Causal Abstractions, Categorically Unified

Markus Englberger
Department of Mathematics and Computer Science
Eindhoven University of Technology
m.englberger@tue.nl

Devendra Singh Dhami
Department of Mathematics and Computer Science
Eindhoven University of Technology

Abstract

We introduce a categorical framework for causal abstraction that unifies and extends existing approaches. Modeling causal systems as Markov functors from free Markov categories generated by directed acyclic graphs, we define a causal abstraction as a deterministic monoidal natural transformation together with an embedding of interventional diagrams of free Markov categories. This separates graphical compatibility under interventions from domain-level clustering of variables or values. Our framework provides an explicit characterization of admissible high-level graphs and recovers prior notions such as constructive τ-abstractions and cluster-DAG abstractions with unobserved confounders. We give concise categorical proofs that interventional distributions factorize over graphical abstractions and that do-calculus applied to the high-level graph yields valid conclusions for the low-level model.

1 Introduction

Causal models are often used to describe the same system at multiple levels of abstraction. A low-level model may represent detailed mechanisms between fine-grained variables, while a high-level model captures coarser, aggregated, or conceptually meaningful structure. A central question in causal modeling is therefore: When does a high-level causal model constitute a valid abstraction of a low-level one?

Several formal notions of causal abstraction have been proposed, ranging from τ-abstractions between structural causal models and cluster DAGs for causal Bayesian networks to category-theoretic formulations; we review them in Section 5. Despite these advances, existing approaches differ significantly in formalism and scope. Some are restricted to deterministic structural models, some focus on node clustering, others on domain coarsening. Further, existing notions of causal abstractions introduce a high amount of notational overhead.

In this paper, we propose a concise categorical framework that unifies and generalizes existing notions of causal abstraction. Our starting point is the observation that a causal model over a directed acyclic graph (DAG) can be represented as a Markov functor from the free Markov category generated by the graph into a suitable ambient Markov category (e.g. Stoch for probabilistic models or Set for deterministic models). Within this setting, we define:

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

A causal abstraction is a deterministic monoidal natural transformation between appropriate Markov functors, together with an embedding of interventional diagrams.

This formulation separates two structural components of abstraction: (1) **Graph-level abstraction**, captured by embeddings of interventional diagrams, which precisely characterize which high-level DAGs are compatible with the low-level interventional structure. (2) **Domain-level abstraction**, captured by deterministic natural transformations, which encode how low-level variables or domains are clustered into high-level ones.

This perspective yields several conceptual and technical benefits: First, it provides a concise and notation-light definition of causal abstraction that uniformly handles deterministic and probabilistic models. Second, it gives an explicit structural characterization of admissible high-level graphs: we show that a DAG is a valid graphical abstraction of a low-level DAG if and only if its interventional diagrams embed into those of the low-level DAG. This makes precise which graph operations preserve compatibility under interventions. Third, our framework recovers and generalizes existing notions of abstraction. In particular: (1) We prove that our categorical definition captures constructive τ-abstractions [Beckers and Halpern, 2019] between causal Bayesian networks whose clustering is acyclic and deletes no common causes (under edge-minimality). (2) We generalize results on cluster-DAGs with unobserved confounders [Anand et al., 2023] by giving concise categorical proofs and extending the validity of do-calculus to arbitrary graphical abstractions between ADMGs. (3) We clarify the relationship to ϕ-abstractions [Otsuka and Saigo, 2022]. Finally, by allowing Markov functors that are strong rather than strict monoidal, the framework also covers abstractions whose high-level variables correspond to directions in the low-level representation space rather than to individual low-level variables, as in the linear representation hypothesis of mechanistic interpretability (Section 4.2).

To separate our contribution from prior work: the Markov-categorical language of Section 2 (free Markov categories and the functorial view of causal Bayesian networks) is due to Fritz [2020], Fritz and Liang [2023], Jacobs et al. [2019]; Proposition 2.1, Proposition 3.1, Theorem 3.1 and the results of Section 3.2 are new; Proposition 4.1 and Theorem 4.1 extend Theorems 2, 5 and 3 of Anand et al. [2023] from cluster DAGs, which arise from merging alone, to all graphical abstractions, which also admit the deletion of non-confounding variables.

2 Markov Categories and Causal Models

We now introduce a categorical formulation of causal models. Fritz [2020] introduced *Markov categories*, representing morphisms in a monoidal category graphically as string diagrams:

Definition 2.1 (Markov Category)**.** A symmetric monoidal category is a Markov category if every object X is equipped with a commutative comonoid structure $(\mathrm{copy}_X, \mathrm{del}_X)$ compatible with the tensor product, and where the monoidal unit I is terminal.

For readers unfamiliar with Markov categories, we provide a brief primer in Appendix A. Throughout, morphisms are drawn as string diagrams, read from bottom to top: wires represent objects, boxes represent morphisms, placing diagrams side by side is the tensor product and stacking them is composition. Relevant Markov categories include the category of stochastic matrices **Stoch** and the category of sets **Set**. For an extensive introduction and list of Markov categories, we refer to Fritz [2020]. For a discussion of how string diagrammatic reasoning in the specific Markov category **Stoch** connects to causal inference, see Jacobs et al. [2019]. For background on monoidal categories more generally, we refer to Yanofsky [2024]. Furthermore, every DAG has an associated Markov category (Fritz [2020]):

Definition 2.2. Given a DAG $L = (\mathbf{V^L}, E^L)$, let Free_L be the Markov category freely generated by the nodes $\mathbf{V^L}$ as objects and boxes  for $A \in \mathbf{V^L}$ as morphisms, where $pa^L(A)$ denotes the parents of A in the graph L.We write box_A for the generating box of A. For disjoint $A, B \subset \mathbf{V^L}$, let $\mathrm{Int}_L(A, B) \subset \mathrm{Free}_L(A, B)$ denote the set of morphisms admitting a string diagram in which only generating boxes of nodes in $\mathbf{V^L} \setminus A$ occur, each at most once. We call the elements

of $\mathrm{Int}_L(A, B)$ **interventional diagrams** of signature $A \to B$ and write Int_L for the family of all of them.

For an explicit construction of freely generated Markov categories, we refer to Fritz and Liang [2023]. Now, we are able to define a causal model over a general Markov category:

Definition 2.3. A **causal model** over a DAG $L = (\mathbf{V^L}, E^L)$ is a Markov functor $F_L : \mathrm{Free}_L \to M$, where M is a Markov category. We write C_M for the class of causal models with values in M.

Given a causal model $F_L : \mathrm{Free}_L \to M$ and a morphism $A \to B$ in Free_L we will denote its image under F_L as $p^{F_L}(B|A)$ and call these images distributions. Further, for a set of nodes $A \subset \mathbf{V^L}$, we simply denote the tensor product in Free_L of these nodes as A.

By a result of Jacobs et al. [2019] (Proposition 3.1), we can identify Causal Bayesian Networks (CBNs) over a DAG L with functors of the form $\mathrm{Free}_L \to Stoch$, assigning a finite set to each node and a conditional probability table (stochastic matrix) to each generating mechanism. Throughout, CBNs are assumed to be *positive*: every value of every variable has positive probability under the observational distribution. The categorical results below (Proposition 2.1, Proposition 3.1 and Proposition 4.1) hold over an arbitrary Markov category M, so that continuous or mixed variables are covered by taking, e.g., M to be the category of standard Borel spaces and Markov kernels; the results stated for CBNs (Theorem 3.1, Lemma 3.1, Proposition 3.2 and Theorem 4.1) are phrased for finite domains in order to invoke this identification, although the proof of Theorem 4.1 only uses the validity of do-calculus for the low-level model together with a categorical transfer of d-separation.

Further, interventional diagrams exactly correspond to all possible types of interventional distributions. For a DAG L and $A \subset \mathbf{V^L}$ we write $L_{\overline{A}}$ for the graph obtained from L by removing all edges into A, and $an_G(B)$ for the set of ancestors of B in a graph G.

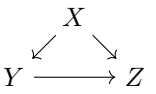
Proposition 2.1. *Let $L = (\mathbf{V^L}, E^L)$ be a DAG and let $A, B \subset \mathbf{V^L}$ be disjoint. Then $\mathrm{Int}_L(A, B)$ consists of exactly one morphism, which we denote by $d_L(A, B)$. If $F_L : \mathrm{Free}_L \to Stoch$ is a CBN over L, then $F_L(d_L(A, B))$ is the interventional distribution*

$$p^{F_L}(B|do(A)) := \int_{\mathbf{V^L}\setminus(A\cup B)} \prod_{C\in\mathbf{V^L}\setminus A} p^{F_L}(C|pa^L(C)).$$

Proof. See Appendix B. □

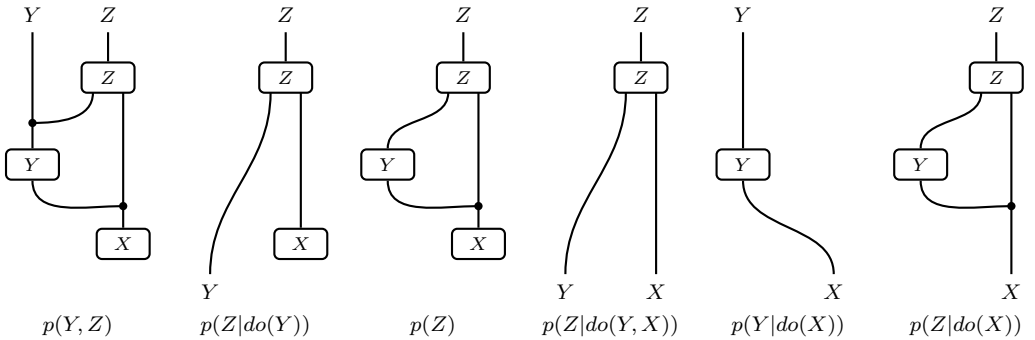
In light of Proposition 2.1, for a causal model $F_L : \mathrm{Free}_L \to M$ over a general Markov category M, we call the images $F_L(d_L(A, B))$ the **interventional distributions of** F_L and denote them by $p^{F_L}(B|do(A))$.

Example 2.1. Consider a CBN over a DAG $L =$



with distribution $p(Y, Z, X) = p(X) \cdot p(Y|X) \cdot p(Z|Y, X)$.

Below we show corresponding string diagrams for different types of interventional distributions:



3 Categorically Unifying Causal Abstractions

We now lay out our categorical framework for causal abstractions.

3.1 Causal abstractions as deterministic natural transformations

We give a concise categorical definition of causal abstraction with which we will recover existing non-categorical notions of causal abstractions. A running example illustrating the definitions and results of this section is given in Appendix C.

Definition 3.1. A causal model $F_H : \mathrm{Free}_H \to M$ is a **causal abstraction** of a causal model $F_L : \mathrm{Free}_L \to M$ if the nodes of H are pairwise disjoint nonempty subsets of $\mathbf{V^L}$, called **clusters**, and there exist

1. a Markov functor $\iota : \mathrm{Free}_H \to \mathrm{Free}_L$ with $\iota(A) = A$ for every node $A \in \mathbf{V^H}$, the tensor product of the nodes of L in A, that **embeds interventional diagrams**, i.e.

$$\iota(d_H(A, B)) = d_L(\iota(A), \iota(B)) \qquad \text{for all disjoint } A, B \subset \mathbf{V^H};$$

2. a monoidal natural transformation $\tau : F_L\iota \Rightarrow F_H$ whose components are deterministic.

Thus a high-level node A is identified with its cluster of low-level nodes; the domains $F_H(A)$ and $F_L(A)$ tell the two readings apart. The nodes of L not contained in any cluster are called **removed nodes**. Monoidality of τ means $\tau_I = \mathrm{id}_I$ and $\tau_{X\otimes Y} = \tau_X \otimes \tau_Y$, so τ is determined by its components $(\tau_A)_{A\in\mathbf{V^H}}$ at the nodes.

Functors ι as in Definition 3.1 and deterministic monoidal natural transformations are closed under composition, so causal abstractions compose.

To disentangle the categorical definition of causal abstractions, we first discuss abstractions on the side of graphs:

Definition 3.2. For a DAG $L = (\mathbf{V^L}, E^L)$, define its **graphical abstractions** as the collection of DAGs that result from applying a sequence of the following three operations to L:

- Deleting a node $A \in \mathbf{V^L}$ with *at most one child* and adding edges from all of A's parents to its child (if any).
- Merging two nodes $A, B \in \mathbf{V^L}$ and combining all incoming edges to either A or B as incoming edges into the merged node and combining all outgoing edges of either A or B as outgoing edges of the merged node and removing potential self-loops. This operation is only valid if the resulting graph is acyclic.
- Adding an edge between two nodes. This operation is only valid if the resulting graph is acyclic.

Example 3.1 (Mediator Removal)**.** Consider a low-level DAG L given by the chain $X \to Y \to Z$. We apply the deletion operation from Definition 3.2 to node Y. Since Y has only one outgoing edge, it is not a confounder, making the operation valid. We remove Y and add an edge from its parent X to its child Z, resulting in the high-level abstraction H given by $X \to Z$.

Example 3.2 (Variable Clustering)**.** Consider a low-level DAG L with edges $X_1 \to Y$ and $X_2 \to Y$. We apply the merging operation to nodes X_1 and X_2. The resulting high-level DAG H contains a merged node $X_{\{1,2\}}$ with a single combined edge $X_{\{1,2\}} \to Y$.

The following proposition translates the graphical abstractions as defined in Definition 3.2 into a categorical notion.

Proposition 3.1. *A DAG H is a graphical abstraction of a DAG L if and only if there exists a Markov functor $\iota : \mathrm{Free}_H \to \mathrm{Free}_L$ as in Definition 3.1. Moreover, such a functor is uniquely determined by its clusters, namely $\iota(\mathrm{box}_A) = d_L(\iota(pa^H(A)), \iota(A))$ for all $A \in \mathbf{V^H}$.*

Proof. See Appendix B. □

Proposition 2.1 and Proposition 3.1 together tell us that the graphical abstractions of a DAG L as defined in Definition 3.2 are the *correct* set of objects when studying causal abstractions as these are exactly those graphs whose induced types of interventional distributions can be related back to those induced by L.

Consider causal models $F_H : \mathrm{Free}_H \to M$, $F_L : \mathrm{Free}_L \to M$, a functor ι as in Proposition 3.1 and a family of deterministic morphisms $(\tau_A : F_L(\iota(A)) \to F_H(A))_{A \in \mathbf{V^H}}$, extended monoidally by $\tau_{A \otimes B} := \tau_A \otimes \tau_B$. Since Free_H is generated under composition and tensor product by the boxes box_A, the copy and discard maps and the symmetries, since deterministic morphisms commute with copying, and since I is terminal, τ is a monoidal natural transformation as soon as, for all $A \in \mathbf{V_H}$, the following diagram commutes:

$$\begin{array}{ccc} F_L(\iota(pa^H(A))) & \xrightarrow{p^{F_L}(A|do(pa^H(A)))} & F_L(\iota(A)) \\ \Big\downarrow{\scriptstyle \tau_{pa^H(A)}} & & \Big\downarrow{\scriptstyle \tau_A} \\ F_H(pa^H(A)) & \xrightarrow{p^{F_H}(A|pa^H(A))} & F_H(A) \end{array}$$

In other words, F_H is a causal abstraction of F_L if the high-level mechanisms are compatible with the mechanisms associated to the corresponding clusters of low-level variables.

When restricting to $M = \mathrm{Stoch}$, we can recover the following notion of causal abstractions between CBNs.

Definition 3.3. Consider two CBNs F_L, F_H over DAGs $L = (\mathbf{V^L}, E^L), H = (\mathbf{V^H}, E^H)$ with distribution p^{F_L}, p^{F_H}, respectively, such that the nodes of H are pairwise disjoint nonempty subsets of $\mathbf{V^L}$ (clusters) and for every $A \in \mathbf{V^H}$ there is a surjective domain map $\tau_A : F_L(A) \to F_H(A)$. F_H is a **causal abstraction between CBNs** of F_L if all interventional distributions coincide in the following sense: for all disjoint $A, B \subset \mathbf{V^H}$, all $a \in F_L(A)$ and all $\tilde{b} \in F_H(B)$:

$$p^{F_L}\left(\tau_B^{-1}(\tilde{b}) | do(a)\right) = p^{F_H}\left(\tilde{b} | do(\tau_A(a))\right) \tag{1}$$

where a set $A \subset \mathbf{V^H}$ of high-level nodes is identified with the set $\bigcup_{A' \in A} A' \subset \mathbf{V^L}$ of low-level nodes, and $\tau_A := \prod_{A' \in A} \tau_{A'}$.

We further demand the clustering of variables to respect acyclicity, i.e. the *cluster graph* with nodes $\mathbf{V^H}$ and an edge $A \to B$ whenever L has a directed path from A to B is acyclic, and to *delete no common causes*: every removed node R that is an ancestor in L of some cluster has a unique *first cluster*, i.e. there is exactly one $C \in \mathbf{V^H}$ such that L has a directed path from R to C whose interior contains no node of any cluster.

Minus the two conditions on acyclicity and not deleting confounders, this notion agrees with the notion of *constructive τ-abstraction*, introduced by Beckers and Halpern [2019]. Definition 3.3 states that it does not matter whether one first applies the causal mechanisms to the low-level model F_L and then maps to the high-level model F_H or vice versa.

We now present our main result, proving that our categorical Definition 3.1 generalizes Definition 3.3. Informally, Part 1 below states that every categorical abstraction is a constructive τ-abstraction, and Part 2 states that, as long as the high-level model carries no redundant edges, every constructive τ-abstraction in the sense of Definition 3.3 becomes a categorical one once the edges induced by the clustering are added to H.

Theorem 3.1. *Consider two CBNs F_L, F_H over DAGs L, H whose nodes are clusters of nodes of L, together with deterministic maps $\tau_A : F_L(A) \to F_H(A)$, $A \in \mathbf{V^H}$.*

1. *If F_H is a causal abstraction of F_L as in the categorical Definition 3.1 in the case $M = Stoch$, then F_H is a causal abstraction of F_L as in Definition 3.3.*
2. *Conversely, assume that F_H is a causal abstraction of F_L as in Definition 3.3 and that F_H is* edge-minimal*, i.e., no edge of H can be deleted such that p^{F_H} still factorizes over the graph. Then there exists a CBN $F_{H'}$ over a DAG H' with $\mathbf{V^{H'}} = \mathbf{V^H}$ and $E^H \subset E^{H'}$ that has the same interventional distributions as F_H and that is a causal abstraction of F_L as in the categorical Definition 3.1, with the same domain maps $(\tau_A)_{A \in \mathbf{V^H}}$.*

Proof. See Appendix B. □

The following example illustrates why in Theorem 3.1, one may have to add edges in the high-level graph H:

Example 3.3. Consider a causal model with graph $L : A \to B$ and an according CBN $F_L : \mathrm{Free}_L \to Stoch$. Coarsening the domain of the variable B may cancel the causal effect from $A \to B$, e.g., by collapsing the whole domain of B to a single value. This leads to a high-level CBN $F_H : \mathrm{Free}_H \to Stoch$ with graph H that has nodes A and B and no edge between them. This constitutes a causal abstraction as in Definition 3.3. However, since H is not a graphical abstraction of L, this does not yet constitute a causal abstraction in the categorical Definition 3.1. We can, however, simply re-add the edge $A \to B$ to obtain a new graph H' without changing any distribution.

By Theorem 3.1, the non-categorical definition of causal abstractions (Definition 3.3) is captured by the categorical definition (Definition 3.1). Since the categorical definition explicitly describes the potential high-level DAGs that can serve as the underlying DAGs of a causal abstractions (those DAGs H admitting a functor ι as in Definition 3.1, which by Proposition 3.1 are the graphical abstractions of Definition 3.2), we get a precise description of suitable high-level DAGs, which is not apparent from the non-categorical Definition. This explicit description of underlying graphs that can serve as causal abstractions is an important consequence of our categorical framework.

3.2 Effect-focused causal abstractions

So far, causal abstractions were defined via a deterministic natural transformation $\tau : F_L \circ \iota \Rightarrow F_H$, which clusters low-level values that have the same *effect* on downstream variables. We now show that our categorical framework supports a dual notion of abstraction, in which stochastic refinement maps $\epsilon : F_H(X) \to F_L(\iota(X))$ in the opposite direction take the place of τ.

Motivation. Consider a CBN over a DAG $L : A \to B \to C$. The distribution factorizes as

$$p(a, b, c) = p(c|b) \cdot p(b|a) \cdot p(a)$$

for a, b, c in the domains of A, B, C, respectively. Suppose we are given a coarsening of the domains of the random variables, i.e. maps τ_A, τ_B, τ_C mapping from the respective domains to clustered domains. We want to know in which cases the resulting distribution on the coarsened domains still factorizes as a CBN over L, i.e. whether

$$p(\tilde{a}, \tilde{b}, \tilde{c}) = p(\tilde{c}|\tilde{b}) \cdot p(\tilde{b}|\tilde{a}) \cdot p(\tilde{a}) \tag{2}$$

for values $\tilde{a}, \tilde{b}, \tilde{c}$ in the coarsened domains of A, B, C, respectively. Equation (2) is equivalent to

$$\sum_{b \in \tau_B^{-1}(\tilde{b})} p(\tilde{c}|b) \cdot p(b|\tilde{a}) \cdot p(\tilde{a}) db =$$
$$\sum_{b \in \tau_B^{-1}(\tilde{b})} p(\tilde{c}|b) \cdot p(b|\tilde{b}) db \cdot \sum_{b \in \tau_B^{-1}(\tilde{b})} p(b|\tilde{a}) \cdot p(\tilde{a}) db$$

Two natural sufficient conditions for this are that $p(b|\tilde{a})$ factorizes as

$$p(b|\tilde{a}) = p(b|\tilde{b}) \cdot p(\tilde{b}|\tilde{a}) \qquad \text{for all } b \in \tau_B^{-1}(\tilde{b}),$$

or that $p(\tilde{c}|b)$ is constant over the cluster $\tau_B^{-1}(\tilde{b})$. In the former case, the domain of variable B is partitioned respecting the shared effect of parent variable A, whereas in the latter case it is partitioned respecting the shared effect on child variable C. The latter is captured by Definition 3.1. We now formalize the former case. First, note that when we restrict to M=Stoch, the deterministic morphisms constituting the natural transformation have right inverses ϵ:

Lemma 3.1. *Consider a CBN $F_H \in C_{Stoch}$ that is a causal abstraction of a CBN $F_L \in C_{Stoch}$ witnessed by a deterministic monoidal natural transformation $\tau : F_L\iota \to F_H$. The component morphisms τ_A have right inverses $\epsilon_A : F_H(A) \to F_L(\iota(A))$.*

Proof. See Appendix B. □

We now let the morphisms ϵ take the place of τ, by requiring naturality with respect to interventional diagrams for the morphisms ϵ :

Definition 3.4. A causal model $F_H : \text{Free}_H \to \text{Stoch}$ is an **effect-focused causal abstraction** of a causal model $F_L : \text{Free}_L \to \text{Stoch}$ if there exist a Markov functor $\iota : \text{Free}_H \to \text{Free}_L$ as in Definition 3.1 and a family of morphisms $(\epsilon_X : F_H(X) \to F_L(\iota(X)))_{X \subset \mathbf{V^H}}$ with deterministic left inverses τ_X, where $\tau_X = \prod_{A \in X} \tau_A$ for deterministic maps $\tau_A : F_L(\iota(A)) \to F_H(A)$, such that the square

$$F_L(\iota(d_H(A, B))) \circ \epsilon_A = \epsilon_B \circ F_H(d_H(A, B))$$

commutes for all disjoint $A, B \subset \mathbf{V^H}$.

To simplify notation, for a low-level value $a \in F_L(\iota(A))$ we will denote its associated cluster of values as $\tilde{a}$, both when viewed on the high-level CBN, i.e. $\tilde{a} := \tau_A(a) \in F_H(A)$, as well as when viewed on the low-level CBN, i.e. $\tilde{a} := \tau_A^{-1}(\tau_A(a))) \subset F_L(\iota(A))$.

The squares of Definition 3.4 express compatibility of mechanisms in the reverse direction. In particular, the square for $(\emptyset, X)$ reads, for every set of high-level nodes $X \subset \mathbf{V^H}$ (in particular for every single node A):

$$\begin{array}{ccc} I & \xrightarrow{p^{F_H}(X)} & F_H(X) \\ \downarrow id & & \downarrow \epsilon_X \\ I & \xrightarrow{p^{F_L}(X)} & F_L(\iota(X)). \end{array}$$

From this we obtain that for $a \in F_L(\iota(A))$, and likewise for tuples $x \in F_L(\iota(X))$ with ϵ_X in place of ϵ_A (using that $\epsilon_A(\cdot|\tilde{a}')$ is supported on $\tau_A^{-1}(\tilde{a}')$ and that τ_A pushes $p^{F_L}(A)$ forward to $p^{F_H}(A)$):

$$p^{F_L}(a) = \sum_{\tilde{a}' \in F_H(A)} \epsilon_A(a|\tilde{a}') \cdot p^{F_H}(\tilde{a}') d\tilde{a}' \tag{3}$$

$$= \epsilon_A(a|\tilde{a}) \cdot p^{F_L}(\tilde{a}) \tag{4}$$

Since $p^{F_L}(\tilde{a}) = p^{F_H}(\tilde{a}) > 0$ by positivity, this implies

$$\epsilon_A(a|\tilde{a}) = p^{F_L}(a|\tilde{a}) \tag{5}$$

Thus, the refinement map ϵ_A coincides with conditioning on the cluster, and for a set X of high-level nodes and every $\tilde{x}$ with $p^{F_H}(\tilde{x}) > 0$, $\epsilon_X(x|\tilde{x}) = p^{F_L}(x|\tilde{x})$ is the joint conditional distribution of the low-level values of the clusters in X given their coarse values. We can further show:

Proposition 3.2. *Let F_H be an effect-focused abstraction of F_L. Then for every high-level node $A \in \mathbf{V^H}$ and low-level values $b \in F_L(\iota(pa^H(A))), a \in F_L(\iota(A))$ such that $p^{F_L}(a), p^{F_L}(b) > 0$,*

$$p^{F_L}(a|\tilde{b}) = \epsilon_A(a|\tilde{a}) \cdot p^{F_L}(\tilde{a}|\tilde{b}),$$

where $p^{F_L}(a|b)$ denotes the mechanism $F_L(\iota(\text{box}_A))(a|b)$ and $p^{F_L}(a|\tilde{b}) := \sum_{b \in \tilde{b}} p^{F_L}(a|b)\, p^{F_L}(b|\tilde{b})$.

Proof. See Appendix B. □

Clustering respects upstream invariance: values are grouped only when their dependence on parent variables is indistinguishable after clustering. Whereas Definition 3.3 constrains τ to only cluster values that have the same effect on children variables, Definition 3.4 constrains τ to only cluster values that are affected the same by parent variables. For an example, see Example D.3 in the appendix.

4 Unifying Prior Perspectives

We now demonstrate how our framework relates to and unifies existing works on causal abstractions.

4.1 Strong and constructive τ-abstractions

In their treatment of causal abstractions, Beckers and Halpern [2019] differentiate between strong τ-abstractions and constructive τ-abstractions. A constructive τ-abstraction can be seen as a deterministic causal abstraction in our framework. Let a deterministic causal model $F_H \in C_{\text{Set}}$ be a causal abstraction (Definition 3.1) of a deterministic causal model $F_L \in C_{\text{Set}}$ witnessed by a natural transformation τ. Since τ is monoidal by Definition 3.1, it preserves the monoidal structure, i.e.

$$\forall A, B \in \mathbf{V^H} : \tau_A \times \tau_B = \tau_{A \otimes B} \tag{6}$$

In other words, the map $\tau : F_L(\mathbf{V^H}) \to F_H(\mathbf{V_H})$ factorizes as $\tau = (\tau_1, ..., \tau_{|V_H|})$. Beckers and Halpern [2019] call such maps constructive. A strong τ-abstraction is a relaxation of a constructive τ-abstraction where there may not be an alignment between clusters of low-level variables and high-level nodes, i.e. τ may not factorize as in eq. (6). Note that eq. (6) refers to the product structure of $F_L(\iota(A \otimes B))$ provided by the strict monoidal functor. If F_L is only strong monoidal (see Section 4.2), with coherence isomorphisms $\mu_{A,B} : F_L(\iota(A)) \times F_L(\iota(B)) \to F_L(\iota(A \otimes B))$, a monoidal natural transformation $\tau : F_L\iota \Rightarrow F_H$ satisfies $\tau_{A\otimes B} \circ \mu_{A,B} = \tau_A \times \tau_B$ instead, so $\tau_{A\otimes B}$ factorizes after undoing the coherence map, not along the given coordinates of $F_L(\iota(A \otimes B))$.

4.2 Causal abstractions in mechanistic interpretability

Geiger et al. [2025] unify several concepts of mechanistic interpretability methods in the language of causal abstractions. Their definition of constructive abstractions corresponds to our concept of a deterministic causal abstraction. In this context, the linear representation hypothesis [Park et al., 2024] postulates that concepts in the activation space of a neural network are encoded as linear directions. These directions need not be the axes induced by the neurons, so intervening on a single concept is in general not an intervention on a single neuron. Such non-aligned interventions are captured by the coherence maps of a strong monoidal causal model: for a strict Markov functor as in Definition 2.3, $F(A) \times F(B) = F(A \otimes B)$ and every intervention is a product of single-variable interventions, whereas a strong monoidal functor relaxes this to isomorphisms $\mu_{A,B} : F(A) \times F(B) \to F(A \otimes B)$ that record a change of coordinates.

Example 4.1. Consider a deterministic causal model $F \in C_{\text{Set}}$ with nodes A, B, Y, edges $A \to Y$, $B \to Y$, and $F(A) = F(B) = \mathbb{R}$, $F(A \otimes B) = \mathbb{R}^2$, where $F(A \otimes B)$ is the activation space of a layer and A, B are two concepts, each encoded along a one-dimensional subspace. If these subspaces are axis aligned, i.e. each concept is encoded by one neuron, F is a strict Markov functor with $F(A) \times F(B) = F(A \otimes B)$. If they are orthogonal but not axis aligned, the coherence map $\mu_{A,B} : F(A) \times F(B) \to F(A \otimes B)$ is still an isomorphism but not the identity, and an intervention on the concept A is an intervention along the direction $\mu_{A,B}(\mathbb{R} \times \{0\})$ of the activation space.

A causal abstraction τ (Definition 3.1) of such a model is then constructive in the sense of Geiger et al. [2025] only after the change of coordinates μ, not along the neurons.

4.3 Cluster-DAGs

Anand et al. [2023] consider abstractions between causal models with unobserved confounders, encoded in acyclic directed mixed graphs (ADMGs). ADMGs can have bidirected edges representing unobserved confounders.

Definition 4.1. A **causal model over an ADMG** $L = (\mathbf{V^L}, E^L)$ is a causal model $F_{L'} \in C_M$ over a DAG L' with node set $\mathbf{V^L} \sqcup \mathbf{U}$ whose *exogenous* nodes $\mathbf{U}$ are parentless and whose latent projection with respect to $\mathbf{U}$ is L; every ADMG arises in this way, e.g. from the DAG with one exogenous node per bidirected edge. An ADMG $H = (\mathbf{V^H}, E^H)$ is a **graphical abstraction between ADMGs** of L *with respect to* L' if there are a DAG H' with node set $\mathbf{V^H} \sqcup \mathbf{U^H}$ whose latent projection with respect to $\mathbf{U^H}$ is H, and a Markov functor $\iota : \mathrm{Free}_{H'} \to \mathrm{Free}_{L'}$ as in Definition 3.1 such that the nodes in $\mathbf{V^H}$ are clusters of endogenous nodes, the nodes in $\mathbf{U^H}$ are clusters of exogenous nodes, and the nodes of $\mathbf{U^H}$ are parentless in H'.

Proposition 4.1. *Let $F_{L'} \in C_M$ be a causal model over an ADMG L and let H be a graphical abstraction between ADMGs of L with respect to L', witnessed by H' and ι. Then $F_{L'} \circ \iota$ is a causal model over the ADMG H whose interventional distributions are those of $F_{L'}$ on the clusters: $p^{F_{L'}\iota}(B|do(A)) = p^{F_{L'}}(B|do(A))$ for all disjoint $A, B \subset \mathbf{V^{H'}}$. In particular, if $F_{L'}$ is a CBN with distribution p, then the interventional distributions of p on the clusters factorize over the ADMG H: for all $A \subset \mathbf{V^H}$,*

$$p(\mathbf{V^H} \setminus A \,|\, do(A)) = \int_{\mathbf{U^H}} \prod_{U \in \mathbf{U^H}} p(U) \prod_{C \in \mathbf{V^H} \setminus A} p\big(C \,\big|\, do(pa_H(C)), do(U_C)\big)\, d\mathbf{U^H},$$

where $U_C \subset \mathbf{U^H}$ is the set of exogenous parents of C in H', so that $U_C \cap U_{C'} \neq \emptyset$ if and only if $C \leftrightarrow C'$ in H, and the parents in A take their intervened values.

Proof. See Appendix B. □

Proposition 4.1 subsumes Theorem 2 (the case $A = \emptyset$) and Theorem 5 (arbitrary A) of Anand et al. [2023]. There, the high-level graph is the cluster DAG of a partition of the endogenous variables; this is the special case of Definition 4.1 in which no node is removed, every exogenous node forms its own cluster and the clusters of endogenous nodes are the cells of the partition. By Proposition 3.1, the cluster DAG is a graphical abstraction of L' whenever it is acyclic, which is the admissibility condition of Anand et al. [2023]; the sets U_C are their latent sets, and the factor $p(C|do(pa_H(C)), do(U_C))$, the mechanism of C in $F_{L'}\iota$, coincides with the conditional distribution $p(C|pa_H(C), U_C)$ wherever the latter is defined, by the Markov property of the CBN $F_{L'}\iota$ over H'.

Given a CBN over an ADMG $L = (\mathbf{V^L}, E^L)$, Anand et al. [2023] show how applying the rules of do-calculus on the high-level graph induced by a partition over V gives valid results on the low-level graph L. We prove a generalized statement for all graphical abstractions between ADMGs (Theorem 3 of Anand et al. [2023] is the case of partitions; see Example D.4 for an example):

Theorem 4.1. *Consider a CBN $F = F_{L'}$ over an ADMG $L = (\mathbf{V^L}, E^L)$ (Definition 4.1) and let $H = (\mathbf{V^H}, E^H)$ be a graphical abstraction between ADMGs of L with respect to L'. Then applying Pearl's do-calculus on H gives valid results on the low-level graph L, i.e. for any disjoint subsets of clusters $X, Y, Z, W \subseteq \mathbf{V^H}$, the following three rules hold wherever the conditionals are defined:*

Rule 1: $p^F(Y|do(X), Z, W) = p^F(Y|do(X), W)$ *if* $(Y \perp\!\!\!\perp Z|X, W)_{H_{\overline{X}}}$,

Rule 2: $p^F(Y|do(X), do(Z), W) = p^F(Y|do(X), Z, W)$ *if* $(Y \perp\!\!\!\perp Z|X, W)_{H_{\overline{X}\underline{Z}}}$,

Rule 3: $p^F(Y|do(X), do(Z), W) = p^F(Y|do(X), W)$ *if* $(Y \perp\!\!\!\perp Z|X, W)_{H_{\overline{X Z_H(W)}}}$,

where $H_{\overline{X}\underline{Z}}$ is obtained from H by removing all directed edges into X, all bidirected edges incident to X, and all directed edges out of Z, and $Z_H(W)$ is the set of nodes in Z that are non-ancestors of any node of W in $H_{\overline{X}}$.

Proof. See Appendix B. □

4.4 Φ-Abstractions

The work by Otsuka and Saigo [2022] is the closest to our proposed framework. Given two CBNs F_L, F_H, viewed as functors $\mathrm{Free}_L \to \mathit{Stoch}$ and $\mathrm{Free}_H \to \mathit{Stoch}$, they call F_H a **Φ-abstraction** of F_L if there exist a graph homomorphism $\phi : L \to H$, i.e. a map on nodes that sends every edge of L to an edge of H, and a natural transformation $\alpha : F_L \Rightarrow F_H\Phi$, where $\Phi : \mathrm{Free}_L \to \mathrm{Free}_H$ is the functor induced by ϕ; the components of α may be stochastic. Since a graph homomorphism

into a DAG cannot identify two adjacent nodes (the edge between them would have to be sent to a self-loop) and is defined on all nodes of L, clusterings of adjacent variables and the removal of variables cannot be modeled as Φ-abstractions (Examples D.1 and D.2 in the appendix). In comparison, the high-level graphs admitted by Definition 3.1 are exactly the graphical abstractions of Definition 3.2 (Proposition 3.1), which include clusterings of adjacent variables as well as the removal of mediators. Otsuka and Saigo [2022] relate Φ-abstractions to the exact transformations of Rubenstein et al. [2017] and, for deterministic α, characterize them by a homogeneity condition on the mechanisms; a counterpart of Theorem 3.1 relating them to constructive τ-abstractions is not available there.

5 Related work

Formal notions of causal abstraction include multi-level cause-effect systems [Chalupka et al., 2016], transformations of structural equation models that commute with interventions [Rubenstein et al., 2017], a compositional account with abstraction error in a category of causal models [Rischel and Weichwald, 2021], and connections to representation learning [D'Acunto et al., 2025] and to graphical consistency [Schooltink and Zennaro, 2025]. The notions of Beckers and Halpern [2019] (Section 4.1), Anand et al. [2023] (Section 4.3), Otsuka and Saigo [2022] (Section 4.4) and Geiger et al. [2025] (Section 4.2) were related to our framework in Section 4. D'Acunto and Battiloro [2025] also relate causal models through functors and natural transformations, but over a source category with only two objects, the joint exogenous and the joint endogenous domain, so that interventional consistency has to be imposed separately, and they study the structure a given abstraction carries. In our framework Free_L is generated by the mechanisms, so consistency of all interventional distributions follows from naturality (Proposition 2.1), and Proposition 3.1 and Theorem 3.1 address which high-level graphs admit an abstraction and how to verify one. The coarsenings of Madaleno et al. [2026] are the graphs generated by merging alone; we also allow deleting non-confounding variables, and Proposition 3.1 explains why these are the operations compatible with interventions. The transit clusters of Tikka et al. [2023] preserve the identifiability of one causal effect, whereas Proposition 3.1 concerns all interventions. Wahl et al. [2024] and Ferreira and Assaad [2025] study Markov properties of grouped variables and macro effects in cluster DMGs, respectively, allowing cycles at the cluster level, which we exclude. Massidda et al. [2024] learn abstractions of linear structural causal models under consistency conditions analogous to our naturality squares, and D'Acunto et al. [2025] via a semantic embedding principle; Proposition 3.1 delimits the hypothesis space of any such method, and its generator criterion reduces verifying a candidate to one local check per high-level mechanism.

6 Conclusion and limitations

We proposed a categorical framework in which a causal abstraction decomposes into an embedding of interventional diagrams, capturing graphical compatibility under interventions, and a deterministic monoidal natural transformation, capturing domain-level clustering. This separation yields a characterization of admissible high-level graphs, recovers constructive τ-abstractions (under edge-minimality) and cluster DAGs as special cases, clarifies the relation to Φ-abstractions, and transfers factorization and do-calculus to the high-level graph by functoriality.

Limitations and future work. Our framework covers hard interventions only: an interventional diagram replaces the mechanism of an intervened variable by an input wire, and soft or stochastic interventions are not represented. Both levels are required to be acyclic, namely DAGs or ADMGs arising as latent projections of DAGs; cyclic cluster graphs as in Ferreira and Assaad [2025] would require different categorical constructions, since Proposition 2.1 relies on acyclicity. The results specific to CBNs are stated for finite domains, since they invoke the identification of CBNs with functors into Stoch, whereas the categorical results hold over any Markov category.

Acknowledgments and Disclosure of Funding

The TU Eindhoven authors received support from their Department of Mathematics and Computer Science and the Eindhoven Artificial Intelligence Systems Institute.

References

Tara V. Anand, Adele H. Ribeiro, Jin Tian, and Elias Bareinboim. Causal effect identification in cluster DAGs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(10):12172–12179, Jun. 2023. doi: 10.1609/aaai.v37i10.26435. URL https://ojs.aaai.org/index.php/AAAI/article/view/26435.

Sander Beckers and Joseph Y. Halpern. 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, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33012678. URL https://doi.org/10.1609/aaai.v33i01.33012678.

A. Carboni and R.F.C. Walters. Cartesian bicategories I. *Journal of Pure and Applied Algebra*, 49 (1):11–32, 1987. ISSN 0022-4049. doi: 10.1016/0022-4049(87)90121-6.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Multi-level cause-effect systems. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 361–369, Cadiz, Spain, 09–11 May 2016. PMLR. URL https://proceedings.mlr.press/v51/chalupka16.html.

Gabriele D'Acunto and Claudio Battiloro. The relativity of causal knowledge. In Silvia Chiappa and Sara Magliacane, editors, *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 863–881. PMLR, 2025. URL https://proceedings.mlr.press/v286/d-acunto25a.html.

Gabriele D'Acunto, Fabio Massimo Zennaro, Yorgos Felekis, and Paolo Di Lorenzo. Causal abstraction learning based on the semantic embedding principle. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 11776–11812. PMLR, 2025. URL https://proceedings.mlr.press/v267/d-acunto25a.html.

Simon Ferreira and Charles K. Assaad. Identifying macro causal effects in a C-DMG over ADMGs. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=905LEugq6R.

Tobias Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, August 2020. ISSN 0001-8708. doi: 10.1016/j.aim.2020.107239. URL http://dx.doi.org/10.1016/j.aim.2020.107239.

Tobias Fritz and Andreas Klingler. The d-separation criterion in categorical probability. *Journal of Machine Learning Research*, 24(46):1–49, 2023. URL http://jmlr.org/papers/v24/22-0916.html.

Tobias Fritz and Wendong Liang. Free gs-Monoidal categories and free Markov categories. *Applied Categorical Structures*, 31(2), April 2023. ISSN 1572-9095. doi: 10.1007/s10485-023-09717-0. URL http://dx.doi.org/10.1007/s10485-023-09717-0.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL http://jmlr.org/papers/v26/23-0058.html.

Bart Jacobs, Aleks Kissinger, and Fabio Zanasi. Causal inference by string diagram surgery. In Mikołaj Bojańczyk and Alex Simpson, editors, *Foundations of Software Science and Computation Structures*, pages 313–329, Cham, 2019. Springer International Publishing. ISBN 978-3-030-17127-8.

Francisco Madaleno, Pratik Misra, and Alex Markham. Coarsening causal DAG models. In Bijan Mazaheri and Niels Richard Hanson, editors, *Proceedings of the Fifth Conference on Causal Learning and Reasoning*, volume 323 of *Proceedings of Machine Learning Research*, pages

1318–1344. PMLR, 2026. URL https://proceedings.mlr.press/v323/madaleno26b.html.

Riccardo Massidda, Sara Magliacane, and Davide Bacciu. Learning causal abstractions of linear structural causal models. In Negar Kiyavash and Joris M. Mooij, editors, *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 2486–2515. PMLR, 2024. URL https://proceedings.mlr.press/v244/massidda24a.html.

Jun Otsuka and Hayato Saigo. On the equivalence of causal models: A category-theoretic approach. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 634–646. PMLR, 11–13 Apr 2022. URL https://proceedings.mlr.press/v177/otsuka22a.html.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Eigil F. Rischel and Sebastian Weichwald. Compositional abstraction error and a category of causal models. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1013–1023. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/rischel21a.html.

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 Gal Elidan and Kristian Kersting, editors, *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, volume R15 of *Proceedings of Machine Learning Research*, pages 151–160. PMLR, 2017. URL https://proceedings.mlr.press/r15/rubenstein-17a.html.

Willem Schooltink and Fabio Massimo Zennaro. Aligning graphical and functional causal abstractions, 2025. URL https://arxiv.org/abs/2412.17080.

Santtu Tikka, Jouni Helske, and Juha Karvanen. Clustering and structural robustness in causal diagrams. *Journal of Machine Learning Research*, 24(195):1–32, 2023. URL http://jmlr.org/papers/v24/21-1322.html.

Jonas Wahl, Urmi Ninad, and Jakob Runge. Foundations of causal discovery on groups of variables. *Journal of Causal Inference*, 12(1):20230041, 2024. doi: 10.1515/jci-2023-0041.

N.S. Yanofsky. *Monoidal Category Theory: Unifying Concepts in Mathematics, Physics, and Computing*. MIT Press, 2024. ISBN 9780262049399. URL https://books.google.nl/books?id=xVTyEAAAQBAJ.

Appendix

A Background on Categorical Probability

In this section, we provide a brief introduction to Markov categories, the mathematical framework used throughout this paper to unify causal abstractions. Since the main text works with causal Bayesian networks represented by *stochastic matrices*, we restrict attention to finite probability spaces and define **Stoch** as the category of stochastic matrices.

For a more extensive treatment of Markov categories, see Fritz [2020].

A.1 The Categories Set and Stoch

To ground the abstract definitions, we first introduce the two most important examples: the category of sets **Set** (for deterministic causal systems) and the category of stochastic Markov kernels **Stoch** (for probabilistic systems).

Definition A.1 (The category **Set**)**.** The category **Set** has

- objects: sets
- morphisms: functions $f : X \to Y$,
- composition: ordinary function composition $(g \circ f)(x) = g(f(x))$,
- identities: $\mathrm{id}_X(x) = x$.

Definition A.2 (The category **Stoch** of stochastic matrices)**.** The category **Stoch** has

- objects: finite sets
- morphisms: stochastic matrices $p : X \to Y$, i.e. functions

$$p : X \times Y \to [0, 1], \qquad (x, y) \mapsto p(y \mid x),$$

  such that for every $x \in X$,

$$\sum_{y \in Y} p(y \mid x) = 1.$$

Composition is given by matrix multiplication (the Chapman–Kolmogorov rule): for $p : X \to Y$ and $q : Y \to Z$,

$$(q \circ p)(z \mid x) = \sum_{y \in Y} q(z \mid y)\, p(y \mid x). \tag{7}$$

The identity $\mathrm{id}_X : X \to X$ is the Kronecker delta:

$$\mathrm{id}_X(x' \mid x) = \begin{cases} 1 & \text{if } x' = x, \\ 0 & \text{otherwise.} \end{cases}$$

A.2 Symmetric Monoidal Structure

Both categories possess a **symmetric monoidal structure** $(\mathcal{C}, \otimes, I, \sigma)$. Intuitively, the tensor product $X \otimes Y$ represents a *composite system* consisting of two subsystems of types X and Y considered side-by-side. A morphism $f \otimes g : X \otimes Y \to X' \otimes Y'$ represents running two processes *in parallel*:

- $f : X \to X'$ acts on the first subsystem and ignores the second,
- $g : Y \to Y'$ acts on the second subsystem and ignores the first,
- the two processes do not exchange information while being executed in parallel.

The symmetry $\sigma_{X,Y} : X \otimes Y \to Y \otimes X$ formalizes that the ordering of wires in a string diagram is irrelevant up to swapping.

Concrete meaning in Set and Stoch. In **Set**, $X \otimes Y = X \times Y$ is the usual cartesian product: a state of the joint system is a pair (x, y). A parallel map $f \otimes g$ is just $(x, y) \mapsto (f(x), g(y))$. In **Stoch**, $X \otimes Y$ is again the product measurable space; a parallel composition $p \otimes q$ is the *product kernel* that samples the two outputs conditionally independently given the corresponding inputs:

$$(p \otimes q)\big((x', y') \mid (x, y)\big) = p(x' \mid x)\, q(y' \mid y). \tag{8}$$

A.3 Markov Categories

A symmetric monoidal category is a **Markov category** if it supports two specific operations: **copying** (data can be duplicated) and **discarding** (data can be ignored).

Definition A.3 (Markov Category)**.** A symmetric monoidal category is a Markov category if every object X is equipped with a commutative comonoid structure $(\mathrm{copy}_X, \mathrm{del}_X)$ compatible with the tensor product, and where the monoidal unit I is terminal.

We now motivate these structures using our running examples.

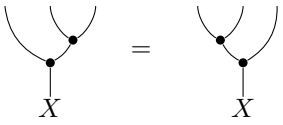
A.3.1 The Copy Map (Information Duplication)

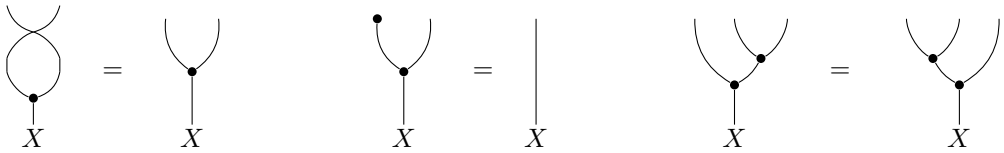
Information can be copied via a morphism $\mathrm{copy}_X : X \to X \otimes X$.

- **In Set:** The diagonal map $\Delta_X : x \mapsto (x, x)$.
- **In Stoch:** copy_X is the deterministic stochastic matrix that outputs (x, x) with probability 1:

$$\mathrm{copy}_X\big((x_1, x_2) \mid x\big) = \begin{cases} 1 & \text{if } x_1 = x_2 = x, \\ 0 & \text{otherwise.} \end{cases}$$

Axioms: The copy map must satisfy *Associativity* and *Commutativity*:





A.3.2 The Discard Map

We can always ignore a system's output. We formalize this via a morphism $\mathrm{del}_X : X \to I$.

- **In Set:** The unique function $X \to \{*\}$.
- **In Stoch:** The unique kernel from X to $\{*\}$.

Let $p : X \to Y$ be a stochastic matrix, written $p(y \mid x)$. Consider the composite $\mathrm{del}_Y \circ p : X \to I$. By (7),

$$(\mathrm{del}_Y \circ p)(* \mid x) = \sum_{y \in Y} \mathrm{del}_Y(* \mid y)\, p(y \mid x) = \sum_{y \in Y} p(y \mid x) = 1. \tag{9}$$

Thus, composing with discard performs *summation over the discarded variable*.

Example A.1 (Marginalizing a joint distribution)**.** A joint distribution on $X \times Y$ can be represented as a morphism $p : I \to X \times Y$, with entries $p(x, y)$. To obtain the marginal distribution on X, compose with $\mathrm{id}_X \otimes \mathrm{del}_Y : X \times Y \to X$:

$$I \xrightarrow{p} X \times Y \xrightarrow{\mathrm{id}_X \otimes \mathrm{del}_Y} X.$$

The resulting distribution $p_X : I \to X$ satisfies

$$p_X(x) = \sum_{y \in Y} p(x, y), \tag{10}$$

so discarding Y corresponds exactly to marginalization over Y.

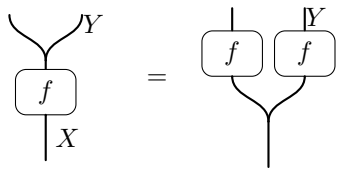
A.4 Deterministic Morphisms

In Markov categories, we can distinguish deterministic functions from stochastic kernels diagrammatically [Carboni and Walters, 1987].

Definition A.4 (Deterministic Morphism)**.** A morphism $f : X \to Y$ in a Markov category is **deterministic** if it commutes with the copy operation:

$$\text{copy}_Y \circ f = (f \otimes f) \circ \text{copy}_X \tag{11}$$

Below are the corresponding string diagrams:



(12)

Interpretation in Stoch. Let $p : X \to Y$ be a stochastic matrix. The left-hand side of (11) corresponds to sampling $y \sim p(\cdot \mid x)$ and then perfectly copying y, yielding a pair (y, y). The right-hand side corresponds to copying x and then sampling two outputs independently from $p(\cdot \mid x)$, yielding a pair (y_1, y_2). These coincide if and only if $p(\cdot \mid x)$ is a Dirac distribution for each x, i.e. if and only if p is induced by a (deterministic) function.

Thus, in Stoch the categorical notion of a deterministic morphism coincides with the usual one.

B Missing Proofs

Proof of Proposition 2.1

Proof. Let $N := (an_{L_{\overline{A}}}(B) \cup B) \setminus A$ be the set of non-intervened nodes that belong to B or are ancestors of B in $L_{\overline{A}}$, and let $A' := A \cap an_{L_{\overline{A}}}(B)$.

Existence. Stack the boxes of the nodes in N in a topological order of $L_{\overline{A}}$, copying each wire labelled by a node $C \in N \cup A'$ once for every consumer it has (its children in N, plus one output wire if $C \in B$), and discard the input wires labelled by $A \setminus A'$. This diagram uses only boxes of nodes in $N \subset \mathbf{V^L} \setminus A$, each once, and has signature $A \to B$. Its image under F_L is $\int_{N \setminus B} \prod_{C \in N} p^{F_L}(C|pa^L(C))$ with the inputs A' held fixed. Since I is terminal in Stoch, integrating out the remaining nodes $\mathbf{V^L} \setminus (A \cup N)$ from the full truncated factorization yields the same kernel, which is $p^{F_L}(B|do(A))$.

Uniqueness. Let D be a diagram representing an element of $\text{Int}_L(A, B)$; since I is terminal, every discard of the output of a box removes that box and every discard of a copy of a wire disappears, so we may assume that discards in D occur only on input wires. Then the output of every box occurring in D is either an output of D or an input of another box of D, so if box_C occurs, there is a chain of boxes from C to a node of B, i.e. $C \in N$. Conversely, $\text{box}_{B'}$ occurs for every $B' \in B$ (as $B' \notin A$ is not an input), and if box_C occurs, every parent P of C must be present as a wire; if $P \notin A$, that wire can only be produced by box_P. By induction, exactly the boxes of N occur. Since each box occurs at most once and the boxes of A are forbidden, every node of $N \cup A'$ has a unique source wire, all copies of which are determined by their consumers, and the input wires of $A \setminus A'$ can only be discarded. Hence D is, up to the comonoid axioms, the diagram constructed above. □

Proof of Proposition 3.1

We first characterize the functor in Definition 3.1 in graph-theoretic terms. For clarity, unlike the main text which identifies a high-level node with its cluster, the proofs keep the clustering as a map: A clustering of L indexed by a set $\mathbf{V^H}$ assigns to every $A \in \mathbf{V^H}$ a nonempty set $\mathrm{cl}(A) \subset \mathbf{V^L}$ such that the sets $\mathrm{cl}(A)$ are pairwise disjoint. The **induced graph** L/cl has node set $\mathbf{V^H}$ and an edge $D \to C$ for $D \neq C$ if and only if L has a directed path from $\mathrm{cl}(D)$ to $\mathrm{cl}(C)$ whose interior consists of removed nodes only (in particular whenever L has an edge between the two clusters). We say that cl **deletes no common causes** if

(U) every removed node R that is an ancestor in L of some cluster has a unique *first cluster*: there is exactly one $C \in \mathbf{V^H}$ such that L has a directed path from R to $\mathrm{cl}(C)$ whose interior consists of removed nodes only.

Throughout, a *directed path from X to $\mathrm{cl}(C)$* ends at its first node in $\mathrm{cl}(C)$, and the path of length zero counts if $X \in \mathrm{cl}(C)$. For a Markov functor $\iota : \mathrm{Free}_H \to \mathrm{Free}_L$ with $\iota(A) = \mathrm{cl}(A)$ on nodes we write $\delta_C := d_L(\iota(pa^H(C)), \iota(C))$. By the proof of Proposition 2.1, the boxes occurring in δ_C are those of the nodes $X \notin \iota(pa^H(C))$ that lie in $\mathrm{cl}(C)$ or have a directed path to $\mathrm{cl}(C)$ whose interior avoids $\iota(pa^H(C))$, and δ_C discards the input wire of a node $X \in \iota(pa^H(C))$ if and only if X has no such path.

Lemma B.1. *Let cl be a clustering of L indexed by $\mathbf{V^H}$ and let $\iota : \mathrm{Free}_H \to \mathrm{Free}_L$ be a Markov functor with $\iota(A) = \mathrm{cl}(A)$ on nodes. Then ι embeds interventional diagrams if and only if*

(i) $\iota(\mathrm{box}_C) = \delta_C$ for every $C \in \mathbf{V^H}$,

(ii) $L/\mathrm{cl} \subseteq H$, i.e. every edge of L/cl is an edge of H, and

(iii) cl deletes no common causes, i.e. (U) holds.

Proof. Step 1: ι embeds interventional diagrams if and only if (i) holds and

(ii′) no generating box of Free_L occurs in $\iota(\mathrm{box}_C)$ for two distinct $C \in \mathbf{V^H}$.

Since $d_H(pa^H(C), C) = \mathrm{box}_C$, embedding interventional diagrams forces (i). Assume (i) and (ii′). The diagram $d_H(A, B)$ uses each box_C at most once and only for $C \notin A$, and $\iota(d_H(A, B))$ is represented by the diagram obtained from it by substituting δ_C for box_C, since ι preserves copies, discards and symmetries. By (ii′) every box of L occurs at most once in this diagram and no box of a node of $\iota(A)$ occurs, since the box of $X \in \mathrm{cl}(A')$ occurs in $\delta_{A'}$ and hence, by (ii′), in no other δ_C. Thus $\iota(d_H(A, B)) \in \mathrm{Int}_L(\iota(A), \iota(B)) = \{d_L(\iota(A), \iota(B))\}$ by Proposition 2.1. On the other hand, if (ii′) does not hold for a box occurring in both δ_C and δ_D, then the diagram representing $\iota(d_H(\emptyset, \{C, D\}))$ contains that box twice, and both occurrences are connected to output wires (through the output bundles of δ_C and δ_D), so they survive normalization; so $\iota(d_H(\emptyset, \{C, D\})) \neq d_L(\emptyset, \iota(C) \otimes \iota(D))$, whose normal form includes the box once.

Step 2: given (i), condition (ii′) is equivalent to

(E) if $X \in \mathrm{cl}(D)$ has a directed path to $\mathrm{cl}(C)$, $C \neq D$, whose interior avoids $\iota(pa^H(C))$, then $D \in pa^H(C)$;

(U_H) for every removed node R there is at most one $C \in \mathbf{V^H}$ such that R has a directed path to $\mathrm{cl}(C)$ whose interior avoids $\iota(pa^H(C))$.

Indeed, all boxes of $\mathrm{cl}(C)$ occur in δ_C, so (ii′) says that the box of a cluster node $X \in \mathrm{cl}(D)$ occurs in no δ_C with $C \neq D$, which by the description of δ_C is (E), and that the box of a removed node occurs in at most one δ_C, which is (U_H).

Step 3: given (i), conditions (E) and (U_H) together are equivalent to (ii) and (iii). Assume (E) and (U_H). An edge $D \to C$ of L/cl is given by a removed-only path from $\mathrm{cl}(D)$ to $\mathrm{cl}(C)$, whose interior avoids $\iota(pa^H(C))$, so $D \in pa^H(C)$ by (E); this is (ii). A removed node that is an ancestor of some cluster reaches at least one cluster by a removed-only path (follow any directed path to a cluster node up to the first cluster node reached), and since a removed-only path avoids every cluster, (U_H) shows that it reaches at most one; this is (U), i.e. (iii). Conversely, assume (ii) and (iii). Every directed path from a node X to $\mathrm{cl}(C)$ whose interior avoids $\iota(pa^H(C))$ is removed-only: otherwise

let $W \in \mathrm{cl}(D'')$ be the last cluster node of its interior; the remainder of the path is a removed-only path from $\mathrm{cl}(D'')$ to $\mathrm{cl}(C)$ with $D'' \neq C$ (the path ends at its first node in $\mathrm{cl}(C)$), so $D'' \to C$ is an edge of $L/\mathrm{cl} \subseteq H$ and $W \in \iota(pa^H(C))$, a contradiction. Hence (E) holds, since a removed-only path from $\mathrm{cl}(D)$ to $\mathrm{cl}(C)$ is an edge of $L/\mathrm{cl} \subseteq H$, and ($\mathrm{U}_H$) reduces to the statement that every removed node reaches at most one cluster by a removed-only path, which follows from (U), as a removed node reaching two clusters by removed-only paths would be an ancestor with two first clusters.

Combining Steps 1 to 3 proves the lemma. □

Proof of Proposition 3.1. Besides the two conditions of the proposition, namely (1.) H is a graphical abstraction of L and (2.) a functor ι as in Definition 3.1 exists, consider

(3.) there exists a clustering cl of L indexed by $\mathbf{V^H}$ that deletes no common causes and satisfies $L/\mathrm{cl} \subseteq H$.

The equivalence of 2. and 3. is Lemma B.1, and the uniqueness statement is (i) of that lemma together with the fact that a Markov functor out of a free Markov category is determined by its values on the generating objects and boxes. It remains to show that 1. and 3. are equivalent.

3. implies 1. Let cl be a clustering that deletes no common causes with $L/\mathrm{cl} \subseteq H$. Call a removed node an *ancestor* if it is an ancestor in L of some cluster node. For an ancestor R, following any directed path from R to a cluster node up to the first cluster node reached yields a removed-only path, so by (U) the cluster $C(R)$ reached does not depend on the path. Consequently, every cluster child of R lies in $\mathrm{cl}(C(R))$, and every removed child R' of R that is an ancestor satisfies $C(R') = C(R)$. Let S_C be the set of ancestors R with $C(R) = C$.

Step 1. Fix a topological order of L. For each $C \in \mathbf{V^H}$, call $B_C := \mathrm{cl}(C) \cup S_C$ the *block* of C, and merge the nodes of B_C into one node by merging them one at a time in the topological order, always merging the next node of B_C into the node obtained from the previous ones. We claim that all intermediate graphs are acyclic, so that these are valid operations. In an intermediate graph G, every node is either an unmerged node of L or the merge of a prefix P_C (in the topological order) of some block B_C; let π map each node of G to its block, or to itself if it is a removed non-ancestor. Suppose G has a cycle γ.

If γ passes through a removed non-ancestor N, then, since all children of N in L are removed non-ancestors, γ consists of removed non-ancestors only and is a cycle of L; a contradiction. If $\pi(\gamma)$ visits two distinct blocks, consider two consecutive distinct blocks D, D' on $\pi(\gamma)$; they are joined by an edge of L from B_D to $B_{D'}$. Its tail is not in S_D: an ancestor $R \in S_D$ has all its cluster children in $\mathrm{cl}(D)$ and all its removed ancestor children in S_D. Hence the tail lies in $\mathrm{cl}(D)$, and prolonging the edge inside $S_{D'}$ (if its head lies in $S_{D'}$) by a removed-only path to $\mathrm{cl}(D')$ gives a removed-only path from $\mathrm{cl}(D)$ to $\mathrm{cl}(D')$, so $D \to D'$ is an edge of L/cl and hence of H. Thus $\pi(\gamma)$ yields a closed walk through at least two distinct nodes of H, i.e. a cycle of H; a contradiction. Finally, if $\pi(\gamma)$ stays inside a single block B_C, then γ passes through the merged prefix node P_C (otherwise it would be a cycle of L), so γ contains an edge $w \to P_C$ with $w \in B_C \setminus P_C$ unmerged, i.e. an edge $w \to u$ of L with $u \in P_C$. But P_C is a prefix of B_C in the topological order, so u precedes w; a contradiction.

Step 2. Delete the removed nodes that are not ancestors, in reverse topological order of L; every child of such a node is again not an ancestor, so each of them is childless at the time of its deletion, and the edges into it simply disappear. Steps 1 and 2 produce a graph H_1 on $\mathbf{V^H}$, and we claim $H_1 = L/\mathrm{cl}$. An edge of H_1 from D to $D' \neq D$ stems from an edge of L from $\mathrm{cl}(D) \cup S_D$ to $\mathrm{cl}(D') \cup S_{D'}$, which by the argument of Step 1 is an edge of L/cl. Conversely, an edge $D \to C$ of L/cl comes from a removed-only path from $\mathrm{cl}(D)$ to $\mathrm{cl}(C)$; its interior nodes are ancestors with first cluster C, hence lie in S_C, so the first edge of the path goes from $\mathrm{cl}(D)$ into $\mathrm{cl}(C) \cup S_C$ and yields the edge $D \to C$ of H_1. Hence $H_1 = L/\mathrm{cl}$.

Step 3. Add the edges of H that are not edges of L/cl one at a time. Every intermediate graph is a subgraph of the DAG H, hence acyclic, so these are valid operations, and the result is H.

1. implies 3. We show by induction along the sequence of operations that every graphical abstraction H of L satisfies $L/\mathrm{cl} \subseteq H$ for a clustering cl that deletes no common causes. For $H = L$ take

the identity clustering. Now let $L/\mathrm{cl} \subseteq H$ with cl deleting no common causes, and let H' be obtained from H by one operation of Definition 3.2. We define a clustering cl' indexed by $\mathbf{V}^{\mathbf{H}'}$ with $L/\mathrm{cl}' \subseteq H'$ that deletes no common causes.

- Merging A, B into a node AB: let $\mathrm{cl}'(AB) := \mathrm{cl}(A) \cup \mathrm{cl}(B)$ and $\mathrm{cl}'(C) := \mathrm{cl}(C)$ otherwise. The removed nodes and the removed-only paths are unchanged, so for $D, C \neq AB$ the edge $D \to C$ belongs to L/cl' iff it belongs to L/cl; $D \to AB$ belongs to L/cl' iff $D \to A$ or $D \to B$ belongs to L/cl, and $AB \to C$ iff $A \to C$ or $B \to C$ does; paths between $\mathrm{cl}(A)$ and $\mathrm{cl}(B)$ yield no edge. Since $L/\mathrm{cl} \subseteq H$, all these edges are edges of the merged graph H'. A removed node whose first cluster was A or B now has first cluster AB, so (U) holds.
- Deleting M with the single child Y: let $\mathrm{cl}'(Y) := \mathrm{cl}(Y) \cup \mathrm{cl}(M)$ and $\mathrm{cl}'(C) := \mathrm{cl}(C)$ otherwise. Again the removed-only paths are unchanged. For $D, C \neq Y$ the edges of L/cl' are those of L/cl; $D \to Y$ belongs to L/cl' iff $D \to Y$ or $D \to M$ belongs to L/cl; and $Y \to C$ belongs to L/cl' iff $Y \to C$ or $M \to C$ belongs to L/cl, where the latter forces $C = Y$ (as Y is the only child of M in $H \supseteq L/\mathrm{cl}$) and yields no edge. Since $L/\mathrm{cl} \subseteq H$ and H' contains the edges from the parents of M to Y, we get $L/\mathrm{cl}' \subseteq H'$. Removed nodes with first cluster M now have first cluster Y, so (U) holds.
- Deleting a childless M: let $\mathrm{cl}'(C) := \mathrm{cl}(C)$ for $C \neq M$, so that the nodes of $\mathrm{cl}(M)$ become removed. Since M has no child in $H \supseteq L/\mathrm{cl}$, there is no removed-only path from $\mathrm{cl}(M)$ to another cluster, and hence (cutting any directed path at the first cluster node it reaches) no node of $\mathrm{cl}(M)$ is an ancestor of a cluster of cl'; the same holds for the removed nodes with first cluster M. For a removed node R with first cluster $C(R) \neq M$, every removed-only path of cl' from R to a cluster avoids $\mathrm{cl}(M)$ (otherwise its initial segment would be a removed-only path of cl from R to $\mathrm{cl}(M)$, forcing $C(R) = M$), so R keeps its unique first cluster, and (U) holds. Likewise a removed-only path of cl' between two clusters cannot pass through $\mathrm{cl}(M)$, as its final segment would be a removed-only path of cl from $\mathrm{cl}(M)$ to a cluster. Hence L/cl' is L/cl with M and its incident edges removed, which is a subgraph of H'.
- Adding an edge: let $\mathrm{cl}' := \mathrm{cl}$; then $L/\mathrm{cl}' = L/\mathrm{cl} \subseteq H \subseteq H'$.

□

Proof of Theorem 3.1

Proof. To simplify notation, for a low-level value $a \in F_L(\mathrm{cl}(A))$ we will denote its associated cluster of values as $\tilde{a}$, both when viewed on the high-level CBN, i.e. $\tilde{a} := \tau_A(a) \in F_H(A)$, as well as when viewed on the low-level CBN, i.e. $\tilde{a} := \tau_A^{-1}(\tau_A(a))) \subset F_L(\mathrm{cl}(A))$.

Part 1: Let F_H be a causal abstraction of F_L as in the categorical Definition 3.1 given by ι and a monoidal natural transformation $\tau : F_L\iota \Rightarrow F_H$. By monoidality, $\tau_A = \prod_{A' \in A} \tau_{A'}$ on $F_L(\mathrm{cl}(A))$ for every $A \subset \mathbf{V}^{\mathbf{H}}$. Let $A, B \subset \mathbf{V}^{\mathbf{H}}$ be disjoint. Naturality of τ with respect to $d_H(A, B)$, together with $\iota(d_H(A, B)) = d_L(\mathrm{cl}(A), \mathrm{cl}(B))$ and Proposition 2.1, gives the commutative square

$$\begin{array}{ccc} F_L(\mathrm{cl}(A)) & \xrightarrow{p^{F_L}(B|do(A))} & F_L(\mathrm{cl}(B)) \\ \downarrow{\scriptstyle \tau_A} & & \downarrow{\scriptstyle \tau_B} \\ F_H(A) & \xrightarrow{p^{F_H}(B|do(A))} & F_H(B) \end{array}$$

Evaluating both composites at $a \in F_L(\mathrm{cl}(A))$ and $\tilde{b} \in F_H(B)$ yields

$$\begin{aligned} p^{F_L}(\tau_B^{-1}(\tilde{b})|do(a)) &= \sum_{b \in \tau_B^{-1}(\tilde{b})} p^{F_L}(b|do(a)) = \big(\tau_B \circ p^{F_L}(B|do(A))\big)(\tilde{b}|a) \\ &= \big(p^{F_H}(B|do(A)) \circ \tau_A\big)(\tilde{b}|a) = p^{F_H}(\tilde{b}|do(\tau_A(a))), \end{aligned}$$

which is eq. (1). It remains to check acyclicity of the cluster graph. A cycle of the cluster graph consists of directed paths of L between clusters; cutting these paths at every cluster node they pass through yields a closed walk of clusters in which consecutive clusters are joined by a path whose interior consists of removed nodes. Since $L/\mathrm{cl} \subseteq H$ by Lemma B.1, each such hop is an edge of H, so H would contain a cycle. Finally, cl deletes no common causes by Lemma B.1 (iii), and the τ_A are surjective, because the square for $d_H(\emptyset, \{A\})$ reads $\tau_A \circ p^{F_L}(A) = p^{F_H}(A)$ and every value $\tilde{a}$ has $p^{F_H}(\tilde{a}) > 0$ by positivity. Hence all conditions of Definition 3.3 hold.

Part 2: Let F_H be a causal abstraction between CBNs of F_L as in Definition 3.3 with clustering cl, so that cl deletes no common causes, i.e. satisfies (U), and let F_H be edge-minimal. Set $H' := L/\mathrm{cl}$. Every edge of H' comes with a directed path of L between the corresponding clusters, so it is an edge of the cluster graph of cl, which is acyclic by Definition 3.3; hence H' is a DAG.

(a) $E^H \subset E^{H'}$. Let $D \to C$ be an edge of H and set $A := pa^H(C) \cup pa^{H'}(C)$, which does not contain C. By eq. (1) applied to $(A, \{C\})$ and the truncated factorization of F_H,

$$p^{F_L}(\tau_C^{-1}(\tilde{c})|do(x_A)) = p^{F_H}(\tilde{c}|do(\tau_A(x_A))) = p^{F_H}\left(\tilde{c} \,\middle|\, \tau_{pa^H(C)}(x_{pa^H(C)})\right)$$

for all $x_A \in F_L(\mathrm{cl}(A))$. By edge-minimality the kernel $p^{F_H}(C|pa^H(C))$ depends on its D-argument, and since τ_D is surjective, the left-hand side depends on the $\mathrm{cl}(D)$-component x_D of x_A for suitable values of the other components. By Proposition 2.1, some node $X \in \mathrm{cl}(D)$ is therefore an ancestor of $\mathrm{cl}(C)$ in $L_{\overline{\mathrm{cl}(A)}}$, i.e. there is a directed path from X to $\mathrm{cl}(C)$ in L that does not enter $\mathrm{cl}(A)$ after its first node. Let W be the last node of this path before its endpoint that lies in a cluster (possibly $W = X$), say $W \in \mathrm{cl}(D'')$. The part of the path after W is a removed-only path from $\mathrm{cl}(D'')$ to $\mathrm{cl}(C)$, so $D'' \to C$ is an edge of $H' = L/\mathrm{cl}$, i.e. $D'' \in pa^{H'}(C) \subset A$. Since the path does not enter $\mathrm{cl}(A)$ after X, this forces $W = X$ and $D'' = D$, so $D \to C$ is an edge of H'.

Now let $\iota' : \mathrm{Free}_{H'} \to \mathrm{Free}_L$ be the Markov functor with $\iota'(C) := \mathrm{cl}(C)$ on nodes and $\iota'(\mathrm{box}_C) := d_L(\iota'(pa^{H'}(C)), \iota'(C))$, and let $F_{H'} : \mathrm{Free}_{H'} \to Stoch$ be the CBN with $F_{H'}(C) := F_H(C)$ and $p^{F_{H'}}(C|pa^{H'}(C)) := p^{F_H}(C|pa^H(C))$, ignoring the additional parents (which is meaningful by (a)).

(b) ι' embeds interventional diagrams. By Lemma B.1, since $H' = L/\mathrm{cl}$ and cl deletes no common causes.

(c) $F_{H'}$ has the same interventional distributions as F_H. Since $H \subset H'$ is acyclic and the mechanisms agree, the truncated factorizations of F_H and $F_{H'}$ coincide for every intervention.

(d) τ is a monoidal natural transformation $F_L\iota' \Rightarrow F_{H'}$. Its components are deterministic, so by the discussion following Proposition 3.1 it suffices to show, for every $C \in \mathbf{V^H}$,

$$\tau_C \circ F_L(\iota'(\mathrm{box}_C)) = p^{F_{H'}}(C|pa^{H'}(C)) \circ \tau_{pa^{H'}(C)}.$$

By Proposition 2.1, $F_L(\iota'(\mathrm{box}_C)) = p^{F_L}(\mathrm{cl}(C)|do(\mathrm{cl}(pa^{H'}(C))))$. For $x \in F_L(\mathrm{cl}(pa^{H'}(C)))$ the left-hand side evaluated at $\tilde{c}$ is

$$\sum_{x_C:\, \tau_C(x_C)=\tilde{c}} p^{F_L}(x_C|do(x)) = p^{F_L}(\tau_C^{-1}(\tilde{c})|do(x)).$$

By eq. (1) applied to $(pa^{H'}(C), \{C\})$ this equals $p^{F_H}(\tilde{c}|do(\tau_{pa^{H'}(C)}(x)))$, and since $pa^{H'}(C) \supset pa^H(C)$, the truncated factorization of F_H gives $p^{F_H}(\tilde{c}|do(\tau_{pa^{H'}(C)}(x))) = p^{F_H}(\tilde{c}|\tau_{pa^H(C)}(x_{pa^H(C)}))$, which is the right-hand side evaluated at x and $\tilde{c}$. □

Proof of Lemma 3.1

Proof. Define for every node A in $\mathbf{V^H}$:

$$\forall a \in F_L(\iota(A)) : \epsilon_A(a|\tilde{a}) := p^{F_L}(a|\tau_A^{-1}(\tilde{a})),$$

which is well defined since $p^{F_L}(\tau_A^{-1}(\tilde{a})) = p^{F_H}(\tilde{a}) > 0$ by naturality of τ and positivity of F_H. Then $\tau_A \circ \epsilon_A = \mathrm{id}_{F_H(A)}$, since $\epsilon_A(\cdot|\tilde{a})$ is supported on $\tau_A^{-1}(\tilde{a})$. □

Proof of Proposition 3.2

Proof. The square of Definition 3.4 for $(pa^H(A), \{A\})$ reads

$$\begin{array}{ccc} F_L(\iota(pa^H(A))) & \xrightarrow{p^{F_L}(A|pa^H(A))} & F_L(\iota(A)) \\ {\scriptstyle \epsilon_{pa^H(A)}}\uparrow & & \uparrow{\scriptstyle \epsilon_A} \\ F_H(pa^H(A)) & \xrightarrow{p^{F_H}(A|pa^H(A))} & F_H(A) \end{array} \tag{13}$$

where $p^{F_L}(A|pa^H(A)) = F_L(\iota(\mathrm{box}_A))$, and where by eq. (5), applied to the set $pa^H(A)$, $\epsilon_{pa^H(A)}(b|\tilde{b}) = p^{F_L}(b|\tilde{b})$ is the joint conditional distribution of the parent clusters given their coarse values. Therefore,

$$p^{F_L}(a|\tilde{b}) = \int_{b\in\tau^{-1}_{pa^H(A)}(\tilde{b})} p^{F_L}(a|b)\cdot p^{F_L}(b|\tilde{b})db \tag{14}$$

$$= \int_{b\in\tau^{-1}_{pa^H(A)})\tilde{b}} p^{F_L}(a|b)\cdot \epsilon_{pa^H(A)}(b|\tilde{b})db \tag{15}$$

$$= \epsilon_A(a|\tilde{a})\cdot p^{F_H}(\tilde{a}|\tilde{b}) \tag{16}$$

$$= p^{F_L}(a|\tilde{a})\cdot p^{F_H}(\tilde{a}|\tilde{b}) \tag{17}$$

$$= p^{F_L}(a|\tilde{a})\cdot \int_{a\in\tau_A^{-1}(\tilde{a})} p^{F_L}(a|\tilde{a})\cdot p^{F_H}(\tilde{a}|\tilde{b})da \tag{18}$$

$$= p^{F_L}(a|\tilde{a})\cdot \int_{a\in\tau_A^{-1}(\tilde{a})} \epsilon_A(a|\tilde{a})\cdot p^{F_H}(\tilde{a}|\tilde{b})da \tag{19}$$

$$= p^{F_L}(a|\tilde{a})\cdot \int_{b\in\tau^{-1}_{pa^H(A)}(\tilde{b})} p^{F_L}(\tilde{a}|b)\cdot \epsilon_{pa^H(A)}(b|\tilde{b})db \tag{20}$$

$$= p^{F_L}(a|\tilde{a})\cdot \int_{b\in\tau^{-1}_{pa^H(A)}(\tilde{b})} p^{F_L}(\tilde{a}|b)\cdot p^{F_L}(b|\tilde{b})db \tag{21}$$

$$= p^{F_L}(a|\tilde{a})\cdot p^{F_L}(\tilde{a}|\tilde{b}) \tag{22}$$

where we used eq. (13) in eq. (16) and eq. (20)(the latter after summing over $a \in \tau_A^{-1}(\tilde{a})$, using that $\epsilon_A(\cdot|\tilde{a})$ is supported on $\tau_A^{-1}(\tilde{a})$), and where we used eq. (5) in eq. (15), eq. (17), eq. (19), and eq. (21). □

Remark. A stochastic natural transformation $\epsilon : F_H \Rightarrow F_L\iota$ on all of Free_H does not exist as soon as some τ_A identifies two values $a \neq a'$ of positive probability: the naturality squares for $\mathrm{copy}_A \circ \mathrm{box}_A$ and for $\mathrm{id}_A \otimes \mathrm{del}_A$ force $\epsilon_{A\otimes A}(a, a'|\tilde{a}, \tilde{a}') = 0$ unless $a = a'$ or $\tilde{a} \neq \tilde{a}'$, with $\epsilon_{A\otimes A}(\cdot,\cdot|\tilde{a}, \tilde{a}')$ supported on $\tau_A^{-1}(\tilde{a}) \times \tau_A^{-1}(\tilde{a}')$, while the square for $\mathrm{box}_A \otimes \mathrm{box}_A$ requires $\sum_{\tilde{a},\tilde{a}'} \epsilon_{A\otimes A}(a, a'|\tilde{a}, \tilde{a}')p^{F_H}(\tilde{a})p^{F_H}(\tilde{a}') = p^{F_L}(a)p^{F_L}(a') > 0$ for $a \neq a'$ in the same fibre; a contradiction. This is why Definition 3.4 only asks for the squares of interventional diagrams, and why the components ϵ_X at tensor objects are part of the data rather than tensor products of the node components. They carry no additional information, however: the square for $(\emptyset, X)$ determines ϵ_X as the conditional distribution of the cluster given its coarse value wherever the latter has positive probability, see eq. (5).

Proof of Proposition 4.1

Proof. $F_{L'} \circ \iota$ is a Markov functor $\mathrm{Free}_{H'} \to M$, hence a causal model over the DAG H', in which the nodes of $\mathbf{U^H}$ are parentless and whose latent projection with respect to $\mathbf{U^H}$ is H; so it is a causal model over the ADMG H. For disjoint $A, B \subset \mathbf{V^{H'}}$, the embedding condition of Definition 3.1 gives

$$(F_{L'}\iota)(d_{H'}(A, B)) = F_{L'}(d_{L'}(\iota(A), \iota(B))),$$

and by Proposition 2.1 (and the definition following it) the left-hand side is $p^{F_{L'\iota}}(B|do(A))$ while the right-hand side is $p^{F_{L'}}(\mathrm{cl}(B)|do(\mathrm{cl}(A)))$.

Now let $F_{L'}$ be a CBN with distribution p. By the first part, $F_{L'\iota}$ is a CBN over the DAG H' whose interventional distributions are those of p on the clusters, so by Proposition 2.1 applied to H' with $B = \mathbf{V^H} \setminus A$, $p(\mathbf{V^H} \setminus A|do(A)) = \int_{\mathbf{U^H}} \prod_{C \in \mathbf{V^{H'}} \setminus A} p^{F_{L'\iota}}(C|pa^{H'}(C))\, d\mathbf{U^H}$. Every exogenous node U of H' is parentless, so its factor is $p^{F_{L'\iota}}(U) = p(U)$; moreover no exogenous node is a child of another node, so the parents of $C \in \mathbf{V^H}$ in H' are its endogenous parents together with U_C, and since no directed path of H' passes through an exogenous node, the endogenous parents of C in H' are exactly $pa_H(C)$ in the latent projection H. Hence the factor of C is $p^{F_{L'\iota}}(C|pa^{H'}(C)) = p(C|do(pa_H(C)), do(U_C))$ by the first part. Finally, since the exogenous nodes of H' are parentless, two endogenous nodes are joined by a bidirected edge in the latent projection H of H' iff they have a common exogenous parent in H', i.e. iff $U_C \cap U_{C'} \neq \emptyset$. □

For the proof of Theorem 4.1 we need that mutilating a graph commutes with passing to the induced graph of a clustering. For a DAG G and sets X, Z of nodes, $G_{\overline{X}\underline{Z}}$ denotes the graph obtained by removing all edges into X and all edges out of Z.

Lemma B.2. *Let* cl *be a clustering of a DAG L' indexed by $\mathbf{V^{H'}}$ that deletes no common causes, with $H' = L'/\mathrm{cl}$, and let $X, Z \subset \mathbf{V^{H'}}$; write $\mathrm{cl}(X) := \bigcup_{C \in X} \mathrm{cl}(C)$. Then* cl*, viewed as a clustering of $L'_{\overline{\mathrm{cl}(X)}\,\underline{\mathrm{cl}(Z)}}$, deletes no common causes, and*

$$L'_{\overline{\mathrm{cl}(X)}\,\underline{\mathrm{cl}(Z)}}/\mathrm{cl} = H'_{\overline{X}\,\underline{Z}}.$$

Moreover, if $C \in \mathbf{V^{H'}} \setminus W$ is not an ancestor of $W \subset \mathbf{V^{H'}}$ in $H'_{\overline{X}}$, then no node of $\mathrm{cl}(C)$ is an ancestor of $\mathrm{cl}(W)$ in $L'_{\overline{\mathrm{cl}(X)}}$.

Proof. A removed-only path from $\mathrm{cl}(D)$ to $\mathrm{cl}(C)$ in L' enters a cluster only with its last edge and leaves a cluster only with its first edge. Hence it uses an edge into $\mathrm{cl}(X)$ iff $C \in X$, and an edge out of $\mathrm{cl}(Z)$ iff $D \in Z$. So the removed-only paths between clusters in $L'_{\overline{\mathrm{cl}(X)}\underline{\mathrm{cl}(Z)}}$ are exactly those of L' from $\mathrm{cl}(D)$ to $\mathrm{cl}(C)$ with $D \notin Z$ and $C \notin X$, which gives the displayed equality. The removed-only paths from a removed node R to a cluster in the mutilated graph form a subset of those in L', so R has at most one first cluster; and if R is an ancestor of some cluster in the mutilated graph, following a directed path to a cluster node up to the first cluster node reached gives a removed-only path, so R has exactly one first cluster. For the last claim, cut a directed path in $L'_{\overline{\mathrm{cl}(X)}}$ from a node of $\mathrm{cl}(C)$ to $\mathrm{cl}(W)$ at all cluster nodes it passes through. Each piece between consecutive cluster nodes of different clusters is a removed-only path, i.e. an edge of $L'_{\overline{\mathrm{cl}(X)}}/\mathrm{cl} = H'_{\overline{X}}$, so C is an ancestor of W in $H'_{\overline{X}}$. □

Proof of Theorem 4.1

Proof. Let H', ι and the clustering cl witness the abstraction, and identify high-level nodes with their clusters, so that the rules concern the sets $\mathrm{cl}(X), \mathrm{cl}(Y), \mathrm{cl}(Z), \mathrm{cl}(W)$ of endogenous low-level nodes. We use two standard facts. First, let G be one of the three mutilations $\overline{X}$, $\overline{X}\,\underline{Z}$ and $\overline{X}\,\overline{Z_H(W)}$, where X, Z and $Z_H(W)$ consist of endogenous nodes, and let H'_G denote H' mutilated at the same sets. Then the latent projection of H'_G is H_G: a directed path or a fork of H' whose interior consists of exogenous nodes loses an edge into X if and only if one of its endpoints lies in X, and loses a directed edge out of Z if and only if it is a directed path starting in Z; so exactly the directed and bidirected edges into X and the directed edges out of Z disappear from the latent projection. Consequently, m-separation of sets of endogenous nodes given endogenous nodes in H_G is equivalent to d-separation of the same sets in H'_G. Second, the three rules of do-calculus hold for the CBN $F_{L'}$ over the DAG L', with the d-separation premises evaluated in the mutilations of L' at the low-level sets. It is therefore sufficient to prove that if Y and Z are d-separated given $X \cup W$ in H'_G, then $\mathrm{cl}(Y)$ and $\mathrm{cl}(Z)$ are d-separated given $\mathrm{cl}(X) \cup \mathrm{cl}(W)$ in the corresponding mutilation of L', for the three mutilations appearing in the rules.

We use the categorical d-separation criterion of Fritz and Klingler [2023] (Definition 19 and Proposition 22 there): for a DAG G and disjoint sets Y, Z, S of its nodes, Y and Z are d-separated given S in G if and only if, in the string diagram $d_G(\emptyset, Y \cup Z \cup S)$ (which is the normalized marginal of the Bayesian network diagram of G in Free_G), after removing all wires labelled by nodes of S, no output wire of Y is joined to an output wire of Z by an undirected path, i.e. a sequence of wires in which consecutive wires are attached to a common box.

Rules 1 and 2. Let G denote $\overline{X}$ or $\overline{X}\,\underline{Z}$, and let L'_G be the corresponding mutilation of L' at $\mathrm{cl}(X)$ and $\mathrm{cl}(Z)$. Let $H'_0 := L'/\mathrm{cl}$, so that $H'_0 \subseteq H'$ by Lemma B.1. By Lemma B.2 applied to H'_0, $L'_G/\mathrm{cl} = (H'_0)_G \subseteq H'_G$ and cl deletes no common causes as a clustering of L'_G, so by Lemma B.1 there is a Markov functor $\iota_G : \mathrm{Free}_{H'_G} \to \mathrm{Free}_{L'_G}$ with $\iota_G(C) = \mathrm{cl}(C)$ on nodes that embeds interventional diagrams; with $S := Y \cup Z \cup X \cup W$ it satisfies $\iota_G(d_{H'_G}(\emptyset, S)) = d_{L'_G}(\emptyset, \mathrm{cl}(S))$. By the proof of Lemma B.1, $d_{L'_G}(\emptyset, \mathrm{cl}(S))$ is represented by the diagram obtained from $d_{H'_G}(\emptyset, S)$ by replacing every box box_C by the diagram $\delta^G_C := d_{L'_G}(\iota_G(pa^{H'_G}(C)), \iota_G(C))$ (the diagram δ_C of Lemma B.1, computed for ι_G) and every wire labelled C by the bundle of wires labelled by the nodes of $\mathrm{cl}(C)$. Every box occurs at most once in the substituted diagram (condition (ii′) in the proof of Lemma B.1), so its normal form is obtained by deleting boxes that are not connected to an output, and every undirected path in the normal form is an undirected path in the substituted diagram. In the substituted diagram, every box belongs to a unique occurrence of some δ^G_C; boxes belonging to different occurrences are attached to a common wire only if that wire belongs to a bundle; and a bundle wire of a cluster D is attached to an occurrence of δ^G_C only if $D \in pa^{H'_G}(C) \cup \{C\}$, i.e. only if the wire D is attached to box_C in $d_{H'_G}(\emptyset, S)$, because the input wires of δ^G_C are the parent bundles and its output wires form the bundle of $\mathrm{cl}(C)$. Consequently, every undirected path with the wires of $\mathrm{cl}(X) \cup \mathrm{cl}(W)$ removed projects to an undirected path in $d_{H'_G}(\emptyset, S)$ with the wires of $X \cup W$ removed: replace each maximal segment running through one occurrence of δ^G_C by the box box_C and each bundle wire by its high-level wire; the result is an undirected path by the preceding sentence, it avoids the wires of $X \cup W$ because a bundle wire is removed if and only if its cluster lies in $X \cup W$, and its endpoints, output wires of nodes of $\mathrm{cl}(Y)$ and $\mathrm{cl}(Z)$, project to output wires of Y and Z. Hence, if Y and Z are d-separated given $X \cup W$ in H'_G, the criterion yields that $\mathrm{cl}(Y)$ and $\mathrm{cl}(Z)$ are d-separated given $\mathrm{cl}(X) \cup \mathrm{cl}(W)$ in L'_G, which is the premise of the respective rule for $F_{L'}$.

Rule 3. Let $Z_L(W)$ be the set of nodes of $\mathrm{cl}(Z)$ that are not ancestors of $\mathrm{cl}(W)$ in $L'_{\overline{\mathrm{cl}(X)}}$; the premise of Rule 3 for $F_{L'}$ is d-separation of $\mathrm{cl}(Y)$ and $\mathrm{cl}(Z)$ given $\mathrm{cl}(X) \cup \mathrm{cl}(W)$ in $L'_{\overline{\mathrm{cl}(X)}\,\overline{Z_L(W)}}$. Applying the argument for Rules 1 and 2 to the mutilation $G = \overline{X}\,\overline{Z_H(W)}$, i.e. to Lemma B.2 with $X \cup Z_H(W)$ in place of X and the empty set in place of Z, shows that the d-separation holds in $L'_{\overline{\mathrm{cl}(X)}\,\overline{\mathrm{cl}(Z_H(W))}}$. For endogenous nodes C and sets W, C is an ancestor of W in $H_{\overline{X}}$ if and only if it is one in $H'_{\overline{X}}$: a directed path of $H'_{\overline{X}}$ splits at its endogenous nodes into pieces with exogenous interior, each of which is an edge of the latent projection $H_{\overline{X}}$, and conversely every edge of $H_{\overline{X}}$ lifts to such a piece. Hence the nodes of $Z_H(W)$ are non-ancestors of W in $H'_{\overline{X}}$ and thus in its subgraph $(H'_0)_{\overline{X}}$, where $H'_0 = L'/\mathrm{cl} \subseteq H'$ as above, and since $Z \cap W = \emptyset$, the last part of Lemma B.2 (applied to H'_0) gives $\mathrm{cl}(Z_H(W)) \subset Z_L(W)$. Removing the edges into $Z_L(W) \setminus \mathrm{cl}(Z_H(W))$ cannot create new d-connections, so the d-separation also holds in $L'_{\overline{\mathrm{cl}(X)}\,\overline{Z_L(W)}}$. □

C Running Example for Section 3

The following example illustrates Definition 3.1, Proposition 3.1 and the generator criterion of Section 3.

Example C.1 (Running example)**.** Let L be the DAG with edges $X_1 \to Y$, $X_2 \to Y$ and $Y \to Z$, and let F_L be a CBN over L with $F_L(X_1) = F_L(X_2) = \{0, 1\}$ and $F_L(Y) = \{0, 1, 2\}$ in which the mechanism of Y depends on its two causes only through their sum, $p^{F_L}(y|x_1, x_2) = q(y|x_1 + x_2)$ for some kernel q. A coarser description keeps only the number $X := X_1 + X_2$ of active causes and the effect Z and forgets the mediator Y: the high-level graph is $H : X \to Z$, with clusters $X = \{X_1, X_2\}$ and $Z = \{Z\}$, while Y is removed. The high-level CBN F_H has $F_H(X) = \{0, 1, 2\}$, $F_H(Z) = F_L(Z)$, $p^{F_H}(\tilde{x}) = \sum_{x_1+x_2=\tilde{x}} p^{F_L}(x_1)p^{F_L}(x_2)$ and $p^{F_H}(z|\tilde{x}) = \sum_y p^{F_L}(z|y)\, q(y|\tilde{x})$.

Graph level. In Example C.1, H arises from L by merging X_1, X_2 and then deleting the mediator Y, which has the single child Z. Had Y a second child W, deleting it would not be allowed: the functor ι would have to send $\mathrm{box}_Z \otimes \mathrm{box}_W$ to a diagram containing box_Y twice, i.e. two independent copies of Y, which does not represent any interventional distribution of F_L.

Mechanism level.

Example C.2 (Running example, continued)**.** In Example C.1, the functor ι sends $X \mapsto X_1 \otimes X_2$, $Z \mapsto Z$, $\mathrm{box}_X \mapsto \mathrm{box}_{X_1} \otimes \mathrm{box}_{X_2}$ and $\mathrm{box}_Z \mapsto \mathrm{box}_Z \circ \mathrm{box}_Y : X_1 \otimes X_2 \to Z$; it embeds interventional diagrams. With $\tau_X(x_1, x_2) := x_1 + x_2$ and $\tau_Z := \mathrm{id}$, the square for box_X commutes by the definition of $p^{F_H}(X)$, and the square for box_Z reads

$$p^{F_L}(z|do(x_1, x_2)) = \sum_y p^{F_L}(z|y)\, q(y|x_1 + x_2) = p^{F_H}(z|\tau_X(x_1, x_2)),$$

so F_H is a causal abstraction of F_L. By contrast, the coarsening $\tau'_X(x_1, x_2) := x_1$, which forgets the second cause, is admissible only if $p^{F_L}(z|do(x_1, x_2)) = \sum_y p^{F_L}(z|y)\, q(y|x_1 + x_2)$ does not depend on x_2, i.e. only if the effect of X_2 on Z is washed out, since the square for box_Z then requires its left-hand side to be a function of x_1 alone.

D Additional Examples

The following two examples complement Section 4.4: they show that clusterings of adjacent variables and the removal of mediators, both admitted by Definition 3.1, are not Φ-abstractions in the sense of Otsuka and Saigo [2022].

Example D.1. Consider a CBN F_L over the DAG $L : A \to B \to C$ and let F_H be the CBN over $H : AB \to C$ obtained by merging A and B, i.e. $F_H(AB) = F_L(A \otimes B)$, $p^{F_H}(a, b) = p^{F_L}(a, b)$ and $p^{F_H}(c|a, b) = p^{F_L}(c|b)$. This is a causal abstraction in the sense of Definition 3.1 with τ the identity. However, there is no graph homomorphism $L \to H$ sending A and B to AB, since the edge $A \to B$ would have to be sent to an edge from AB to itself. The same applies to merging B and C, or all three nodes.

Example D.2. Consider the same CBN F_L and let F_H be the CBN over the graph $H : A \to C$ induced by F_L, i.e. $p^{F_H}(A) = p^{F_L}(A)$ and $p^{F_H}(C|A) = p^{F_L}(C|A)$, which is a causal abstraction in the sense of Definition 3.1 with τ the identity. Since ϕ is defined on all nodes of L, the node B would have to be sent to A or to C, and in either case one of the edges $A \to B$, $B \to C$ would be sent to a self-loop; so there is no graph homomorphism $L \to H$, and F_H is not a Φ-abstraction of F_L.

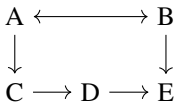
Example D.3. We now discuss an example, adapted from Beckers and Halpern [2019], where this alternative definition may be useful. Consider a voting scenario with 100 voters who can either vote for or against a proposition. The campaign for the proposition can air some subset of two advertisements to try to influence how the voters vote. The low-level model is characterized by binary variables A_1, A_2, and X_i, $i = 1, \ldots, 100$. X_i denotes voter i's vote, so $X_i = 1$ if voter i votes for the proposition, and $X_i = 0$ if voter i votes against. A_i denotes whether ad i is run.

One would like to cluster the voters into groups that are equally affected by the ads. One way to do so, discussed in Beckers and Halpern [2019], is to cluster voters into groups for which the probability $p(x_i|a_1, a_2)$ coincides. However, we can further coarsen the partition of the voters by noting that two voters may have an initial bias towards the proposition independent of the effect the ad (the cause) has on them, i.e. one may instead cluster voters into groups $\mathbf{X_c} = \{X_{c_1}, ..., X_{c_{|X_c|}}\} \subset \mathbf{X}$ such that $\forall x_c := (x_{c_1}, ..., x_{c_{|\mathbf{X_C}|}}) \in \mathbf{X_c}$:

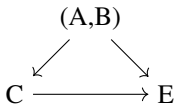
$$p(x_c|a_1, a_2) = p(x_c|\sum_{j=1}^{|\mathbf{X_c}|} x_{c_j}) \cdot p(\sum_{j=1}^{|\mathbf{X_c}|} x_{c_j}|a_1, a_2)$$

The causal mechanism can be captured without loss in a high-level model that does not model every voter but only groups of voters who are affected equally by the ads; within such a group of voters, only the sum of votes needs to be captured in the high-level domain. This is the type of causal abstraction captured by Definition 3.4.

Example D.4. Consider an ADMG L



and a CBN with distribution p that factorizes over L. A graphical abstraction between ADMG (Definition 4.1) of L is given by the ADMG H (in this case, even a DAG) with nodes $(A, B), C, E$ and edges



(i.e., we merged nodes A, B and removed node D). Proposition 4.1 then states that p is also a CBN over this simpler ADMG H. For example, the interventional distribution $p(E|do(C))$ factorizes as

$$\int_{(A,B)} p(A, B) \cdot p(E|(A, B), C).$$

Theorem 4.1 states that applying Pearl's do-calculus on H gives valid statements about the CBN p. For example, applying the second rule of Pearl's do-calculus tells us that $p(E|do(A, B)) = p(E|A, B)$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and the contribution paragraph of Section 1 list exactly the results proved in Sections 3 and 4, and the standing assumptions (acyclic graphs, hard interventions, positivity) are stated in Section 2.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses the restriction to hard interventions and to acyclic graphs (DAGs and ADMGs arising as latent projections of DAGs), and the hypotheses of the converse direction of Theorem 3.1 (edge-minimality, no deleted common causes) are stated with the result.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All results are stated with their assumptions in Sections 2–4 and proved in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [N/A]

Justification: No experiments were conducted.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The paper does not include experiments requiring code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed and adhered to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is purely theoretical and has no direct societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.