcal{J}|$. Even more, a parental assessment set exploits parental relationships to set the size $|\mathcal{J}_p| = |\mathcal{X}'| < |\mathcal{J}_c|$. Custom assessment sets $\mathcal{J}_u$ may also limit the number of pairs to be considered. Moreover, Prop. 2 provides an efficient test to evaluate whether any interventional measure of abstraction approximation will be finite with complexity $O(|E|)$ proportional to the number of edges in the DAG $\mathcal{G}_{\mathcal{M}}$; if not, no further computation is required. We can then evaluate the overall error more efficiently as in Alg. 2. The main challenge in further shrinking the assessment set $\mathcal{J}$ follows from the negative properties (NE1, NE2): without compositionality, it is not possible to reduce the evaluation of composed interventions to the one of its composing parts.

Algorithm 1 Overall IC error evaluation

```
In: $\mathcal{M}, \mathcal{M}', \boldsymbol{\alpha} = \langle R, a, \alpha \rangle$
Out: $e_{IC}(\boldsymbol{\alpha})$
1: Initialize $\mathbf{E} = \{\}$
2: Let $\mathcal{J}$ be the set of all non-empty disjoint pairs $(\mathbf{X}', \mathbf{Y}')$
3: for $(\mathbf{X}', \mathbf{Y}') \in \mathcal{J}$ do                    $\triangleright O(2^{2|\mathcal{X}'|})$
4:     Compute $E_{IC}(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}')$ as in Eq. 1 and add to $\mathbf{E}$
5: return sup $\mathbf{E}$
```

Algorithm 2 Abstraction evaluation

```
In: $\mathcal{M}, \mathcal{M}', \boldsymbol{\alpha} = \langle R, a, \alpha \rangle$, $I \in \{IC, IIL, ISIL, ISC\}$, $\mathcal{J}$
Out: $e_I(\boldsymbol{\alpha})$
1: if $a$ is not order-preserving then                    $\triangleright O(|E|)$
2:     return $\infty$
3: Initialize $\mathbf{E} = \{\}$
4: for $(\mathbf{X}', \mathbf{Y}') \in \mathcal{J}$ do                    $\triangleright O(|\mathcal{J}|)$
5:     Compute $E_I(\boldsymbol{\alpha}, \mathbf{X}', \mathbf{Y}')$ as in Eq. 3 and add to $\mathbf{E}$
6: return sup $\mathbf{E}$
```

Beyond evaluating abstractions, an agent may be interested in improving or learning new abstractions. If given an unsatisfactory abstraction $\boldsymbol{\alpha}$ with a high error $e_I(\boldsymbol{\alpha})$, properties (O1) and (O2) point to the possibility of sequentially improving the abstraction by searching for a new abstraction $\boldsymbol{\beta}$ such $e_I(\boldsymbol{\beta\alpha}) < e_I(\boldsymbol{\alpha})$. If given an incomplete abstraction $\boldsymbol{\alpha}$ for which one or more elements among $\mathcal{M}', R, a, \alpha_{X'}$ are not completely specified, it is possible to learn a complete specification of the abstraction by minimizing a chosen measure of abstraction approximation. In this case, properties (O3) and (O4) highlight minima the optimization may achieve. In both cases, Alg. 2 can be used as a building block to find an optimal abstraction by computing the error for each candidate abstraction in a set $\mathcal{K}$ (see Alg. **??** in App. **??**). This algorithm has a computational complexity of $O(|\mathcal{K}||\mathcal{J}|)$.

	Opt IC	Opt IIL
$\hat{P}(\mathbf{a}\|do(Sm=0))$	0.607±0.003	0.413±0.006
$\hat{P}(\mathbf{b}\|do(\alpha_{Sm'}(Sm)=0))$	0.600±0.003	0.498±0.006
$\hat{P}(\mathbf{a}\|do(Sm=1))$	0.797±0.003	0.681±0.005
$\hat{P}(\mathbf{b}\|do(\alpha_{Sm'}(Sm)=1))$	0.797±0.004	0.799±0.004

Table 4: Comparison between distributions $\hat{P}(\alpha_{\mathbf{Y}'}(\mathbf{Y})|do(\mathbf{X}))$ and $\hat{P}(\mathbf{Y}'|do(\alpha_{\mathbf{X}'}(\mathbf{X})))$, where $\mathbf{Y}' = Hea'$ and $\mathbf{X}' = Sm'$, when learning by optimizing for IC or IIL. $\mathbf{a}, \mathbf{b}$ are placeholders for $\alpha_{Hea'}(Hea) = 1$ and $Hea' = 1$, respectively.

6 Empirical Evaluation

We run empirical simulations for the two scenarios in Fig. 2: (i) in the *health scenario* we perform *abstraction learning* and show how our metrics would produce different results fit for distinct downstream tasks; and, (ii) in the *lung cancer scenario* we run *abstraction evaluation* to show the effects of the choice of assessment sets. Base model details are provided in App. **??**. These scenarios are designed to encompass a variety of configurations such as different structures in the low- and high-level model (chains, colliders, and forks), different a-maps among nodes (one-to-one, many-to-one), different number of variables and domain cardinality in the high-level model. Empirical distributions are computed from 10^4 samples; means and standard deviations are computed out of 10 repetitions. All simulations are available online[1].

Health scenario. Let's consider the health scenario in Fig. 2. The abstracted model (blue) is not fully defined (three candidate stochastic matrices for $Sm' \to Hea'$ have been proposed) and the abstraction itself is defined only in terms of R and a. Lab B wants to learn the best stochastic matrix and abstraction, wrt interventions performed by Lab A, in light of the downstream tasks described in Ex. 4 and 5. Lab B performs *abstraction learning* using Alg. **??**. Two different solutions are learned by minimizing either IC or ILL. Definition of models, abstractions, and solutions are in App. **??**.

Considering the downstream task described in Ex. 4, Tab. 4 shows that the closest agreement between the low- and high-level model in classifying patients for further exam is achieved by minimizing IC. Instead, considering the downstream task described in Ex. 5, Tab. 5 shows that the best match between the predictive distribution computed in the low-level model and in the low-level model when replacing a sub-part of it with a high-level abstraction, is obtained by minimizing IIL. In conclusion, an agent should choose carefully which measure to minimize according to its aim.

Lung cancer scenario. Let's consider the lung cancer scenario in Fig. 2. All models are completely defined, while the abstraction is given only in terms of R and a. Lab C wants to find the best abstraction minimizing ISIL, in light of the downstream task described in Ex. 6 and while considering three different assessment sets (causal, parental and custom). Lab C performs *abstraction evaluation* using Alg. 2. Three different solutions are learned by minimizing ISIL with the three assessment sets. Exact definition of models, abstractions and solutions are provided in App. **??**.

	$Sm = 0$	$Sm = 1$
$\hat{P}(CA=1\|do(Sm))$	0.679±0.004	0.766±0.005
$\hat{P}_{IC}(CA=1\|do(Sm))$	0.256±0.006	0.341±0.005
$\hat{P}_{IIL}(CA=1\|do(Sm))$	0.427±0.005	0.680±0.005

Table 5: Comparison between the empirical distribution computed only on the low-level model $\hat{P}$ and distributions using the abstraction minimizing IC ($\hat{P}_{IC}$) or IIL ($\hat{P}_{IC}$).

	$Env'' = 0$	$Env'' = 1$	$Env'' = 2$
$\hat{P}(\mathbf{a})$	0.445±0.003	0.555±0.003	0.655±0.004
$\hat{P}_{\mathcal{J}_c}(\mathbf{b})$	0.194±0.003	0.271±0.005	0.438±0.005
$\hat{P}_{\mathcal{J}_p}(\mathbf{b})$	0.563±0.005	0.730±0.005	0.807±0.003
$\hat{P}_{\mathcal{J}_u}(\mathbf{b})$	0.557±0.005	0.730±0.005	0.806±0.004

Table 6: Comparison between the empirical distribution computed only on the high-level model ($\hat{P}$) and using the abstraction minimizing ISIL wrt causal set ($\hat{P}_{\mathcal{J}_c}$), parental set ($\hat{P}_{\mathcal{J}_p}$), or custom set ($\hat{P}_{\mathcal{J}_u}$). $\mathbf{a}, \mathbf{b}$ are placeholders for $LC'' = 1|do(Env'')$ and $\alpha_{LC''}(LC) = 1|do(\alpha^{+}_{Env''}(Env''))$, respectively.

Tab. 6 confirms that, if the aim is to predict lung cancer under environmental experiments, then the best result is obtained when minimizing wrt a targeted assessment sets ($\mathcal{J}_p, \mathcal{J}_u$); larger sets require more computation (see Tab. **??**), and end up selecting a solution that, by mediating among many interventions, underperforms wrt the intervention of interest. This demonstrates the importance for an agent to optimize wrt a set of interventions that is relevant to its aim.

7 Conclusion

We introduced a family of interventional measures of abstraction approximation to quantify consistency and information loss. Our empirical simulations show that optimizing such measures lead to learning different optimal abstractions, which fit different constraints and downstream tasks. The proposed framework empowers modellers and agents by providing them with a set of measures that will help them learn abstractions that better suit their specific needs.

While this work focuses on four key measures (IC, IIL, ISC, ISIL), the proposed framework can accommodate new custom measures by modifying one or more of the parameters discussed in Sec. 3, combinations of existing measures, or shifting the focus to observational/counterfactual properties.

Future work could consider enhancing our learning algorithm. In one direction, we want to exploit the structure of the SCMs and properties of graph morphisms to reduce the size of assessment sets $\mathcal{J}$. Formal bounds may also be found for the tradeoff between the size of $\mathcal{J}$ and the error in estimating $E_I(\boldsymbol{\alpha})$. In another direction, we want to consider the algebraic properties of factoring stochastic matrices and sparse abstraction matrices in order to simplify the search space.

[1] https://github.com/FMZennaro/CausalAbstraction/tree/main/papers/2023-quantifying-consistency-and-infoloss

Acknowledgments

TD acknowledges support from a UKRI Turing AI acceleration Fellowship [EP/V02678X/1]. The authors thank the anonymous reviewers for their suggestions in improving this work.

References

[Beckers and Halpern, 2019] Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2678–2685, 2019.

[Beckers *et al.*, 2020] Sander Beckers, Frederick Eberhardt, and Joseph Y Halpern. Approximate causal abstractions. In *Uncertainty in Artificial Intelligence*, pages 606–615. PMLR, 2020.

[Chalupka *et al.*, 2017] Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, 2017.

[Chang and Fung, 1990] Kuo-Chu Chang and Robert M. Fung. Refinement and coarsening of bayesian networks. In *Conference on Uncertainty in Artificial Intelligence*, 1990.

[Dietterich, 2000] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[Eberhardt and Lee, 2022] Frederick Eberhardt and Lin Lin Lee. Causal emergence: When distortions in a map obscure the territory. *Philosophies*, 7(2):30, 2022.

[Elidan and Gould, 2008] Gal Elidan and Stephen Gould. Learning bounded treewidth bayesian networks. *Advances in neural information processing systems*, 21, 2008.

[Fritz, 2020] Tobias Fritz. A synthetic approach to markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, 2020.

[Geiger *et al.*, 2021] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

[Guyon *et al.*, 2008] Isabelle Guyon, Constantin Aliferis, Greg Cooper, André Elisseeff, Jean-Philippe Pellet, Peter Spirtes, and Alexander Statnikov. Design and analysis of the causation and prediction challenge. In *Causation and Prediction Challenge*, pages 1–33. PMLR, 2008.

[Hoel, 2017] Erik P Hoel. When the map is better than the territory. *Entropy*, 19(5):188, 2017.

[Kroer and Sandholm, 2018] Christian Kroer and Tuomas Sandholm. A unified framework for extensive-form game abstraction with bounds. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 613–624, 2018.

[Kusner *et al.*, 2017] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[Levin, 1992] Simon A Levin. The problem of pattern and scale in ecology: the robert h. macarthur award lecture. *Ecology*, 73(6):1943–1967, 1992.

[Mitchell, 2021] Melanie Mitchell. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101, 2021.

[Otsuka and Saigo, 2022] Jun Otsuka and Hayato Saigo. On the equivalence of causal models: A category-theoretic approach. *arXiv preprint arXiv:2201.06981*, 2022.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[Rischel and Weichwald, 2021] Eigil F Rischel and Sebastian Weichwald. Compositional abstraction error and a category of causal models. *arXiv preprint arXiv:2103.15758*, 2021.

[Rischel, 2020] Eigil Fjeldgren Rischel. The category theory of causal models. Master's thesis, University of Copenhagen, 2020.

[Rokach, 2019] Lior Rokach. *Ensemble learning: pattern classification using ensemble methods*. World Scientific, 2019.

[Rubenstein *et al.*, 2017] Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*, pages 808–817. Curran Associates, Inc., 2017.

[Sandholm, 2015] Tuomas Sandholm. Abstraction for solving large incomplete-information games. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 4127–4131. AAAI Press, 2015.

[Schmutz *et al.*, 2020] Valentin Schmutz, Wulfram Gerstner, and Tilo Schwalger. Mesoscopic population equations for spiking neural networks with synaptic short-term plasticity. *The Journal of Mathematical Neuroscience*, 10(1):1–32, 2020.

[Spivak, 2014] David I Spivak. *Category theory for the sciences*. MIT Press, 2014.

[Zennaro *et al.*, 2023] Fabio Massimo Zennaro, Máté Drávucz, Geanina Apachitei, W. Dhammika Widanage, and Theodoros Damoulas. Jointly learning consistent causal abstractions over multiple interventional distributions. In *2nd Conference on Causal Learning and Reasoning*, 2023.