Root-Intervention Flattening: Counterexamples to Quotient-Graph Identifiability in Causal Representation Learning

Kishan
KyroDB
Kishan@KyroDB.com

Abstract

Paired observations before and after a perfect intervention retain same-unit information through the reuse of non-target exogenous noise. Li, Kaba, and Ravanbakhsh (AISTATS 2025, Theorem 3.1) claim that, under observational faithfulness, absolute continuity, continuously differentiable mechanisms, and diffeomorphic mixing, an arbitrary finite family of such interventions identifies a quotient causal graph generated by the intervention targets' non-descendant sets. We give counterexamples showing that the claim is false as published.

There are two independent obstructions. First, the paper invokes a strict edge-preserving graph homomorphism. That convention is incompatible with quotient blocks that contain an edge; a two-node faithful Gaussian chain already contradicts the theorem when the equal-law rival is the ground-truth model itself. Second, even after replacing strict homomorphisms by edge-contracting quotient maps, a substantive identifiability failure remains. We construct an edgeless three-variable Gaussian model and a faithful chain-plus-isolate model with two positive-probability atomic root interventions. Their observed pre/post pairs agree pathwise conditional on each target, while the prescribed quotient is the three-node edgeless graph and cannot be an abstraction of the rival chain.

The decisive failure in the proof is geometric: a diffeomorphism transports the paired relation, but need not preserve the class of coordinate partial diagonals. A shear sends $(d, 0, 0)$ to $(d, d, 0)$, and the two coordinate-set Boolean algebras have respectively eight and four elements. We prove a general root-intervention flattening theorem and an n-variable family using the minimum possible number, $\lceil \log_2 n \rceil$, of grouped root-intervention regimes needed to induce n singleton atoms. We conclude with precise repair options and a coordinate-free weakening of the identifiable object.

1 Introduction

Causal representation learning asks when latent causal variables and their structural relations can be recovered from high-dimensional observations [7]. Observational data alone leave a large nonlinear reparameterization ambiguity, so positive identifiability results introduce additional variation: temporal structure, multiple environments, sparse perturbations, or interventions. In the same-unit paired setting, one observes $(X, \widetilde{X})$ before and after a random perfect intervention while the non-target exogenous noises are reused. Under sufficiently rich atomic intervention coverage, this coupling can identify the latent causal model up to componentwise reparameterization [3].

Li et al. [4] study incomplete and grouped intervention families. Their Theorem 3.1 states, in essence, that the paired law identifies a structural-causal-model abstraction whose graph is obtained

by quotienting the ground-truth DAG by the atom partition of the finite Boolean algebra generated by target non-descendant sets. The proposed conclusion is natural: overlapping invariance sets would reveal their intersections and complements, and hence the finest causally meaningful blocks supported by the available interventions.

This paper shows that the conclusion does not follow from the published assumptions. The failure is not caused by hidden confounding, soft interventions, singular latent distributions, noninvertible mixing, unfaithfulness, target-label ambiguity, or a distributional coincidence. The main counterexample is affine and Gaussian, satisfies faithfulness, and yields pathwise equality conditional on every intervention target.

Our contributions are as follows.

- We isolate a literal incompatibility between the quotient construction and the strict edge-preserving convention for graph homomorphisms invoked by the published definition of SCM abstraction. This gives a two-node self-counterexample.
- We give a three-variable counterexample that survives the natural repair in which internal edges may be contracted. It directly falsifies the implication asserted by Theorem 3.1.
- We locate the failed proof step. Arbitrary latent diffeomorphisms preserve set-theoretic intersections and complements, but they do not preserve the *class of coordinate partial diagonals*. Consequently, "which named coordinates stayed equal" is not a diffeomorphism-invariant statistic of a paired support.
- We prove a root-intervention flattening theorem: whenever only roots are intervened upon, a smooth SCM can be absorbed into an edgeless exogenous-coordinate model with a composite observation map.
- We construct counterexamples in every dimension $n \geq 2$ with $n-1$ causal edges and only $\lceil \log_2 n \rceil$ grouped root-intervention regimes. This number of regimes is minimal for generating an n-atom Boolean partition.

The result is deliberately scoped. We do not analyze Theorem 3.2 of Li et al. [4], and we do not claim that causal abstractions are intrinsically ill-defined. The negative result concerns Theorem 3.1 under its published hypothesis class and its published abstraction relation. Richer intervention coverage or a restricted reparameterization class can restore positive identifiability statements.

2 Setup and the audited claim

2.1 Paired perfect-intervention models

Let $G = (V, E)$ be a finite DAG with $V = \{1, \dots, n\}$. A Markovian SCM has mutually independent exogenous variables E_i and structural equations

$$Z_i = f_i(Z_{\mathrm{pa}_G(i)}, E_i), \qquad i \in V. \tag{1}$$

A random intervention target I takes values in a finite family $\mathcal{I} \subseteq 2^V$. Conditional on $I = S$, a perfect intervention replaces each target mechanism by

$$\widetilde{Z}_i = \widetilde{f}_i^{(S)}(\widetilde{E}_i), \qquad i \in S, \tag{2}$$

where the fresh target noise $\widetilde{E}_S$ is independent of the pre-intervention exogenous vector E. Every non-target variable is recomputed using its original mechanism and the same non-target noise:

$$\widetilde{Z}_j = f_j(\widetilde{Z}_{\mathrm{pa}_G(j)}, E_j), \qquad j \notin S. \tag{3}$$

A diffeomorphism $g : \mathcal{Z} \to \mathcal{X}$ yields the observed pair

$$(X, \widetilde{X}) = (g(Z), g(\widetilde{Z})). \tag{4}$$

The target may be hidden. Our counterexamples satisfy the stronger equality of target-conditional laws, so they remain valid if I is observed.

We use the target-excluding non-descendant convention forced by the invariance equation in the audited paper:

$$\mathrm{nd}_G(S) := V \setminus (S \cup \mathrm{de}_G(S)), \tag{5}$$

where $\mathrm{de}_G(S)$ denotes strict descendants of S. These vertices are guaranteed by the graph semantics to remain unchanged. In the full-support counterexamples below, no additional endogenous coordinate remains equal almost surely.

2.2 The finite Boolean algebra and quotient graph

For a target family $\mathcal{I}$, let

$$\mathcal{B}_G(\mathcal{I}) := \sigma(\{\mathrm{nd}_G(S) : S \in \mathcal{I}\})$$

be the finite Boolean algebra generated by the non-descendant sets. Its atoms form a partition $\mathcal{P}_G(\mathcal{I})$ of V. Given a vertex partition $\mathcal{P}$, the quotient $G/\mathcal{P}$ has the blocks as vertices and, for distinct blocks $A, B \in \mathcal{P}$, an edge $A \to B$ whenever some $i \in A$ and $j \in B$ satisfy $i \to j$ in G. Internal edges are discarded.

A directed graph homomorphism $\phi : H \to K$ is a vertex map satisfying

$$u \to v \in E(H) \quad \Longrightarrow \quad \phi(u) \to \phi(v) \in E(K). \tag{6}$$

This is the edge-preserving definition used by the abstraction framework cited in Li et al. [4]; see Otsuka and Saigo [5]. In particular, if H and K are loopless, adjacent source vertices cannot be identified unless the target contains the corresponding self-loop.

2.3 Theorem 3.1 under audit

Let $\theta^\star$ be a ground-truth model in the hypothesis class of Li et al. [4]: its observational distribution is faithful to its DAG; the exogenous and intervention-noise distributions are absolutely continuous; its structural and intervention mechanisms are continuously differentiable; and its mixing function is a diffeomorphism. Define

$$Q^\star := G^\star / \mathcal{P}_{G^\star}(\mathcal{I}^\star). \tag{7}$$

Theorem 3.1 claims that every model θ' in the same class satisfying

$$\mathcal{L}_{\theta'}(X, \widetilde{X}) = \mathcal{L}_{\theta^\star}(X, \widetilde{X}) \tag{8}$$

admits an SCM abstraction with graph $Q^\star$. Its graph component is a surjective homomorphism $G' \to Q^\star$. Since a failure of the graph component already falsifies the abstraction relation, the distributional component need not be examined in the counterexamples below.

3 First obstruction: quotient contraction is not a strict graph homomorphism

The published quotient operation discards edges internal to a block, while the published abstraction relation requires every source edge to map to a target edge. These operations are incompatible before any statistical argument is invoked.

Proposition 3.1 (Two-node self-counterexample)**.** *Under the definitions in section 2, Theorem 3.1 of Li et al. [4] is false even when the equal-law rival is the ground-truth model itself.*

Proof. Let $G^\star$ be the two-node chain $1 \to 2$ and take

$$Z_1 = E_1, \qquad Z_2 = Z_1 + E_2, \qquad E_1, E_2 \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

Use the identity mixing function and the single positive-probability target $S = \{1\}$. Replace Z_1 by an independent $C \sim \mathcal{N}(1,1)$ and recompute $\widetilde{Z}_2 = C + E_2$. The model is faithful, affine, smooth, and nondegenerate, and it satisfies every published assumption.

The intervention changes the target and its only descendant, so $\mathrm{nd}_{G^\star}(\{1\}) = \varnothing$. Hence the generated Boolean algebra is $\{\varnothing, V\}$, its atom partition is $\{V\}$, and $Q^\star$ is a single loopless vertex. Apply the theorem with $\theta' = \theta^\star$. Equality in (8) is tautological, but a graph homomorphism $G^\star \to Q^\star$ would have to send the edge $1 \to 2$ to a self-loop. No such loop exists. Therefore the required SCM abstraction does not exist. □

Remark 3.2 (A definition-level repair). The canonical quotient map becomes valid if (6) is replaced by the weaker condition

$$u \to v \quad \Longrightarrow \quad \phi(u) = \phi(v) \text{ or } \phi(u) \to \phi(v). \tag{9}$$

This allows internal edges to contract. Such a change repairs proposition 3.1, but it changes the published abstraction definition. More importantly, it does not repair the substantive counterexample in the next section, where source and target have the same number of vertices and no contraction is possible.

4 A smooth Gaussian counterexample beyond edge contraction

Let

$$A, B, D \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \qquad C_1, C_3 \overset{\text{iid}}{\sim} \mathcal{N}(1,1),$$

all mutually independent. Let

$$\mathbb{P}(I = \{1\}) = p, \qquad \mathbb{P}(I = \{3\}) = 1 - p, \qquad 0 < p < 1.$$

Both targets are atomic roots and have positive probability.

4.1 The edgeless ground truth

Let G^{F} be the edgeless graph on three vertices and define

$$W = (W_1, W_2, W_3) = (A, B, D).$$

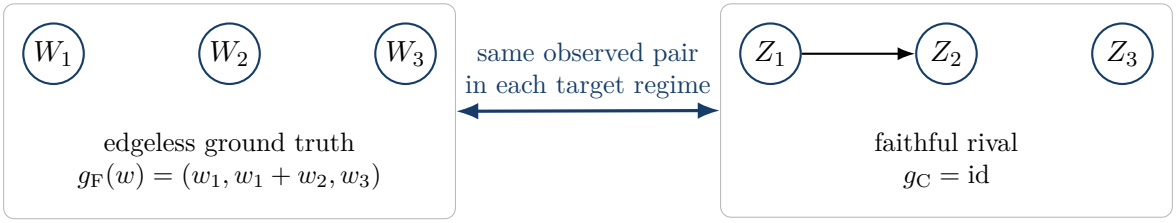

Figure 1: Two incompatible latent graphs generate the same target-conditional observed pair. The triangular observation map in the flat model is the solution map of the chain.

Use the triangular mixing map

$$g_F(w_1, w_2, w_3) = (w_1, w_1 + w_2, w_3). \tag{10}$$

Its Jacobian determinant is one. The two perfect interventions are

$$\widetilde{W}^{\{1\}} = (C_1, B, D),$$
$$\widetilde{W}^{\{3\}} = (A, B, C_3).$$

4.2 The equal-law chain rival

Let G^C have the single edge $1 \to 2$, with vertex 3 isolated, and define

$$Z_1 = A, \qquad Z_2 = Z_1 + B, \qquad Z_3 = D.$$

Use the identity mixing map $g_C = \mathrm{id}$. The same two root interventions give

$$\widetilde{Z}^{\{1\}} = (C_1, C_1 + B, D),$$
$$\widetilde{Z}^{\{3\}} = (A, A + B, C_3).$$

The target mechanism is completely severed, and every non-target mechanism reuses its original noise.

Proposition 4.1 (Target-conditional pathwise equality)**.** *For each $S \in \{\{1\}, \{3\}\}$,*

$$(g_F(W), g_F(\widetilde{W}^S)) = (g_C(Z), g_C(\widetilde{Z}^S)) \quad \textit{almost surely.} \tag{11}$$

Consequently, the target-marked and target-marginalized observed laws are identical.

Proof. For $S = \{1\}$, both sides of (11) equal

$$((A, A + B, D), (C_1, C_1 + B, D)).$$

For $S = \{3\}$, both equal

$$((A, A + B, D), (A, A + B, C_3)).$$

No integration or mixture-identifiability argument is required. □

Lemma 4.2 (Premise verification)**.** *Both models in proposition 4.1 belong to the hypothesis class stated for Theorem 3.1 of Li et al. [4].*

Proof. All latent and noise spaces are $\mathbb{R}$; all observational and intervention noises have smooth positive densities; all mechanisms and mixing functions are affine and C^∞; and both mixing maps are diffeomorphisms. The interventions are perfect root interventions with fresh target noise independent of (A, B, D), while every non-target noise is reused exactly.

The flat observational law is a product distribution and is faithful to the edgeless DAG. The rival covariance is

$$\mathrm{Cov}(Z) = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Thus Z_1 and Z_2 are dependent, while $Z_3 \perp\!\!\!\perp (Z_1, Z_2)$. These are exactly the conditional independences of the graph $1 \to 2$ plus an isolated vertex, so the Gaussian rival is faithful. □

4.3 The prescribed quotient and the contradiction

In the flat ground-truth graph,

$$\mathrm{nd}_{G^{\mathrm{F}}}(\{1\}) = \{2, 3\}, \qquad \mathrm{nd}_{G^{\mathrm{F}}}(\{3\}) = \{1, 2\}.$$

These sets generate all singletons:

$$\{1\} = \{2, 3\}^c, \qquad \{3\} = \{1, 2\}^c, \qquad \{2\} = \{2, 3\} \cap \{1, 2\}.$$

Hence $\mathcal{B}_{G^{\mathrm{F}}}(\mathcal{I}) = 2^V$, its atom partition is

$$\mathcal{P}_{G^{\mathrm{F}}}(\mathcal{I}) = \{\{1\}, \{2\}, \{3\}\},$$

and the prescribed quotient is the three-vertex edgeless graph itself.

Theorem 4.3 (Counterexample to Theorem 3.1 as published)**.** *The models in proposition 4.1 satisfy every stated premise of Theorem 3.1 of Li et al. [4] and have identical observed paired laws, but the claimed SCM-abstraction conclusion is false.*

Proof. By proposition 4.1 and lemma 4.2, the premise holds. The conclusion requires a surjective graph homomorphism from the rival graph G^{C} to the prescribed three-vertex edgeless quotient. Any surjection between two three-element vertex sets is a bijection. Therefore the endpoints of the edge $1 \to 2$ have distinct images, and both the strict condition (6) and the contraction-friendly condition (9) require an edge between those images. The target graph is edgeless. No such map exists. □

The argument is robust to target disclosure, to arbitrary target probabilities in $(0, 1)$, and to arbitrarily small nonzero shears. Replacing (10) by

$$g_\lambda(w) = (w_1, \lambda w_1 + w_2, w_3), \qquad \lambda \neq 0,$$

produces a faithful rival $Z_2 = \lambda Z_1 + B$ and converges to the identity map as $\lambda \to 0$.

Remark 4.4 (Gaussianity is inessential). The construction does not use Gaussian rotational symmetry. Replace A, B, D by independent centered logistic variables and C_1, C_3 by independent shifted logistic variables, all with finite nonzero variance and smooth positive densities. The pathwise identities are unchanged, $Z_3 \perp\!\!\!\perp (Z_1, Z_2)$, and

$$\mathrm{Cov}(Z_1, Z_2) = \lambda \,\mathrm{Var}(A) \neq 0.$$

Thus the flat model remains faithful to the edgeless graph and the rival remains faithful to the chain-plus-isolate graph. Gaussianity is used only for notational transparency.

5 Where the published proof fails

The observed-law equality legitimately induces a diagonal transport of the latent pair measure. In the notation of Li et al. [4], the pushforward identity in their Eq. (24) is valid. Even granting the component matching constructed around their Eqs. (28)–(30), the unsupported step is its promotion to complement and intersection preservation immediately before their Eq. (33). In the present notation, the latent change of coordinates is

$$q := g_{\mathrm{C}}^{-1} \circ g_{\mathrm{F}}, \qquad q(w_1, w_2, w_3) = (w_1, w_1 + w_2, w_3). \tag{12}$$

Thus the rival pair measure is the pushforward of the flat pair measure by $q \times q$. The error is not in this transport. The error is the subsequent replacement of transported geometric relations by lists of equal *named coordinates*.

5.1 Coordinate partial diagonals are not diffeomorphism invariant

For $A \subseteq V$, define the coordinate partial diagonal

$$D_A := \{(z, \tilde{z}) : z_A = \tilde{z}_A\}. \tag{13}$$

A diffeomorphism preserves intersections and complements of ordinary subsets of pair space, but it need not map D_A to D_B for any coordinate index set B. Moreover, Boolean operations on coordinate labels are not the ordinary set operations on these relations: $D_A \cap D_B = D_{A \cup B}$, while D_{A^c} is not the pair-space complement of D_A. Ordinary set-theoretic equivariance therefore gives no coordinate-lattice equivariance.

In the flat model, the target-conditional supports are exactly

$$\mathcal{R}^{\mathrm{F}}_{\{1\}} = D_{\{2,3\}}, \qquad \mathcal{R}^{\mathrm{F}}_{\{3\}} = D_{\{1,2\}}. \tag{14}$$

Under $q \times q$,

$$(q \times q)(D_{\{2,3\}}) = \{(z, \tilde{z}) : z_3 = \tilde{z}_3,\ z_2 - z_1 = \tilde{z}_2 - \tilde{z}_1\}, \tag{15}$$

$$(q \times q)(D_{\{1,2\}}) = D_{\{1,2\}}. \tag{16}$$

The first image is an oblique partial diagonal, strictly smaller than $D_{\{3\}}$, and it is not D_B for any $B \subseteq V$. In difference coordinates this is the elementary shear

$$(d, 0, 0) \longmapsto (d, d, 0). \tag{17}$$

For nonlinear q, the situation is even less coordinate-like because generally $q(z) - q(\tilde{z}) \neq q(z - \tilde{z})$.

5.2 The Boolean-algebra extension is impossible

The maximal coordinate-equality sets in the rival are

$$\mathrm{nd}_{G^{\mathrm{C}}}(\{1\}) = \{3\}, \qquad \mathrm{nd}_{G^{\mathrm{C}}}(\{3\}) = \{1, 2\}.$$

They generate only

$$\mathcal{B}_{G^{\mathrm{C}}}(\mathcal{I}) = \{\varnothing, \{3\}, \{1, 2\}, V\}, \tag{18}$$

which has four elements and atoms $\{1, 2\}$ and $\{3\}$. The ground algebra has eight elements and three atoms. Therefore no Boolean-algebra isomorphism can connect them.

More directly, after extracting the maximal rival coordinate-equality set from each transported component, the resulting generator matching is

$$\{2,3\} \longmapsto \{3\}, \qquad \{1,2\} \longmapsto \{1,2\}. \tag{19}$$

The only alternative is to swap the two rival generators; since those generators are complements, their intersection is still empty. Thus no generator matching can extend to a Boolean isomorphism. For the matching in (19), intersection preservation would require the nonempty ground atom $\{2\}$ to satisfy

$$\phi(\{2\}) = \phi(\{2,3\} \cap \{1,2\}) = \{3\} \cap \{1,2\} = \varnothing,$$

which is impossible for a Boolean-algebra isomorphism.

Proposition 5.1 (Failure of the complement-disentangling inference)**.** *For $N = \{2,3\}$ and $\phi(N) = \{3\}$, the conditional independence used to infer complement-wise disentanglement in the proof of Theorem 3.1 is false:*

$$q(W)_{\phi(N)^c} \not\perp\!\!\!\perp W_N \mid W_{N^c}. \tag{20}$$

Proof. Here

$$q(W)_{\phi(N)^c} = (W_1, W_1 + W_2), \qquad W_N = (W_2, W_3), \qquad W_{N^c} = W_1.$$

Conditional on W_1, the second component $W_1 + W_2$ reveals W_2 exactly. For example,

$$\mathrm{Var}(W_2 \mid W_1, q(W)_{\{1,2\}}) = 0, \qquad \mathrm{Var}(W_2 \mid W_1) = 1.$$

Thus (20) fails. □

The complement projection relevant to this inference uses the target-$\{3\}$ component, which keeps W_1 and W_2 fixed while varying W_3. It never varies W_2 while holding only W_1 fixed. Equality on that restricted coupling therefore cannot establish functional or statistical independence from W_2. The unsupported passage from component-wise invariances to complements and intersections is precisely where the coordinate Boolean algebra enters the proof.

6 The root-intervention flattening theorem

The three-variable construction is an instance of a general symmetry. If all available targets are roots, a perfect root replacement can be performed in the observational exogenous coordinate system and then propagated through the original solution map. The resulting latent graph is edgeless.

Theorem 6.1 (Root-intervention flattening)**.** *Let a finite Markovian SCM have independent exogenous vector $E = (E_1, \ldots, E_n)$ with an absolutely continuous product law, DAG G, and observational solution map*

$$s : \mathcal{E}_1 \times \cdots \times \mathcal{E}_n \longrightarrow \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n.$$

Assume s and the observation map g are diffeomorphisms. Let R be the set of roots, and suppose every available target S lies in 2^R.

For each targeted root $i \in S$, suppose its replacement value can be written

$$\widetilde{Z}_i = f_i(U_i^S), \tag{21}$$

where f_i is the observational root mechanism, U_S^S is independent of E, and its law is absolutely continuous. Define

$$E_j^S = \begin{cases} U_j^S, & j \in S, \\ E_j, & j \notin S. \end{cases} \tag{22}$$

Then the original target-conditional observed pair is pathwise equal to that of an edgeless SCM $W_i = E_i$ with observation map

$$h := g \circ s \tag{23}$$

and perfect coordinate replacements $\widetilde{W}_i = U_i^S$ for $i \in S$.

Proof. For a targeted root, (21) makes the corresponding coordinate of $s(E^S)$ equal to the delivered intervention value. For every non-target vertex, the original mechanism receives the same non-target exogenous noise and the recursively updated parent values. Induction in a topological ordering of G therefore gives

$$\widetilde{Z}^S = s(E^S).$$

With $W = E$ and $\widetilde{W}^S = E^S$,

$$\begin{aligned} (g(Z), g(\widetilde{Z}^S)) &= (g(s(E)), g(s(E^S))) \\ &= (h(W), h(\widetilde{W}^S)) \end{aligned}$$

pathwise. Since s and g are diffeomorphisms, so is h. The flat observational mechanisms and intervention mechanisms may be taken to be identities, and the assumed absolute continuity of E and U_S^S places the flat model in the same smooth hypothesis class. □

Remark 6.2. If each root mechanism f_i is a diffeomorphism, then any smooth perfect root replacement $\widetilde{Z}_i = \widetilde{f}_i^{(S)}(\widetilde{E}_i)$ has the pullback

$$U_i^S = f_i^{-1}(\widetilde{f}_i^{(S)}(\widetilde{E}_i)).$$

Thus (21) is automatic whenever the resulting pullback law is absolutely continuous.

Corollary 6.3 (Structural non-identification under root-only coverage)**.** *Whenever a faithful SCM and its flattened realization from theorem 6.1 both lie in the declared hypothesis class, the target-conditional paired law cannot distinguish their causal graphs. Any graph functional that differs between the two models is not identifiable from that paired law without an additional restriction.*

The theorem explains the mechanism behind theorem 4.3. The triangular map (10) is exactly the solution map of the chain, and replacing a root exogenous coordinate before applying that map is exactly a perfect root intervention in the chain.

7 An infinite, intervention-minimal family

The ambiguity is not a low-dimensional accident. We now construct faithful linear-Gaussian counterexamples in every dimension, with a target family that is minimal for producing a singleton atom partition.

Lemma 7.1 (Atom-count bound)**.** *A Boolean algebra generated by k subsets of a finite set has at most 2^k atoms. Consequently, generating the singleton partition of an n-element set requires*

$$k \geq \lceil \log_2 n \rceil.$$

Proof. Each element is assigned its k-bit membership signature across the generators. Two elements lie in the same atom exactly when their signatures agree. There are at most 2^k signatures. □

Theorem 7.2 (Code-separated root-intervention counterexamples)**.** *For every $n \geq 2$, there exist two n-variable faithful linear-Gaussian SCMs and*

$$k = \lceil \log_2 n \rceil$$

positive-probability grouped perfect root-intervention regimes such that:

1. *the target-conditional observed pairs agree pathwise;*
2. *the ground-truth Boolean algebra has n singleton atoms, so the quotient prescribed by Theorem 3.1 is the n-vertex edgeless graph;*
3. *the rival graph has $n-1$ edges; and*
4. *no surjective strict or contraction-friendly graph homomorphism maps the rival graph to the prescribed quotient.*

The number k of regimes is minimal for conclusion 2.

Proof. Let $V = \{1, \ldots, n\}$ and set $k = \lceil \log_2 n \rceil$. Assign distinct codes $c_i \in \{0,1\}^k$ to the vertices, with $c_n = 0$, $c_i \neq 0$ for $i < n$, and with every standard basis vector e_t appearing among $c_1, \ldots, c_{n-1}$. This is possible because $k \leq n - 1 \leq 2^k - 1$. For $t = 1, \ldots, k$, define the target

$$S_t := \{i < n : (c_i)_t = 1\}. \tag{24}$$

The basis-code condition makes the S_t nonempty and pairwise distinct. Choose every S_t with positive probability.

Let $E_1, \ldots, E_n$ be independent standard Gaussians and fix nonzero coefficients $a_1, \ldots, a_{n-1}$. The ground model is edgeless with $W_i = E_i$ and mixing map

$$g_{\mathrm{F}}(w)_i = w_i \quad (i < n), \qquad g_{\mathrm{F}}(w)_n = w_n + \sum_{i=1}^{n-1} a_i w_i. \tag{25}$$

The rival has roots $1, \ldots, n-1$, child n, edges $i \to n$, identity mixing, and equations

$$Z_i = E_i \quad (i < n), \qquad Z_n = E_n + \sum_{i=1}^{n-1} a_i Z_i. \tag{26}$$

Under target S_t, replace each targeted root noise E_i by a fresh $U_i^t \sim \mathcal{N}(1,1)$, mutually independent within the regime and independent of E, and reuse all non-target noises. The flat model applies (25) after the same replacements; the rival recomputes its child by (26). Therefore the observed pairs agree pathwise in every regime by theorem 6.1.

The ground graph is edgeless, so

$$\mathrm{nd}_{G^{\mathrm{F}}}(S_t) = V \setminus S_t.$$

Vertex i has membership signature $\big(1 - (c_i)_t\big)_{t=1}^k$ across these non-descendant sets. The codes are distinct, hence all signatures are distinct and the atom partition is singleton. By lemma 7.1, k is minimal.

The rival Gaussian is faithful when all $a_i \neq 0$; a direct partial-covariance proof is given in section A. Its graph has $n-1$ edges. A surjection from its n vertices to the n singleton quotient vertices is a bijection, so no edge can be contracted. Every source edge would have to map to an edge of the edgeless target, which is impossible. □

Remark 7.3 (Arbitrarily weak causal coupling). The coefficients a_i in theorem 7.2 may be chosen arbitrarily small but nonzero. Thus the failure occurs in every neighborhood of the identity mixing map and is not caused by a large or pathological coordinate change.

Remark 7.4 (The three-variable collider). For $n = 3$, choose codes $10, 01, 00$. The two targets are $\{1\}$ and $\{2\}$, while the rival graph is the collider $1 \to 3 \leftarrow 2$. Both the ground and rival generated coordinate algebras have eight elements, yet the prescribed edgeless quotient is still incompatible with the rival. Hence the failure is not merely an eight-versus-four cardinality artifact of the chain example.

8 Boundaries and repair options

The theorem has two distinct failure modes, so its structural and statistical components require separate repairs.

8.1 Use an edge-contracting structural morphism

Replacing strict graph homomorphisms by (9), or by another explicitly defined contraction morphism, repairs the internal-edge inconsistency in proposition 3.1. It does not repair theorems 4.3 and 7.2, because their vertex maps are forced to be bijections. Saving those conclusions would require deleting edges between distinct macro-vertices, which no longer expresses edge preservation or contraction.

8.2 Add intervention coverage that breaks root flattening

The ambiguity in section 4 relies on the absence of an intervention on the residual coordinate W_2, which becomes the child residual in the chain. Add the flat atomic intervention

$$\widetilde{W} = (A, C_2, D), \qquad C_2 \perp\!\!\!\perp (A, B, D).$$

After the shear, the post-intervention rival coordinates would be

$$(A, A + C_2, D). \tag{27}$$

This cannot be a perfect intervention on the rival child: conditional on the unchanged parent A, its replacement value retains A, whereas a perfect child intervention must sever the parent edge and be generated by fresh noise independent of the pre-intervention exogenous vector. Any target containing root 1 is also impossible, because the first coordinate remains A pathwise under a fresh absolutely continuous reset. Finally, any target omitting child 2 reuses the original child noise B rather than the fresh C_2. Thus no rival perfect-intervention target reproduces (27) with the required same-unit coupling.

The established full-recovery theorem of Brehmer et al. [3] uses the empty regime and all atomic perfect interventions, together with stronger pointwise-diffeomorphic mechanism assumptions. That theorem provides a sufficient repair in its own setting. We do not claim that all atomic targets are necessary in every model class; the relevant requirement is enough intervention variation to rule out the latent gauge transformations under consideration.

8.3 Restrict the admissible latent reparameterizations

A direct structural repair is to require the equal-law transition map q to preserve the relevant coordinate lattice. Let D_A be as in (13). One may assume that there exists a Boolean-algebra isomorphism Φ satisfying

$$(q \times q)(D_A) = D_{\Phi(A)} \qquad \text{for every } A \in \mathcal{B}_{G^\star}(\mathcal{I}^\star). \tag{28}$$

A stronger and cleaner sufficient condition is block-separability after a permutation of the quotient atoms, with each block map q_B a diffeomorphism:

$$q(z)_{\pi(B)} = q_B(z_B) \qquad \text{for each atom } B. \tag{29}$$

Either condition supplies the missing lattice-equivariance step. Neither follows from the raw paired law in our examples; it is an additional restriction on the ambiguity class.

8.4 Weaken the identifiable object

Under unrestricted diffeomorphic mixing, equality of observed paired laws implies that latent paired measures are related by the diagonal action

$$\mu \longmapsto (q \times q)_{\#}\mu, \qquad q \in \mathrm{Diff}(\mathcal{Z}). \tag{30}$$

Therefore any universally identifiable latent object must be invariant under this action. Without additional structure, the full paired measure is determined at most up to its diagonal-diffeomorphism orbit. When intervention targets are observed, or when the mixture components are otherwise separately identified, the corresponding family of target-conditional measures and supports may likewise be considered only up to a simultaneous diagonal diffeomorphism. The list of coordinate indices on which equality holds is not invariant. A corrected theorem can return such an orbit, or an invariant derived from it, without claiming a privileged quotient of named endogenous coordinates.

9 Relation to prior work

The positive result most directly adjacent to this note is Brehmer et al. [3], which uses complete atomic intervention coverage under stronger invertibility assumptions. Sparse perturbations provide another route to weakly supervised representation identification under additional structural assumptions [1]. Other work obtains causal-representation identifiability from intervention diversity or nonparametric conditions that break generic latent symmetries [8].

The abstraction relation audited here combines distributional aggregation with a graph homomorphism inspired by categorical treatments of causal models [5]; related abstraction formalisms use different structural and interventional consistency requirements [2, 6]. Our result does not transfer automatically to those definitions. It shows that the particular quotient-identifiability implication in Li et al. [4] is invalid under its stated assumptions and that root-only paired interventions possess a general flattening symmetry that any corrected theorem must address.

10 Conclusion

Theorem 3.1 of Li et al. [4] is false as published. On a literal reading, its quotient construction can discard an internal edge that its strict abstraction morphism is required to preserve. After repairing

that definition, the statistical claim still fails: a smooth, faithful, linear-Gaussian edgeless model and a chain model produce exactly the same observed pair in each of two perfect root-intervention regimes, yet the prescribed quotient and the rival graph are incompatible.

The proof transports the paired law correctly through a diffeomorphism but then treats coordinate equality as though it were a coordinate-free relation. It is not. A shear preserves the full oblique paired constraint while changing which coordinate differences vanish, and the attempted complement/intersection extension fails both algebraically and through an explicit false conditional independence.

Root-intervention flattening identifies the underlying symmetry, and the code-separated family shows that the failure persists in every dimension with the minimum number of grouped regimes needed to generate a singleton partition. A valid positive theorem must therefore change the structural morphism, add symmetry-breaking interventions, restrict the latent transition maps, weaken the conclusion, or combine these moves.

A Faithfulness of the Gaussian star family

Let $R = (R_1, \dots, R_m) \sim \mathcal{N}(0, I_m)$, let $E \sim \mathcal{N}(0,1)$ be independent, and define

$$Y = \sum_{i=1}^{m} a_i R_i + E, \qquad a_i \neq 0.$$

The associated DAG has edges $R_i \to Y$ and no other edges.

Lemma A.1. *The joint Gaussian law of $(R_1, \dots, R_m, Y)$ is faithful to the star-collider DAG whenever all a_i are nonzero.*

Proof. For a nondegenerate Gaussian law, conditional independence of vectors is equivalent to vanishing of the corresponding conditional cross-covariance matrix. It is therefore enough to check every pair of unconditioned vertices against every conditioning set. Distinct roots R_i, R_j are independent conditional on any subset K of the remaining roots, since the roots are mutually independent; this is exactly the collider-blocked d-separation when Y is not conditioned upon.

If one also conditions on Y, then after fixing R_K the residual child is

$$Y_K = \sum_{\ell \notin K} a_\ell R_\ell + E.$$

The Gaussian conditional covariance is

$$\operatorname{Cov}(R_i, R_j \mid Y, R_K) = -\frac{a_i a_j}{1 + \sum_{\ell \notin K} a_\ell^2}, \tag{31}$$

which is nonzero. Conditioning on the collider therefore opens every root–root path, exactly as d-separation predicts.

Finally, for any $K \subseteq \{R_\ell : \ell \neq i\}$,

$$\operatorname{Cov}(R_i, Y \mid R_K) = a_i \neq 0,$$

so no direct parent–child edge is accidentally cancelled. Hence the Gaussian conditional independences coincide exactly with the DAG's d-separations. □

B A minimal substantive shear counterexample

The two-node self-counterexample in proposition 3.1 uses the internal-edge incompatibility. A separate two-node construction also exhibits the coordinate-shear ambiguity.

Let $A, B \sim \mathcal{N}(0,1)$ and $C \sim \mathcal{N}(1,1)$ be independent. The flat ground truth is

$$W = (A, B), \qquad g_{\mathrm{F}}(w_1, w_2) = (w_1, w_1 + w_2),$$

with target $\{1\}$ and post-state (C, B). The rival is the chain

$$Z_1 = A, \qquad Z_2 = Z_1 + B,$$

with identity mixing and post-state $(C, C + B)$. Both produce

$$\big((A, A + B), (C, C + B)\big)$$

pathwise. In the flat graph, $\mathrm{nd}(\{1\}) = \{2\}$, so the atom partition is singleton and the prescribed quotient is the two-node edgeless graph. A surjection from the two-node chain to that quotient is a bijection and cannot preserve its edge. This example survives the edge-contraction repair for the same reason as theorem 4.3.

References

[1] Kartik Ahuja, Jason S. Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. In *Advances in Neural Information Processing Systems*, volume 35, pages 15516–15528, 2022.

[2] Sander Beckers and Joseph Y. Halpern. Abstracting causal models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1):2678–2685, 2019. doi: 10.1609/aaai.v33i01.33012678.

[3] Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco S. Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 38319–38331, 2022.

[4] Xiusi Li, Sékou-Oumar Kaba, and Siamak Ravanbakhsh. On the identifiability of causal abstractions. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3241–3249. PMLR, 2025.

[5] Jun Otsuka and Hayato Saigo. On the equivalence of causal models: A category-theoretic approach. In James Cussens and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 634–646. PMLR, 2022.

[6] Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In Gal Elidan, Kristian Kersting, and Alexander T. Ihler, editors, *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2017.

[7] 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.

[8] Julius von Kügelgen, Michel Besserve, Wendong Liang, Luigi Gresele, Armin Kekić, Elias Bareinboim, David M. Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems*, volume 36, pages 48603–48638, 2023.