CAUSAL ABSTRACTION OF LINEAR SYSTEMS: SUFFICIENT CONDITIONS AND METHODS

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal systems often admit multiple levels of description, where micro-level variables are aggregated into macro-level representations. A valid causal abstraction (CA) must preserve interventional semantics while producing a more compact, interpretable model. We study the problem of learning CAs for linear structural causal models (SCMs), formulating it as a two-stage pipeline: graph partitioning followed by linear aggregation weight learning. For the first stage, we establish sufficient graphical conditions for exact CA and propose three partitioning algorithms including one that guarantees these conditions. For the second stage, we extend algebraic characterizations of CA to a subclass of cyclic SCMs and introduce a differentiable objective for learning aggregation weights as a constrained optimization problem. Together, our end-to-end framework reduces a p-variable system to a target dimension $k < p$. Experiments on synthetic graphs and empirical economic input-output networks demonstrate that our approach is able to recover non-trivial abstractions and outperforms existing unsupervised baselines while offering a principled trade-off between exactness and informativeness.

1 INTRODUCTION

In many scientific and data-driven disciplines, structural causal models (SCMs) are the standard way to represent cause and effect relationships and reason about systems (Haavelmo, 1943; Goldberger, 1972). For complex systems with a large number of interacting variables, it may be practical or insightful to compress a high-dimensional (*low-level*) causal model $\mathcal{M}_{\mathcal{L}}$ that captures all the variables into a low-dimensional (*high-level*) causal model $\mathcal{M}_{\mathcal{H}}$ which aggregates them. Causal abstraction (CA) goes beyond compression, providing both a micro-to-macro state map τ and corresponding intervention map ω such that interventions can be consistently evaluated across levels (Beckers & Halpern, 2019).

Figure 1 gives an illustrative example of a low-level model of six sectors in the UK economy and how it may be abstracted into a high-level model over three macro-sectors. The high-level model simplifies the supply chain and crucially allows us to accurately estimate the effects of, e.g., interventions on the food and beverages sectors on the meat processing macro-sector. There are well-established formal definitions of CA (Beckers & Halpern, 2019; Rubenstein et al., 2017; Rischel, 2020) and CA of linear SCMs have been studied extensively by Massidda et al. (2024). This is limited to acyclic models, and therefore would not apply in our economic example. Previous works have generally assumed prior knowledge of low- and high-level models, studying formal aspects of the abstraction relation rather than constructive learning procedures, with some notable exceptions (Zhu et al., 2024). By contrast, we propose a two-stage CA framework where we first learn a graph partition and then the linear weights, which aggregate the clusters into high-level variables. We extend the theoretical results of Massidda et al. (2024) and propose unsupervised methods for learning the partition and corresponding linear weights. Our contributions include:

1. We provide an algebraic characterisation of constructive abstractions that allows for cyclic models with non-negative weight matrix $\mathbf{A}$ and spectral radius $\rho(\mathbf{A}) < 1$. Our formulation allows us to learn the linear aggregation weights as a constrained optimisation problem *given a graph partition and low-level weight matrix* $\mathbf{A}$.
2. We provide a sufficient condition on partitions of causal graphs (an acyclic high-level graph structure as well as what we call the *All-In* and *One-Child* conditions) that guarantees

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

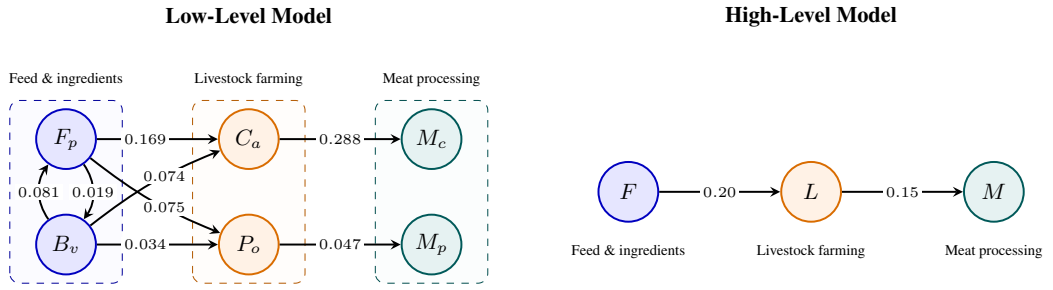


Figure 1: *(Left)* A low-level causal graph over six UK sectors from the EXIOBASE dataset (Stadler et al., 2021; 2018) where edge weights a_{ij} represent the monetary input from industry i required to make 1 unit of output in industry j. (Edge weights $\leq 10^{-3}$ are omitted.) The low-level variables are partitioned into clusters: *Feed & ingredients* comprises a cycle involving food products F_p and beverages B_v; *Livestock farming* comprises cattle C_a and poultry P_o while *Meat processing* comprises cattle meat M_c and poultry meat M_p. This partition satisfies conditions which we prove to be sufficient for CA. *(Right)* The clusters are aggregated into high-level variables by an abstraction map $\mathbf{T}$. The map ω ensures that corresponding interventions in each model are consistent.

> the existence of a constructive abstraction map. The one-child condition, in particular, constrains the high-level structure we are able to represent and often means collapsing much of the low-level graph into a single cluster. We propose a family of partition-learning methods that allows a user to select the abstraction with the lowest information loss whilst controlling the abstraction error within some tolerance level α.

In Section 2 we review core definitions and results from the CA literature. Section 3 establishes a sufficient graphical condition for the existence of a linear CA and extends algebraic characterizations to cyclic SCMs with nonnegative weights and bounded spectral radius. Section 4 introduces our pipeline for learning graph partitions and linear aggregation weights. We validate our framework in Section 5 across synthetic networks with ground-truth abstractions and empirical economic input-output matrices, demonstrating that our approach outperforms competitors while enabling a controllable trade-off between exactness and model informativeness.

1.1 Related Work

Formal frameworks for CA establish how a high-level SCM can serve as a simplified, faithful representation of a detailed low-level SCM. Rubenstein et al. (2017) introduced *exact transformations*, requiring that the pushforward of low-level interventional distributions under a mapping function commutes with interventions performed directly on the high-level model. Beckers & Halpern (2019) formalised a sequence of increasingly restrictive, distribution-free notions of abstraction, culminating in *strong abstractions*, which guarantee interventional consistency across all possible hard interventions independently of any background probability distribution, and *constructive abstractions*, which additionally require high-level variables to form a strict partition over the low-level variables. This framework has since been reformulated category-theoretically (Rischel, 2020; D'Acunto et al., 2025), extended to *approximate abstractions* to account for bounded commutation error (Beckers & Halpern, 2020), and adapted to accommodate soft interventions (Massidda et al., 2023). While these foundational definitions support general functional forms, this work focuses on linear SCMs paired with linear abstraction maps τ as defined by Beckers & Halpern (2019). Because trivial mappings (e.g., collapsing all low-level variables into a single high-level node) formally satisfy these abstraction axioms, an auxiliary goal is to minimise the reconstruction error between the original low-level model and a reconstructed model after abstraction while keeping the commutative error bounded by some threshold.

Recent work has shifted toward learning causal abstractions rather than assuming them *a priori*. Massidda et al. (2024) derive a graphical characterisation of strong abstractions for linear SCMs, though their results do not provide a mechanism to learn an informative abstraction directly from low-level data or adjacency matrices. In the Targeted Causal Reduction paradigm (Kekić et al.,

2024), a high-level target variable of interest Y is assumed, and the objective is to learn an abstraction of the remaining low-level variables focused solely on their downstream effect on Y. Zennaro et al. (2023) study the problem of learning the abstraction map given the low-level and high-level models. When no prior knowledge of the high-level SCM is available, Zhu et al. (2024) proposed an unsupervised method using a KL-divergence term to enforce exact transformations, combined with a PCA-inspired projection loss to prevent trivial solutions. In contrast to Zhu et al. (2024), our work aims to learn *strong abstractions* rather than exact transformations, guaranteeing distribution-independent interventional consistency.

Beyond pure theory, causal abstraction has emerged as a cornerstone of explainable AI (XAI) and causal interpretability. Geiger et al. (2022; 2024) adapted CA to evaluate whether the internal representations of deep neural networks align with a given high-level causal graph over semantically meaningful variables. They also use *interchange interventions* to train the low-level neural activations to align with the high-level causal variables. Parallel lines of research explore relaxed or non-functional variants of CA. For example, Anand et al. (2023) defined *cluster DAGs* which form equivalence classes over low-level causal models that preserve structural properties (such as d-separation and Pearl's *do*-calculus rules) without enforcing exact functional fidelity. Similarly, Chalupka et al. (2016) considered partitioning high-dimensional input spaces based on their interventional effects on an output system to construct macro-level causal variables.

Finally, CA relates to broader multi-scale causal modeling paradigms. Causal Representation Learning (CRL) aims to discover latent high-level causal variables and their relations directly from low-level, unstructured data such as pixels (Locatello et al., 2019; Schölkopf et al., 2021; Ahuja et al., 2023). Unlike CRL, which typically assumes no prior low-level causal graph, Hierarchical Causal Modelling (Weinstein & Blei, 2026) represents low-level and high-level variables simultaneously on the same graph with directed edges connecting macro- and micro-nodes. CA differs from both approaches by maintaining two distinct SCMs operating at different granularities and establishing a formal abstraction mapping between their structural components.

2 BACKGROUND

Notation. A random variable is denoted by a capital letter (e.g., X) and its realized value by a lowercase letter (e.g., x). Multivariate collections are denoted by bold letters. For example, $\mathbf{X}$ may refer to a vector $(X_1, ..., X_n)$ of random variables or a set $\{X_1, ..., X_n\}$ depending on context. If $\Theta \subseteq \{1, ..., n\}$ is an index set then $\mathbf{X}_\Theta := \{X_i \mid i \in \Theta\}$. $\mathcal{R}(\mathbf{X})$ is used to denote the range of values of $\mathbf{X}$ and for $\mathbf{V} \subseteq \mathbf{X}$ and $\mathbf{v} \in \mathcal{R}(\mathbf{V})$, $\text{Rst}(\mathbf{X}, \mathbf{v}) := \{\mathbf{x} \in \mathcal{R}(\mathbf{X}) \mid \mathbf{v} \text{ is the restriction of } \mathbf{x} \text{ to } \mathbf{V}\}$. Matrices are also denoted by bold letters (e.g., $\mathbf{T}$, $\mathbf{S}$). Given a Euclidean space $\mathbb{R}^n$, we define selector matrices $\mathbf{F}_{ii} = \mathbf{e}_i\mathbf{e}_i^T$ for each $i = 1, ..., n$ (where $\mathbf{e}_i$ is the i'th standard basis vector) and $\mathbf{F}_\Theta := \sum_{i\in\Theta} \mathbf{F}_{ii}$ for an index set $\Theta \subseteq \{1, ..., n\}$.

In keeping with Pearl (2009); Peters et al. (2017), we define a linear SCM as follows,

Definition 1 (Linear SCM)**.** A linear *structural causal model* (SCM) is a tuple $\mathcal{M} = (\mathbf{U}, \mathbf{X}, \mathbf{A}, \mathbb{P}(\mathbf{u}))$ where $\mathbf{U}$ is a set of *exogenous* variables; $\mathbf{X}$ is a set of *endogenous* variables; $\mathbf{A}$ is an *adjacency* (weight) matrix such that a_{ij} is the direct effect of X_i on X_j; and $\mathbb{P}(\mathbf{u})$ is a joint probability distribution over $\mathcal{R}(\mathbf{U})$. The values of $\mathbf{X}$ are determined by the equation

$$\mathbf{X} = \mathbf{A}^T\mathbf{X} + \mathbf{U}.$$

A linear SCM is *solvable* if, for any choice of $\mathbf{u} \in \mathcal{R}(\mathbf{U})$, $\mathbf{X}$ is uniquely determined. For solvable linear SCMs, $(\mathbf{I} - \mathbf{A}^T)$ is invertible, yielding the reduced form equation:

$$\mathcal{M}(\mathbf{u}) = (\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{u} = \mathbf{E}^T\mathbf{u},$$

where $\mathbf{E} := (\mathbf{I} - \mathbf{A})^{-1}$ is the *total effect matrix*. For $\mathbf{V} \subseteq \mathbf{X}$ and $\mathbf{v} \in \mathcal{R}(\mathbf{V})$, let $\mathbf{V} \leftarrow \mathbf{v}$ denote a *hard intervention* that sets the values of $\mathbf{V}$ to $\mathbf{v}$ no matter the values of $\mathbf{V}$'s parents. This results in the modified model:

$$\mathcal{M}(\mathbf{u}, \mathbf{V} \leftarrow \mathbf{v}) = \mathbf{E}^T_{\text{do}(\Theta_\mathbf{v})}((\mathbf{I} - \mathbf{F}_{\Theta_\mathbf{v}})\mathbf{u} + \mathbf{F}_{\Theta_\mathbf{v}}\mathbf{v}),$$

where $\Theta_\mathbf{v}$ is the index set corresponding to $\mathbf{V}$, we embed $\mathbf{v}$ in $\mathcal{R}(\mathbf{X})$ and assume that $\mathbf{E}_{\text{do}(\Theta_\mathbf{v})} := (\mathbf{I} - \mathbf{A}(\mathbf{I} - \mathbf{F}_{\Theta_\mathbf{v}}))^{-1}$ exists.

We equip models with a set $\mathcal{I}$ of allowed interventions, considering the full power set $\mathcal{I} = 2^{\mathbf{X}}$ of hard interventions while excluding soft interventions (Eberhardt & Scheines, 2007). An SCM $\mathcal{M}$ induces a directed graph $\mathcal{G} = (\mathbf{X}, \mathcal{E})$, where a directed edge $X_i \to X_j$ exists iff $a_{ij} \neq 0$. In this case, we say X_i is a *parent* of X_j and write $X_i \in \text{Pa}(X_j)$ (equivalently, X_j is a *child* of X_i). Note that an acyclic graph corresponds to a solvable linear SCM, but the graph of a solvable SCM may contain cycles (Bongers et al., 2021). Note also that a solvable linear SCM need not be solvable under an intervention, but we show in Lemma 1 that some cyclic linear SCMs are solvable under all possible interventions. Finally, we assume standard causal sufficiency and faithfulness conditions (Peters et al., 2017; Pearl, 2009).

Let $\mathcal{M}_{\mathcal{L}} = (\mathbf{X}, \mathbf{U}, \mathbf{A}, \mathbb{P}_{\mathcal{L}}(\mathbf{U}))$ be a low-level SCM with graph $\mathcal{G}_{\mathcal{L}}$ and dimension $|\mathbf{X}| = p$, and $\mathcal{M}_{\mathcal{H}} = (\mathbf{Y}, \tilde{\mathbf{U}}, \tilde{\mathbf{A}}, \mathbb{P}_{\mathcal{H}}(\tilde{\mathbf{U}}))$ a high-level SCM with graph $\mathcal{G}_{\mathcal{H}}$ and dimension $|\mathbf{Y}| = k < p$. The framework of Beckers & Halpern (2019) formalizes the core requirements for CA. First observe that a map $\tau : \mathcal{R}(\mathbf{X}) \to \mathcal{R}(\mathbf{Y})$, naturally induces a map from interventions on $\mathbf{X}$ to interventions on $\mathbf{Y}$ defined, for $\mathbf{W} \subseteq \mathbf{Y}$ and $\mathbf{w} \in \mathcal{R}(\mathbf{W})$, by

$$\omega_\tau(\mathbf{V} \leftarrow \mathbf{v}) := \mathbf{W} \leftarrow \mathbf{w} \quad \text{such that there is an equality of sets } \tau(\text{Rst}(\mathbf{X}, \mathbf{v})) = \text{Rst}(\mathbf{Y}, \mathbf{w}).$$

In Figure 1, interventions are only defined on $\{F_p, B_v\}$,$\{C_a, P_o\}$ and $\{M_c, M_p\}$ in the low-level model, and these correspond to interventions on F, L and M (respectively) in the high-level model. The set of subsets $\mathbf{V} \subseteq \mathbf{X}$ on which $\omega_\tau(\mathbf{V} \leftarrow \mathbf{v})$ is defined are denoted by $\mathcal{I}^\tau_{\mathcal{L}}$ and $\mathcal{I}^\tau_{\mathcal{H}} := \omega_\tau(\mathcal{I}^\tau_{\mathcal{L}})$. Below we provide formal definitions of CA (Beckers & Halpern, 2019).

Definition 2. $\mathcal{M}_{\mathcal{H}}$ is a *strong τ-abstraction* of $\mathcal{M}_{\mathcal{L}}$ if τ is a surjective map on endogenous variables; $\omega_\tau(\mathcal{I}^\tau_{\mathcal{L}})$ is the set of all simultaneous high-level interventions; and there is a surjective map γ on exogenous variables that satisfies the following *consistency equations*:

$$\tau(\mathcal{M}_{\mathcal{L}}(\mathbf{u}, \mathbf{V} \leftarrow \mathbf{v})) = \mathcal{M}_{\mathcal{H}}(\gamma(\mathbf{u}), \omega_\tau(\mathbf{V} \leftarrow \mathbf{v})) \quad \forall\, [\mathbf{V} \leftarrow \mathbf{v}] \in \mathcal{I}^\tau_{\mathcal{L}}, \mathbf{u} \in \mathcal{R}(\mathbf{U}) \tag{1}$$

A *constructive τ-abstraction* is a strong τ-abstraction that induces a partition $\{C_1, ..., C_k, C_{k+1}\}$ of $\mathbf{X}$ (with C_i disjoint subsets of $\mathbf{X}$) such that $\tau = (\tau_1, ..., \tau_k)$ with $\tau_i : \mathcal{R}(C_i) \to \mathcal{R}(Y_i)$ for $i = 1, ..., k$ and C_{k+1} a possibly empty set of marginalised variables. We say such an abstraction is *marginalisation-free* if C_{k+1} is empty.

Note that the definition above is independent of the distribution $\mathbb{P}_{\mathcal{L}}(\mathbf{U})$. When we have the distribution, we can define *approximate abstractions* (Beckers & Halpern, 2020).

Definition 3. $\mathcal{M}_{\mathcal{H}}$ is a *τ-α approximate causal abstraction* of $\mathcal{M}_{\mathcal{L}}$ if $d_\tau(\mathcal{M}_{\mathcal{H}}, \mathcal{M}_{\mathcal{L}}) \leq \alpha$ where

$$d_\tau(\mathcal{M}_{\mathcal{H}}, \mathcal{M}_{\mathcal{L}}) := \min_{\gamma} \max_{\mathbf{V} \leftarrow \mathbf{v} \in \mathcal{I}^\tau_{\mathcal{L}}} \mathbb{E}_{\mathbf{u} \sim \mathbb{P}_{\mathcal{L}}}[d_{\mathcal{H}}(\tau(\mathcal{M}_{\mathcal{L}}(\mathbf{u}, \mathbf{V} \leftarrow \mathbf{v})), \mathcal{M}_{\mathcal{H}}(\gamma(\mathbf{u}), \omega_\tau(\mathbf{V} \leftarrow \mathbf{v})))] \tag{2}$$

and $d_{\mathcal{H}}(\cdot, \cdot)$ is a *metric* on $\mathcal{R}(\mathbf{Y})$.

3 Linear Abstractions

Massidda et al. (2024) provide algebraic and graphical characterisations of linear abstractions of linear SCMs with acyclic graphs under some reasonable assumptions. In this section, we show that the algebraic characterisation extends to linear SCMs with cyclic graphs provided the weight matrix $\mathbf{A}$ is non-negative and has spectral radius $\rho(\mathbf{A}) < 1$. We then provide a sufficient condition on graph partitions for the existence of a linear CA. Proofs of key results are in Appendix A.

3.1 Problem Setup

In the linear regime, abstraction maps τ and γ are represented by $p \times k$ transformation matrices $\mathbf{T}, \mathbf{S}$, whose entries in the r-th row and j-th column are denoted t_{rj} and s_{rj}, respectively. The columns of $\mathbf{T}$, which aggregate the low-level endogenous variables, could plausibly lie on the Boolean hypercube, the unit hypersphere, the simplex, or some other constrained space. Each of these might be appropriate in different contexts, but in general we assume $\mathbf{T}, \mathbf{S} \in \mathbb{R}^{p \times k}$. For a constructive $\mathbf{T}$-abstraction, each row of $\mathbf{T}$ contains at most one non-zero entry and under causal sufficiency, $\mathbf{S}$ is likewise constructive. Constructive $\mathbf{T}, \mathbf{S}$ each define partitions on $\mathbf{X}$ and $\mathbf{U}$ respectively.

Definition 4. For constructive $\mathbf{T}, \mathbf{S} \in \mathbb{R}^{p\times k}$ define *endogenous* and *exogenous index partitions*

$$\pi(j) := \{r \mid t_{rj} \neq 0\} \qquad \Pi(j) := \{r \mid s_{rj} \neq 0\} \quad \text{for } j = 1, ..., k$$

Throughout this paper, we assume constructiveness, so that an *abstraction* should be read as a *constructive, strong* $\mathbf{T}$*-abstraction*. For linear abstractions, Equation 1 may be written as

$$\mathbf{T}^T\mathbf{E}^T_{do(\pi(\Theta))}((\mathbf{I} - \mathbf{F}_{\pi(\Theta)})\mathbf{u} + \mathbf{F}_{\pi(\Theta)}\mathbf{v}) = \tilde{\mathbf{E}}^T_{do(\Theta)}((\mathbf{I} - \tilde{\mathbf{F}}_\Theta)(\mathbf{S}^T\mathbf{u}) + \tilde{\mathbf{F}}_\Theta(\mathbf{T}^T\mathbf{v}))$$

When Θ is empty, this yields the *observational consistency* equation $\mathbf{S}\tilde{\mathbf{E}} = \mathbf{E}\mathbf{T}$ (assuming $\mathcal{M}_\mathcal{L}$ and $\mathcal{M}_\mathcal{H}$ are solvable). For non-empty Θ, we have the *interventional consistency* equations which additionally assume solvability under interventions. The following lemma gives conditions under which the *consistency equations* are well-defined.

Lemma 1. *If $\mathbf{A} \in \mathbb{R}^{n\times n}$ satisfies either (i) $\mathbf{A}$ is strictly upper-triangular or (ii) $\mathbf{A}$ is non-negative with spectral radius $\rho(\mathbf{A}) < 1$, then $\mathbf{E} = (\mathbf{I} - \mathbf{A})^{-1}$, $\mathbf{E}_{do(\Theta)} = (\mathbf{I} - \mathbf{A}(\mathbf{I} - \mathbf{F}_\Theta))^{-1}$ exist and the submatrix $\mathbf{E}_{\Theta\Theta}$ (restricted to rows and columns in Θ) is invertible for all $\Theta \subseteq \{1, \dots, n\}$.*

Our formulation of the *consistency equations* together with Lemma 1 allows us to extend our study of linear CA from acyclic SCMs to SCMs with non-negative weight matrix $\mathbf{A}$ and spectral radius $\rho(\mathbf{A}) < 1$. We now address the problem of structural cancellation by defining the following.

Definition 5 ($\mathbf{T}$-faithfulness)**.** Let $\mathcal{M}_\mathcal{H}$ be a $\mathbf{T}$-abstraction of $\mathcal{M}_\mathcal{L}$. We say that the abstraction is $\mathbf{T}$*-faithful* if low-level effects in $\mathcal{M}_\mathcal{L}$ are not cancelled out by aggregation via $\mathbf{T}$ i.e.,

$$(a_{rs} \neq 0 \wedge t_{sj} \neq 0) \implies (\mathbf{AT})_{rj} \neq 0 \quad \text{for all } r, s \in \{1, \dots, p\},\ j \in \{1, \dots, k\}.$$

Unlike Massidda et al. (2024), we do not assume $\mathbf{T}$-faithfulness (although we think it is a desirable property) nor acyclicity. On the other hand, we assume $\mathbf{T}$ is constructive (which they show follows from acyclicity) as well as causal faithulness and causal sufficiency.

3.2 An Algebraic Characterization

Let $B(j) := \pi(j) \cup \Pi(j)$ (for $j = 1, \dots, k$) define a set of *blocks*. With respect to these blocks, we define block matrices $\mathbf{T}, \mathbf{S} \in \mathbb{R}^{p\times k}$ and $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{p\times p}$ according to

$$\mathbf{t}_j := \mathbf{T}_{B(j),j}, \quad \mathbf{s}_j := \mathbf{S}_{B(j),j}, \quad \mathbf{A}_{ij} := \mathbf{A}_{B(i),B(j)}, \quad \mathbf{E}_{ij} := \mathbf{E}_{B(i),B(j)}.$$

The matrices are only well-defined if B partitions $\mathbf{X}$. The following lemma gives a necessary and sufficient condition for this: if X_r maps to Y_j under $\mathbf{T}$ then U_r must map to $\tilde{U}_j$ (or to 0) under $\mathbf{S}$.

Lemma 2. *Let $\mathbf{T}, \mathbf{S}$ be constructive linear transformations. Then the following are equivalent: (i) $\mathbf{S}_{\pi(j),j'} = 0$ for $j \neq j'$ and (ii) $B(j) := \pi(j) \cup \Pi(j)$ partitions the rows of $\mathbf{T}$ and $\mathbf{S}$.*

Massidda et al. (2024) show that $\pi(j) \subseteq \Pi(j)$ but their proof relies on $\mathbf{T}$-faithfulness. We can instead show the following:

Lemma 3. *Let $\mathcal{M}_\mathcal{H}$ be a constructive $\mathbf{T}$-abstraction of $\mathcal{M}_\mathcal{L}$ with exogenous map $\mathbf{S}$. Then $\mathbf{S}_{\pi(j),j'} = \mathbf{0}$ for $j \neq j'$ and if the abstraction is marginalization-free then $\Pi(j) \subseteq \pi(j)$ for $j = 1, ..., k$.*

Lemma 4. *Let $\mathcal{M}_\mathcal{L}$ and $\mathcal{M}_\mathcal{H}$ be linear SCMs with total effect matrices $\mathbf{E}$ and $\tilde{\mathbf{E}}$ respectively. Suppose $\mathbf{T}, \mathbf{S}$ are constructive maps with $\mathbf{S}_{\pi(j),j'} = \mathbf{0}$ for $j \neq j'$. Then interventional consistency for high-level interventions on $\mathbf{Y}_\Theta$ is equivalent to*

$$\mathbf{E}_{do(\pi(\Theta))}\mathbf{T} = (\mathbf{S} + (\mathbf{T} - \mathbf{S})\tilde{\mathbf{F}}_\Theta)\tilde{\mathbf{E}}_{do(\Theta)}$$

The following theorem recovers what we call the *Massidda Equations* (Massidda et al., 2024).

Theorem 5. *Let $\mathcal{M}_\mathcal{L}$ and $\mathcal{M}_\mathcal{H}$ be linear SCMs with adjacency weight matrices $\mathbf{A}$ and $\tilde{\mathbf{A}}$ respectively. $\mathcal{M}_\mathcal{H}$ is a constructive $\mathbf{T}$-abstraction of $\mathcal{M}_\mathcal{L}$ with exogenous map $\mathbf{S}$ if and only if the blocks $B(i) := \pi(i) \cup \Pi(i)$ partition $\mathbf{X}$, and for $i, j = 1, \dots, k$*

$$(\mathbf{I} - \mathbf{A}_{ii})\mathbf{s}_i = (1 - \tilde{a}_{ii})\mathbf{t}_i \quad \textit{and} \quad \mathbf{A}_{ij}\mathbf{s}_j = \tilde{a}_{ij}\mathbf{t}_i \tag{3}$$

We emphasize that we have so far not assumed $\mathbf{T}$-faithfulness or acyclicity; only that if $\mathcal{M}_\mathcal{L}$ or $\mathcal{M}_\mathcal{H}$ are cyclic, then they satisfy the second condition in Lemma 1.

3.3 A Sufficient Graphical Condition

Theorem 5 fully characterizes constructive $\mathbf{T}$-abstractions; yet solving for $\mathbf{T}$ and $\tilde{\mathbf{A}}$ is non-trivial. The following corollary expresses Equation 3 as a set of collinearity conditions:

Corollary 6. *Let $\mathcal{M}_{\mathcal{L}}$ and $\mathcal{M}_{\mathcal{H}}$ be linear SCMs with total effect matrices $\mathbf{A}$ and $\tilde{\mathbf{A}}$ respectively. Suppose $\mathbf{T} \in \mathbb{R}^{p\times k}$ is a constructive matrix and $\mathbf{S} \in \mathbb{R}^{p\times k}$ is such that $\mathbf{S}_{\pi(j),j'} = \mathbf{0}$ for $j \neq j'$. Then $\mathcal{M}_{\mathcal{H}}$ is a constructive $\mathbf{T}$-abstraction of $\mathcal{M}_{\mathcal{L}}$ if and only if for all $i, j = 1, ..., k$*

$$\mathbf{A}_{ij}(\mathbf{I} - \mathbf{A}_{jj})^{-1}\mathbf{t}_j \propto \mathbf{t}_i. \tag{4}$$

If $\mathcal{M}_{\mathcal{H}}$ has no self-loops (i.e. diag$(\tilde{\mathbf{A}}) = \mathbf{0}$), we can recover $\tilde{\mathbf{A}}$ as the proportionality constant.

Solutions to these collinearity conditions turn out to be constrained by the partition π induced by $\mathbf{T}$, motivating our approach of first learning a valid partition. We denote by $\mathcal{P}(\mathcal{G})$, the set of partitions clustering nodes in $\mathcal{G}$ into *disjoint, non-empty* clusters and define, as in Anand et al. (2023),

Definition 6 (Cluster Graph)**.** Let $\mathcal{G} = (\mathbf{X}, \mathcal{E})$ be a graph and $\pi \in \mathcal{P}(\mathcal{G})$, a partition. A *cluster graph* is a graph $\mathcal{G}_\pi = (V_\pi, \mathcal{E}_\pi)$ where V_π is the set of clusters and an edge $\mathbf{X}_{\pi(i)} \to \mathbf{X}_{\pi(j)}$ exists if and only if $i \neq j$ and there exists an edge $X_{i_1} \to X_{j_1}$ from some $i_1 \in \pi(i)$ to some $j_1 \in \pi(j)$.

We present conditions on $\pi \in \mathcal{P}(\mathcal{G}_{\mathcal{L}})$ which, together, are sufficient for *marginalization-free,* $\mathbf{T}$*-faithful abstractions* (in practice, our partition-learning methods do not support learning a marginalized cluster). First, consider the following special case of a result from Massidda et al. (2024):

Proposition 7. *Let $\mathcal{M}_{\mathcal{H}}$ be a marginalization-free, constructive $\mathbf{T}$-abstraction of $\mathcal{M}_{\mathcal{L}}$. For any i, j if there is an edge $Y_i \to Y_j$ then $\forall r \in \pi(i)$, $\exists s \in \pi(j)$ such that there is an edge $X_r \to X_s$. Moreover if the abstraction is also $\mathbf{T}$-faithful then an edge $X_r \to X_s$ implies an edge $Y_i \to Y_j$.*

The All-In condition: $\pi \in \mathcal{P}(\mathcal{G})$ satisfies the *All-In condition* iff for all high-level clusters i, j with $i \neq j$, an edge $X_{r_1} \to X_{s_1}$ for some $r_1 \in \pi(i)$ and $s_1 \in \pi(j)$ implies that for all $r_2 \in \pi(i)$ there is an edge $X_{r_2} \to X_{s_2}$ to some $s_2 \in \pi(j)$ (possibly $s_2 = s_1$).

The corollary of Proposition 7 is that for marginialization-free, $\mathbf{T}$-faithful abstractions, π necessarily satisfies the All-In condition. In Appendix B.2, we show why it is not sufficient while motivating the *One-Child* condition defined below.

The One-Child Condition: *A partition $\pi \in \mathcal{P}(\mathcal{G})$ satisfies the one child condition iff all nodes in the cluster graph $\mathcal{G}_\pi$ have at most one child.*

Theorem 8 (Acyclic)**.** *Let $\mathcal{M}_{\mathcal{L}}$ be a linear SCM with graph $\mathcal{G}_{\mathcal{L}}$ and suppose $\pi \in \mathcal{P}(\mathcal{G}_{\mathcal{L}})$ satisfies the All-In and One-Child conditions. If $\mathcal{G}_\pi$ is acyclic (which holds when $\mathcal{G}_{\mathcal{L}}$ is acyclic), then there exists a marginalization-free, $\mathbf{T}$-faithful constructive $\mathbf{T}$-abstraction of $\mathcal{M}_{\mathcal{L}}$ that aligns with π.*

Theorem 8 admits cyclic $\mathcal{G}_{\mathcal{L}}$ provided cycles are contained within a cluster so that $\mathcal{G}_\pi$ is acyclic. We also emphasize that this is not a necessary condition, merely a sufficient one.

4 Learning Abstractions

We consider the task of learning abstractions of linear SCMs $\mathcal{M}_{\mathcal{L}}$ as consisting of two parts: first learning a partition π and then learning a constructive matrix $\mathbf{T}$ that aligns with π and aggregates the clusters into high-level variables. In this section, we present algorithms for each part of the abstraction pipeline, aiming to learn approximate abstractions within an error tolerance α that maximally *preserve information about $\mathcal{M}_{\mathcal{L}}$* (notions we formalize below).

4.1 Learning $\mathbf{T}$

Corollary 6 allows us to solve for $\mathbf{T}$, given $\pi \in \mathcal{P}(\mathcal{G}_{\mathcal{L}})$, by casting the collinearity condition as a constrained optimization objective,

$$\min_{\mathbf{T}} \quad \sum_{i,j} \|\mathbf{M}_{ij}\mathbf{t}_j - \tilde{a}_{ij}\mathbf{t}_i\|_2^2 \quad \text{subject to} \quad \|\mathbf{t}_i\|_2 = 1 \quad \text{for all } i. \tag{5}$$

Method	All-In ↓	1-child ↓	Intra ↓
AOC Closure	**0.000**	**0.000**	0.626
OC Closure	0.305	**0.000**	0.428
Greedy All-In	**0.000**	0.791	**0.000**
ground truth	0.000	0.650	0.000

Table 1: Our three partition-learning algorithms evaluated on the following metrics: *All-In* is the fraction of node pairs that violate the All-In condition; *1-child* is the number of child-edges in excess of one per cluster divided by the total number of high-level edges and *Intra* is the fraction of low-level edges within clusters. For each of these columns, lower is better. Values are medians over $n = 40$ instances of low-level graphs with a ground truth abstraction. $p = 100$, $k = 10$, high-level density $= 0.5$ and each algorithm is given a budget of 20 clusters.

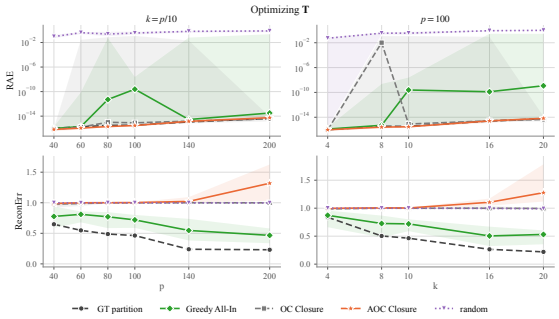


Figure 2: We optimize $\mathbf{T}$ (Alg 4) on a range of partitions: ground-truth (GT) partitions; random partitions of size k and partitions learned by our three greedy algorithms (with budget k). For each setting, we independently sample 20 low-level graphs and report the median and interquartile range of RAE and ReconErr for each partition. RAE appears to be essentially zero in all settings, on the partitions for which an abstraction is known to exist (GT and AOC Closure). All methods appear to beat random partitions on average although the variance over graph instances is high. Greedy All-In, notably achieves low ReconErr indicating it's learning non-trivial abstractions.

where $\mathbf{M}_{ij} := \mathbf{A}_{ij}(\mathbf{I} - \mathbf{A}_{jj})^{-1}$ and $\tilde{a}_{ij} := \mathbf{t}_i^T \mathbf{M}_{ij} \mathbf{t}_j / \|\mathbf{t}_i\|_2^2$ (we assume no self-loops in the high-level model). In Appendix C, we provide details of an optimization routine for this problem. The idea is to observe that for each i, $\mathbf{t}_i$ is completely determined by $\{\mathbf{t}_j \mid j \text{ is a descendant of } i\}$ and so the only free parameters are $\{\mathbf{t}_j \mid \pi(j) \text{ is a sink in } \mathcal{G}_\pi\}$. For nodes $\pi(i)$ in a cycle, the Massidda Equations are overdetermined and imply an eigenvalue equation which we can solve analytically. A degenerate solution of this objective would be to learn $\mathbf{T}$ so that $\mathbf{M}_{ij}\mathbf{t}_j = \mathbf{0}$ even when $\mathbf{A}_{ij} \neq \mathbf{0}$. This corresponds exactly to abstractions that are not $\mathbf{T}$-faithful. We have no guarantees of $\mathbf{T}$-faithful solutions but we find that with a few restarts, the it will often converge on $\mathbf{T}$-faithful solutions (Fig 8). We propose the following metrics for abstractions.

Definition 7. Let $\mathcal{M}_\mathcal{H}, \mathcal{M}_\mathcal{L}$ be linear SCMs such that $\mathcal{M}_\mathcal{H}$ is a constructive $\mathbf{T}$-abstraction of $\mathcal{M}_\mathcal{L}$ with exogenous map $\mathbf{S}$ such that $\mathbf{S}_{\pi(j),j'} = \mathbf{0}$ for $j \neq j'$. The *Relative Abstraction Error* is,

$$\text{RAE} := \max_{\Theta \subseteq \{1,\dots,k\}} \frac{\mathbb{E}[\|\mathbf{u}^T(\mathbf{E}_{do(\pi(\Theta))}\mathbf{T} - (\mathbf{S} + (\mathbf{T} - \mathbf{S})\tilde{\mathbf{F}}_\Theta)\tilde{\mathbf{E}}_{do(\Theta)})\|_2]}{\mathbb{E}[\|\mathbf{u}^T\mathbf{E}_{do(\pi(\Theta))}\mathbf{T}\|_2]},$$

and the *Reconstruction Error* as

$$\text{ReconErr} := \frac{\|\mathbf{A} - \mathbf{T}\tilde{\mathbf{A}}\mathbf{T}^\dagger\|_F}{\|\mathbf{A}\|_F}$$

RAE is related to approximate abstractions (Definition 3) since the numerator gives the expected difference between the LHS and RHS of the consistency equations (Lemma 4). However, taking a maximum error over the interventional value $\mathbf{v}$ is impractical since $\mathbf{v}$ is in general unbounded and therefore, unless the abstraction is exact, so is the error (we do not assume access to an interventional distribution). In practice, sampling over all possible subsets Θ is intractable for large k, and so we consider only the empty set and singleton sets as a proxy. Finally, in our setting, $\mathbf{S}$ is determined directly from $\mathbf{T}$ and so we do not minimize over the exogenous map.

4.2 Learning π

We present three agglomerative clustering algorithms with details described in Appendix C. We saw from Proposition 7 that the All-In condition is a necessary condition on partitions π for learning a marginalization-free, $\mathbf{T}$-faithful abstraction. We also saw that, in addition, the One-Child condition guarantees an exact abstraction if $\mathcal{G}_\pi$ is acyclic. These algorithms aim to satisfy one or both of these conditions while minimizing the fraction of *intra-cluster edges* as a proxy for ReconErr.

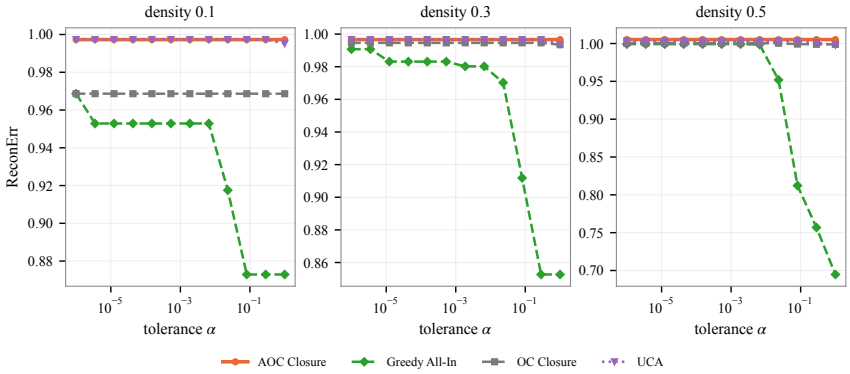


Figure 3: Apart from UCA, all methods are partition-learning methods with $\mathbf{T}$ optimized on the partition they learn (UCA learns $\mathbf{T}$ directly). For each tolerance α fixed in advance, we accept the learned abstraction if $\mathrm{RAE} \leq \alpha$ *and* its ReconErr is strictly below AOC Closure's, otherwise we fall back to AOC Closure which is guaranteed to admit a valid abstraction by Theorem 8. Each point on the plot shows the median ReconErr over the 20 instances per panel. $p = 100$, $k = 10$ and all methods had a budget of 10 clusters.

One-Child Closure (OC Closure). This begins with singleton clusters and iteratively merges clusters with common parents until the One-Child condition is satisfied. As this is order-dependent, we draw several random orders and select the one with the lowest fraction of intra-cluster edges.

All-In and One-Child Closure (AOC Closure). This implements OC Closure to obtain an initial partition. We then merge offending clusters until the All-In condition is satisfied: If $X_r, X_s \in \mathbf{X}_{\pi(i)}$ are such that X_s has an edge into $\mathbf{X}_{\pi(j)}$ $(i \neq j)$ but X_r doesn't, we say that the pair X_r, X_s violate the All-In condition and we merge $\mathbf{X}_{\pi(i)}$ and $\mathbf{X}_{\pi(j)}$.

Greedy All-In. This begins with singleton clusters and iteratively merges clusters, greedily minimizing the number of pairs X_r, X_s that violate the All-In condition.

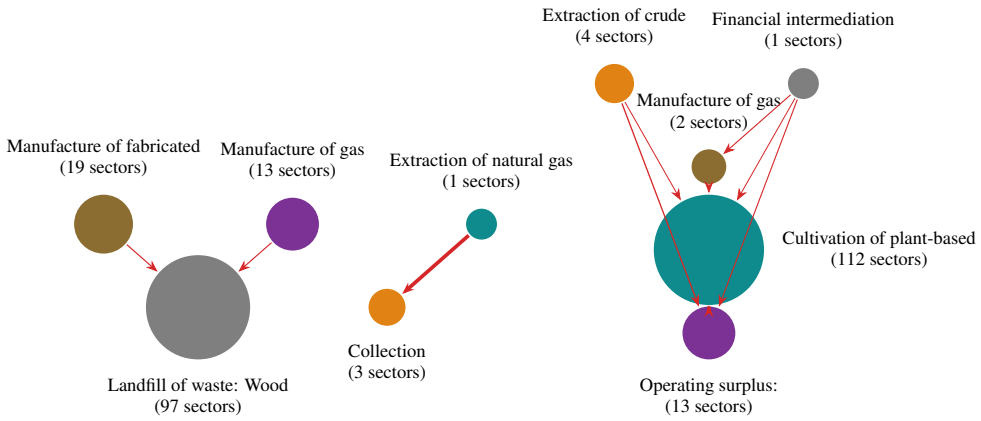
We also investigated a gradient-based partition learning algorithm where we encoded the conditions directly as differentiable loss functions and learned a soft distribution over clusters. Across the different settings we tested it in, it was dominated on average by Greedy All-In on RAE and ReconErr. Although it is competitive with the other baselines on some settings, because of its unreliability, we omit it from the main comparison and provide further details in Appendix C.

5 Experiments

We compare our two-stage pipeline against the only other existing method for unsupervised causal abstraction we are aware of: a method we call *UCA* proposed by Zhu et al. (2024). It learns $\mathbf{T}$ directly via gradient-descent by minimizing a KL-divergence term and a reconstruction term.

Random Graphs. We evaluate our approach on randomly generated directed acyclic graphs (DAGs) with a ground truth abstraction (generation details can be found in Appendix D). We propose fixing an RAE tolerance $\alpha > 0$ in advance and then, for each method, accept the learned abstraction only iff $\mathrm{RAE} \leq \alpha$ and its ReconError is lower than what AOC Closure attains. AOC Closure will always admit an exact abstraction for DAGs by Theorem 5 so and so our goal is to minimize ReconError while remaining within a α tolerance for RAE. Results are presented in Figure 7.

Leontief Input-Output Models Leontief Input-Output models represent an economy as a collection of interdependent sectors, where each sector produces goods consumed by other sectors (internal demand) and final consumers (external demand). Formally, let $\mathbf{x} \in \mathbb{R}^p$ denote the total output vector across p sectors, $\mathbf{M} \in \mathbb{R}^{p \times p}$ denote the consumption matrix, where entry m_{ij} represents the units of product i required to produce one unit of product j, and $\mathbf{d} \in \mathbb{R}^p$ denote the exogenous final demand vector. The macroeconomic equilibrium is governed by $\mathbf{x} = \mathbf{M}\mathbf{x} + \mathbf{d}$. This resembles the equations governing linear SCMs where $\mathbf{M}$ acts as the structural weight matrix. Realistic economies



(a) AOC Closure (b) Greedy All-In

Figure 4: **Abstracting a Leontief input-output model of the UK economy (Stadler et al., 2018; 2021)** With the exception of a handful of subsidised "sectors" included in the dataset (e.g. R&D or various taxes), the consumption matrix $\mathbf{M}$ is non-negative and $\rho(\mathbf{M}) < 1$. As the weights span several orders of magnitude, we prune the consumption matrix at a threshold of 0.05, keeping only the strongest connections (density drops from 0.9 to 0.02 but total absolute weight falls by about 25%) and leaving a low-level graph with 31 self-loops. We label each high-level node with the top contributor from its cluster, determined by the largest absolute value in that column of $\mathbf{T}$. *(Left)* AOC Closure learns a partition which, after optimising $\mathbf{T}$, attains RAE $= 6 \times 10^{-17}$ (essentially 0) and ReconErr $= 0.971$, with about 61% of edges contained within clusters. *(Right)* Greedy All-In learns an abstraction with RAE $= 3 \times 10^{-9}$, ReconErr $= 0.912$ and about 19% of edges landing within clusters.

are such that there are cyclic inter-sector dependencies, but $\mathbf{M}$ is typically non-negative and satisfies $\rho(\mathbf{M}) < 1$ (Leontief, 1986) i.e. the second condition in Lemma 1. Figure 4 shows the result of our abstraction pipeline.

6 Discussion

We have shown that the algebraic characterization of linear abstractions can be extended to linear SCMs with cyclic graphs provided they have non-negative weight matrix $\mathbf{A}$ with spectral radius $\rho(\mathbf{A}) < 1$. In particular, this applies to Leontief input-output models of typical economies. We have also shown that the All-In and One-Child conditions are, together, sufficient conditions on a partition π of an acyclic graph $\mathcal{G}$ such that an exact abstraction $\mathbf{T}$ exists that is aligned with π. We propose three agglomorative clustering algorithms that target one or both of the All-In and One-Child conditions: *OC Closure*, *AOC Closure* and *Greedy All-In* as well as an optimization objective for learning $\mathbf{T}$ given a partition π. Our study on random DAGs suggest that *Greedy All-In* + our $\mathbf{T}$-optimization algorithm dominates the other methods including UCA (the only prior method we are aware of) on the RAE – ReconErr trade–off (though OC Closure appears competitive in low density settings). The One-Child condition (paired with the All-In condition) provides a useful guarantee of an abstraction, but is typically too restrictive on the high-level model. On the other hand, the All-In condition allows for more high-level structure but often at the cost of high RAE (particularly in denser graphs). We would be interested to see if tighter sufficient conditions can be found that are also amenable to efficient learning methods.

In future work, we would be interested to see how our understanding of linear abstractions can be extended to non-linear SCMs and whether they might have applications in XAI. We suspect we will need novel methods for learning $\mathbf{T}$ that scale well with respect to p and k.

AI use statement

In this work, we used generative AI tools for implementing our algorithms and iterating fast on different early variants. For our mathematical proofs and ideas, generative AI tools were used in a limited way, mostly to sanity-check them. However we do owe a few insights to AI: the use of

Gelfand's formula in our proof of Lemma 1 and the simplification of the observational consistency equation from $\mathbf{S}\tilde{\mathbf{E}} = \mathbf{E}\mathbf{T}$ to $(\mathbf{I} - \mathbf{A})\mathbf{S} = \mathbf{T}(\mathbf{I} - \tilde{\mathbf{A}})$ (used in the proof of Theorem 5). We have not used generative AI tools for drafting sections of the paper except for generating pseudocode, tables and figures. We have not used generative AI tools for retrieval and discovery nor for generating synthetic datasets. LLM-generated code has been reviewed by one of the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

Proofs and assumptions are provided in the Appendix as well as additional detail on experiments and algorithms. We provide our source code in the supplementary material.

Table 2: Comparison of assumptions made in this paper and by Massidda et al. (2024).

Assumption	Massidda et al. (2024)	This Paper
Causal Faithfulness	Yes	Yes
Causal Sufficiency	Yes	Yes
$\mathbf{T}$-faithfulness	Yes	No
Acyclicity	Yes	No
Constructive $\mathbf{T}$	No[†]	Yes

[†] They show that under acyclicity assumptions, $\mathbf{T}$-abstractions are necessarily constructive.

A Mathematical Proofs

A summary of our assumptions is shown in Table 2.

A.1 Proof of Lemma 1

Lemma 1. *If $\mathbf{A} \in \mathbb{R}^{n\times n}$ satisfies either (i) $\mathbf{A}$ is strictly upper-triangular or (ii) $\mathbf{A}$ is non-negative with spectral radius $\rho(\mathbf{A}) < 1$, then $\mathbf{E} = (\mathbf{I}-\mathbf{A})^{-1}$, $\mathbf{E}_{do(\Theta)} = (\mathbf{I} - \mathbf{A}(\mathbf{I}-\mathbf{F}_\Theta))^{-1}$ exist and the submatrix $\mathbf{E}_{\Theta\Theta}$ (restricted to rows and columns in Θ) is invertible for all $\Theta \subseteq \{1,\dots,n\}$.*

Proof. To see that $(\mathbf{I}-\mathbf{A})$ is invertible, note that the characteristic polynomial of $\mathbf{A}$ is given by $\det(t\mathbf{I}-\mathbf{A}) = \prod_{i=1}^{n}(t-\lambda_i)$ where λ_i are the complex eigenvalues of $\mathbf{A}$. Under each condition, the largest eigenvalue of $\mathbf{A}$ (by modulus) has modulus less than 1 (i.e. the spectral radius $\rho(\mathbf{A}) < 1$) and so no eigenvalue λ_i equals 1. Therefore $\det(\mathbf{I}-\mathbf{A}) \neq 0$ and $(\mathbf{I}-\mathbf{A})$ is invertible.

Let $\bar{\Theta} := \{1,...,n\} \setminus \Theta$. Ordering indices so that Θ precedes $\bar{\Theta}$, we have wlog

$$\mathbf{I} - \mathbf{A}(\mathbf{I}-\mathbf{F}_\Theta) = \begin{pmatrix} \mathbf{I}_{|\Theta|} & (\mathbf{I}-\mathbf{A})_{\Theta\bar{\Theta}} \\ \mathbf{0} & (\mathbf{I}-\mathbf{A})_{\bar{\Theta}\bar{\Theta}} \end{pmatrix}$$
$$\Rightarrow \det(\mathbf{I}-\mathbf{A}(\mathbf{I}-\mathbf{F}_\Theta)) = \det((\mathbf{I}-\mathbf{A})_{\bar{\Theta}\bar{\Theta}})$$

So it suffices to show that $\det(\mathbf{I}-\mathbf{A}_{\bar{\Theta}\bar{\Theta}}) \neq 0$ or that $\rho(\mathbf{A}_{\bar{\Theta}\bar{\Theta}}) < 1$. If $\mathbf{A}$ is strictly upper-triangular then $(\mathbf{I}-\mathbf{A})_{\bar{\Theta}\bar{\Theta}}$ is upper-triangular with 1's along its main diagonal and is thus invertible. Now suppose $\mathbf{A} \geq 0$ entry-wise and its spectral radius is $\rho(\mathbf{A}) < 1$. By considering directed paths of length $m \geq 0$ between $r, s \in \bar{\Theta}$ we have

$$(\mathbf{A}^m_{\bar{\Theta}\bar{\Theta}})_{rs} \leq (\mathbf{A}^m)_{rs}$$

since the LHS sums over all paths within $\bar{\Theta}$; the RHS sums over all paths in $\{1,...,n\}$ and the edge weights are non-negative. Then by Gelfand's Formula,

$$\rho(\mathbf{A}_{\bar{\Theta}\bar{\Theta}}) = \lim_{m\to\infty} \|\mathbf{A}^m_{\bar{\Theta}\bar{\Theta}}\|_\infty^{1/m} \leq \lim_{m\to\infty} \|\mathbf{A}^m\|_\infty^{1/m} = \rho(\mathbf{A}) < 1$$

Finally, write $\mathbf{F}_\Theta = \mathbf{P}\mathbf{P}^T$ where $\mathbf{P} \in \mathbb{R}^{n\times|\Theta|}$ and the j'th column of $\mathbf{P}$ is given by the θ_j'th standard basis vector $\mathbf{e}_{\theta_j}$ (where θ_j is the j'th element of Θ). The matrix determinant lemma gives

$$\begin{aligned}
\det((\mathbf{I}-\mathbf{A}) + (\mathbf{A}\mathbf{P})\mathbf{P}^T) &= \det(\mathbf{I}-\mathbf{A}) \cdot \det(\mathbf{I}_{|\Theta|} + \mathbf{P}^T(\mathbf{I}-\mathbf{A})^{-1}\mathbf{A}\mathbf{P}) \\
&= \det(\mathbf{I}-\mathbf{A}) \cdot \det(\mathbf{I}_{|\Theta|} + \mathbf{P}^T(\mathbf{E}-\mathbf{I})\mathbf{P}) \\
&= \det(\mathbf{I}-\mathbf{A}) \cdot \det(\mathbf{I}_{|\Theta|} + \mathbf{P}^T\mathbf{E}\mathbf{P} - \mathbf{I}_{|\Theta|}) \\
&= \det(\mathbf{I}-\mathbf{A}) \cdot \det(\mathbf{E}_{\Theta\Theta})
\end{aligned}$$

Since $\det((\mathbf{I}-\mathbf{A}) + (\mathbf{A}\mathbf{P})\mathbf{P}^T) \neq 0$, $\det(\mathbf{E}_{\Theta\Theta}) \neq 0$ and $\mathbf{E}_{\Theta\Theta}$ is invertible. □

A.2 Proof of Lemma 2

Lemma 2. *Let $\mathbf{T}, \mathbf{S}$ be constructive linear transformations. Then the following are equivalent: (i) $\mathbf{S}_{\pi(j),j'} = 0$ for $j \neq j'$ and (ii) $B(j) := \pi(j) \cup \Pi(j)$ partitions the rows of $\mathbf{T}$ and $\mathbf{S}$.*

Proof. We first show that if $\mathbf{S}_{\pi(j),j'} = \mathbf{0}$ for $j \neq j'$ then $B(j) \cap B(j') = \emptyset$ for $j \neq j'$. Indeed, if $r \in B(j) \cap B(j')$ then $r \in \pi(j)$ or $r \in \Pi(j) \setminus \pi(j)$. In the first case, constructiveness of $\mathbf{T}$ implies $r \notin \pi(j')$ and the hypothesis on $\mathbf{S}$ implies $r \notin \Pi(j')$. In the second case, constructiveness of $\mathbf{S}$ implies $r \notin \Pi(j')$ and the hypothesis on $\mathbf{S}$ implies $r \notin \pi(j')$.

Now suppose $B(j)$ partitions the rows of $\mathbf{T}$ and $\mathbf{S}$. If $r \in \pi(j)$ and $r \in \Pi(j')$ for $j \neq j'$ then $r \in B(j) \cap B(j')$ contradicting the fact that $B(j)$ partitions the rows of $\mathbf{T}$ and $\mathbf{S}$. Therefore we must have $\mathbf{S}_{\pi(j),j'} = \mathbf{0}$ for $j \neq j'$. □

A.3 Proof of Lemma 3

Lemma 3. *Let $\mathcal{M}_\mathcal{H}$ be a constructive $\mathbf{T}$-abstraction of $\mathcal{M}_\mathcal{L}$ with exogenous map $\mathbf{S}$. Then $\mathbf{S}_{\pi(j),j'} = \mathbf{0}$ for $j \neq j'$ and if the abstraction is marginalization-free then $\Pi(j) \subseteq \pi(j)$ for $j = 1, ..., k$.*

Proof. Suppose $j \neq j'$ and $r \in \pi(j)$ with $s_{rj'} \neq 0$. By constructiveness of $\mathbf{T}$, $r \notin \pi(j')$ and by causal sufficiency, $s_{rj} = 0$. Now consider the low-level intervention $[\mathbf{X}_{-\pi(j')} \leftarrow \mathbf{x}]$ (which fixes every variable except $\mathbf{X}_{\pi(j')}$) and the corresponding high-level intervention $[\mathbf{Y}_{-j'} \leftarrow \mathbf{y}]$ (which fixes every high-level variable except $Y_{j'}$). For this pair of interventions, the consistency equation requires that

$$\tau(\mathcal{M}_\mathcal{L}(\mathbf{u}, \mathbf{X}_{-\pi(j')} \leftarrow \mathbf{x})) = \mathcal{M}_\mathcal{H}(\gamma(\mathbf{u}), \mathbf{Y}_{-j'} \leftarrow \mathbf{y}) \quad \forall\, \mathbf{u} \in \mathcal{R}(\mathcal{U}_\mathcal{L})$$

$\mathcal{M}_\mathcal{L}(\mathbf{u}, \mathbf{X}_{-\pi(j')} \leftarrow \mathbf{x})$ is a function of $\mathbf{U}_{\pi(j')}$, so the LHS is a function of a subset of $\mathbf{U}_{\pi(j')}$ (a subset since τ could cancel the effects of some of these variables). The RHS is a function of $\tilde{U}_{j'}$ which is a function of U_r since $s_{rj'} \neq 0$. But $U_r \notin \mathbf{U}_{\pi(j')}$ and since the equation must hold for all values of $\mathbf{U}$, $\mathcal{M}_\mathcal{H}$ cannot be a constructive $\mathbf{T}$-abstraction of $\mathcal{M}_\mathcal{L}$.

Furthermore if the abstraction is marginalisation-free then $t_{rj'} = 0 \Rightarrow \exists j \neq j'$ such that $t_{rj} \neq 0$ and thus $s_{rj'} = 0$ since $r \in \pi(j)$. Equivalently $s_{rj'} \neq 0 \Rightarrow t_{rj'} \neq 0$. □

A.4 Proof of Lemma 4

Lemma 4. *Let $\mathcal{M}_\mathcal{L}$ and $\mathcal{M}_\mathcal{H}$ be linear SCMs with total effect matrices $\mathbf{E}$ and $\tilde{\mathbf{E}}$ respectively. Suppose $\mathbf{T}, \mathbf{S}$ are constructive maps with $\mathbf{S}_{\pi(j),j'} = \mathbf{0}$ for $j \neq j'$. Then interventional consistency for high-level interventions on $\mathbf{Y}_\Theta$ is equivalent to*

$$\mathbf{E}_{do(\pi(\Theta))}\mathbf{T} = (\mathbf{S} + (\mathbf{T} - \mathbf{S})\tilde{\mathbf{F}}_\Theta)\tilde{\mathbf{E}}_{do(\Theta)}$$

Proof. By definition of π and the assumption on $\mathbf{S}$, we have for all values of $\mathbf{u}$ and $\mathbf{v}$,

$$\tilde{\mathbf{F}}_\Theta \mathbf{T}^T \mathbf{v} = \tilde{\mathbf{F}}_\Theta \mathbf{T}^T((\mathbf{I} - \mathbf{F}_{\pi(\Theta)})\mathbf{u} + \mathbf{F}_{\pi(\Theta)}\mathbf{v}),$$
$$(\mathbf{I} - \tilde{\mathbf{F}}_\Theta)\mathbf{S}^T\mathbf{u} = (\mathbf{I} - \tilde{\mathbf{F}}_\Theta)\mathbf{S}^T((\mathbf{I} - \mathbf{F}_{\pi(\Theta)})\mathbf{u} + \mathbf{F}_{\pi(\Theta)}\mathbf{v}).$$

For simultaneous interventions on $\mathbf{Y}_\Theta$, the consistency equations require

$$\tau(\mathcal{M}_\mathcal{L}(\mathbf{u}, \mathbf{X}_{\pi(\Theta)} \leftarrow \mathbf{v})) = \mathcal{M}_\mathcal{H}(\gamma(\mathbf{u}), \mathbf{Y}_\Theta \leftarrow \tau_\Theta(\mathbf{v}))$$
$$\Leftrightarrow \mathbf{T}^T\mathbf{E}^T_{do(\pi(\Theta))}((\mathbf{I} - \mathbf{F}_{\pi(\Theta)})\mathbf{u} + \mathbf{F}_{\pi(\Theta)}\mathbf{v}) = \tilde{\mathbf{E}}^T_{do(\Theta)}((\mathbf{I} - \tilde{\mathbf{F}}_\Theta)(\mathbf{S}^T\mathbf{u}) + \tilde{\mathbf{F}}_\Theta(\mathbf{T}^T\mathbf{v}))$$

Substituting the equations above yields a factor of $((\mathbf{I} - \mathbf{F}_{\pi(\Theta)})\mathbf{u} + \mathbf{F}_{\pi(\Theta)}\mathbf{v})$ on both sides and since this must hold for all values of $\mathbf{u}$ and $\mathbf{v}$ we have, after taking the transpose,

$$\mathbf{E}_{do(\pi(\Theta))}\mathbf{T} = (\mathbf{S} + (\mathbf{T} - \mathbf{S})\tilde{\mathbf{F}}_\Theta)\tilde{\mathbf{E}}_{do(\Theta)}$$

□

A.5 Proof of Theorem 5

Theorem 5. *Let $\mathcal{M}_{\mathcal{L}}$ and $\mathcal{M}_{\mathcal{H}}$ be linear SCMs with adjacency weight matrices $\mathbf{A}$ and $\tilde{\mathbf{A}}$ respectively. $\mathcal{M}_{\mathcal{H}}$ is a constructive $\mathbf{T}$-abstraction of $\mathcal{M}_{\mathcal{L}}$ with exogenous map $\mathbf{S}$ if and only if the blocks $B(i) := \pi(i) \cup \Pi(i)$ partition $\mathbf{X}$, and for $i, j = 1, \ldots, k$*

$$(\mathbf{I} - \mathbf{A}_{ii})\mathbf{s}_i = (1 - \tilde{a}_{ii})\mathbf{t}_i \quad \textit{and} \quad \mathbf{A}_{ij}\mathbf{s}_j = \tilde{a}_{ij}\mathbf{t}_i \tag{3}$$

Proof. Consider the following hypotheses

I Observational Consistency is satisfied

II Interventional Consistency is satisfied

III The blocks $B(i)$ partition $\mathbf{X}$

IV $\tilde{e}_{ij}\mathbf{s}_i = \mathbf{E}_{ij}\mathbf{t}_j$ for all $i, j = 1, ..., k$

We show that (**I** and **II**) $\Leftrightarrow$ (**III** and **IV**) by showing the following

- **II** $\Rightarrow$ **III**
- (**I** and **III**) $\Rightarrow$ **IV**
- (**III** and **IV**) $\Rightarrow$ **I**
- (**I** and **III**) $\Rightarrow$ **II**

Lemma 3 says that **II** $\Rightarrow$ **III**. Next we show that **I** and **III** $\Rightarrow$ **II**. Let $\Theta \subseteq \{1, ..., k\}$ be a non-empty subset. Lemma 4 says that if **III** holds then **II** is equivalent to

$$\begin{aligned}
&\mathbf{E}_{do(\pi(\Theta))}\mathbf{T} = (\mathbf{S} + (\mathbf{T} - \mathbf{S})\tilde{\mathbf{F}}_\Theta)\tilde{\mathbf{E}}_{do(\Theta)} \\
\Leftrightarrow\ &\mathbf{T}\tilde{\mathbf{E}}^{-1}_{do(\Theta)} = \mathbf{E}^{-1}_{do(\pi(\Theta))}(\mathbf{S} + (\mathbf{T} - \mathbf{S})\tilde{\mathbf{F}}_\Theta) \\
\Leftrightarrow\ &\mathbf{T}(\mathbf{I} - \tilde{\mathbf{A}} + \tilde{\mathbf{A}}\tilde{\mathbf{F}}_\Theta) = (\mathbf{I} - \mathbf{A} + \mathbf{A}\mathbf{F}_{\pi(\Theta)})(\mathbf{S} + (\mathbf{T} - \mathbf{S})\tilde{\mathbf{F}}_\Theta) \\
\Leftrightarrow\ &\mathbf{T}(\mathbf{I} - \tilde{\mathbf{A}} + \tilde{\mathbf{A}}\tilde{\mathbf{F}}_\Theta) = \mathbf{S} + (\mathbf{T} - \mathbf{S})\tilde{\mathbf{F}}_\Theta - \mathbf{A}\mathbf{S} + \mathbf{A}\mathbf{S}\tilde{\mathbf{F}}_\Theta + \mathbf{A}\mathbf{F}_{\pi(\Theta}\mathbf{S}(\mathbf{I} - \tilde{\mathbf{F}}_\Theta)
\end{aligned}$$

where we use $\mathbf{F}_{\pi(\Theta)}\mathbf{T} = \mathbf{T}\tilde{\mathbf{F}}_\Theta$ which follows by definition of π. Using $(\mathbf{I} - \mathbf{A})\mathbf{S} = \mathbf{T}(\mathbf{I} - \tilde{\mathbf{A}})$ from observational consistency, we have

$$\begin{aligned}
\mathbf{T}(\mathbf{I} - \tilde{\mathbf{A}}) + \mathbf{T}\tilde{\mathbf{A}}\tilde{\mathbf{F}}_\Theta &= (\mathbf{I} - \mathbf{A})\mathbf{S} + (\mathbf{T} - \mathbf{S})\tilde{\mathbf{F}}_\Theta + \mathbf{A}\mathbf{S}\tilde{\mathbf{F}}_\Theta + \mathbf{A}\mathbf{F}_{\pi(\Theta)}\mathbf{S}(\mathbf{I} - \tilde{\mathbf{F}}_\Theta) \\
\Leftrightarrow \mathbf{T}\tilde{\mathbf{A}}\tilde{\mathbf{F}}_\Theta &= (\mathbf{T} - \mathbf{S})\tilde{\mathbf{F}}_\Theta + \mathbf{A}\mathbf{S}\tilde{\mathbf{F}}_\Theta + \mathbf{A}\mathbf{F}_{\pi(\Theta)}\mathbf{S}(\mathbf{I} - \tilde{\mathbf{F}}_\Theta) \\
\Leftrightarrow \mathbf{T}(\tilde{\mathbf{A}} - \mathbf{I})\tilde{\mathbf{F}}_\Theta &= (\mathbf{A} - \mathbf{I})\mathbf{S}\tilde{\mathbf{F}}_\Theta + \mathbf{A}\mathbf{F}_{\pi(\Theta)}\mathbf{S}(\mathbf{I} - \tilde{\mathbf{F}}_\Theta) \\
\Leftrightarrow \mathbf{0} &= \mathbf{A}\mathbf{F}_{\pi(\Theta)}\mathbf{S}(\mathbf{I} - \tilde{\mathbf{F}}_\Theta) \\
\Leftrightarrow \mathbf{0} &= \sum_{j \in \Theta} \mathbf{A}\mathbf{F}_{\pi(j)}\mathbf{S} - \sum_{j \in \Theta}\sum_{j' \in \Theta} \mathbf{A}\mathbf{F}_{\pi(j)}\mathbf{S}\tilde{\mathbf{F}}_{j'j'}
\end{aligned}$$

But for $j \neq j'$, $\mathbf{F}_{\pi(j)}\mathbf{S}\tilde{\mathbf{F}}_{j'j'} = \mathbf{0}$ since $\mathbf{S}_{\pi(j),j'} = \mathbf{0}$ so this expression reduces to

$$\sum_{j \in \Theta} \mathbf{A}\mathbf{F}_{\pi(j)}\mathbf{S}(\mathbf{I} - \tilde{\mathbf{F}}_{jj}) = \mathbf{0}$$

which follows from the same argument.

Finally we show that given **III**, $\mathbf{I} \Leftrightarrow \mathbf{IV}$. Without loss of generality, apply permutations to $\mathbf{X}$ and $\mathbf{Y}$ so that $\mathbf{T}$ and $\mathbf{S}$ are block diagonal (this is possible because of **III**):

$$\mathbf{T} = \begin{pmatrix} \mathbf{t}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{t}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{t}_k \end{pmatrix} \quad \mathbf{S} = \begin{pmatrix} \mathbf{s}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{s}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{s}_k \end{pmatrix}$$

With the same respective permutations, we can write $\mathbf{A}$ and $\tilde{\mathbf{A}}$ as

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \dots & \mathbf{A}_{1k} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \dots & \mathbf{A}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{k1} & \mathbf{A}_{k2} & \dots & \mathbf{A}_{kk} \end{pmatrix} \quad \tilde{\mathbf{A}} = \begin{pmatrix} \tilde{a}_{11} & \tilde{a}_{12} & \dots & \tilde{a}_{1k} \\ \tilde{a}_{21} & \tilde{a}_{22} & \dots & \tilde{a}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{a}_{k1} & \tilde{a}_{k2} & \dots & \tilde{a}_{kk} \end{pmatrix}$$

The observational consistency equation, $\mathbf{S}\tilde{\mathbf{E}} = \mathbf{E}\mathbf{T} \Leftrightarrow (\mathbf{I} - \mathbf{A})\mathbf{S} = \mathbf{T}(\mathbf{I} - \tilde{\mathbf{A}})$, gives

$$(\mathbf{I} - \mathbf{A}_{ii})\mathbf{s}_i = (1 - \tilde{a}_{ii})\mathbf{t}_i \quad \text{and} \quad \mathbf{A}_{ij}\mathbf{s}_j = \tilde{a}_{ij}\mathbf{t}_i$$

for $i, j = 1, \dots, k$ □

A.6 Proof of Corollary 6

Corollary 6. *Let $\mathcal{M}_{\mathcal{L}}$ and $\mathcal{M}_{\mathcal{H}}$ be linear SCMs with total effect matrices $\mathbf{A}$ and $\tilde{\mathbf{A}}$ respectively. Suppose $\mathbf{T} \in \mathbb{R}^{p \times k}$ is a constructive matrix and $\mathbf{S} \in \mathbb{R}^{p \times k}$ is such that $\mathbf{S}_{\pi(j),j'} = \mathbf{0}$ for $j \neq j'$. Then $\mathcal{M}_{\mathcal{H}}$ is a constructive $\mathbf{T}$-abstraction of $\mathcal{M}_{\mathcal{L}}$ if and only if for all $i, j = 1, ..., k$*

$$\mathbf{A}_{ij}(\mathbf{I} - \mathbf{A}_{jj})^{-1}\mathbf{t}_j \propto \mathbf{t}_i. \tag{4}$$

If $\mathcal{M}_{\mathcal{H}}$ has no self-loops (i.e. diag$(\tilde{\mathbf{A}}) = \mathbf{0}$), we can recover $\tilde{\mathbf{A}}$ as the proportionality constant.

Proof. This follows from Theorem 5 by substituting $\mathbf{s}_j = (1 - \tilde{a}_{jj})(\mathbf{I} - \mathbf{A}_{jj})^{-1}\mathbf{t}_j$ into $\mathbf{A}_{ij}\mathbf{s}_j = \tilde{a}_{ij}\mathbf{t}_i$. □

A.7 Proof of Proposition 7

Proposition 7. *Let $\mathcal{M}_{\mathcal{H}}$ be a marginalization-free, constructive $\mathbf{T}$-abstraction of $\mathcal{M}_{\mathcal{L}}$. For any i, j if there is an edge $Y_i \to Y_j$ then $\forall r \in \pi(i)$, $\exists s \in \pi(j)$ such that there is an edge $X_r \to X_s$. Moreover if the abstraction is also $\mathbf{T}$-faithful then an edge $X_r \to X_s$ implies an edge $Y_i \to Y_j$.*

Proof. This is a special case of Theorem 1 from Massidda et al. (2024). Suppose $Y_i \to Y_j$ and consider two interventions on $\mathbf{Y}$: ($Y_i \leftarrow w$,$\mathbf{W} \leftarrow \mathbf{0}$) and ($Y_i \leftarrow w'$,$\mathbf{W} \leftarrow \mathbf{0}$) where $\mathbf{W} = \mathbf{Y} \setminus \{Y_i, Y_j\}$ and $w \neq w'$. Since Y_i has a direct linear effect on Y_j, the value of Y_j under the two interventions is different.

Now suppose $X_r \in \mathbf{X}_{\pi(i)}$ has no edge into $\mathbf{X}_{\pi(j)}$. Consider the following low-level interventions: ($X_r \leftarrow \frac{w}{t_{ri}}, \mathbf{V} \leftarrow 0$) and ($X_r \leftarrow \frac{w'}{t_{ri}}, \mathbf{V} \leftarrow 0$) where $\mathbf{V} = \mathbf{X} \setminus \{X_r, \mathbf{X}_{\pi(j)}\}$. These are exactly in correspondence with the high-level interventions above, however here $\mathbf{Y}_j = \mathbf{t}_j^T \mathbf{X}_{\pi(j)}$ has the same value under the two low-level interventions since X_r has no direct effect on it and any indirect effect is mediated by variables outside of $\mathbf{X}_{\pi(j)}$ which are intervened upon.

Suppose $X_r \to X_s$ with $X_r \in \mathbf{X}_{\pi(i)}$ and $X_s \in \mathbf{X}_{\pi(j)}$. Consider again two low-level interventions: ($\mathbf{X}_{\pi(i)} \leftarrow \mathbf{c}, \mathbf{V} \leftarrow \mathbf{0}$) and ($\mathbf{X}_{\pi(i)} \leftarrow \mathbf{c}', \mathbf{V} \leftarrow \mathbf{0}$) where $\mathbf{V} = \mathbf{X} \setminus \mathbf{X}_{\pi(j)}$ and $\mathbf{c} \neq \mathbf{c}'$ (i.e. we intervene on all low-level variables except $\mathbf{X}_{\pi(j)}$). Under these interventions $Y_j = \mathbf{t}_j^T \mathbf{X}_{\pi(j)}$ takes different values since $\mathbf{X}_{\pi(i)}$ has a direct effect on $\mathbf{X}_{\pi(j)}$. Under the corresponding high-level interventions, Y_j must take different values too and so must be a function of Y_i. □

A.8 Proof of Theorem 8

Theorem 8 (Acyclic)**.** *Let $\mathcal{M}_{\mathcal{L}}$ be a linear SCM with graph $\mathcal{G}_{\mathcal{L}}$ and suppose $\pi \in \mathcal{P}(\mathcal{G}_{\mathcal{L}})$ satisfies the All-In and One-Child conditions. If $\mathcal{G}_{\pi}$ is acyclic (which holds when $\mathcal{G}_{\mathcal{L}}$ is acyclic), then there exists a marginalization-free, $\mathbf{T}$-faithful constructive $\mathbf{T}$-abstraction of $\mathcal{M}_{\mathcal{L}}$ that aligns with π.*

Proof. Suppose $\mathcal{G}_{\pi}$ is acyclic. For each cluster $\mathbf{X}_{\pi(i)}$ the corresponding value of $\mathbf{t}_i$ is determined by $\{\mathbf{t}_j \mid \mathbf{X}_{\pi(j)} \text{ is a child of } \mathbf{X}_{\pi(i)}\}$ since from the Massidda Equations (Eq 3) we have, assuming no self-loops in the high-level model,

$$\tilde{a}_{ij}\mathbf{t}_i = \mathbf{A}_{ij}(\mathbf{I} - \mathbf{A}_{jj})^{-1}\mathbf{t}_j$$

Write $\mathbf{M}_{ij} := \mathbf{A}_{ij}(\mathbf{I} - \mathbf{A}_{jj})^{-1}$ for the total effect of $\mathbf{X}_{\pi(i)}$ on $\mathbf{X}_{\pi(j)}$ mediated through $\mathbf{X}_{\pi(j)}$. By the All-In condition, each row in $\mathbf{A}_{ij}$ has at least one non-zero entry (i.e. it is row-wise non-zero) and by causal faithfulness the same is true of each row of $\mathbf{M}_{ij}$. Now let $\mathbf{X}_{\pi(j)}$ be a sink in $\mathcal{G}_{\pi}$ and consider a chain $\mathbf{X}_{\pi(i)} \to \mathbf{X}_{\pi(i_1)} \to \cdots \to \mathbf{X}_{\pi(i_{l-1})} \to \mathbf{X}_{\pi(j)}$ of length l. Define for each $r = 0, ..., l-1$ (with $i_0 = i$ and $i_l = j$)

$$b^{(r)} := \tilde{a}_{i_r i_{r+1}} \tilde{a}_{i_{r+1} i_{r+2}} \cdots \tilde{a}_{i_{l-1} j}$$
$$\mathbf{B}^{(r)} := \mathbf{M}_{i_r i_{r+1}} \mathbf{M}_{i_{r+1} i_{r+2}} \cdots \mathbf{M}_{i_{l-1} j}$$

Applying the Massidda equations recursively gives

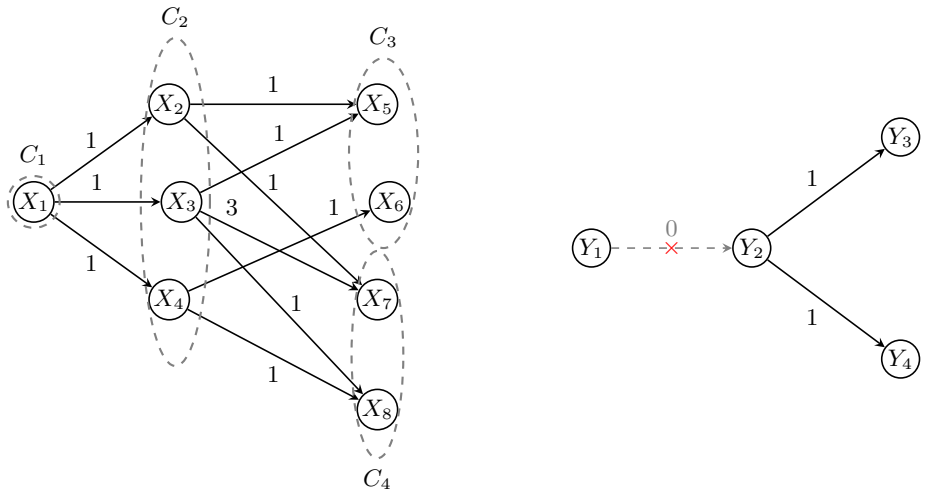
$$b^{(r)}\mathbf{t}_{i_r} = \mathbf{B}^{(r)}\mathbf{t}_j$$

We must have $\mathbf{t}_{i_r}$ entry-wise non-zero for each $r = 0, ..., l$. In fact, we can always pick $\mathbf{t}_j$ so that $\mathbf{t}_j$ and $\{\mathbf{B}^{(r)}\mathbf{t}_j \mid 0 \le r \le l-1\}$ are simultaneously entry-wise non-zero vectors. First note that by causal faithfulness $\mathbf{B}^{(r)}$ is row-wise non-zero since it represents the total effect of $\mathbf{X}_{\pi(i_r)}$ on $\mathbf{X}_{\pi(j)}$ along the path $\mathbf{X}_{\pi(i_r)} \to \cdots \to \mathbf{X}_{\pi(j)}$. Then for any finite l, the set of values of $\mathbf{t}_j$ that violate at least one of the non-zero constraints is a finite union of hyperplanes in $\mathbb{R}^{|\pi(j)|}$. This violating set cannot cover the entire space and so a solution is guaranteed to exist. The same argument generalises to more complex chains as the violating set will still be a finite union of hyperplanes. □

B Examples

B.1 Example 1: $\mathbf{T}$-unfaithful abstraction

Consider this example where a natural partition satisfying the All-In condition, admits no $\mathbf{T}$-faithful abstractions.



$$\mathbf{A}_{12} = (1, 1, 1), \quad \mathbf{A}_{23} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{A}_{24} = \begin{pmatrix} 1 & 0 \\ 3 & 1 \\ 0 & 1 \end{pmatrix}$$
$$\mathbf{t}_2 = (t_{21},\ t_{22},\ t_{23})^T, \quad \mathbf{t}_3 = (t_{31},\ t_{32})^T, \quad \mathbf{t}_4 = (t_{41},\ t_{42})^T$$

Constructive $\mathbf{T}$-abstractions are characterised by the Massidda Equations (Eq 3) which can be written as

$$\tilde{a}_{12}t_1 = \mathbf{A}_{12}(\mathbf{I}-\mathbf{A}_{22})^{-1}\mathbf{t}_2 = (1 \quad 1 \quad 1)\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}^{-1}\begin{pmatrix}t_{21}\\t_{22}\\t_{23}\end{pmatrix} = t_{21}+t_{22}+t_{23}$$

$$\tilde{a}_{23}\mathbf{t}_2 = \mathbf{A}_{23}(\mathbf{I}-\mathbf{A}_{33})^{-1}\mathbf{t}_3 = \begin{pmatrix}1&0\\1&0\\0&1\end{pmatrix}\begin{pmatrix}1&0\\0&1\end{pmatrix}^{-1}\begin{pmatrix}t_{31}\\t_{32}\end{pmatrix} = (t_{31},\ t_{31},\ t_{32})^T$$

$$\tilde{a}_{24}\mathbf{t}_2 = \mathbf{A}_{24}(\mathbf{I}-\mathbf{A}_{44})^{-1}\mathbf{t}_4 = \begin{pmatrix}1&0\\3&1\\0&1\end{pmatrix}\begin{pmatrix}1&0\\0&1\end{pmatrix}^{-1}\begin{pmatrix}t_{41}\\t_{42}\end{pmatrix} = (t_{41},\ 3t_{41}+t_{42},\ t_{42})^T$$

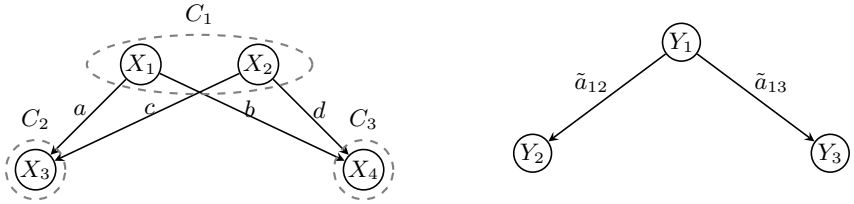
By assumption $t_1, \mathbf{t}_2, \mathbf{t}_3, \mathbf{t}_4 \neq \mathbf{0} \Rightarrow \tilde{a}_{23}, \tilde{a}_{24} \neq 0$ and so we must have

$$\mathbf{t}_2 \in \text{span}((t_{31},\ t_{31},\ t_{32})^T) \cap \text{span}((t_{41},\ 3t_{41}+t_{42},\ t_{42})^T) = \text{span}((1,\ 1,\ -2)^T)$$

But then $\tilde{a}_{12}t_1 = 0$ and therefore $\tilde{a}_{12} = 0$. So a constructive $\mathbf{T}$-abstraction exists with e.g. $t_1 = 1$; $\mathbf{t}_2 = (1,\ 1,\ -2)^T$ and $\mathbf{t}_3 = \mathbf{t}_4 = (1,\ -2)^T$ and in particular any constructive $\mathbf{T}$-abstraction must have $\tilde{a}_{12} = 0$ i.e. no edge $Y_1 \to Y_2$. Indeed $Y_2 = X_2 + X_3 - 2X_4 = U_2 + U_3 - 2U_4$ is independent of $Y_1 = X_1$ since $\mathbf{t}_2$ cancels the effect of X_1 on Y_2.

B.2 Example 2: Insufficiency of the All-In Condition

Consider the following low-level model and proposed abstraction,



$$\mathbf{A}_{12} = \begin{pmatrix}a\\c\end{pmatrix}, \quad \mathbf{A}_{13} = \begin{pmatrix}b\\d\end{pmatrix}, \quad \mathbf{t}_1 = (t_{11},\ t_{12})^T$$

Observe that the partition satisfies the All-In condition and respects the topological ordering of $\mathbf{A}$. For $(i,j) = (1,2)$ and $(1,3)$, the Massidda equations require,

$$\tilde{a}_{12}\mathbf{t}_1 = \mathbf{A}_{12}(\mathbf{I}-\mathbf{A}_{22})^{-1}\mathbf{t}_2 = \begin{pmatrix}a\\c\end{pmatrix}(1-0)^{-1}t_2 = \begin{pmatrix}at_2\\ct_2\end{pmatrix}$$

$$\tilde{a}_{13}\mathbf{t}_1 = \mathbf{A}_{13}(\mathbf{I}-\mathbf{A}_{33})^{-1}\mathbf{t}_3 = \begin{pmatrix}b\\d\end{pmatrix}(1-0)^{-1}t_3 = \begin{pmatrix}bt_3\\dt_3\end{pmatrix}$$

By assumption $t_2, t_3 \neq 0$ and so an abstraction exists if and only if $ad = bc$, a condition on the low-level adjacency matrix $\mathbf{A}$ that is not satisfied in general for arbitrary values a, b, c, d. The children of Y_1 require that there be a non-trivial intersection of $\text{Im}(\mathbf{A}_{12})$ and $\text{Im}(\mathbf{A}_{13})$.

In general, $\mathbf{t}_i$ is constrained by its descendants. Let $\mathbf{M}_{ij} := \mathbf{A}_{ij}(\mathbf{A}_{jj})^{-1}$ and $\mathbf{M}_{(i\to j)} := \mathbf{M}_{ii_1}\mathbf{M}_{i_1i_2}\ldots\mathbf{M}_{i_lj}$ where $(i, i_1, \ldots, i_l, j)$ is the specific path given by $i \to j$. Properly satisfying the Massidda Equations requires that there be a non-trivial intersection of $\text{Im}(\mathbf{M}_{(i\to j)})$ over all paths $i \to j$ and all descendants j of i. Without access to the edge weights $\mathbf{A}$, one solution is to ensure that no high-level cluster has more than one child.

C Algorithms

C.1 Gradient-Based Causal Abstraction Partition-Learning (GradCAP)

We propose the following gradient-based partition-learning algorithm, which we call *GradCAP*. Begin with some choice of k and let $\mathbf{Z} \in \mathbb{R}^{p\times k}$ be a matrix of logits. Let $\mathbf{\Phi} \in [0,1]^{p\times k}$ be a soft partition matrix, implemented by a row-wise softmax function on $\mathbf{Z}$. Given a *binary adjacency matrix* $\mathbf{B}$, define soft reachability matrices $\mathbf{H}, \mathbf{H}' \in [0,1]^{p\times k}$ by

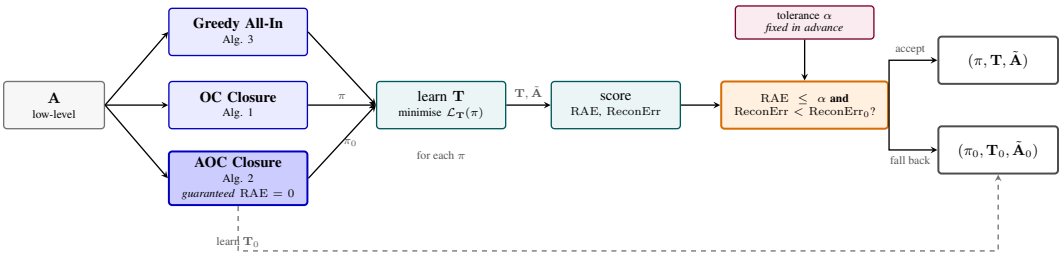


Figure 5: **Overview of the abstraction pipeline.** Flowchart illustrating high-level abstraction fitting, scoring, and condition-checking.

$$h_{rj} = \max_{1 \leq l \leq p} b_{lr}\phi_{lj}, \qquad h'_{rj} = (1 - \phi_{rj}) \cdot \max_{1 \leq l \leq p} b_{rl}\phi_{lj}.$$

h_{rj} is a measure of the strength of edges from cluster j to X_r while h'_{rj} measures the strength of edges from X_r to cluster j and zeros out connections to one's own cluster. We also define a mask $\mathbf{M} \in [0,1]^{p \times p}$, which gives the co-assignment of two low-level variables X_r and X_s and zeros out the main diagonal, and pair-wise cross-entropy matrices [1] $\mathbf{C}, \mathbf{C}' \in \mathbb{R}^{p \times p}$ by

$$m_{rs} := \sum_j \phi_{rj}\phi_{sj}(1 - \delta_{rs}), \quad c_{rs} := \sum_j \mathrm{CE}(h_{rj}, h_{sj}), \quad c'_{rs} := \sum_j \mathrm{CE}(h'_{rj}, h'_{sj}). \tag{6}$$

Finally, define $\mathcal{L}(\lambda_{\text{in}}, \lambda_{\text{out}}, \lambda_{\text{one-child}}) = \lambda_{\text{in}}\mathcal{L}_{\text{in}} + \lambda_{\text{out}}\mathcal{L}_{\text{out}} + \lambda_{\text{one-child}}\mathcal{L}_{\text{one-child}}$, a family of losses parameterized by $\lambda_{\text{in}}, \lambda_{\text{out}}, \lambda_{\text{one-child}} \in \mathbb{R}_{\geq 0}$ where

$$\mathcal{L}_{\text{in}} := \tfrac{k}{p^2}\langle \mathbf{M}, \mathbf{C} \rangle_F, \quad \mathcal{L}_{\text{out}} := \tfrac{k}{p^2}\langle \mathbf{M}, \mathbf{C}' \rangle_F, \quad \mathcal{L}_{\text{one-child}} := \tfrac{1}{p}\sum_r (|\mathbf{h}'_r|_1^2 - |\mathbf{h}'_r|_2^2), \tag{7}$$

$\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product and $\mathbf{h}'_r$ is the r'th row vector of $\mathbf{H}'$. $\mathcal{L}(\lambda_{\text{in}}, \lambda_{\text{out}}, \lambda_{\text{one-child}})$ is differentiable with respect to $\mathbf{Z}$ almost everywhere and depends only on the low-level (binary) adjacency matrix. This allows us to, beginning with an initial $\mathbf{Z}_0$, run gradient-descent on for a fixed number of epochs and then assign low-level variable r to a cluster in $\arg\max_j(\phi_{rj})$. Note that $\mathcal{L}_{\text{out}}$ encodes the All-Out condition while $\mathcal{L}_{\text{in}}$ encodes the analogous *All-In* condition which requires nodes in a cluster to have the same set of cluster parents.

D Data Generation

We generate our random DAGs with ground truth abstractions by first sampling a high-level model with the Erdos-Renyi model at density d. We then select our low-level dimension p, sample a random partition $\{\pi(1), \ldots, \pi(k)\}$ and a ground truth abstraction map $\mathbf{T}$. We pass these as inputs to the following algorithm from Massidda et al. (2024) to generate the low-level model. We reject low-level graphs that are disconnected and we shuffle the low-level indices so that the topological ordering is not known in advance. The *blockiness* parameter $\beta \in [0,1]$ in our concretization algorithm (Alg 5) controls how structurally homogenous the abstraction blocks are. At $\beta = 0$, they minimally satisfy the All-In condition (i.e. each node in cluster $\pi(i)$ has one edge into one other node in $\pi(j)$) whereas at $\beta = 1$ they maximally satisfy the condition so that every node is one cluster is connected to every node in another. Reconstruction Error is controlled by the ground truth β since even the ground truth partition that generated the low-level graph cannot achieve 0 ReconErr unless $\beta = 1$. In our experiments throughout this paper, we have fixed $\beta = 0.8$ to strike a balance between being realistic while allowing ReconErr to discriminate freely between abstractions.

[1] Note that the cross-entropy function has no information-theoretic significance here; it is merely a useful mathematical function that, in conjunction with the mask, encodes All-Out violations while simultaneously encouraging the soft partition Φ to harden.

Algorithm 1: OC CLOSURE

Input: $\mathbf{A} \in \mathbb{R}^{p \times p}$; d
Output: $\pi^\star$
$\pi^\star \leftarrow \varnothing$; $\ell^\star \leftarrow \infty$
for $t = 1, \ldots, d$ **do**
  $\pi \leftarrow$ `OneChildClosure(A)`
  if $\mathrm{Intra}(\pi) < \ell^\star$ **then**
    $\pi^\star \leftarrow \pi$; $\ell^\star \leftarrow \mathrm{Intra}(\pi)$
  end
end
return $\pi^\star$

Function `OneChildClosure(A)`**:**
  $\pi \leftarrow \{\{1\}, \ldots, \{p\}\}$
  repeat
    foreach $\mathbf{X} \in \pi$ **do**
      $\mathcal{C} \leftarrow \mathrm{children}(\mathbf{X})$
      if $|\mathcal{C}| > 1$ **then** merge $\mathcal{C}$
    end
  until *no merge*
  return π

Algorithm 2: AOC CLOSURE

Input: $\mathbf{A}$; d; budget k
Output: π_0
$\pi_{\mathrm{valid}}, \pi_{\mathrm{wide}} \leftarrow \varnothing$
for $t = 1, \ldots, d$ **do**
  $\pi \leftarrow \mathrm{OCC}(\mathbf{A})$
  if $|\pi| > |\pi_{\mathrm{wide}}|$ **then** $\pi_{\mathrm{wide}} \leftarrow \pi$
  if $\mathrm{AllIn}(\pi) \wedge |\pi| > |\pi_{\mathrm{valid}}|$ **then**
    $\pi_{\mathrm{valid}} \leftarrow \pi$
end
if $\pi_{\mathrm{valid}} \neq \varnothing$ **then**
  $\pi_0 \leftarrow \pi_{\mathrm{valid}}$
else
  $\pi_0 \leftarrow \pi_{\mathrm{wide}}$
  while $\exists$ *violating* $\mathbf{X} \to \mathbf{Y}$ **do**
    $\pi_0 \leftarrow \mathrm{merge}(\mathbf{X}, \mathbf{Y})$
  end
end
if k given **then** $\pi_0 \leftarrow \textsc{Coarsen}(\pi_0, k)$
return π_0

Algorithm 3: GREEDY ALL-IN

Input: $\mathbf{A}$; budget k; c
Output: π with $|\pi| = k$
$\pi \leftarrow \{\{1\}, \ldots, \{p\}\}$
$\mathbf{\Sigma} \leftarrow$ Jaccard similarity
while $|\pi| > k$ **do**
  $\mathcal{P} \leftarrow$ top c pairs under $\mathbf{\Sigma}$
  foreach $(\mathbf{X}, \mathbf{Y}) \in \mathcal{P}$ **do**
    $\pi' \leftarrow \mathrm{merge}(\mathbf{X}, \mathbf{Y})$
    $s_{\mathbf{XY}} \leftarrow \big(V(\pi'), \Delta\mathrm{Intra}\big)$
  end
  $\pi \leftarrow \mathrm{merge}(\arg\min s_{\mathbf{XY}})$
end
return π

Figure 6: **Partition learning algorithms.** $\mathrm{Intra}(\cdot)$ is the fraction of edges that are within clusters. $V(\pi)$ is the All-In violation count for a partition π (the number of node pairs that violate the condition)

Algorithm 4: Abstraction fit: cycles pinned by an eigenproblem, sinks optimised

Input: low-level adjacency $\mathbf{A} \in \mathbb{R}^{p\times p}$; hard partition $\{\pi(1), \ldots, \pi(k)\}$; restarts R
Output: high-level adjacency $\tilde{\mathbf{A}} \in \mathbb{R}^{k\times k}$; abstraction $\mathbf{T} \in \mathbb{R}^{p\times k}$

$\mathcal{E}_\pi \leftarrow \{(i,j) : i \neq j,\ \mathbf{A}_{\pi(i)\pi(j)} \neq \mathbf{0}\}$ `// cluster graph edges`
foreach $(i,j) \in \mathcal{E}_\pi$ **do** $\mathbf{M}_{ij} \leftarrow \mathbf{A}_{\pi(i)\pi(j)}\big(\mathbf{I} - \mathbf{A}_{\pi(j)\pi(j)}\big)^{-1}$
$\mathcal{C}_1, \ldots, \mathcal{C}_r \leftarrow$ StronglyConnectedComponents$(\mathcal{E}_\pi)$
order $\mathcal{C}_1, \ldots, \mathcal{C}_r$ so each follows every component it points into

`// Phase 1 -- Massidda Equations become an overdetermined system of eigenvector equatinos in SCCs. We solve one of these equations analytically.`
foreach $\mathcal{C}$ with $|\mathcal{C}| > 1$ ***or*** a self-loop **do**
  foreach $i \in \mathcal{C}$ **do**
    $(i = c_1, c_2, \ldots, c_m) \leftarrow$ shortest simple cycle through i inside $\mathcal{C}$
    $\mathbf{\Pi} \leftarrow \mathbf{M}_{c_1c_2}\mathbf{M}_{c_2c_3}\cdots\mathbf{M}_{c_mc_1}$
    $\mathbf{t}_i \leftarrow$ unit eigenvector of $\mathbf{\Pi}$ for the largest-magnitude eigenvalue
  end
end

`// Phase 2 -- the sink clusters are the only ones whose` $\mathbf{t}_i$ `is a free parameter.`
$\mathcal{S} \leftarrow \{i : \mathcal{C} = \{i\}$ acyclic, $\neq \mathbf{x}ists\ j\ (i,j) \in \mathcal{E}_\pi\}$; $\quad n \leftarrow \sum_{i\in\mathcal{S}} |\pi_i|$
if $n = 0$ **then**
  $(t, \rho) \leftarrow$ `Propagate(∅)` `// no sinks: closed form,` R `unused`
else
  $\mathbf{x}^\star \leftarrow \arg\min_{\mathbf{x}\in\mathbb{R}^n} \rho(\mathbf{x})$, where $\rho(\mathbf{x})$ is returned by `Propagate(x)`, by SLSQP from R random starts, keeping the best
  $(t, \rho) \leftarrow$ `Propagate(`$\mathbf{x}^\star$`)`
end

$\tilde{a}_{ij} \leftarrow \langle \mathbf{M}_{ij}\mathbf{t}_j,\ \mathbf{t}_i\rangle / \langle \mathbf{t}_i, \mathbf{t}_i\rangle$ for $(i,j) \in \mathcal{E}_\pi$, and 0 otherwise
$\mathbf{T} \leftarrow$ the $p \times k$ matrix with $\mathbf{t}_i$ placed on rows $\pi(i)$ of column i
return $\tilde{\mathbf{A}}, \mathbf{T}$

Function `Propagate(x):`
  foreach $i \in \mathcal{S}$ **do** $\mathbf{t}_i \leftarrow \mathbf{x}_{\pi(i)} / \|\mathbf{x}_{\pi(i)}\|$ `// sink block of` $\mathbf{x}$
  $\rho \leftarrow 0$
  foreach acyclic singleton $\mathcal{C} = \{i\}$, $i \notin \mathcal{S}$, in order **do**
    $\mathcal{K} \leftarrow \{j : (i,j) \in \mathcal{E}_\pi\}$
    $v_j \leftarrow \mathbf{M}_{ij}\mathbf{t}_j / \|\mathbf{M}_{ij}\mathbf{t}_j\|$ for $j \in \mathcal{K}$, sign-aligned to v_{j_1}
    $\mathbf{t}_i \leftarrow \overline{v}/\|\overline{v}\|$; $\quad \rho \leftarrow \rho + \sum_j \|v_j - \mathbf{t}_i\|^2$ `// collinearity disagreement`
  end
  return t, ρ

Algorithm 5: $\mathbf{T}$-concretisation sampling

Input: high-level $\tilde{\mathbf{A}} \in \mathbb{R}^{k\times k}$; abstraction $\mathbf{T} \in \mathbb{R}^{p\times k}$ with blocks $\pi(1), \dots, \pi(k)$; internal density δ; blockiness β; heterogeneity η

Output: $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $\mathbf{S}$ with $(\mathbf{I} - \mathbf{A})\mathbf{S} = \mathbf{T}(\mathbf{I} - \tilde{\mathbf{A}})$

foreach $j = 1, \dots, k$ **do**
  $\mathbf{A}_{\pi(j)\pi(j)} \leftarrow$ random DAG on $|\pi(j)|$ nodes at density δ
  $\mathbf{s}_j \leftarrow (\mathbf{I} - \mathbf{A}_{\pi(j)\pi(j)})^{-1}\mathbf{t}_j$
end
foreach (i, j) with $i \neq j$, $\tilde{a}_{ij} \neq 0$ **do**
  foreach the ℓ-th member r of $\pi(i)$ **do**
    $\mathcal{I} \leftarrow \{n^\star\} \cup \{n : b_n = 1\}, \quad n^\star \sim \mathcal{U}\{n : s_j[n] \neq 0\}, b_n \sim \mathrm{Bern}(\beta)$
    $c \leftarrow \mathbf{0}; \quad c_{\mathcal{I}} \leftarrow \bar{s}/(\bar{s}^\top \bar{s}) + \eta\,\epsilon, \quad \bar{s} := s_j[\mathcal{I}], \epsilon \perp \bar{s}, \|\epsilon\| = 1$ // so $c^\top s_j = 1$
    $\mathbf{A}_{r,\pi(j)} \leftarrow \tilde{a}_{ij}\,\mathbf{t}_i[\ell]\,c^\top$ // hence $\mathbf{A}_{\pi(i)\pi(j)}\mathbf{s}_j = \tilde{a}_{ij}\,\mathbf{t}_i$
  end
end
return $\mathbf{A}, \mathbf{S}$

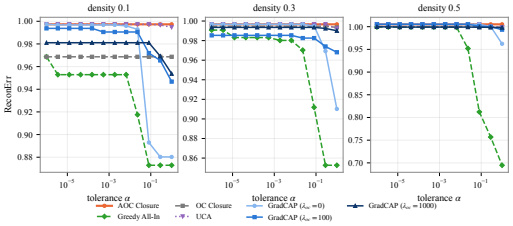


Figure 7: **Reconstruction Error against tolerated Relative Abstraction Error (RAE), on random DAGs.** Apart from UCA, all methods are partition-learning methods with $\mathbf{T}$ optimized on the partition they learn (UCA learns $\mathbf{T}$ directly). For each tolerance α fixed in advance, we accept the learned abstraction if $\mathrm{RAE} \leq \alpha$ *and* its ReconErr is strictly below AOC Closure's, otherwise we fall back to AOC Closure (Theorem 8). Median over 20 instances ($p = 100, k^* = k = 10$).

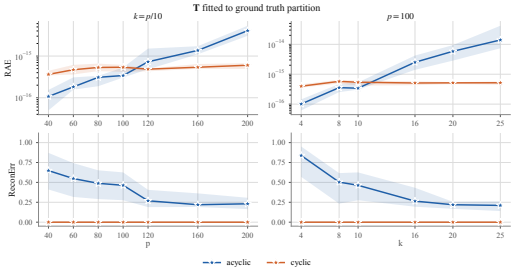


Figure 8: We apply Algorithm 4 to our optimization problem (Eq. 5) on ground truth partitions for which an exact abstraction is known to exist. We sample 20 cyclic and 20 acyclic partitions and report the median and interquartile ranges. We run this across a range of values of p and k with density fixed at 0.5 and blockiness at $\beta = 1$. RAE appears to be essentially zero in all settings (fairly stable for cyclic partitions but degrading slightly with k for acyclic partitions). ReconErr is typically driven by bad partitions, marginalization or $\mathbf{T}$-unfaithfulness. Here, Low ReconErr is indicative of $\mathbf{T}$-faithfulness.

REFERENCES

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 372–407. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/ahuja23a.html.

Tara V. Anand, Adele H. Ribeiro, Jin Tian, and Elias Bareinboim. Causal effect identification in cluster DAGs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 12172–12179, 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 33rd AAAI Conference on Artificial Intelligence*, pp. 2678–2685, 2019. doi: 10.1609/aaai.v33i01.33012678. URL https://aaai.org/papers/02678-abstracting-causal-models/.

Sander Beckers and Joseph Y. Halpern. Approximate causal abstraction. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence*, volume 115 of *Proceedings of Machine Learning Research*, pp. 606–615. PMLR, 2020. URL https://proceedings.mlr.press/v115/beckers20a.html.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):pp. 2885–2915, 2021. ISSN 00905364, 21688966. URL https://www.jstor.org/stable/27170673.

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

Gabriele D'Acunto, Fabio Massimo Zennaro, Yorgos Felekis, and Paolo Di Lorenzo. Causal abstraction learning based on the semantic embedding principle. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11776–11812. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/d-acunto25a.html.

Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007. ISSN 00318248, 1539767X. URL http://www.jstor.org/stable/10.1086/525638.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In Francesco Locatello and Vanessa Didelez (eds.), *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Arthur S. Goldberger. Structural equation methods in the social sciences. *Econometrica*, 40(6): 979–1001, 1972. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1913851.

Trygve Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1):1–12, 1943. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1905714.

Armin Kekić, Bernhard Schölkopf, and Michel Besserve. Targeted reduction of causal models. In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 1953–1980. PMLR, 2024. URL https://proceedings.mlr.press/v244/kekic24a.html.

Wassily Leontief. *Input-Output Economics*. Oxford University Press, New York, 2nd edition, 1986.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Riccardo Massidda, Atticus Geiger, Thomas Icard, and Davide Bacciu. Causal abstraction with soft interventions. In Mihaela van der Schaar, Cheng Zhang, and Dominik Janzing (eds.), *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pp. 68–87. PMLR, 11–14 Apr 2023. URL https://proceedings.mlr.press/v213/massidda23a.html.

Riccardo Massidda, Sara Magliacane, and Davide Bacciu. Learning causal abstractions of linear structural causal models. In Negar Kiyavash and Joris M. Mooij (eds.), *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 2486–2515. PMLR, 15–19 Jul 2024. URL https://proceedings.mlr.press/v244/massidda24a.html.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2nd edition, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

Eigel Fjeldgren Rischel. The category theory of causal models. Master's thesis, University of Copenhagen, Copenhagen, Denmark, 2020.

Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models, 2017. URL https://arxiv.org/abs/1707.00819.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Konstantin Stadler, Richard Wood, Tatyana Bulavskaya, Martin Bruckner, Stefan Giljum, Arjan de Koning, Jeroen Kuenen, José Acosta-Fernández, Arkaitz Usubiaga, Moana Simas, Olga Ivanova, Jan Weinzettel, Jannick H. Schmidt, Stefano Merciai, and Arnold Tukker. Exiobase 3: Developing a time series of detailed environmentally extended multi-regional input-output tables. *Journal of Industrial Ecology*, 22(3):502–515, 2018. doi: 10.1111/jiec.12715.

Konstantin Stadler, Richard Wood, Tatyana Bulavskaya, Martin Bruckner, Stefan Giljum, Arjan de Koning, Jeroen Kuenen, José Acosta-Fernández, Arkaitz Usubiaga, Moana Simas, Olga Ivanova, Jan Weinzettel, Jannick H. Schmidt, Stefano Merciai, and Arnold Tukker. Exiobase 3, 2021. URL https://doi.org/10.5281/zenodo.5589597.

Eli N. Weinstein and David M. Blei. Hierarchical causal models. *Journal of Machine Learning Research*, 27(37):1–73, 2026. URL http://jmlr.org/papers/v27/25-0899.html.

Fabio Massimo Zennaro, Máté Drávucz, Geanina Apachitei, W. Dhammika Widanage, and Theodoros Damoulas. Jointly learning consistent causal abstractions over multiple interventional distributions. In Mihaela van der Schaar, Cheng Zhang, and Dominik Janzing (eds.), *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pp. 88–121. PMLR, 11–14 Apr 2023. URL https://proceedings.mlr.press/v213/zennaro23a.html.

Yuchen Zhu, Sergio Hernan Garrido Mejia, Bernhard Schölkopf, and Michel Besserve. Unsupervised causal abstraction. In *NeurIPS 2024 Workshop on Causal Representation Learning*, 2024. URL https://openreview.net/forum?id=BP0e8RvFwd.