CHARACTERIZING CONSTRUCTIVE CAUSAL ABSTRACTION VIA INTERVENTION COMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal abstraction relates causal models across levels of description. Constructive abstractions are widely used and provide useful structural properties. Beckers and Halpern conjectured that strong τ-abstraction would imply constructivity under minor technical conditions. Existing positive results concern linear structural causal models under additional assumptions. We address the general case. First, a fixed edgeless low-level model admits strong abstractions with arbitrary high-level DAGs, even with analytic mechanisms, independent full-support noise, and a full-rank abstraction mapping. Second, for strong τ-abstractions, preservation of intervention composition on the full canonical domain is sufficient for constructivity. This condition is stronger than constructivity of the mapping: natural aggregations admit canonical interventions that violate composition preservation. Such interventions also cause failures in published abstraction results. We restrict the low-level domain while preserving every realizable high-level intervention. For a surjective mapping with full high-level intervention coverage, composition preservation on either restricted domain is then equivalent to constructivity. Finally, a controlled nonlinear learning experiment uses composition constraints to improve prediction over unregularized learning when input correlations change.

1 INTRODUCTION

Causal abstraction studies how high-level and low-level structural causal models (SCMs) should correspond under interventions. Rubenstein et al. (2017) introduced exact transformations between probabilistic causal models. Beckers & Halpern (2019) developed the notions of τ-abstraction, strong τ-abstraction, and constructive τ-abstraction, each adding requirements to the previous one. In particular, constructivity requires different high-level variables to depend on disjoint groups of low-level variables. This requirement is philosophically well motivated, since variables with overlapping ontological bases may fail to represent genuinely distinct causal factors.

It has also been frequently adopted in recent work. Constructive abstraction and its distributed extensions now provide a basis for neural explanations, intervention-based training, and analyses of neural representations (Geiger et al., 2021; 2022; 2025; Sutter et al., 2025). Beyond mechanistic interpretability, constructive mappings are used in causal inference across levels and in learning abstractions from data (Xia & Bareinboim, 2024; 2025; D'Acunto et al., 2025). They also help relate structural and potential-outcome models (Ibeling & Icard, 2023). Their model classes or alignments assume a partition, sometimes in transformed coordinates, while their intervention domains and consistency requirements differ. Their use of this structure makes it important to understand when constructivity follows from more basic intervention requirements.

Beckers & Halpern (2019) conjectured that strong abstraction would imply constructivity under minor technical conditions. For linear SCMs and a linear abstraction mapping, Massidda et al. (2024, Corollary 1) derive constructivity from strong abstraction under the assumptions of faithfulness, causal sufficiency, and no cancellation through the abstraction mapping. However, a general characterization remains open. We revisit the conjecture through three questions: *(i) Does strong τ-abstraction imply constructivity under natural regularity conditions? (ii) What property of the canonical intervention mapping forces constructivity? (iii) On which intervention domain is this property also necessary for constructivity?* We address these questions in a general, non-parametric setting.

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.

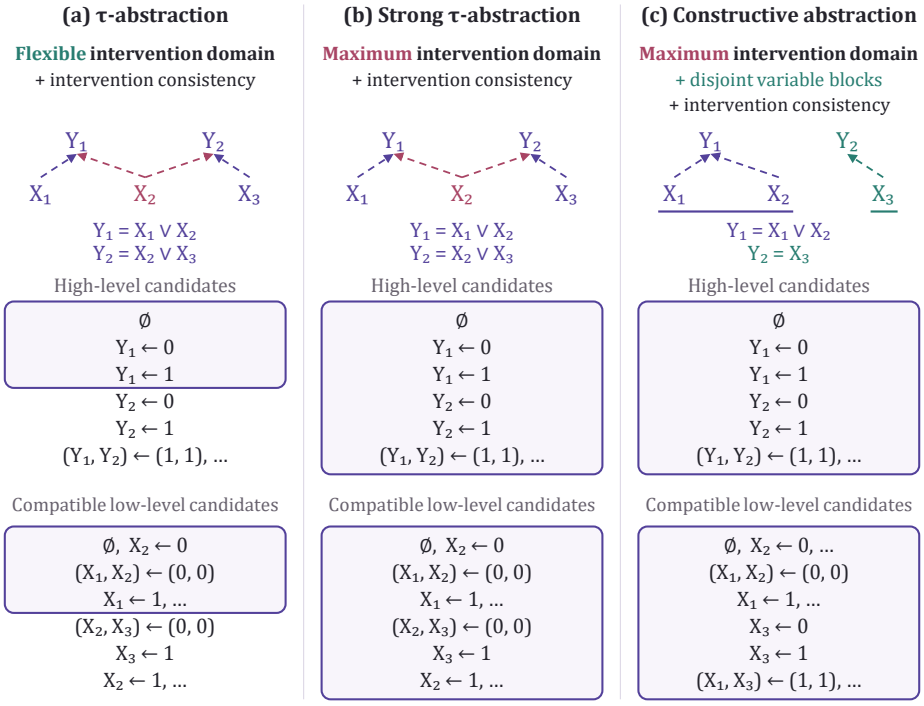


Figure 1: **Basic notions of causal abstraction.** Dashed arrows show dependence in τ, and shaded boxes mark the intervention domains. (a) τ-abstraction permits restricted domains. (b) Strong τ-abstraction uses the full canonical domain and covers all high-level interventions. (c) Constructive τ-abstraction additionally requires disjoint variable blocks. All three require intervention consistency.

Even with analytic mechanisms, independent full-support noise, and a full-rank abstraction mapping, a fixed edgeless low-level SCM admits a strong abstraction with every high-level DAG (Theorem 1). Thus these regularity conditions allow non-constructive strong abstractions. We therefore seek a stronger condition on the intervention mapping.

Our condition concerns intervention composition (Lorenz & Tull, 2026): mapping a composite low-level intervention must agree with composing its high-level images. A strong τ-abstraction whose canonical intervention mapping preserves composition is constructive (Theorem 2).

Constructivity of the abstraction mapping alone does not imply composition preservation on the full canonical intervention domain. Moreover, requiring model consistency on this full domain can exclude natural aggregations before composition is considered. For example, let $S = X_1 + X_2$ over $\mathbb{R}$. Intervening on X_1 alone leaves the range of S still equal to $\mathbb{R}$, so this low-level intervention corresponds to an empty high-level intervention. But in a simple SCM where X_1 and X_2 are both roots, changing X_1 changes the realized sum. The model therefore fails full-domain τ-abstraction consistency. Beckers & Halpern (2019, Example 3.15) identify a similar example.

We therefore consider two smaller domains, $\mathcal{J}_\tau$ and $\mathcal{J}_\tau^{\mathrm{rel}}$. On these domains, a relevant low-level variable may be fixed only when one of its high-level dependents is intervened on. (The difference between the two is that $\mathcal{J}_\tau$ but not $\mathcal{J}_\tau^{\mathrm{rel}}$ permits fixing variables ignored by τ.) Both domains preserve every high-level intervention realizable by the full domain (Equation 5). In this sense, the low-level interventions we exclude are redundant anyway.

These restrictions have three benefits. First, sums, averages, and other natural aggregations can satisfy interventional consistency on these domains. Second, the restrictions help repair literature results challenged by the excluded interventions.

Third, on either restricted domain, composition preservation is equivalent to constructivity if τ is surjective and all high-level interventions are covered (Theorem 3). Corollary 4 applies this result to

τ-abstractions whose consistency is checked on that domain, characterizing constructivity of their abstraction mappings.

We also use composition equalities as training constraints for a nonlinear abstraction mapping. A controlled experiment tests prediction after input correlations change and compares these constraints with ordinary regularization and constructive architectures.

2 PRELIMINARIES

We assume familiarity with the SCM framework. Let M_L and M_H be acyclic SCMs with finite sets of endogenous variables $V_L = \{X_1, \dots, X_n\}$ and $V_H = \{Y_1, \dots, Y_m\}$. Their state spaces are the full Cartesian products $\mathcal{X} = \prod_{r=1}^{n} \mathcal{X}_r$ and $\mathcal{Y} = \prod_{j=1}^{m} \mathcal{Y}_j$, with $|\mathcal{Y}_j| \geq 2$. Dependence is evaluated on these full state spaces, including unobserved states. We write $[d] = \{1, \dots, d\}$ and identify variables with their indices. Definitions 2.1–2.4 follow Beckers & Halpern (2019), with minor notation changes.

Definition 2.1 (Hard interventions)**.** A hard intervention $i = (S \leftarrow s)$, with support $\mathrm{supp}(i) = S$, replaces the equations for the variables in S by constants. We write explicit joint assignments as $(X_1, X_2) \leftarrow (x_1, x_2)$ and single-variable assignments as $X_r \leftarrow x_r$. Given a context e (an assignment to the exogenous or noise variables), we write $M_L^i(e)$ for the resulting solution in that context.

Definition 2.2 (Canonical intervention mapping)**.** For $i = (S \leftarrow s)$, let $\mathrm{Rst}(V_L, i) = \mathrm{Rst}(V_L, S \leftarrow s) := \{x \in \mathcal{X} : x|_S = s\}$; define $\mathrm{Rst}(V_H, j)$ likewise. Here the second Rst argument is an intervention. A surjective abstraction mapping $\tau : \mathcal{X} \to \mathcal{Y}$ defines the canonical intervention mapping by

$$\omega_\tau(i) = j \iff \tau(\mathrm{Rst}(V_L, i)) = \mathrm{Rst}(V_H, j), \qquad \mathcal{I}_L^\tau = \mathrm{dom}\, \omega_\tau. \tag{1}$$

This set equality fixes $\mathcal{I}_L^\tau$ before any intervention-consistency check.

We next define τ-abstraction, strong τ-abstraction, and constructivity of the abstraction mapping (Figure 1).

Definition 2.3 (τ-abstraction)**.** For allowed domains $D \subseteq \mathcal{I}_L^\tau$ and $\mathcal{I}_H$, $(M_H, \mathcal{I}_H)$ is a τ-abstraction of (M_L, D) if $\mathcal{I}_H = \omega_\tau(D)$ and one surjective context map $\tau_{\mathcal{U}}$ satisfies

$$\tau\big(M_L^i(e)\big) = M_H^{\omega_\tau(i)}\big(\tau_{\mathcal{U}}(e)\big) \quad \text{for every allowed } i \text{ and every } e. \tag{2}$$

Definition 2.4 (Strong τ-abstraction)**.** A strong τ-abstraction is a τ-abstraction with $D = \mathcal{I}_L^\tau$ and $\mathcal{I}_H = \mathcal{I}_H^*$, the set of all high-level hard interventions.

Definition 2.5 (Constructive abstraction mappings)**.** The relevant set $R_j = \mathrm{Rel}(\tau_j)$ contains r if changing only x_r can change $\tau_j(x)$. The abstraction mapping is *constructive* exactly when the R_j are pairwise disjoint. Equivalently, each τ_j depends only on a block B_j in a partition $B_0 \sqcup B_1 \sqcup \cdots \sqcup B_m = [n]$, where B_0 is ignored.

The constructive τ-abstraction of Beckers & Halpern (2019, Definition 3.17) also requires the original strong abstraction condition.

Definition 2.6 (Intervention composition)**.** The composition $i \circ k$ applies k first and i second. Assignments from k remain in place unless i assigns the same variable, in which case the value from i is used. For a domain $D \subseteq \mathcal{I}_L^\tau$, define

$$\mathrm{Comp}_D : \qquad i \circ k \in D, \qquad \omega_\tau(i \circ k) = \omega_\tau(i) \circ \omega_\tau(k) \quad (\forall i, k \in D). \tag{3}$$

Note that closure under composition is part of the requirement.

3 WHEN STRONG ABSTRACTION IMPLIES CONSTRUCTIVITY

Theorem 1 (One edgeless model, every DAG)**.** *For every $m \geq 1$, there is one edgeless SCM $M_{L,m}$ on $2m$ real variables such that every DAG G on m nodes is the graph of a strong τ_G-abstraction $M_{H,G}$ of $M_{L,m}$. Whenever G contains an edge, τ_G is non-constructive.*

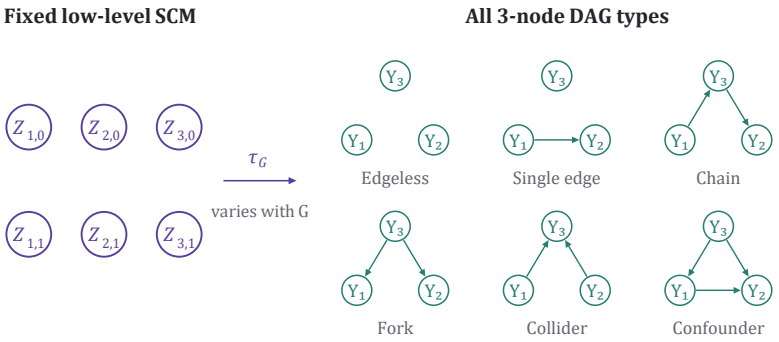


Figure 2: **One edgeless low-level model admits every high-level DAG.** The six independent low-level variables are fixed. Varying τ_G yields a strong abstraction for every three-node DAG, shown here up to relabeling (Theorem 1).

Construction. Figure 2 illustrates the case $m = 3$. Before choosing G, fix $M_{L,m}$ by $Z_{j,s} = E_{j,s}$ with independent standard Gaussian noise variables. For any DAG G on $[m]$, define $M_{H,G}$ in topological order: each Y_j is a function of its parents in G and the noise pair $(E_{j,0}, E_{j,1})$. These functions depend on every specified parent, while equal noise inputs (a, a) give $Y_j = a$ for all parent values. The abstraction mapping τ_G uses the same recursion with each noise pair replaced by the low-level state pair $(z_{j,0}, z_{j,1})$.

Thus assigning $(Z_{j,0}, Z_{j,1}) \leftarrow (a, a)$ implements $Y_j \leftarrow a$, and doing so for any set of nodes realizes every high-level intervention. Appendix A gives the explicit functions and proves consistency for every intervention in $\mathcal{I}_L^\tau$. For any edge $r \to j$, the low-level pair for r affects both $(\tau_G)_r$ and $(\tau_G)_j$, so τ_G is non-constructive.

Both models have real-analytic mechanisms and independent full-support noise blocks. The mapping $\tau_G : \mathbb{R}^{2m} \twoheadrightarrow \mathbb{R}^m$ is real analytic and has Jacobian rank m everywhere. These regularity conditions therefore do not make strong abstraction imply constructivity. What additional condition guarantees constructivity? One answer is given by the next result.

Theorem 2 (Composition forces constructivity)**.** *If M_H is a strong τ-abstraction of M_L and $\mathrm{Comp}_{\mathcal{I}_L^\tau}$ holds, then $R_j \cap R_k = \varnothing$ for $j \neq k$.*

Proof sketch. Suppose a low-level intervention assigns a variable relevant to τ_k, but its high-level image leaves Y_k unassigned. Choose two low-level states that differ only in that variable and have different τ_k values. Assigning all low-level variables to either state and then applying the intervention gives identical low-level assignments. The resulting low-level states have the same image under τ and hence determine the same high-level assignment. Yet the high-level compositions retain different assignments to Y_k. Composition preservation excludes this possibility. Thus, whenever $\omega_\tau(i)$ leaves Y_k unassigned, we must have $\mathrm{supp}(i) \cap R_k = \varnothing$. Full intervention coverage then forces disjoint relevant sets.

Example 4.1 gives a simpler, concrete illustration of why composition rules out overlapping relevant sets. The proof is in Appendix B.

4 Characterizing constructivity on restricted intervention domains

The previous result concerns the full canonical intervention domain, as required by strong τ-abstraction. In this section, we show that there are good reasons to restrict this domain and that, once it is restricted properly, the sufficient condition for constructivity becomes an exact characterization.

4.1 RESTRICTING THE INTERVENTION DOMAIN

Why is it reasonable to restrict the intervention domain?

In the definition of Beckers & Halpern (2019), a constructive τ-abstraction must also be a strong τ-abstraction. It therefore requires consistency on the full canonical intervention domain. This requirement can exclude natural aggregations before composition is considered. For example, Proposition 4.1 shows that the mapping $\tau(x_1, x_2) = x_1 + x_2$ on $\mathbb{R}^2$ cannot define a strong τ-abstraction between acyclic SCMs.

Beckers & Halpern (2019, Example 3.15) already note that strong abstraction can exclude useful models. In their energy example, a mass intervention maps to the empty high-level intervention but changes the represented energy. They obtain a τ-abstraction by restricting the allowed interventions.[1] This issue affects an entire class of aggregation mappings. For a surjective affine mapping between real Cartesian state spaces, strong τ-abstraction between acyclic SCMs forces each high-level variable to depend on exactly one low-level variable, with different high-level variables using different low-level variables.

Proposition 4.1 (Affine abstraction mappings)**.** *Let M_L, M_H be acyclic SCMs with state spaces $\mathbb{R}^n, \mathbb{R}^m$, and let $\tau(x) = Ax + b$ be surjective. If M_H is a strong τ-abstraction of M_L, then*

$$\tau_j(x) = a_j x_{r_j} + b_j, \qquad a_j \neq 0,$$

where $r_1, \ldots, r_m$ are distinct. (Proof: Appendix E.)

This domain requirement also affects published examples. Examples 1, 3, and 5 of Massidda et al. (2024) do not satisfy the full-domain consistency condition stated in their Section 2; Appendix F.3 gives the calculations.

Application-oriented work also uses smaller intervention domains. Geiger et al. (2025, Remark 32 and Definition 33) define constructive abstraction using interventions on complete cells of a given alignment, excluding its ignored cell. This differs from the full canonical domain of Beckers & Halpern (2019). We therefore recommend stating the intervention domain explicitly when using constructive abstraction.

The canonical intervention mapping is defined by images of state cylinders. For $\tau(x_1, x_2) = x_1 + x_2$ over $\mathbb{R}$, the intervention $X_1 \leftarrow a$ has the full high-level state space as its cylinder image because $a + \mathbb{R} = \mathbb{R}$. It therefore maps to the empty high-level intervention. The same holds for binary XOR because $a \oplus \{0, 1\} = \{0, 1\}$. These interventions can still change the represented outcome of a model. Appendix F uses these mappings to construct counterexamples to the sufficiency claim in Theorem 3 of Massidda et al. (2024) and the lifting claim in Theorem 5.4 of Beckers et al. (2020), respectively. Restricting the intervention domain recovers the lifting and block-sufficiency conclusions under the assumptions stated below.

Proper intervention domains for retaining all realizable high-level interventions. Write $R_K = \bigcup_{j \in K} R_j$ and $B_0 = [n] \setminus R_{[m]}$. Define

$$\boxed{\begin{aligned} \mathcal{J}_\tau &= \{i \in \mathcal{I}_L^\tau : \mathrm{supp}(i) \subseteq B_0 \cup R_{\mathrm{supp}(\omega_\tau(i))}\}, \\ \mathcal{J}_\tau^{\mathrm{rel}} &= \{i \in \mathcal{I}_L^\tau : \mathrm{supp}(i) \subseteq R_{\mathrm{supp}(\omega_\tau(i))}\}. \end{aligned}} \tag{4}$$

These domains are determined by τ, without a given partition or a change to ω_τ. We regard an assignment to a relevant variable as reasonable only if that variable is relevant to a targeted high-level variable. Both domains impose this rule; $\mathcal{J}_\tau^{\mathrm{rel}}$ also excludes interventions on B_0. This distinction matters when omitted variables affect retained ones.

If $\omega_\tau(i) = (K \leftarrow b)$, deleting assignments outside R_K gives $i_0 \in \mathcal{J}_\tau^{\mathrm{rel}}$ with the same image. The larger low-level cylinder still fixes the targeted values to b, and its image contains the original high-level cylinder. Hence

$$\omega_\tau(\mathcal{J}_\tau^{\mathrm{rel}}) = \omega_\tau(\mathcal{J}_\tau) = \omega_\tau(\mathcal{I}_L^\tau). \tag{5}$$

Every realizable high-level intervention is retained, while assignments outside the stated supports are removed. Neither domain is necessarily closed under composition. Appendix C.2 compares these domains with the complete-cell rule.

[1] The same example is numbered 3.16 in arXiv:1812.03789v4.

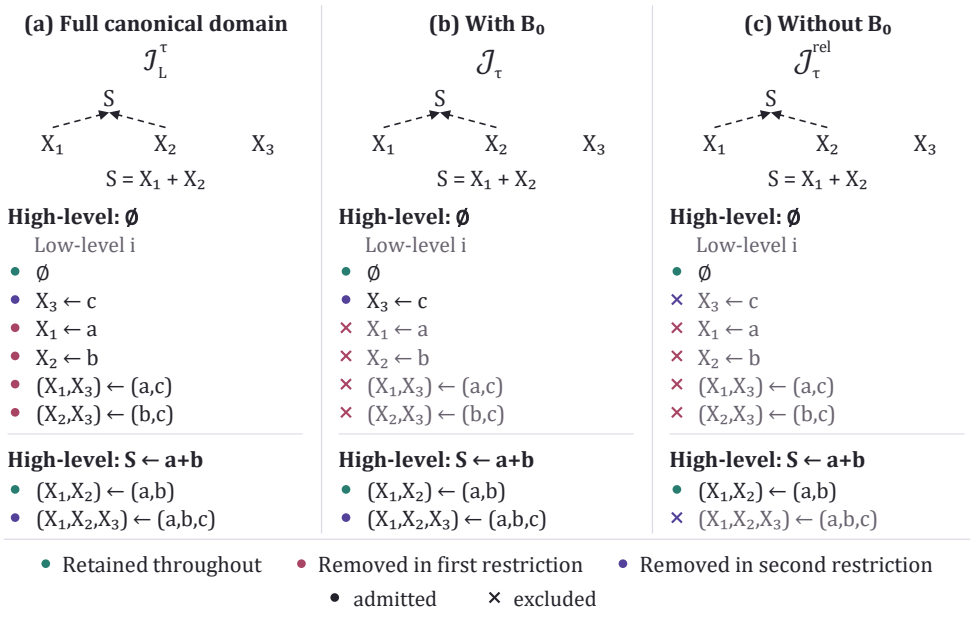


Figure 3: **Restricting low-level interventions while preserving high-level coverage.** For $\tau(x_1, x_2, x_3) = x_1 + x_2$ over $\mathbb{R}$, $R_S = \{1, 2\}$ and $B_0 = \{3\}$. The same eight intervention families are grouped by their image under ω_τ; all parameters are real. Dots mark admitted families and crosses mark excluded families. The domain $\mathcal{J}_\tau$ excludes assignments to X_1 or X_2 that map to the empty high-level intervention; $\mathcal{J}_\tau^{\mathrm{rel}}$ additionally excludes the remaining interventions assigning X_3. All three domains realize $\varnothing$ and every $S \leftarrow s$, $s \in \mathbb{R}$. Dashed arrows indicate dependence in the abstraction mapping.

Repairs to existing abstraction results. On $\mathcal{J}_\tau$, we recover the finite-state constructive lifting conclusion of Beckers et al. (2020, Theorem 5.4), under their deterministic-context convention. On $\mathcal{J}_\tau^{\mathrm{rel}}$, we recover the block-matrix sufficiency conclusion of Massidda et al. (2024, Theorem 3). This does not establish the unrestricted linear biconditional. Appendix F states the assumptions and limits, including an equivalence for ordered blocks and the specified context map.

4.2 A NECESSARY AND SUFFICIENT CONDITION FOR CONSTRUCTIVITY

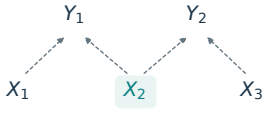
Surprisingly, on these restricted intervention domains, our composition condition becomes both necessary and sufficient for constructivity.

Theorem 3 (Constructivity on the target-relevant domain)**.** *On full Cartesian state spaces with finitely many coordinates and $|\mathcal{Y}_j| \geq 2$, let $\tau : \mathcal{X} \twoheadrightarrow \mathcal{Y}$ satisfy $\omega_\tau(\mathcal{J}_\tau) = \mathcal{I}_H^*$, where $\mathcal{I}_H^*$ contains all high-level hard interventions. Then*

$$\boxed{\tau \text{ is constructive} \iff \mathrm{Comp}_{\mathcal{J}_\tau}.} \tag{6}$$

The same statement holds with $\mathcal{J}_\tau^{\mathrm{rel}}$.

Example 4.1 (Overlapping sums)**.**



Let $\tau : \mathbb{R}^3 \to \mathbb{R}^2$ be given by

$$\tau_1(x) = x_1 + x_2, \qquad \tau_2(x) = x_2 + x_3.$$

The dashed arrows indicate dependence in τ, not causal edges. Both outputs depend on X_2.

Consider low-level and high-level models with real-valued endogenous variables and structural equations

$$X_2 = E_2, \quad X_1 = E_1 - X_2, \quad X_3 = E_3 - X_2; \qquad Y_1 = U_1, \quad Y_2 = U_2,$$

with context spaces $\mathbb{R}^3, \mathbb{R}^2$ and $\tau_{\mathcal{U}}(e) = (e_1, e_3)$. First consider $h = (X_1 \leftarrow 2)$. Its cylinder image is $\mathbb{R}^2$, so $\omega_\tau(h) = \varnothing$, yet its decoded outcome is $(2 + e_2, e_3)$ rather than (e_1, e_3). This violates full-domain consistency. The restricted domain excludes h because it assigns a relevant variable without targeting either high-level variable. On $\mathcal{J}_\tau = \mathcal{J}_\tau^{\mathrm{rel}}$, consistency holds: fixing either sum's input pair leaves the other sum equal to its noise, and assigning all three variables fixes both sums. Every high-level intervention remains covered, so these models form a τ-abstraction on this domain.

Now take the admitted interventions

$$i = (X_1, X_2) \leftarrow (2, 3), \qquad \omega_\tau(i) = (Y_1 \leftarrow 5),$$
$$k = (X_2, X_3) \leftarrow (4, 5), \qquad \omega_\tau(k) = (Y_2 \leftarrow 9).$$

As Figure 4 shows, applying k and then i overwrites the shared X_2 from 4 to 3. Consequently,

$$\omega_\tau(i \circ k) = \omega_\tau\big((X_1, X_2, X_3) \leftarrow (2, 3, 5)\big) = (Y_1, Y_2) \leftarrow (5, 8),$$
$$\omega_\tau(i) \circ \omega_\tau(k) = (Y_1, Y_2) \leftarrow (5, 9).$$

At the high level, the later assignment to Y_1 preserves $Y_2 = 9$; its low-level implementation changes Y_2 to 8 through X_2. Composition preservation forbids this interference. Theorem 3 shows that, with full high-level intervention coverage, it forces the relevant sets to be disjoint.

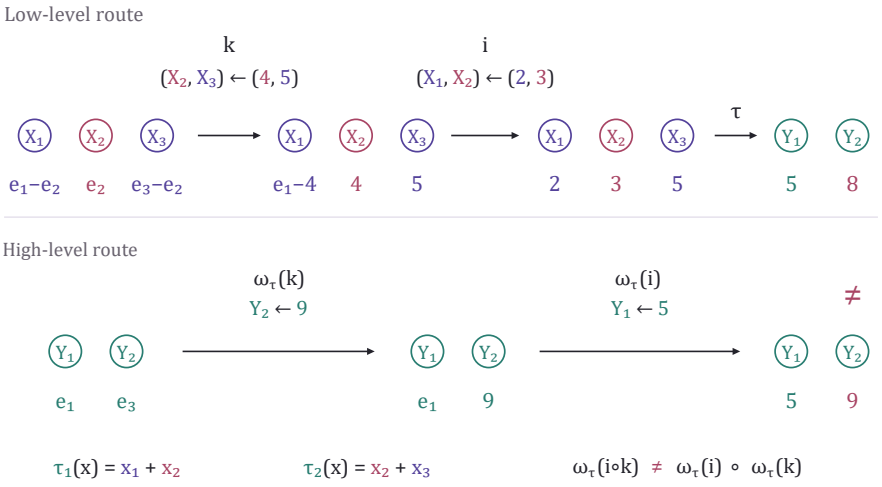


Figure 4: **Composition fails for overlapping sums.** In Example 4.1, applying k and then i gives $(Y_1, Y_2) = (5, 8)$ under τ, whereas composing their images under ω_τ gives $(5, 9)$.

Corollary 4 (Application to causal models)**.** *Suppose $(M_H, \mathcal{I}_H^*)$ is a τ-abstraction of $(M_L, \mathcal{J}_\tau)$. Then its abstraction mapping is constructive if and only if $\mathrm{Comp}_{\mathcal{J}_\tau}$ holds. The same statement holds with $\mathcal{J}_\tau^{\mathrm{rel}}$.*

Proof. The abstraction assumes consistency, a surjective τ, and $\omega_\tau(\mathcal{J}_\tau) = \mathcal{I}_H^*$. Apply Theorem 3; the argument for $\mathcal{J}_\tau^{\mathrm{rel}}$ is the same. □

Appendix D proves both domain versions. The corollary requires model consistency on the specified domain; composition characterizes the mapping's constructivity under this assumption and full high-level intervention coverage.

5 A CONTROLLED LEARNING EXAMPLE

We illustrate how the composition condition gives a training constraint for a nonlinear abstraction mapping. Let $x = ((a_j, b_j))_{j=1}^4 \in \mathbb{R}^8$ and learn

$$\tau_{\theta,j}(x) = \log\big((1 - w_{\theta,j}(x))e^{a_j} + w_{\theta,j}(x)e^{b_j}\big), \qquad w_{\theta,j}(x) = \sigma(g_{\theta,j}(x)).$$

Here $\sigma(t) = (1 + e^{-t})^{-1}$. Each gate network $g_{\theta,j}$ can read all eight inputs. The target mapping has the same form, but each gate depends only on its own pair, so it is constructive. All methods receive the pair assignment.

From composition to a loss. For every parameter value, $i = (a_k, b_k) \leftarrow (c, c)$ has canonical image $\omega_{\tau_\theta}(i) = Y_k \leftarrow c$ and belongs to both restricted domains. Assigning equal entries to several pairs realizes every high-level hard intervention. By Lemma C.3, composition with an intervention on all low-level variables requires $\tau_{\theta,j}(x') = \tau_{\theta,j}(x)$ for $j \neq k$, where $(V_L \leftarrow x') = i \circ (V_L \leftarrow x)$. The targeted output already equals c, so the loss checks only the other outputs.

At each step, we draw $B = 128$ states $\widetilde{x}^{(r)}$ by independently sampling each pair, with replacement, from the observed training states. Both entries of a pair come from the same state. For each r, draw k_r uniformly from $\{1, \ldots, 4\}$ and $c_r \sim N(0, 1)$. Let $i_r = (a_{k_r}, b_{k_r}) \leftarrow (c_r, c_r)$ and $(V_L \leftarrow x'^{(r)}) = i_r \circ (V_L \leftarrow \widetilde{x}^{(r)})$. The three penalties are

$$\begin{aligned}
\mathcal{L}^{\text{out}}_{\text{Comp}}(\theta) &= \frac{1}{3B} \sum_{r=1}^{B} \sum_{j \neq k_r} \left[\tau_{\theta,j}(x'^{(r)}) - \tau_{\theta,j}(\widetilde{x}^{(r)})\right]^2, \\
\mathcal{L}^{\text{gate}}_{\text{Comp}}(\theta) &= \frac{1}{3B} \sum_{r=1}^{B} \sum_{j \neq k_r} \left[w_{\theta,j}(x'^{(r)}) - w_{\theta,j}(\widetilde{x}^{(r)})\right]^2, \\
\mathcal{L}^{\text{logit}}_{\text{Comp}}(\theta) &= \frac{1}{3B} \sum_{r=1}^{B} \sum_{j \neq k_r} \left[g_{\theta,j}(x'^{(r)}) - g_{\theta,j}(\widetilde{x}^{(r)})\right]^2.
\end{aligned} \tag{7}$$

The factor $3B$ averages over the three unchanged outputs in each sample. Away from $a_j = b_j$, equality of outputs, gates, and logits is equivalent for these checks, although the squared losses differ. For N labeled states $(x^{(n)}, y^{(n)})$, each variant minimizes

$$\mathcal{L}^{q}(\theta) = \frac{1}{4N} \sum_{n=1}^{N} \left\| \tau_\theta(x^{(n)}) - y^{(n)} \right\|_2^2 + \lambda \mathcal{L}^{q}_{\text{Comp}}(\theta), \qquad q \in \{\text{out}, \text{gate}, \text{logit}\}. \tag{8}$$

Both evaluations in each penalty are differentiated. The regularizer uses input states and intervention values without new target labels. Validation selects $\lambda = 1, 10, 0.1$ for the output, gate, and logit variants, respectively; the selected Comp method is Gate Comp with $\lambda = 10$.

Setting and results. Training uses 64 labeled states with correlation 0.95 between input pairs. Testing removes this correlation while keeping each pair's distribution fixed. Selection uses 1,024 labeled validation states per development seed, and final evaluation uses eight fresh training seeds. The methods in Figure 5 each train for 3,000 steps. Appendix G gives the data, selection rules, budgets, and complete results.

Gate Comp reduces mean prediction MSE from 5.52×10^{-3} to 2.64×10^{-4}, compared with the same network without regularization. Ordinary pair changes and a Jacobian penalty give similar errors, 2.62×10^{-4} and 2.41×10^{-4}; the constructive width-24 model gives the lowest error, 1.55×10^{-4}. Thus the characterization supplies a useful training constraint, while the known partition also supports stronger architectural restrictions. These finite tests illustrate the condition; they do not establish exact constructivity of the trained networks.

6 Related work

Constructivity and graphical abstraction. Beckers & Halpern (2019) conjectured that strong abstraction implies constructivity under minor technical conditions. Massidda et al. (2024) gave linear results under faithfulness and no-cancellation assumptions; Appendix F states the scope of our repair to their Theorem 3. Cluster DAGs support causal identification (Anand et al., 2023). Schooltink & Zennaro (2024) relate graphical and functional abstractions under additional conditions, including bijective range maps; our criterion does not remove those conditions.

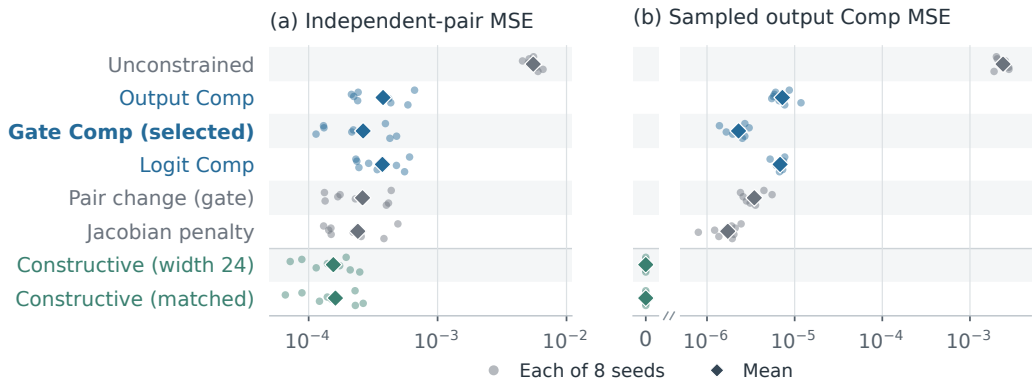


Figure 5: **Learning with composition constraints.** Left: prediction MSE after input pairs become independent. Right: output composition MSE on sampled legal interventions. Dots show eight training seeds; diamonds show means. All methods know the input pairs. Gate Comp was selected on validation data; the two constructive controls restrict each gate to its own pair. Their composition errors are exactly zero.

Composition and non-aligned interventions. Lorenz & Tull (2026, Appendix B and Proposition 70) study intervention monoids and homomorphisms and establish composition preservation for constructive abstractions on induced domains. We instead characterize constructivity using the canonical intervention mapping on restricted domains that preserve high-level coverage. Englberger & Dhami (2025) discuss non-constructive abstractions with non-aligned interventions; our universality result keeps the same low-level coordinate interventions.

Expressive representations. Sutter et al. (2025) prove universality for input-restricted distributed abstractions under assumptions including input injectivity and strict output surjectivity. Their layer-wise alignment bijections change intervention coordinates. Our theorem uses one fixed edgeless SCM and the full canonical domain, with analytic surjective mappings of full row rank. Factored-space representations also connect independent factors to graphical models (Garrabrant et al., 2024).

Interventions in learned representations. CPMs use counterfactual concept supervision (Wu et al., 2023); RAVEL separates target changes from preserving other attributes (Huang et al., 2024); DAS learns distributed alignments with multi-source training (Geiger et al., 2024). Our experiment uses canonical composition equalities as training constraints. Projected abstractions support causal inference under lossy representations (Xia & Bareinboim, 2025), while mechanistic independence gives additional identifiability conditions (Matthes et al., 2026).

Limitations. Complete composition checks can be costly. The learning example uses a synthetic mapping family and known input pairs. Its prediction gains do not establish an advantage over ordinary regularization or explicit constructivity. Finite tests on a continuous domain can find failures but cannot establish constructivity.

7 Conclusion

One fixed edgeless model admits strong abstractions with arbitrary high-level graphs, despite strong regularity conditions. On either restricted domain, composition characterizes constructivity for a surjective abstraction mapping with full high-level intervention coverage. A controlled learning experiment uses these composition constraints to improve prediction over unregularized learning when input correlations change.

REPRODUCIBILITY STATEMENT

The appendices give complete proofs, generated-data definitions, learning objectives, selection rules, and training budgets. Saved configurations, source and data hashes, and per-seed results accompany the experiment. The figure and result table can be rebuilt from these records without model training.

AI USE STATEMENT

Generative AI assisted with literature search, mathematical derivations, writing proofs, designing and giving feedback on methodology and experiments, generating synthetic data, and implementing and executing code. It also provided critical ingredients for proving mathematical claims. The authors take full responsibility for the correctness and final content of this work, including any text, claims, derivations, proofs, or code produced with the assistance of generative AI.

REFERENCES

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.

Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pp. 2678–2685, 2019.

Sander Beckers, Frederick Eberhardt, and Joseph Y Halpern. Approximate causal abstractions. In *Uncertainty in artificial intelligence*, pp. 606–615. PMLR, 2020.

Gabriele D'Acunto, Fabio Massimo Zennaro, Yorgos Felekis, and Paolo Di Lorenzo. Causal abstraction learning based on the semantic embedding principle. *arXiv preprint arXiv:2502.00407*, 2025.

Markus Englberger and Devendra Singh Dhami. Causal abstractions, categorically unified. *arXiv preprint arXiv:2510.05033*, 2025.

Scott Garrabrant, Matthias Georg Mayer, Magdalena Wache, Leon Lang, Sam Eisenstat, and Holger Dell. Factored space models: Towards causality between levels of abstraction. *arXiv preprint arXiv:2412.02579*, 2024.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in neural information processing systems*, 34:9574–9586, 2021.

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 *International Conference on Machine Learning*, pp. 7324–7338. PMLR, 2022.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pp. 160–187. PMLR, 2024.

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.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. Ravel: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687, 2024.

Duligur Ibeling and Thomas Icard. Comparing causal frameworks: Potential outcomes, structural models, graphs, and abstractions. *Advances in Neural Information Processing Systems*, 36:80130–80141, 2023.

Robin Lorenz and Sean Tull. Causal and compositional abstraction. *arXiv preprint arXiv:2602.16612*, 2026.

Riccardo Massidda, Sara Magliacane, and Davide Bacciu. Learning causal abstractions of linear structural causal models. *arXiv preprint arXiv:2406.00394*, 2024.

Stefan Matthes, Zhiwei Han, and Hao Shen. Mechanistic independence: A principle for identifiable disentangled representations. In *International Conference on Learning Representations*, volume 2026, pp. 109930–109969, 2026.

Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. *arXiv preprint arXiv:1707.00819*, 2017.

Willem Schooltink and Fabio Massimo Zennaro. Aligning graphical and functional causal abstractions. *arXiv preprint arXiv:2412.17080*, 2024.

Denis Sutter, Julian Minder, Thomas Hofmann, and Tiago Pimentel. The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability? In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 149455–149503. Curran Associates, Inc., 2025. doi: 10.52202/085713-5000. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/dbb98528c9870377f3f0d133aae6050b-Paper-Conference.pdf.

Zhengxuan Wu, Karel D'Oosterlinck, Atticus Geiger, Amir Zur, and Christopher Potts. Causal proxy models for concept-based model explanations. In *International conference on machine learning*, pp. 37313–37334. PMLR, 2023.

Kevin Xia and Elias Bareinboim. Neural causal abstractions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 20585–20595, 2024.

Kevin Xia and Elias Bareinboim. Causal abstraction inference under lossy representations. *arXiv preprint arXiv:2509.21607*, 2025.

A Proof of Theorem 1

Proof. Fix $m \geq 1$. Use the following low-level SCM for every DAG on $[m]$:

$$M_{L,m}: \qquad Z_{j,s} = E_{j,s}, \qquad E_{j,s} \overset{\text{iid}}{\sim} N(0,1), \qquad j \in [m],\ s \in \{0,1\}. \tag{9}$$

It has $2m$ real-valued endogenous variables and no edges.

Let G be any DAG on $[m]$. For $p = (p_r)_{r \in \mathrm{pa}_G(j)}$ and $u = (u_0, u_1)$, define

$$w_j(p) = \frac{1}{1 + \exp(-\sum_{r \in \mathrm{pa}_G(j)} p_r)}, \tag{10}$$

$$F_j(p, u) = \log\big((1 - w_j(p))e^{u_0} + w_j(p)e^{u_1}\big). \tag{11}$$

An empty parent sum is zero. For all arguments,

$$0 < w_j(p) < 1, \qquad F_j(p, (a, a)) = a, \qquad \frac{\partial F_j}{\partial u_0} > 0, \quad \frac{\partial F_j}{\partial u_1} > 0. \tag{12}$$

Put $U_j = (E_{j,0}, E_{j,1})$ and let $\tau_{\mathcal{U}}$ group the $2m$ scalar contexts into these m pairs. This map is a bijection, and it sends the context law to the product of m bivariate standard normal laws.

In any topological ordering of G, define the high-level SCM and the abstraction mapping recursively by

$$M_{H,G}:\ Y_j = F_j(Y_{\mathrm{pa}_G(j)}, U_j), \qquad (\tau_G)_j(z) = F_j((\tau_G)_{\mathrm{pa}_G(j)}(z), z_j), \tag{13}$$

where $z_j = (z_{j,0}, z_{j,1})$. In this proof, write $M_{L,m}$, $M_{H,G}$, and τ_G as M_L, M_H, and τ.

First, τ is surjective. Given $y \in \mathbb{R}^m$, choose $z_{j,0} = z_{j,1} = y_j$ for every j. Induction along the topological ordering and (12) give $\tau(z) = y$. For a high-level intervention $k = (K \leftarrow a)$, use

$$i = (Z_{j,0}, Z_{j,1})_{j \in K} \leftarrow (a_j, a_j)_{j \in K}. \tag{14}$$

Every state in $\mathrm{Rst}(V_L, i)$ maps into $\mathrm{Rst}(V_H, k)$. Conversely, for every $y \in \mathrm{Rst}(V_H, k)$ the diagonal preimage just constructed belongs to $\mathrm{Rst}(V_L, i)$. Therefore $\tau(\mathrm{Rst}(V_L, i)) = \mathrm{Rst}(V_H, k)$, so every high-level intervention is in the image of ω_τ.

We next prove consistency for every canonical intervention. Fix $i \in \mathcal{I}_L^\tau$ and write $\omega_\tau(i) = (K \leftarrow a)$. If $j \in K$, then i must set both $Z_{j,0}$ and $Z_{j,1}$. Otherwise, varying an unset member of this pair while holding all other coordinates fixed would preserve all parent outputs and change τ_j strictly by (12). This contradicts $\tau_j = a_j$ on $\mathrm{Rst}(V_L, i)$.

If $j \notin K$, then i must set neither member of the pair. Indeed, suppose that it sets one member to c. Fix parent values p that are compatible with $y_K = a$. The contribution of the fixed member to the positive sum in (11) gives the finite lower bound

$$y_j \geq c + \log \alpha, \qquad \alpha \in \{1 - w_j(p), w_j(p)\} > 0.$$

But the cylinder $\mathrm{Rst}(V_H, K \leftarrow a)$ permits arbitrarily negative y_j with these same parent values. This is a contradiction.

For a context e, let $y = \tau(M_L^i(e))$. Since $M_L^i(e) \in \mathrm{Rst}(V_L, i)$, we have $y_j = a_j$ for $j \in K$. For $j \notin K$, the intervention leaves both low-level equations in the pair unchanged. Hence

$$y_j = a_j \quad (j \in K), \qquad y_j = F_j(y_{\mathrm{pa}_G(j)}, U_j(e)) \quad (j \notin K).$$

These are exactly the intervened high-level equations. Their unique solution proves (2) for all $i \in \mathcal{I}_L^\tau$ and all e. Together with surjectivity of τ, ω_τ, and the context map, this proves strong abstraction on the full canonical domain.

Finally, w_j is strictly increasing in each parent argument, and F_j varies strictly with w_j whenever $u_0 \neq u_1$. Thus every edge of G occurs in the dependence graph of the high-level structural functions, and there are no other edges: $G_{M_{H,G}} = G$. All structural functions and τ are real analytic. In topological order, the Jacobian submatrix consisting of the columns for $Z_{j,0}$, $j \in [m]$, is triangular with strictly positive diagonal. Thus $\mathrm{rank}\, J_\tau(z) = m$ for every $z \in \mathbb{R}^{2m}$. The low-level scalar noises and the high-level noise blocks are independent and have full support. The high-level observational law also has full support: the preimage under the continuous surjection τ of every nonempty open set is nonempty and open, and therefore has positive Gaussian probability.

For an edge $r \to j$, set every pair other than those for r and j to diagonal values, and choose $z_{j,0} \neq z_{j,1}$. Varying $z_{r,0}$ changes both τ_r and τ_j, so their relevant sets overlap. Thus τ is non-constructive whenever G contains an edge. □

B Composition preservation and full-domain rigidity

We use the setting and definitions of Section 2. On a Cartesian product with finitely many coordinates, each τ_j depends only on R_j: two states that agree on R_j can be joined by changing irrelevant coordinates one at a time, without changing τ_j. Since $|\mathcal{Y}_j| \geq 2$, distinct high-level hard interventions have distinct cylinders, so the canonical image is unique. In particular,

$$\omega_\tau(V_L \leftarrow x) = (V_H \leftarrow \tau(x)).$$

The full-domain composition condition is

$$i \circ k \in \mathcal{I}_L^\tau, \qquad \omega_\tau(i \circ k) = \omega_\tau(i) \circ \omega_\tau(k) \qquad (i, k \in \mathcal{I}_L^\tau). \tag{15}$$

Both rigidity proofs use the following lemma.

Lemma B.1 (Relevant coordinates of untargeted outputs)**.** *Let $\mathcal{D} \subseteq \mathcal{I}_L^\tau$ contain $(V_L \leftarrow x)$ for every $x \in \mathcal{X}$. Suppose composition is preserved for each pair $(i, (V_L \leftarrow x))$ with $i \in \mathcal{D}$. If $\omega_\tau(i)$ does not set Y_j, then*

$$\mathrm{supp}(i) \cap R_j = \varnothing. \tag{16}$$

Proof. If i sets X_r and x, x' differ only in coordinate r, then $i \circ (V_L \leftarrow x) = i \circ (V_L \leftarrow x')$. Applying ω_τ gives

$$\omega_\tau(i) \circ (V_H \leftarrow \tau(x)) = \omega_\tau(i) \circ (V_H \leftarrow \tau(x')).$$

Since $\omega_\tau(i)$ does not set Y_j, this equality implies $\tau_j(x) = \tau_j(x')$. Thus $r \notin R_j$. □

Proof of Theorem 2. Apply Lemma B.1 with $\mathcal{D} = \mathcal{I}_L^\tau$. Fix $j \in [m]$, $a \in \mathcal{X}$, and $h = (R_j \leftarrow a_{R_j})$. Dependence only on R_j gives

$$\tau(\mathrm{Rst}(V_L, h)) \subseteq \{y \in \mathcal{Y} : y_j = \tau_j(a)\}. \tag{17}$$

For the reverse inclusion, take y in the set on the right. By surjectivity of ω_τ, there is an $i \in \mathcal{I}_L^\tau$ with

$$\omega_\tau(i) = ([m] \setminus \{j\} \leftarrow y_{-j}).$$

Let a' be the state assigned by $i \circ (V_L \leftarrow a)$. By Lemma B.1, i sets no coordinate in R_j. Thus $a' \in \mathrm{Rst}(V_L, h)$. Composition preservation gives

$$\begin{aligned}(V_H \leftarrow \tau(a')) &= \omega_\tau(i \circ (V_L \leftarrow a)) \\ &= \omega_\tau(i) \circ (V_H \leftarrow \tau(a)) = (V_H \leftarrow y).\end{aligned}$$

Hence $\tau(a') = y$, which proves equality in (17). Thus h is canonical and $\omega_\tau(h) = (j \leftarrow \tau_j(a))$. For $k \neq j$, applying Lemma B.1 to h gives $R_j \cap R_k = \varnothing$. This proves constructivity. □

C Properties of the target-relevant domains

Lemma C.1 (Coverage and interventions on all variables)**.** *For every surjective abstraction mapping τ, both $\mathcal{J}_\tau$ and $\mathcal{J}_\tau^{\mathrm{rel}}$ contain $\varnothing$, and*

$$\omega_\tau(\mathcal{J}_\tau^{\mathrm{rel}}) = \omega_\tau(\mathcal{J}_\tau) = \omega_\tau(\mathcal{I}_L^\tau). \tag{18}$$

Moreover, for every $x \in \mathcal{X}$,

$$(V_L \leftarrow x) \in \mathcal{J}_\tau, \qquad q_x = (R_{[m]} \leftarrow x_{R_{[m]}}) \in \mathcal{J}_\tau^{\mathrm{rel}},$$

and

$$\omega_\tau(V_L \leftarrow x) = \omega_\tau(q_x) = (V_H \leftarrow \tau(x)).$$

If $B_0 \neq \varnothing$, then $(V_L \leftarrow x) \notin \mathcal{J}_\tau^{\mathrm{rel}}$.

Proof. The empty intervention satisfies both filters. Both $(V_L \leftarrow x)$ and q_x have cylinder image $\{\tau(x)\}$ because τ does not depend on B_0. Their images under ω_τ have support $[m]$. Their low-level supports are $[n]$ and $R_{[m]}$, respectively. These supports satisfy the claimed domain conditions.

For coverage, fix $i = (S \leftarrow s) \in \mathcal{I}_L^\tau$ with $\omega_\tau(i) = (K \leftarrow b)$. Keep only the assignments to coordinates relevant to the outputs in K:

$$i_K = (S \cap R_K \leftarrow s_{S \cap R_K}).$$

The inclusion $\mathrm{Rst}(V_L, i) \subseteq \mathrm{Rst}(V_L, i_K)$ gives $\mathrm{Rst}(V_H, K \leftarrow b) \subseteq \tau(\mathrm{Rst}(V_L, i_K))$. Conversely, given $x \in \mathrm{Rst}(V_L, i_K)$, make the assignments in $S \setminus R_K$ to obtain $x' \in \mathrm{Rst}(V_L, i)$. These changes affect none of the coordinates τ_j for $j \in K$, so $\tau_K(x) = b$. Thus $\tau(\mathrm{Rst}(V_L, i_K)) = \mathrm{Rst}(V_H, K \leftarrow b)$ and $i_K \in \mathcal{J}_\tau^{\mathrm{rel}}$. The inclusions $\mathcal{J}_\tau^{\mathrm{rel}} \subseteq \mathcal{J}_\tau \subseteq \mathcal{I}_L^\tau$ prove the claim. □

Lemma C.2 (Removing assignments to B_0)**.** *Let $p(i)$ delete all assignments of i to B_0. Then*

$$\tau(\mathrm{Rst}(V_L, i)) = \tau(\mathrm{Rst}(V_L, p(i))), \qquad p(i \circ k) = p(i) \circ p(k).$$

Consequently $i \in \mathcal{J}_\tau$ if and only if $p(i) \in \mathcal{J}_\tau^{\mathrm{rel}}$, and $\omega_\tau(i) = \omega_\tau(p(i))$ whenever i is canonical. Closure under composition on one of the two domains is equivalent to closure on the other, and composition preservation on all pairs in one domain is equivalent to that on all pairs in the other.

Proof. Assignments to B_0 do not affect the image under τ, and removing these assignments commutes with intervention composition. These facts prove both displayed identities and the membership equivalence. Every intervention in $\mathcal{J}_\tau^{\mathrm{rel}}$ is also an intervention in $\mathcal{J}_\tau$ and may be extended by arbitrary assignments to B_0. Applying the same identities to each composite proves the final claims. □

This lemma concerns the abstraction mapping and its canonical intervention domain. The SCM consistency requirements on the two domains can still differ: a coordinate outside $R_{[m]}$ may affect coordinates in $R_{[m]}$ through the low-level equations. Theorem 3 applies to both domains.

C.1 The support filter alone does not ensure closure

For binary coordinates, let $\tau(a, z_0, z_1) = (a, z_a)$. Then $R_X = \{A\}$, $R_Y = \{A, Z_0, Z_1\}$, and $B_0 = \varnothing$. The interventions

$$i = (A \leftarrow 1), \qquad k = (A, Z_0) \leftarrow (0, 0)$$

have canonical images $X \leftarrow 1$ and $(X, Y) \leftarrow (0, 0)$, so both belong to $\mathcal{J}_\tau = \mathcal{J}_\tau^{\mathrm{rel}}$. However,

$$i \circ k = (A, Z_0) \leftarrow (1, 0), \qquad \omega_\tau(i \circ k) = (X \leftarrow 1).$$

Its support $\{A, Z_0\}$ is not contained in $R_X = \{A\}$, so the composite belongs to neither restricted domain.

Lemma C.3 (Composition preservation forces closure)**.** *For a surjective abstraction mapping τ and $\mathcal{D} \in \{\mathcal{J}_\tau, \mathcal{J}_\tau^{\mathrm{rel}}\}$, the following statements are equivalent:*

1. *$\omega_\tau(i \circ k) = \omega_\tau(i) \circ \omega_\tau(k)$ whenever $i, k, i \circ k \in \mathcal{D}$.*
2. *For every $i \in \mathcal{D}$ and $x \in \mathcal{X}$,*
$$\omega_\tau(i \circ (V_L \leftarrow x)) = \omega_\tau(i) \circ (V_H \leftarrow \tau(x)).$$
3. *$\mathcal{D}$ is closed under composition and $\omega_\tau(i \circ k) = \omega_\tau(i) \circ \omega_\tau(k)$ for all $i, k \in \mathcal{D}$.*

If τ is constructive, all three statements hold on both domains.

Proof. Statement 3 implies statement 1. To prove that 1 implies 2, fix $i \in \mathcal{D}$ and $x \in \mathcal{X}$. Use $k = (V_L \leftarrow x)$ when $\mathcal{D} = \mathcal{J}_\tau$, and $k = q_x$ from Lemma C.1 when $\mathcal{D} = \mathcal{J}_\tau^{\mathrm{rel}}$. Let x' be the state assigned by $i \circ (V_L \leftarrow x)$. If $\mathcal{D} = \mathcal{J}_\tau$, both k and $i \circ k$ assign all low-level endogenous variables. If $\mathcal{D} = \mathcal{J}_\tau^{\mathrm{rel}}$, both assign exactly the coordinates in $R_{[m]}$. Their cylinder images are the singletons $\{\tau(x)\}$ and $\{\tau(x')\}$, since τ does not depend on B_0. Their images under ω_τ set all high-level variables, so both interventions belong to $\mathcal{D}$. Statement 1 gives

$$\omega_\tau(i \circ (V_L \leftarrow x)) = \omega_\tau(i \circ k) = \omega_\tau(i) \circ (V_H \leftarrow \tau(x)),$$

which is statement 2.

Assume statement 2 and take $i, k \in \mathcal{D}$. Intervention composition gives

$$\mathrm{Rst}(V_L, i \circ k) = \{x' \in \mathcal{X} : \exists x \in \mathrm{Rst}(V_L, k), \quad i \circ (V_L \leftarrow x) = (V_L \leftarrow x')\}.$$

This identity, statement 2, and canonicality of k give

$$\begin{aligned}\tau(\mathrm{Rst}(V_L, i \circ k)) &= \{y' \in \mathcal{Y} : \exists x \in \mathrm{Rst}(V_L, k), \quad \omega_\tau(i) \circ (V_H \leftarrow \tau(x)) = (V_H \leftarrow y')\} \\ &= \{y' \in \mathcal{Y} : \exists y \in \mathrm{Rst}(V_H, \omega_\tau(k)), \quad \omega_\tau(i) \circ (V_H \leftarrow y) = (V_H \leftarrow y')\} \\ &= \mathrm{Rst}(V_H, \omega_\tau(i) \circ \omega_\tau(k)).\end{aligned}$$

Thus $i \circ k$ is canonical and has the required image under ω_τ. Let S, T be the supports of i, k, and K, Q the supports of $\omega_\tau(i), \omega_\tau(k)$. The support of $i \circ k$ is $S \cup T$, and that of $\omega_\tau(i \circ k)$ is $K \cup Q$. For $\mathcal{D} = \mathcal{J}_\tau$,

$$S \cup T \subseteq B_0 \cup R_K \cup R_Q = B_0 \cup R_{K \cup Q};$$

the same inclusion without B_0 applies for $\mathcal{D} = \mathcal{J}_\tau^{\mathrm{rel}}$. Hence $i \circ k \in \mathcal{D}$, proving statement 3.

Finally, suppose τ is constructive and $\omega_\tau(i) = (K \leftarrow b)$. For $x \in \mathcal{X}$, let x' be the state assigned by $i \circ (V_L \leftarrow x)$. For $j \in K$, we have $x' \in \mathrm{Rst}(V_L, i)$, so $\tau_j(x') = b_j$. For $j \notin K$, the disjoint relevance sets and the support filter give $\mathrm{supp}(i) \cap R_j = \varnothing$. Thus $\tau_j(x') = \tau_j(x)$. These coordinate equalities prove statement 2 on both domains. □

C.2 Comparison with interventions on complete alignment cells

Geiger et al. (2025, Definition 31, Remark 32, and Definition 33) start from an alignment: disjoint cells Π_j for the high-level variables, an optional ignored cell, and partial surjective cell maps. Their intervention mapping is defined for interventions that assign complete cells, with cell values in the domains of the corresponding maps, and do not assign the ignored cell. This is a different domain from the full canonical domain in Beckers & Halpern (2019, Definition 3.14).

For a direct comparison with our setting, assume the cell maps are total and the state spaces are full products. Then $R_j \subseteq \Pi_j$. An intervention on complete cells indexed by K is canonical: it fixes the outputs in K, while all other outputs remain unrestricted. Its support is contained in $B_0 \cup R_K$, so it belongs to $\mathcal{J}_\tau$. It need not belong to $\mathcal{J}_\tau^{\mathrm{rel}}$, since a cell may contain coordinates on which its map does not depend. For example, $\Pi_1 = \{X_1, X_2\}$ and $\tau_1(x) = x_1$ give $X_2 \in B_0$, although a complete-cell intervention assigns both coordinates.

If $\Pi_j = R_j$ for every j, the complete-cell rule assigns exactly R_K and is contained in $\mathcal{J}_\tau^{\mathrm{rel}}$. The inclusion can be strict. For $\tau(x_1, x_2) = x_1 \vee x_2$ on $\{0,1\}^2$,

$$R_Y = \{X_1, X_2\}, \qquad \omega_\tau(X_1 \leftarrow 1) = (Y \leftarrow 1).$$

This intervention belongs to $\mathcal{J}_\tau^{\mathrm{rel}}$ but does not assign the complete cell. These comparisons concern the domains; consistency of the models must be checked on the chosen domain.

D Proof of Theorem 3

The target-domain condition is

$$i \circ k \in \mathcal{J}_\tau, \qquad \omega_\tau(i \circ k) = \omega_\tau(i) \circ \omega_\tau(k) \qquad (i, k \in \mathcal{J}_\tau). \tag{19}$$

Proof of Theorem 3. We use the surjective abstraction mapping τ, the full coverage $\omega_\tau(\mathcal{J}_\tau) = \mathcal{I}_H^*$, and the state spaces from Section 2. Suppose first that (19) holds. Lemma C.1 ensures that $\mathcal{J}_\tau$ contains all interventions $(V_L \leftarrow x)$, so Lemma B.1 applies with $\mathcal{D} = \mathcal{J}_\tau$. Fix $j, a \in \mathcal{X}$, and $h = (R_j \leftarrow a_{R_j})$. Since τ_j depends only on R_j, the cylinder image of h is contained in $\{y \in \mathcal{Y} : y_j = \tau_j(a)\}$. For any y in this set, full high-level intervention coverage gives $i \in \mathcal{J}_\tau$ with

$$\omega_\tau(i) = ([m] \setminus \{j\} \leftarrow y_{-j}).$$

Let a' be the state assigned by $i \circ (V_L \leftarrow a)$. Lemma B.1 says that i sets no coordinate in R_j, so $a' \in \mathrm{Rst}(V_L, h)$. Moreover,

$$\begin{aligned}(V_H \leftarrow \tau(a')) &= \omega_\tau(i \circ (V_L \leftarrow a)) \\ &= \omega_\tau(i) \circ (V_H \leftarrow \tau(a)) = (V_H \leftarrow y).\end{aligned}$$

Thus $\tau(a') = y$, and $\tau(\mathrm{Rst}(V_L, h)) = \mathrm{Rst}(V_H, j \leftarrow \tau_j(a))$. Hence h is canonical and belongs to $\mathcal{J}_\tau$: its support is R_j, and the support of $\omega_\tau(h)$ is $\{j\}$. Applying Lemma B.1 to h shows $R_j \cap R_k = \varnothing$ for all $k \neq j$. Thus τ is constructive.

Conversely, if τ is constructive, the final assertion of Lemma C.3 proves closure and composition preservation on $\mathcal{J}_\tau$. This proves the equivalence.

For $\mathcal{J}_\tau^{\mathrm{rel}}$, Lemma C.1 gives $\omega_\tau(\mathcal{J}_\tau^{\mathrm{rel}}) = \omega_\tau(\mathcal{J}_\tau)$. Thus full high-level intervention coverage on one domain is equivalent to full coverage on the other. Lemma C.2 also makes their closure and composition conditions equivalent. The same characterization therefore holds on $\mathcal{J}_\tau^{\mathrm{rel}}$. □

E Proof of Proposition 4.1

Proof of Proposition 4.1. Write $A_{:r}$ for column r of A. A low-level cylinder that assigns coordinates in S has affine image with direction space $\mathrm{span}\{A_{:r} : r \notin S\}$. For each k, coverage of the high-level intervention that fixes every variable except Y_k therefore gives a low-level cylinder whose free columns span $\mathrm{span}(e_k)$. In particular, A has a nonzero column parallel to each coordinate vector e_k.

Suppose row j has at least two nonzero entries. Fix a context u and choose r last among these entries in a low-level causal order. Deleting column r preserves a nonzero column parallel to e_k for every $k \neq j$, since $A_{jr} \neq 0$. It also preserves another column with a nonzero jth entry. Thus

$$\mathrm{rank}\, A_{:,-r} = m, \qquad \tau(\{x : x_r = a\}) = \mathbb{R}^m, \qquad \omega_\tau(X_r \leftarrow a) = \varnothing \quad (a \in \mathbb{R}).$$

All other coordinates on which τ_j depends precede X_r and are unchanged by this intervention. For some value $c(u)$ independent of a, consistency would require

$$A_{jr}a + c(u) = \tau_j(M_L^{X_r \leftarrow a}(u)) = M_{H,j}^{\varnothing}(\tau_{\mathcal{U}}(u)) \quad \text{for every } a \in \mathbb{R}.$$

This is impossible because $A_{jr} \neq 0$. Every row therefore has at most one nonzero entry. The full row rank of A implies that every row has exactly one, and that their columns are distinct. □

The argument uses acyclicity and model consistency, rather than intervention composition, and allows nonlinear structural equations.

Constructivity need not preserve full-domain composition. For example, the abstraction mapping $\tau(x_1, x_2) = x_1 + x_2$ on $\mathbb{R}^2$ is constructive and affine, and ω_τ is surjective. Yet $\omega_\tau(X_1 \leftarrow 1) = \varnothing$, whereas

$$\omega_\tau((X_1 \leftarrow 1) \circ (V_L \leftarrow (0,0))) = (S \leftarrow 1),$$
$$\omega_\tau(X_1 \leftarrow 1) \circ \omega_\tau(V_L \leftarrow (0,0)) = (S \leftarrow 0).$$

The full-domain condition therefore fails. The target-domain condition holds by Theorem 3.

F Counterexamples and repairs to published results

The two counterexamples below use the same fact about value ranges. For real addition and binary XOR, fixing one input leaves the full output range available through the other input:

$$a + \mathbb{R} = \mathbb{R}, \qquad a \oplus \{0,1\} = \{0,1\}.$$

In the first identity $a \in \mathbb{R}$; in the second $a \in \{0,1\}$. Thus the image under τ of the state cylinder for a single-variable intervention is the whole high-level state space. By the definition of ω_τ, that intervention maps to the empty intervention. This is a statement about *state cylinders*, not about the actual solutions of the intervened SCM. The represented solution can still change.

The conclusions being tested are different. The linear theorem claims that its matrix conditions imply exact abstraction for the given models. The approximate-abstraction theorem claims that changing the low-level equations can produce an exact abstraction. The binary example rules out *every* such recursive low-level model.

We use the definitions in Section 2 and the domains in Equation 4. Here γ denotes the context map in Equation 2.

F.1 Constructive lifting: Beckers–Eberhardt–Halpern

The published claim. Theorem 5.4 of Beckers et al. (2020, p. 614), numbered Theorem 5.6 in the supplement (p. 10), includes the conclusion

> "then we can find a model $(M'_L, \mathcal{I}_L)$ that is similar to $(M_L, \mathcal{I}_L)$".

For completeness, the full mathematical claim is restated below in our model notation. The prose is restated, rather than quoted. For finite signatures and recursive deterministic models, its premises are:

1. $\tau : \mathcal{R}_L(V_L) \to \mathcal{R}_H(V_H)$ is constructive;
2. $(M_L, \mathcal{I}_L)$ and $(M_H, \mathcal{I}_H)$ are τ-consistent in the sense of its Definition 2.4;
3. $(M_H, \mathcal{I}_H)$ is a τ-α-approximate abstraction of $(M_L, \mathcal{I}_L)$ in the sense of its Definition 3.2.

Its conclusion is that one can find a similar model $(M'_L, \mathcal{I}_L)$ for which $(M_H, \mathcal{I}_H)$ is an exact τ-abstraction, in time polynomial in $|\mathcal{I}_L|\,|\mathcal{R}'_L(U_L)|$. Here *similar* means that only the equations may change: the signature and allowed interventions stay fixed. The source's τ-consistency is a condition on signatures, intervention sets, and context cardinalities. It is not exact interventional consistency.

The binary XOR example. Let the low-level signature have endogenous variables X_1, X_2 with range $\{0, 1\}$ and one exogenous variable E with range $\{0\}$. Let the high-level signature have endogenous variable S with range $\{0, 1\}$ and one exogenous variable U with range $\{0\}$. Set

$$M_L : \quad X_1 = 0, \quad X_2 = 0; \qquad M_H : \quad S = 0; \qquad \tau(x_1, x_2) = x_1 \oplus x_2.$$

Allow all nine low-level hard interventions and all three high-level hard interventions. We now check each premise.

1. **Finite signatures and recursive models.** Each signature is finite. Each model has one context, constant equations, and an edgeless graph. Both models are therefore recursive deterministic SCMs.
2. **Constructive abstraction mapping.** The mapping is surjective because $\tau(0, s) = s$ for $s \in \{0, 1\}$. The partition has one represented block $\{X_1, X_2\}$ and an empty ignored block. This satisfies the source's Definition 2.6.
3. **τ-consistency.** The full list of intervention types and their images is

i	$\tau(\mathrm{Rst}(V_L, i))$	$\omega_\tau(i)$
$\varnothing$	$\{0, 1\}$	$\varnothing$
$X_1 \leftarrow a$	$a \oplus \{0, 1\} = \{0, 1\}$	$\varnothing$
$X_2 \leftarrow b$	$\{0, 1\} \oplus b = \{0, 1\}$	$\varnothing$
$(X_1, X_2) \leftarrow (a, b)$	$\{a \oplus b\}$	$S \leftarrow a \oplus b$

   where $a, b \in \{0, 1\}$. Thus all $1 + 2 + 2 + 4 = 9$ low-level interventions belong to $\mathcal{I}_L^\tau$, and $\omega_\tau(\mathcal{I}_L) = \{\varnothing, S \leftarrow 0, S \leftarrow 1\} = \mathcal{I}_H$. The context cardinalities are both one, so $|\mathcal{R}_L(U_L)| \geq |\mathcal{R}_H(U_H)|$ also holds. These are all three requirements of Definition 2.4.
4. **τ-1-approximate abstraction.** Use discrete metrics at both levels; in particular, $d_S(s, s') = \mathbf{1}\{s \neq s'\}$ at the high level. The only context map is $\gamma(0) = 0$, and it is surjective. The errors for the four intervention types above are, respectively,

$$0, \qquad a, \qquad b, \qquad 0.$$

   For example, $X_1 \leftarrow 1$ gives the low-level solution $(1, 0)$, whose XOR is one, while the high-level solution under the empty intervention is zero. Hence the maximum error is exactly one. There is only one context map over which to minimize, so the distance in Definition 3.2 is $d_\tau(M_L, M_H) = 1 \leq \alpha$ for $\alpha = 1$. The theorem imposes no further bound on α.

Why the conclusion fails for every similar recursive model. Suppose a similar recursive M'_L makes M_H an exact τ-abstraction. Since $\omega_\tau(X_k \leftarrow a) = \varnothing$,

$$\tau\big((M'_L)^{X_k \leftarrow a}(0)\big) = M_H^{\varnothing}(0) = 0 \qquad (k \in \{1, 2\},\ a \in \{0, 1\}).$$

Each intervention fixes one coordinate, and XOR zero forces the other coordinate to match:

$$(M'_L)^{X_1\leftarrow 0}(0) = (0,0), \quad (M'_L)^{X_1\leftarrow 1}(0) = (1,1) \quad \Longrightarrow \quad X_1 \to X_2,$$
$$(M'_L)^{X_2\leftarrow 0}(0) = (0,0), \quad (M'_L)^{X_2\leftarrow 1}(0) = (1,1) \quad \Longrightarrow \quad X_2 \to X_1.$$

There are only two endogenous variables, and each intervention leaves the other equation unchanged. Thus both edges are forced, contradicting acyclicity.

What changes on the restricted domains. Here $B_0 = \varnothing$ and $R_S = \{X_1, X_2\}$, so

$$\mathcal{J}_\tau = \mathcal{J}_\tau^{\mathrm{rel}} = \{\varnothing\} \cup \{(X_1, X_2) \leftarrow (a,b) : a, b \in \{0,1\}\}.$$

The four single-variable interventions are excluded. All macro-interventions are still represented, and the original constant model is exact on every remaining intervention. The next theorem gives the general repair.

Theorem 5 (Constructive lifting on $\mathcal{J}_\tau$)**.** *Let the abstraction mapping $\tau : \mathcal{X} \twoheadrightarrow \mathcal{Y}$ be constructive on finite Cartesian state spaces. Let M_H be a recursive deterministic SCM, and let $\gamma : \mathcal{E} \twoheadrightarrow \mathcal{U}$ be any fixed surjection of contexts. There exists a recursive low-level SCM M'_L on the given micro-signature such that M_H is its exact abstraction on the entire $\mathcal{J}_\tau$, with context map γ and intervention mapping ω_τ. If the high-level order is independent of context, the low-level order can also be chosen independent of context. With explicit finite tables, the construction takes polynomial time in the table size and $n|\mathcal{E}||\mathcal{X}|$.*

Proof. Constructivity gives disjoint, nonempty minimal blocks R_j and surjections $\bar{\tau}_j : \mathcal{X}_{R_j} \twoheadrightarrow \mathcal{Y}_j$. Choose a right inverse r_j of each $\bar{\tau}_j$, and constants for the ignored coordinates B_0. Write the macro-equation as $Y_j = h_j(Y_{\mathrm{pa}(j)}, u)$. For each coordinate $a \in R_j$ define

$$X_a = \left[r_j\big(h_j\big((\bar{\tau}_k(X_{R_k}))_{k\in\mathrm{pa}(j)}, \gamma(e)\big)\big)\right]_a.$$

Give every coordinate in B_0 its chosen constant equation. No coordinate depends on another coordinate in the same represented block. Order the blocks by a topological order of M_H, and put ignored coordinates first. This gives a recursive model, context by context if necessary.

Fix $i \in \mathcal{J}_\tau$, put $\omega_\tau(i) = (K \leftarrow b)$, and let $x = (M'_L)^i(e)$. For $j \in K$, the canonical cylinder identity gives $\bar{\tau}_j(x_{R_j}) = b_j$ for every value of the unassigned coordinates in that block. For $j \notin K$, the definition of $\mathcal{J}_\tau$ and disjointness imply that i assigns no coordinate of R_j. All equations in that block therefore remain unchanged, giving

$$\bar{\tau}_j(x_{R_j}) = h_j\big((\bar{\tau}_k(x_{R_k}))_{k\in\mathrm{pa}(j)}, \gamma(e)\big).$$

Thus $\tau(x)$ solves the intervened high-level equations. Their unique solution proves consistency with the same γ for every intervention. Assigning all relevant coordinates of selected blocks gives every macro-intervention, proving coverage. Right inverses can be found by enumerating the finite micro-state table; the displayed equations then give the stated polynomial construction. □

Replacing the published theorem's allowed low-level domain by $\mathcal{J}_\tau$ therefore makes its existence conclusion hold. Existence does not require the approximate-consistency premise. The context map that satisfies that premise can be kept unchanged. Since all total micro-assignments belong to $\mathcal{J}_\tau$, the explicit state table has at most $|\mathcal{J}_\tau|$ entries. Thus the tabular construction is also polynomial in the size of the full restricted intervention family and the context space. This does not give a polynomial bound for succinct input encodings with exponentially large state tables.

The proof satisfies each hard assignment even when an intervention assigns only part of a targeted block. The published construction uses a fixed right inverse to choose a representative of each macro-value. That representative need not satisfy the given micro-assignment. Choosing another representative cannot fix the XOR counterexample on $\mathcal{I}_L^\tau$. Our theorem keeps the source's general deterministic-context convention. It does not claim that the same lift exists for an arbitrary prescribed micro-signature with a separate independent exogenous input for each micro-equation.

F.2 Linear block sufficiency: Massidda–Magliacane–Bacciu

The published claim. Theorem 3 of Massidda et al. (2024, p. 2492) is titled "Block Abstraction". It states when M_H is a linear $\mathbf{T}$-abstraction of M_L. The following restatement uses our model symbols and retains the source's quantifiers and displayed conditions; the prose is restated. Let $M_H = (\boldsymbol{Y}, \boldsymbol{U}, \mathbf{M}, P_U)$ and $M_L = (\boldsymbol{X}, \boldsymbol{E}, \mathbf{W}, P_E)$ be linear SCMs with graphs G_H, G_L. The theorem asserts

$$
\begin{aligned}
&M_H \text{ is a linear } \mathbf{T}\text{-abstraction of } M_L \\
&\quad\iff \quad \text{for every topological order } \prec_H \text{ of } G_H, \text{ there is a topological order } \prec_L \text{ of } G_L \\
&\qquad\qquad \text{such that, for every } Y_i, Y_j \in \boldsymbol{Y}, \\
&\qquad Y_i \prec_H Y_j \iff \Pi(Y_i) \prec_L \Pi(Y_j), \qquad \text{[source Eq. (17)]} \\
&\qquad \mathbf{W}_{ij} s_j = m_{ij} t_i. \qquad \text{[source Eq. (18)]}
\end{aligned}
$$

Here $\Pi(Y_j)$ is the source's concrete block, t_j is its column of abstraction coefficients, and $s_j = (I - \mathbf{W}_{jj})^{-1} t_j$ is its column of exogenous coefficients. The notation $\mathbf{W}_{ij}$ refers to the block in source Eq. (14). The source assumes acyclicity, causal sufficiency, faithfulness, and no cancellation of concrete-variable effects through the abstraction mapping (Section 2). It requires consistency for every hard intervention and every context, not only for interventions on complete relevant sets (Section 2, Eq. (4), and Definition 1).

The real-sum example. Take independent noises $E_1, E_2 \sim \mathrm{Laplace}(0, 1)$ with declared context space $\mathbb{R}^2$. Define

$$
M_L: \quad X_1 = E_1, \quad X_2 = E_2; \qquad M_H: \quad S = U; \qquad \tau(x_1, x_2) = x_1 + x_2.
$$

The high-level context space is $\mathbb{R}$. Give U the law of $E_1 + E_2$. Both levels allow all hard interventions. We check the standing assumptions and every condition on the right-hand side of the theorem.

1. **Linear acyclic models and independent noises.** The structural matrices are $\mathbf{W} = 0_{2\times 2}$ and $\mathbf{M} = 0_{1\times 1}$. Both graphs are edgeless. The low-level noises are independent; there is only one high-level noise. Thus both models are causally sufficient. The choice of Laplace noise also shows that the example does not depend on Gaussian noise.
2. **Faithfulness and no cancellation.** The low-level distribution has exactly the independence $X_1 \perp X_2$ required by its edgeless graph. The single-variable high-level graph has no nontrivial conditional independence to test. Each micro-variable contributes to τ with coefficient $+1$, and there is no other directed path whose effect could cancel it. Thus the source's no-cancellation assumption also holds.
3. **Surjective linear abstraction mapping and compatible contexts.** The abstraction matrix is $\mathbf{T} = (1, 1)^\top$, of rank one, and $\tau(0, s) = s$ for every real s. The map $\gamma(e_1, e_2) = e_1 + e_2$ is surjective and gives observational consistency. The chosen law of U is exactly $\gamma_{\#} P_E$. Hence the example does not rely on a mismatch between the noise laws.
4. **The stated intervention domain and all macro-interventions.** For all real a, b, the intervention mapping is

i	$\tau(\mathrm{Rst}(V_L, i))$	$\omega_\tau(i)$
$\varnothing$	$\mathbb{R}$	$\varnothing$
$X_1 \leftarrow a$	$a + \mathbb{R} = \mathbb{R}$	$\varnothing$
$X_2 \leftarrow b$	$\mathbb{R} + b = \mathbb{R}$	$\varnothing$
$(X_1, X_2) \leftarrow (a, b)$	$\{a + b\}$	$S \leftarrow a + b$

   Every low-level hard intervention is therefore in $\mathcal{I}_L^\tau$. Every high-level intervention $S \leftarrow s$ has the implementation $(X_1, X_2) \leftarrow (0, s)$.
5. **The block-order condition, source Eq. (17).** There is one concrete block $\Pi(S) = \{X_1, X_2\}$. Indeed the reduced-form matrices are $\mathbf{F} = I_2$ and $\mathbf{G} = 1$, so the exogenous coefficients are $\mathbf{F}\mathbf{T}\mathbf{G}^{-1} = (1, 1)^\top$, with both entries nonzero. The high-level graph has one topological order. Either order of X_1, X_2 is valid for the low-level graph, and there is no pair of distinct blocks to order. For the repeated pair (S, S), both strict order statements are false. Hence Eq. (17) holds for every pair.

6. **The matrix condition, source Eq. (18).** The only block has

$$t_1 = \begin{pmatrix}1\\1\end{pmatrix}, \qquad s_1 = (I - \mathbf{W}_{11})^{-1} t_1 = t_1, \qquad \mathbf{W}_{11} s_1 = \begin{pmatrix}0\\0\end{pmatrix} = m_{11} t_1.$$

Thus even the diagonal instance of the printed condition holds. If Eq. (18) is read only for distinct blocks, as in the source proof, there is no such pair and the condition still holds.

Why the sufficiency conclusion fails. Any context map that could give exact abstraction must first give observational consistency. Since $M_H(u) = u$, this forces $\gamma(e) = e_1 + e_2$ at every context. Under $X_1 \leftarrow a$,

$$\tau(M_L^{X_1 \leftarrow a}(e)) = a + e_2, \qquad M_H^{\omega_\tau(X_1 \leftarrow a)}(\gamma(e)) = M_H^{\varnothing}(\gamma(e)) = e_1 + e_2.$$

For example, $a = 1$ and $e = (0, 0)$ give $1 \neq 0$. More generally, the identity fails whenever $e_1 \neq a$, which occurs with probability one for each fixed a. Thus this is not an issue confined to a single zero-probability context. No alternative context map can repair the failure, because the empty intervention already fixes that map. The distributions also differ: $\mathrm{Var}(a + E_2) = 2$, whereas $\mathrm{Var}(U) = 4$. The models satisfy the theorem's right-hand side and standing assumptions, but they do not satisfy its $\mathbf{T}$-abstraction conclusion.

The shared reason and the domain repair. The singleton intervention maps to the empty intervention for exactly the same reason as in the XOR example: its state-cylinder image is the full macro-range. In the linear example it changes the represented solution; in the binary example some such intervention must change it in every recursive model with the given signature. Here again $B_0 = \varnothing$, $R_S = \{X_1, X_2\}$, and

$$\mathcal{J}_\tau = \mathcal{J}_\tau^{\mathrm{rel}} = \{\varnothing\} \cup \{(X_1, X_2) \leftarrow (a, b) : a, b \in \mathbb{R}\}.$$

On this restricted domain the given real-sum models are exact abstractions with the same context map and noise laws. The source's sufficiency proof in Appendix B.14 checks interventions on complete relevant sets and misses the single-variable interventions above. The results below state precisely which sufficiency claim is recovered by restricting the domain. In the repaired matrix criterion, $W_{ij} s_j = m_{ij} t_i$ concerns distinct blocks $i < j$; s_j already accounts for internal edges W_{jj}.

Lemma F.1 ($\mathcal{J}_\tau^{\mathrm{rel}}$ for a block-linear abstraction mapping)**.** *Let $P_1, \ldots, P_m$ be disjoint micro-blocks, and $\tau_j(x) = t_j^\top x_{P_j}$ with $t_j \neq 0$. Put $R_j = \{a \in P_j : (t_j)_a \neq 0\}$. On real Cartesian spaces,*

$$\mathcal{J}_\tau^{\mathrm{rel}} = \{(R_K \leftarrow a) : K \subseteq [m],\ a \in \mathbb{R}^{R_K}\}, \qquad \omega_\tau(R_K \leftarrow a) = (K \leftarrow \tau_K(a)).$$

Proof. A linear coordinate τ_j is constant on a micro-cylinder if and only if every coordinate in R_j has been assigned. Otherwise its image is all of $\mathbb{R}$. Disjointness makes the joint image the product of these individual images. Hence every micro-intervention is canonical, with target set $K = \{j : R_j \subseteq \mathrm{supp}(i)\}$. Membership in $\mathcal{J}_\tau^{\mathrm{rel}}$ additionally requires $\mathrm{supp}(i) \subseteq R_K$, forcing $\mathrm{supp}(i) = R_K$. The reverse inclusion is immediate. □

Theorem 6 (Exact block criterion on $\mathcal{J}_\tau^{\mathrm{rel}}$)**.** *Consider real linear SCMs $X = W^\top X + E$ and $Y = M^\top Y + U$. Partition the microvariables into disjoint blocks $P_1, \ldots, P_m$, with $W_{ij} = 0$ for $i > j$, each W_{jj} acyclic, and M strictly upper triangular. Let $T = \mathrm{diag}(t_1, \ldots, t_m)$, where $t_j \neq 0$ and zero entries within t_j are allowed. Define*

$$s_j = (I - W_{jj})^{-1} t_j, \qquad S = \mathrm{diag}(s_1, \ldots, s_m), \qquad \gamma(e) = S^\top e.$$

Then exact consistency on the entire $\mathcal{J}_\tau^{\mathrm{rel}}$ with this context map holds if and only if

$$W_{ij} s_j = m_{ij} t_i \qquad (1 \leq i < j \leq m). \tag{20}$$

The abstraction mapping and context map are surjective. If the noises E_a are independent and $P_U = \gamma_\# P_E$, the macro-noises are independent and the corresponding distributional identities also hold.

Proof. Each s_j is nonzero because $I - W_{jj}$ is invertible. Hence the abstraction mapping and context map are both surjective. For sufficiency, take any $i = (R_K \leftarrow a) \in \mathcal{J}_\tau^{\mathrm{rel}}$, fix a context e, and write its micro-solution blockwise as x_j. Put $y_j = t_j^\top x_j$. For $j \in K$, the cylinder identity gives the required macro-value. For $j \notin K$, no coordinate of P_j was assigned: ignored coordinates are excluded from $\mathcal{J}_\tau^{\mathrm{rel}}$. Consequently

$$x_j = (I - W_{jj}^\top)^{-1} \left(\sum_{i<j} W_{ij}^\top x_i + e_j \right),$$

and therefore

$$y_j = \sum_{i<j} x_i^\top W_{ij} s_j + e_j^\top s_j = \sum_{i<j} m_{ij} y_i + \gamma_j(e).$$

These are the macro-equations left unchanged by the intervention. The acyclic macro-model has a unique solution, which proves consistency for every intervention in $\mathcal{J}_\tau^{\mathrm{rel}}$.

For necessity, use the empty intervention, which belongs to $\mathcal{J}_\tau^{\mathrm{rel}}$. Because the contexts range over $\mathbb{R}^n$, observational consistency is the matrix identity

$$T^\top (I - W^\top)^{-1} = (I - M^\top)^{-1} S^\top, \quad \text{equivalently} \quad T(I - M) = (I - W)S.$$

Its off-diagonal (i, j) block is $-m_{ij} t_i = -W_{ij} s_j$; its diagonal block holds by definition of s_j. Finally, if the micro-noises are independent, the disjoint input blocks of S make the macro-noises independent. Pushing forward the context law gives the distributional identities. □

This theorem repairs the published sufficiency implication under the stated block assumptions. Under those assumptions, it also gives an equivalence for the induced blockwise context map. Variables downstream of all represented blocks can be added without changing the result. The theorem does not show that every abstraction on a restricted domain has disjoint, ordered noise blocks.

Why $\mathcal{J}_\tau$ does not suffice when B_0 is nonempty. Take the faithful chain $A = E_A$, $B = A + E_B$, $C = B + E_C$, with independent standard Gaussian noises, and abstraction mapping $\tau(a, b, c) = (b, c)$. Use blocks $P_1 = \{A, B\}$, $P_2 = \{C\}$, $t_1 = (0, 1)^\top$, $t_2 = 1$, $s_1 = (1, 1)^\top$, $s_2 = 1$, and macro-model $Y_1 = U_1$, $Y_2 = Y_1 + U_2$, with $(U_1, U_2) = (E_A + E_B, E_C)$. The nontrivial cross-block condition is $W_{12} s_2 = (0, 1)^\top = t_1$. Here $B_0 = \{A\}$, so $A \leftarrow a$ belongs to $\mathcal{J}_\tau$ and its image under ω_τ is the empty intervention. Its represented solution is $(a + e_B, a + e_B + e_C)$ instead of $(e_A + e_B, e_A + e_B + e_C)$. Thus the matrix conditions do not ensure $\mathcal{J}_\tau$ consistency. The smaller domain $\mathcal{J}_\tau^{\mathrm{rel}}$ contains exactly the interventions on B, C and satisfies the preceding theorem.

F.3 The intervention domain in the published linear examples

Section 2 of Massidda et al. (2024) uses full hard interventions, the cylinder-image definition of ω_τ, and pointwise consistency (their Eq. (4), p. 2488). Their Definition 1 requires strong abstraction. The following calculations show why their Examples 1, 3, and 5 require a smaller intervention domain. They concern the stated definitions and examples, rather than the learning method on a smaller domain. For any fixed context e, consistency implies

$$\omega_\tau(i) = \omega_\tau(i') \quad \Longrightarrow \quad \tau(M_L^i(e)) = \tau(M_L^{i'}(e)).$$

Each example violates this implication before composition is tested.

Example 1: an omitted variable affects retained variables. Their Eq. (10) and Appendix B.5, Eqs. (52)–(54), specify

$$X_1 = E_1, \quad X_2 = X_1 + E_2, \quad X_3 = -X_1 + E_3, \quad X_4 = X_2 + X_3 + E_4, \quad X_5 = X_1 + X_4 + E_5, \quad \tau(x) = (x_1, x_4, x_5).$$

Since X_2 is omitted, $\omega_\tau(X_2 \leftarrow a) = \varnothing$ for every a. At $e = (1, 1, 1, 1, 1)$,

$$M_L(e) = (1, 2, 0, 3, 5), \qquad \tau(M_L(e)) = (1, 3, 5),$$
$$M_L^{X_2 \leftarrow 3}(e) = (1, 3, 0, 4, 6), \quad \tau(M_L^{X_2 \leftarrow 3}(e)) = (1, 4, 6).$$

The represented outcomes differ despite the same empty high-level intervention. The example relaxes faithfulness, but consistency is still required. Here $X_2 \in B_0$, so $\mathcal{J}_\tau$ retains this intervention and $\mathcal{J}_\tau^{\mathrm{rel}}$ removes it. On $\mathcal{J}_\tau^{\mathrm{rel}}$, the published high-level model $Y_1 = U_1$, $Y_2 = U_2$, $Y_3 = Y_1 + Y_2 + U_3$ does satisfy consistency with $\gamma(e) = (e_1, e_2 + e_3 + e_4, e_5)$: substituting the unassigned low-level equations gives these high-level equations, and assignments of X_1, X_4, X_5 directly fix their corresponding outputs.

Example 3: an omitted intermediate variable. For this calculation, label the six nodes in the source's diagram (p. 2491) as $X_1, \dots, X_6$ in the order upper left, lower left, upper middle, upper right, lower middle, lower right. Its edges are

$$X_1 \to X_3 \to X_4 \to X_5 \to X_6, \qquad X_2 \to X_5.$$

The indicated relevant sets give an abstraction mapping of the form

$$\tau(x) = (c_1 x_1, c_2 x_2, c_4 x_4 + c_6 x_6), \qquad c_1 c_2 c_4 c_6 \neq 0.$$

In particular, the third relevant set has two variables. The last structural equation is $X_6 = w_{56} X_5 + E_6$, with $w_{56} \neq 0$. Since X_5 is omitted and X_4 is its ancestor,

$$\omega_\tau(X_5 \leftarrow a) = \omega_\tau(X_5 \leftarrow a') = \varnothing,$$
$$\tau_3(M_L^{X_5 \leftarrow a}(e)) - \tau_3(M_L^{X_5 \leftarrow a'}(e)) = c_6 w_{56}(a - a') \neq 0 \quad (a \neq a').$$

These interventions belong to $\mathcal{J}_\tau$ because $X_5 \in B_0$; they are excluded by $\mathcal{J}_\tau^{\mathrm{rel}}$. There is also an aggregation issue: assigning the last node X_6 alone has empty high-level image but changes the third output. Both restricted domains exclude that intervention. Thus $\mathcal{J}_\tau$ addresses the latter issue but not the omitted intermediate variable.

Example 5: two sums. Equation (19) specifies $\tau(x) = (x_1 + x_2, x_3 + x_4)$. Both models claimed to abstract in the example have $X_1 = E_1$ and $X_2 = X_1 + E_2$. Therefore

$$\omega_\tau(X_2 \leftarrow a) = \varnothing, \qquad \tau_1(M_L^{X_2 \leftarrow a}(e)) = e_1 + a.$$

Two distinct values of a contradict consistency. Using the displayed coefficients and $e = (1, 1, 1, 1)$, both models give

$$\tau(M_L(e)) = (3, 5), \qquad \tau(M_L^{X_2 \leftarrow 3}(e)) = (4, 6).$$

Here $B_0 = \varnothing$. Both restricted domains remove the intervention used above. The block-matrix criterion of Theorem 6 applies on this smaller domain.

G Nonlinear abstraction learning

This experiment learns an abstraction mapping from labeled states. It studies the training constraints supplied by intervention composition; it does not fit a pair of SCMs. The input-pair assignment is available to every method.

G.1 Target mapping and data

Write $x = ((a_j, b_j))_{j=1}^m$. The target uses the pooling formula in Section 5, with gate logits

$$g_j^*(x) = 0.6 \sin(a_j + 0.5 b_j) + 0.35(a_j - b_j) + \beta_j, \qquad (\beta_1, \dots, \beta_4) = (-0.7, -0.2, 0.2, 0.7).$$

Each output depends only on its own pair. To draw inputs, independently sample a shared vector $u \sim N(0, I_2)$ and vectors $\epsilon_j \sim N(0, I_2)$, and set

$$(a_j, b_j) = \sqrt{\rho}\, u + \sqrt{1 - \rho}\, \epsilon_j.$$

Thus each pair has distribution $N(0, I_2)$, and matching coordinates in different pairs have correlation ρ. The main setting has $m = 4$, 64 labeled training states, and $\rho = 0.95$. Each development seed has 1,024 additional labeled validation states at the training correlation. Independent-pair testing uses $\rho = 0$ and 8,192 new states. The labels are the target mapping's outputs. Training, validation, and test draws use separate random generators.

Each learned gate is a two-layer network with a tanh hidden layer. A gate that can read all $2m$ inputs has width 24. The constructive controls read only their own pair, using width 24 or a width chosen to match the total parameter count. For four outputs these widths give 388 and 964 parameters, respectively; the unrestricted model has 964 parameters. The matched width is 60. Training uses Adam and prediction MSE. All unrestricted methods share initialization and labeled data for each seed.

G.2 Canonical interventions and the training constraint

For a set K and values c_k, consider

$$i = ((a_k, b_k) \leftarrow (c_k, c_k))_{k \in K}.$$

Every gate lies strictly between zero and one. Assigning (c_k, c_k) therefore fixes output k to c_k, regardless of all other inputs. Conversely, to realize any remaining output vector, assign each remaining pair two copies of its desired value. Hence the cylinder image is exactly the high-level cylinder for $Y_K \leftarrow c_K$. This proves surjectivity and full high-level intervention coverage for every parameter value. Both coordinates in pair k are relevant to output k: changing either coordinate from an equal pair makes the output lie strictly between the unequal entries. The intervention therefore belongs to both $\mathcal{J}_{\tau_\theta}$ and $\mathcal{J}^{\mathrm{rel}}_{\tau_\theta}$.

Let $(V_L \leftarrow x') = i \circ (V_L \leftarrow x)$. The target outputs are already fixed correctly. The remaining composition equalities are

$$\tau_{\theta,j}(x') = \tau_{\theta,j}(x), \qquad j \notin K.$$

For fixed $a_j \neq b_j$, the pooling output is strictly monotone in its finite gate logit. Thus equality of the outputs, equality of the sigmoid gates, and equality of the logits have the same zero set for this check. At $a_j = b_j$, the output is independent of the gate. The continuous sampling used here assigns probability zero to this case for an unchanged pair. A gate or logit residual is specific to this parameterization; its size is not the size of the output residual.

Equations 7 and 8 give the exact losses for the selected four-output configurations. For m outputs and one targeted pair, the denominators are $B(m-1)$ and mN. When samples target different numbers of pairs, the code averages over all non-target entries in the batch. The selected gate configuration uses $\lambda = 10$, learning rate 0.01, 3,000 steps, and 128 fresh constraints per step. Its base states are formed by independently recombining input pairs from the observed training states. It changes one pair at a time to (c, c) with $c \sim N(0, 1)$. This procedure requests no new target labels.

G.3 Controls, selection, and evaluation

Controls include ordinary supervised learning, L1, weight decay, three ordinary pair-change penalties, a Jacobian penalty, label augmentation, and two constructive architectures. Ordinary pair changes replace selected pairs by independent normal entries and penalize changes in untargeted outputs, sigmoid gates, or gate logits. The selected logit variant changes all but one pair; the other two change one pair. The Jacobian control penalizes derivatives of $\tau_{\theta,j}$ with respect to inputs outside output j's own pair:

$$\mathcal{L}_{\mathrm{Jac}}(\theta) = \frac{1}{2Bm(m-1)} \sum_{r=1}^{B} \sum_{j=1}^{m} \sum_{\ell \notin \{2j-1, 2j\}} \left(\frac{\partial \tau_{\theta,j}}{\partial x_\ell}(\widetilde{x}^{(r)}) \right)^2.$$

Label augmentation uses the known target value of a diagonal-pair assignment and the constructive target's unchanged outputs. It uses the existing labels and operation semantics. All methods receive the same input-pair information.

The broad search contains 478 configurations at development seed 10. The search varies the loss form, penalty strength, learning rate, constraint sampling, and target size. Shortlisted configurations are checked at fresh seeds 11 and 12, with 1,000- and 3,000-step budgets, for 140 further runs. Two selection rules were fixed: validation prediction MSE, and that MSE plus output composition MSE. The latter uses 1,024 independent unlabeled Gaussian states for each of three target modes. Both rules selected the same configuration for each method. Figure 5 therefore shows each fitted model once. Gate Comp was the Comp family selected by validation; the Jacobian penalty was the selected ordinary regularization control.

Final configurations were fixed before evaluating test outcomes. The reported setting uses seeds 100–107. The two selection-rule labels reuse the same runs and are not separate replications.

Prediction MSE compares the learned and target mappings. Output composition MSE instead measures the learned mapping's own consistency under the equalities above. We evaluate 8,192 new states for each of three target modes: one pair, a nonempty proper subset, and all but one pair. The reported composition metric gives the three modes equal weight. It does not use target-model labels after an intervention and is not an average over all canonical interventions.

Method	Prediction MSE (10^{-4})	Comp MSE	Parameters	Steps
Unconstrained	55.199 ± 6.252	2.39×10^{-3}	964	3000
L1	11.398 ± 5.096	3.62×10^{-4}	964	3000
Weight decay	39.986 ± 5.001	1.31×10^{-3}	964	3000
Output Comp	3.777 ± 1.764	7.24×10^{-6}	964	3000
Gate Comp	2.644 ± 1.475	2.29×10^{-6}	964	3000
Logit Comp	3.733 ± 1.517	6.82×10^{-6}	964	3000
Pair change, output	3.867 ± 1.989	9.19×10^{-6}	964	3000
Pair change, gate	2.618 ± 1.331	3.46×10^{-6}	964	3000
Pair change, logit	3.410 ± 1.473	6.27×10^{-6}	964	3000
Label augmentation	8.727 ± 2.533	1.89×10^{-4}	964	3000
Jacobian penalty	2.406 ± 1.323	1.73×10^{-6}	964	3000
Constructive, width 24	1.554 ± 0.622	0	388	3000
Constructive, matched count	1.613 ± 0.723	0	964	3000

Table 1: Selected methods in the main setting. Prediction errors are means $\pm$ sample standard deviations across eight training seeds. Composition errors are seed means. All methods use 64 training labels and the same known pair assignment; tuning also uses 1,024 labeled validation states per development seed.

Interpretation and cost. All methods receive known pair assignments. The comparison measures the effect of sampled composition constraints, with ordinary regularization and explicit constructivity as controls. Sampled errors do not certify exact constructivity.

Gate Comp and its pair-change control each evaluate 960,000 training examples, counting repeated states. The Jacobian method evaluates 576,000 examples and performs 12,000 additional input-gradient passes. These costs exclude development search. All reported methods use 3,000 training steps.