# Appendices

## Table of Contents

# A Extended Discussion of Related Work

Prior work on causal representation learning with general nonlinear relationships (both among latents and between latents and observations) and without an explicit task or label typically relies on some form of *weak supervision*. One example of weak supervision is "multi-view" data consisting of tuples of related observations. von Kügelgen et al. [123] consider counterfactual pairs of observations arising from imperfect interventions through data augmentation, and prove identifiability for the invariant non-descendants of intervened-upon variables. Brehmer et al. [18] also use counterfactual pre- and post-intervention views and show that the latent SCM can be identified given all single-node perfect stochastic interventions. Another type of weak-supervision is temporal structure [2], possibly combined with nonstationarity [132, 133], interventions on known targets [80, 81], or observed actions inducing sparse mechanism shifts [74, 75, 110]. Other works use more explicit supervision in the form of annotations of the ground truth causal variables or a known causal graph [111, 126, 130].

A different line of work instead approaches causal representation learning from the perspective of causal discovery in the presence of latent variables [115]. Doing so from *purely observational i.i.d. data* requires additional constraints on the generative process, such as restrictions on the graph structure or particular parametric and distributional assumptions, and typically leverages the identifiability of linear ICA [27, 38, 79]. For *linear, non-Gaussian* models, Silva et al. [113] show that the causal DAG can be recovered up to Markov equivalence if all observed variables are "pure" in that they have a unique latent causal parent. Cai et al. [23] and Xie et al. [128, 129] extend this result to identify the full graph given two pure observed children per latent, and Adams et al. [1] provide sufficient and necessary graphical conditions for full identification. For *discrete* causal variables, Kivva et al. [68] introduce a similar "no twins" condition to reduce the task of learning the number and cardinality of latents and the Markov-equivalence class of the graph to mixture identifiability.

Other lines of work have investigated the relationship between causal models at different levels of coarse-graining or abstraction [6, 11, 12, 24–26, 35, 105, 134, 135], or learning invariant representations in a supervised setting [3, 8, 32, 33, 73, 86, 92, 124, 125], often for domain generalization.

Other concurrent works address, e.g., learning from soft interventions with polynomial mixing [136], or inferring both causal graph and the number of latents subject to graphical constraints [63].

## A.1 Comparison of Related Identifiability Results

To complement the presentation of related multienvironment CRL works in § 1, we provide a structured overview of and comparison with existing identifiability results for causal representation learning in Tab. 1. The table categorizes work along different dimensions. First, we make a distinction based on the type of data (observational, interventional, or counterfactual) results rely on (colour coded in green, yellow, and red, respectively). These are also referred to as different rungs in the "ladder of causation" [97] or layers in the Pearl Causal Hierarchy [10]. Second, we categorize work depending on the assumptions placed on the latent causal model and the mixing function. As can be seen, works relying solely on observational data often require restrictive graphical assumptions on the mixing function. On the other hand, access to much more informative counterfactual data has allowed identification even for nonparametric causal models and mixing functions. Our work can be viewed as a step towards addressing the lack of nonparametric identifiability results in the interventional regime.

We emphasize that Tab. 1 is not exhaustive: certain relevant works did not easily fit into our categorization or the causal representation learning framework adopted in the present work. For example, not listed are works that leverage temporal structure [2, 74, 75, 80, 81], rely on heterogenous data and distribution shifts which are not directly expressed in terms of or linked to interventions [82, 132, 133], allow for edges from observed to latent variables [1], or require more direct supervision [111, 130].

Table 1: **Comparison of Existing Identifiability Results for Causal Reresentation Learning.** All of the listed works assume invertibility (or injectivity) of the mixing function, as well as causal sufficiency (Markovianity) for the causal latent variables. Most or all of the listed results require additional technical assumptions, and may provide additional results, which we omit for sake of readability; see the references for more details.

| **Work** | **Layer** | **Causal Model** | **Mixing Function** | **Main Identifiability Result** |
|---|---|---|---|---|
| Cai et al. [23], Xie et al. [128, 129] | observational | linear, non-Gaussian | linear with non-Gaussian noise s.t. each $V_i$ has 2 pure (obs. or unobs.) children | number of latents + $G$ |
| Kivva et al. [68] | observational | discrete, nonparametric | indentifiable mixture model s.t. obs. children of $V_i \not\subseteq$ obs. children of $V_j$ | number, cardinality & dist. of discrete latents + $G$ up to Markov equivalence |
| Ahuja et al. [5, Thm. 4] | observational | nonlinear w. independent support [103, 125] | finite-degree polynomial | $\boldsymbol{V}$ up to permutation, shift, & linear scaling |
| Squires et al. [117, Thms. 1 & 2] | interventional | linear | linear | $G$ and $\boldsymbol{V}$ up to partial-order preserving permutations from obs. dist. & all single-node *perfect* interventions |
| Squires et al. [117, Thm. 1] | interventional | linear | linear | $G$ up to transitive closure from obs. dist. & all single-node *imperfect* interventions |
| Varici et al. [122, Thm. 16] | interventional | nonparametric | linear | $G$ and $\boldsymbol{V}$ up to partial-order preserving permutations from obs. dist. & all single-node *perfect* interventions |
| Ahuja et al. [5, Thm. 2] | interventional | nonparametric | finite-degree polynomial | $\boldsymbol{V}$ up to permutation, shift, and linear scaling from all single-node *perfect hard* interventions |
| Buchholz et al. [22] | interventional | linear Gaussian | nonparametric | $G$ and $\boldsymbol{V}$ up to permutation from obs. dist. & all single-node *perfect* interventions |
| **This Work** (Thm. 3.2) | interventional | nonparametric | nonparametric | for $n = 2$: $G$ and $\boldsymbol{V}$ up to $\sim_{\text{CRL}}$ from all single-node *perfect* interventions, subject to genericity (3.2) |
| **This Work** (Thm. 3.4) | interventional | nonparametric | nonparametric | $G$ and $\boldsymbol{V}$ up to $\sim_{\text{CRL}}$ from two distinct, paired single-node *perfect* interventions per node |
| von Kügelgen et al. [123] | counterfactual | nonparametric | nonparametric | block of non-descendants $\boldsymbol{V}_{\text{nd}(\mathcal{I})}$ up to invertible function from fat-hand *imperfect* interventions on $\boldsymbol{V}_{\mathcal{I}}$ |
| Brehmer et al. [18] | counterfactual | nonparametric | nonparametric | $G$ and $\boldsymbol{V}$ up to $\sim_{\text{CRL}}$ from all single-node *perfect* interventions |

## B Proof of Minimality of the CRL Equivalence Class $\sim_{\text{CRL}}$

First, let us recall the main statements from § 2.2.

**Definition 2.6** ($\sim_{\text{CRL}}$-identifiability)**.** Let $\mathcal{H}$ be a space of unmixing functions $h : \mathcal{X} \to \mathbb{R}^n$ and let $\mathcal{G}$ be the space of DAGs over $n$ vertices. Let $\sim_{\text{CRL}}$ be the equivalence relation on $\mathcal{H} \times \mathcal{G}$ defined as

$$(h_1, G_1) \sim_{\text{CRL}} (h_2, G_2) \quad \iff \quad (h_2, G_2) = (\boldsymbol{P}_{\pi^{-1}} \circ \phi \circ h_1, \pi(G_1)) \tag{2.3}$$

for some element-wise diffeomorphism $\phi(\boldsymbol{v}) = (\phi_1(v_1), \dots, \phi_n(v_n))$ of $\mathbb{R}^n$ and a permutation $\pi$ of $[n]$ such that $\pi : G_1 \mapsto G_2$ is a graph isomorphism and $\boldsymbol{P}_\pi$ the corresponding permutation matrix.

**Proposition 3.1** (Minimality of $\sim_{\text{CRL}}$; informal)**.** *Let $\boldsymbol{Z}$ be any representation that is $\sim_{\text{CRL}}$ equivalent to $\boldsymbol{V}$, with $G' = \pi(G)$ the associated DAG. Then for any intervention on $\boldsymbol{V}_\mathcal{I} \subseteq \boldsymbol{V}$ in $G$, there exists an equally sparse intervention on $\boldsymbol{Z}_{\pi(\mathcal{I})} \subseteq \boldsymbol{Z}$ in $G'$ inducing the same observed distribution on $\boldsymbol{X}$.*

We now restate this result more formally.

**Proposition B.1** (Minimality of $\sim_{\text{CRL}}$)**.** *Let $(h, G') \sim_{\text{CRL}} (f^{-1}, G)$ with $\pi$ denoting the graph isomorphism mapping $G$ to $G'$ (i.e., a permutation that preserves the partial topological order of $G$). Let $\boldsymbol{Z} = h(\boldsymbol{X})$ be the inferred representation with distribution $Q_{\boldsymbol{Z}} = h_*(P_{\boldsymbol{X}})$ Markov w.r.t. $G'$ and associated density $q$. Let $\mathcal{I}^e \subseteq [n]$ denote a set of intervention targets, and consider an intervention that changes $p_i(v_i \mid \boldsymbol{v}_{\text{pa}(i)})$ to some intervened mechanism $\tilde{p}_i(v_i \mid \boldsymbol{v}_{\text{pa}(i)})$ for all $i \in \mathcal{I}^e$, giving rise to the interventional distributions $P^e_{\boldsymbol{V}}$ and $P^e_{\boldsymbol{X}} = f_*(P^e_{\boldsymbol{V}})$. Then there exist appropriately chosen $\tilde{q}_{\pi(i)}(z_{\pi(i)} \mid \boldsymbol{z}_{\text{pa}(\pi(i),G')})$ for $i \in \mathcal{I}^e$ such that the resulting interventional distribution $Q^e_{\boldsymbol{Z}}$ gives rise to the same observed distributions, that is, $P^e_{\boldsymbol{X}} = h_*^{-1}(Q^e_{\boldsymbol{Z}})$.*

*Proof.* Since $(h, G') \sim_{\text{CRL}} (f^{-1}, G)$, by Defn. 2.6 we have

$$\boldsymbol{Z} = \boldsymbol{P}_{\pi^{-1}} \circ \phi(\boldsymbol{V}) \tag{B.1}$$

for some element-wise diffeomorphism $\phi$ with inverse $\psi = \phi^{-1}$. Then (B.1) implies that for all $i \in [n]$

$$V_i = \psi_i(Z_{\pi(i)}) \tag{B.2}$$

According to (B.2), each conditional in the Markov factorization of $Q_{\boldsymbol{Z}}$ is given in terms of $p$ by

$$q_{\pi(i)}\left(z_{\pi(i)} \mid \boldsymbol{z}_{\text{pa}(\pi(i);G')}\right) = p_i\left(\psi_i\left(z_{\pi(i)}\right) \mid \psi_{\text{pa}(i)}\left(\boldsymbol{z}_{\text{pa}(\pi(i);G')}\right)\right)\left|\frac{\mathrm{d}\psi_i}{\mathrm{d}z_{\pi(i)}}\left(z_{\pi(i)}\right)\right| \tag{B.3}$$

where we have used the change of variables in (B.2) and the fact that $\pi(\text{pa}(i)) = \text{pa}(\pi(i); G')$ since $\pi : G \mapsto G'$ is a graph isomorphism.

Consider an intervention that changes $p_i(v_i \mid \boldsymbol{v}_{\text{pa}(i)})$ to some intervened mechanism $\tilde{p}_i(v_i \mid \boldsymbol{v}_{\text{pa}(i)})$ for all $i \in \mathcal{I}^e$. Denote the corresponding intervened joint distribution by $P^e_{\boldsymbol{V}}$ with joint density $p^e$ given by

$$p^e(\boldsymbol{v}) = \prod_{i \in \mathcal{I}^e} \tilde{p}_i\left(v_i \mid \boldsymbol{v}_{\text{pa}(i)}\right) \prod_{j \in [n] \setminus \mathcal{I}^e} p_j\left(v_j \mid \boldsymbol{v}_{\text{pa}(j)}\right). \tag{B.4}$$

Denote by $Q^e_{\boldsymbol{Z}} = (\boldsymbol{P}_{\pi^{-1}} \circ \phi)_*(P^e_{\boldsymbol{V}})$ the corresponding distribution over $\boldsymbol{Z}$ with joint density given by $q^e$

$$q^e(\boldsymbol{z}) = p^e(\psi \circ \boldsymbol{P}_\pi(\boldsymbol{z}))\left|\det \boldsymbol{J}_{\psi \circ \boldsymbol{P}_\pi}(\boldsymbol{z})\right| \tag{B.5}$$

$$= \prod_{i \in \mathcal{I}^e} \tilde{p}_i\left(\psi_i\left(z_{\pi(i)}\right) \mid \psi_{\text{pa}(i)}\left(\boldsymbol{z}_{\text{pa}(\pi(i);G')}\right)\right)\left|\frac{\mathrm{d}\psi_i}{\mathrm{d}z_{\pi(i)}}\left(z_{\pi(i)}\right)\right| \prod_{j \in [n] \setminus \mathcal{I}^e} q_{\pi(j)}\left(z_{\pi(j)} \mid \boldsymbol{z}_{\text{pa}(\pi(j);G')}\right), \tag{B.6}$$

where we have used (B.3), (B.4), and the fact that $\boldsymbol{J}_\psi$ is diagonal.

By defining

$$\tilde{q}_{\pi(i)}\left(z_{\pi(i)} \mid \boldsymbol{z}_{\text{pa}(\pi(i);G')}\right) := \tilde{p}_i\left(\psi_i\left(z_{\pi(i)}\right) \mid \psi_{\text{pa}(i)}\left(\boldsymbol{z}_{\text{pa}(\pi(i);G')}\right)\right)\left|\frac{\mathrm{d}\psi_i}{\mathrm{d}z_{\pi(i)}}\left(z_{\pi(i)}\right)\right| \tag{B.7}$$

we finally arrive at

$$q^e(\boldsymbol{z}) = \prod_{i \in \mathcal{I}^e} \tilde{q}_{\pi(i)} \left( z_{\pi(i)} \mid \boldsymbol{z}_{\mathrm{pa}(\pi(i);G')} \right) \prod_{j \in [n] \setminus \mathcal{I}^e} q_{\pi(j)} \left( z_{\pi(j)} \mid \boldsymbol{z}_{\mathrm{pa}(\pi(j);G')} \right) . \tag{B.8}$$

This shows that any intervention on $\{V_i\}_{i \in \mathcal{I}^e} \subseteq \boldsymbol{V}$ which replaces

$$\left\{ p_i(v_i \mid \boldsymbol{v}_{\mathrm{pa}(i)}) \right\}_{i \in \mathcal{I}^e} \mapsto \left\{ \tilde{p}_i(v_i \mid \boldsymbol{v}_{\mathrm{pa}(i)}) \right\}_{i \in \mathcal{I}^e} , \tag{B.9}$$

can equivalently be captured by an intervention on $\{Z_{\pi(i)}\}_{i \in \mathcal{I}^e} \subseteq \boldsymbol{Z}$ which replaces

$$\left\{ q_{\pi(i)} \left( z_{\pi(i)} \mid \boldsymbol{z}_{\mathrm{pa}(\pi(i);G')} \right) \right\}_{i \in \mathcal{I}^e} \mapsto \left\{ \tilde{q}_{\pi(i)} \left( z_{\pi(i)} \mid \boldsymbol{z}_{\mathrm{pa}(\pi(i);G')} \right) \right\}_{i \in \mathcal{I}^e} . \tag{B.10}$$

with $\tilde{q}_i$ defined according to (B.7). □

# C Identifiability Proofs

## C.1 Auxiliary Lemmata

**Lemma C.1** (Lemma 2 of Brehmer et al. [18])**.** *Let $A = C = \mathbb{R}$ and $B = \mathbb{R}^n$. Let $f : A \times B \to C$ be differentiable. Define two differentiable measures $p_A$ on $A$ and $p_C$ on $C$. Let $\forall b \in B$, $f(\cdot, b) : A \to C$ be measure-preserving, in the sense that the pushforward of $p_A$ is always $p_C$. Then $f(\cdot, b)$ is constant in $b$ on $B$.*

*Proof.* See Appendix A.2 of Brehmer et al. [18]. □

**Lemma C.2** (Preservation of conditional independence under invertible transformation.)**.** *Let $X$ and $Y$ be continuous real-valued random variables, and let $\boldsymbol{Z}$ be a continuous random vector taking values in $\mathbb{R}^n$. Suppose that $(X, Y, \boldsymbol{Z})$ have a joint density w.r.t. the Lebesgue measure. Let $f : \mathbb{R} \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$, and $h : \mathbb{R}^n \to \mathbb{R}^n$ be diffeomorphisms. Then $X \perp\!\!\!\perp Y \mid \boldsymbol{Z} \implies f(X) \perp\!\!\!\perp g(Y) \mid h(\boldsymbol{Z})$.*

*Proof.* Denote by $p(x, y, \boldsymbol{z})$ the density of $(X, Y, \boldsymbol{Z})$. Then $X \perp\!\!\!\perp Y \mid \boldsymbol{Z}$ implies that for all $(x, y, \boldsymbol{z})$, $p$ can be factorized as follows:

$$p(x, y, \boldsymbol{z}) = p_{\boldsymbol{z}}(\boldsymbol{z}) p_x(x \mid \boldsymbol{z}) p_y(y \mid \boldsymbol{z}) . \tag{C.1}$$

Let $(A, B, \boldsymbol{C}) = (f(X), g(Y), h(\boldsymbol{Z}))$, and write $\tilde{f} = f^{-1}$, $\tilde{g} = g^{-1}$, and $\tilde{h} = h^{-1}$.

Then $(A, B, \boldsymbol{C})$ also has a density $q(a, b, \boldsymbol{c})$, which for all $(a, b, \boldsymbol{c})$ is given by the change of variable formula:

$$q(a, b, \boldsymbol{c}) = p \left( \tilde{f}(a), \tilde{g}(b), \tilde{h}(\boldsymbol{c}) \right) \left| \frac{\mathrm{d}\tilde{f}}{\mathrm{d}a}(a) \frac{\mathrm{d}\tilde{g}}{\mathrm{d}b}(b) \det \boldsymbol{J}_{\tilde{h}}(\boldsymbol{c}) \right| \tag{C.2}$$

$$= p_{\boldsymbol{z}} \left( \tilde{h}(\boldsymbol{c}) \right) \left| \det \boldsymbol{J}_{\tilde{h}}(\boldsymbol{c}) \right| \; p_x \left( \tilde{f}(a) \mid \tilde{h}(\boldsymbol{c}) \right) \left| \frac{\mathrm{d}\tilde{f}}{\mathrm{d}a}(a) \right| \; p_y \left( \tilde{g}(b) \mid \tilde{h}(\boldsymbol{c}) \right) \left| \frac{\mathrm{d}\tilde{g}}{\mathrm{d}b}(b) \right| \tag{C.3}$$

where in (C.2) we have used the fact that $(X, Y, \boldsymbol{Z}) \mapsto (f(X), g(Y), h(\boldsymbol{Z}))$ has block-diagonal Jacobian, and in (C.3) that $p$ factorises as in (C.1). Next, define

$$q_{\boldsymbol{c}}(\boldsymbol{c}) := p_{\boldsymbol{z}} \left( \tilde{h}(\boldsymbol{c}) \right) \left| \det \boldsymbol{J}_{\tilde{h}}(\boldsymbol{c}) \right| , \tag{C.4}$$

$$q_a(a \mid \boldsymbol{c}) := p_x \left( \tilde{f}(a) \mid \tilde{h}(\boldsymbol{c}) \right) \left| \frac{\mathrm{d}\tilde{f}}{\mathrm{d}a}(a) \right| , \tag{C.5}$$

$$q_b(b \mid \boldsymbol{c}) := p_y \left( \tilde{g}(b) \mid \tilde{h}(\boldsymbol{c}) \right) \left| \frac{\mathrm{d}\tilde{g}}{\mathrm{d}b}(b) \right| . \tag{C.6}$$

Since $p_{\boldsymbol{z}}$, $p_x$, and $p_y$ are valid densities (non-negative and integrating to one), so are $q_{\boldsymbol{c}}$, $q_a$, and $q_b$. Substitution into (C.3) yields for all $(a, b, \boldsymbol{c})$,

$$q(a, b, \boldsymbol{c}) = q_{\boldsymbol{c}}(\boldsymbol{c}) q_a(a \mid \boldsymbol{c}) q_b(b \mid \boldsymbol{c}) , \tag{C.7}$$

which implies that $A \perp\!\!\!\perp B \mid \boldsymbol{C}$, concluding the proof. □

### C.2 Proof of Thm. 3.2

**Theorem 3.2** (Bivariate identifiability up to $\sim_{\text{CRL}}$ from one perfect stochastic intervention per node)**.** *Suppose that we have access to multiple environments $\{P^e_{\boldsymbol{X}}\}_{e\in\mathcal{E}}$ generated as described in § 2 under Asms. 2.2, 2.5, 2.8 and 2.9 with $n = 2$. Let $(h, G')$ be any candidate solution such that the inferred latent distributions $Q^e_{\boldsymbol{Z}} = h_*(P^e_{\boldsymbol{X}})$ of $\boldsymbol{Z} = h(\boldsymbol{X})$ and the inferred mixing function $h^{-1}$ satisfy the above assumptions w.r.t. the candidate causal graph $G'$. Assume additionally that*

*(A1) all densities $p^e$ and $q^e$ are continuously differentiable and fully supported on $\mathbb{R}^n$;*

*(A2) we have access to a* known observational environment $e_0$ *and one* single node perfect intervention for each node*, with* unknown targets*: there exist $n+1$ environments $\mathcal{E} = \{e_i\}_{i=0}^n$ such that $\mathcal{I}^{e_0} = \varnothing$ and for each $i \in [n]$ we have $\mathcal{I}^{e_i} = \{\pi(i)\}$ for an unknown permutation $\pi$ of $[n]$;*

*(A3) for all $i \in [n]$, the intervened mechanisms $\tilde{p}_i(v_i)$ differ from the corresponding base mechanisms $p_i(v_i \mid \boldsymbol{v}_{\text{pa}(i)})$ everywhere, in the sense that*

$$\forall \boldsymbol{v}: \qquad \frac{\partial}{\partial v_i} \frac{\tilde{p}_i(v_i)}{p_i(v_i \mid \boldsymbol{v}_{\text{pa}(i)})} \neq 0\,; \tag{3.1}$$

*(A4) **("genericity")** the base and intervened mechanisms are not fine-tuned to each other, in the sense that there exists a continuous function $\varphi : \mathbb{R}^+ \to \mathbb{R}$ for which*

$$\mathbb{E}_{\boldsymbol{v}\sim P^{e_0}_{\boldsymbol{V}}}\left[\varphi\left(\frac{\tilde{p}_2(v_2)}{p_2(v_2 \mid v_1)}\right)\right] \neq \mathbb{E}_{\boldsymbol{v}\sim P^{e_1}_{\boldsymbol{V}}}\left[\varphi\left(\frac{\tilde{p}_2(v_2)}{p_2(v_2 \mid v_1)}\right)\right] \tag{3.2}$$

*Then the ground truth is identified in the sense of Defn. 2.6, that is, $(f^{-1}, G) \sim_{\text{CRL}} (h, G')$.*

*Proof.* From the assumption of a shared mixing $f$ and shared encoder $h$ across all environments, we have that

$$\boldsymbol{Z} = h(\boldsymbol{X}) = h(f(\boldsymbol{V})) = h \circ f(\boldsymbol{V})\,. \tag{C.8}$$

Let $\psi = f^{-1} \circ h^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$\boldsymbol{V} = \psi(\boldsymbol{Z})\,.$$

By Asm. 2.5, $f$, $h$, and thus also $h \circ f$ are diffeomorphisms. Hence, $\psi$ is well-defined and also diffeomorphic.

By the change of variable formula, for all $e$ and all $\boldsymbol{z}$ the density $q^e(\boldsymbol{z})$ is given by

$$q^e(\boldsymbol{z}) = p^e(\psi(\boldsymbol{z}))\left|\det \boldsymbol{J}_\psi(\boldsymbol{z})\right| \tag{C.9}$$

where $(\boldsymbol{J}_\psi(\boldsymbol{z}))_{ij} = \frac{\partial \psi_i}{\partial z_j}(\boldsymbol{z})$ denotes the Jacobian of $\psi$.

We now consider two separate cases, depending on whether the intervention targets in $q^{e_i}$ for $e_i \in \{e_1, e_2\}$ match those in $p^{e_i}$ (Case 1) or not (Case 2).

**Case 1: Aligned Intervention Targets.** According to Asm. 2.8 and (A2), (C.9) applied to the known observational environment $e_0$ and the interventional environments $e_1, e_2$ leads to the system of equations:

$$q_1(z_1)q_2(z_2 \mid z_{\text{pa}(2;G')}) = p_1\left(\psi_1(\boldsymbol{z})\right) p_2\left(\psi_2(\boldsymbol{z}) \mid \psi_{\text{pa}(2)}(\boldsymbol{z})\right)\left|\det \boldsymbol{J}_\psi(\boldsymbol{z})\right| \tag{C.10}$$

$$\tilde{q}_1(z_1)q_2(z_2 \mid z_{\text{pa}(2;G')}) = \tilde{p}_1\left(\psi_1(\boldsymbol{z})\right) p_2\left(\psi_2(\boldsymbol{z}) \mid \psi_{\text{pa}(2)}(\boldsymbol{z})\right)\left|\det \boldsymbol{J}_\psi(\boldsymbol{z})\right| \tag{C.11}$$

$$q_1(z_1)\tilde{q}_2(z_2) = p_1\left(\psi_1(\boldsymbol{z})\right) \tilde{p}_2\left(\psi_2(\boldsymbol{z})\right)\left|\det \boldsymbol{J}_\psi(\boldsymbol{z})\right| \tag{C.12}$$

where $z_{\text{pa}(2;G')}$ denotes the parents of $z_2$ in $G'$, and $\text{pa}(2)$ denotes the parents of $V_2$ in $G$.

Note that neither side of the previous equations can be zero due to the full support assumption[12] (A1) and $\psi$ being diffeomorphic implying the determinant is non-zero.

[12] This can also be relaxed to fully supported on a Cartesian product of intervals.

Taking quotients of (C.11) and (C.10), yields

$$\frac{\tilde{q}_1}{q_1}(z_1) = \frac{\tilde{p}_1}{p_1}(\psi_1(\boldsymbol{z}))\,. \tag{C.13}$$

Next, we take the partial derivative w.r.t. $z_2$ on both sides and use the chain rule to obtain:

$$0 = \left(\frac{\tilde{p}_1}{p_1}\right)'(\psi_1(\boldsymbol{z}))\,\frac{\partial \psi_1}{\partial z_2}(\boldsymbol{z})\,. \tag{C.14}$$

Now, by assumption (A3), the first term on the RHS of (C.14) is non-zero everywhere. Hence,

$$\forall \boldsymbol{z}: \qquad \frac{\partial \psi_1}{\partial z_2}(\boldsymbol{z}) = 0\,. \tag{C.15}$$

It follows that $\psi_1$ is, in fact, a scalar function, and

$$V_1 = \psi_1(Z_1)\,. \tag{C.16}$$

Since $\psi$ is a diffeomorphism, $\psi_1$ must also be diffeomorphic. Hence, by the change of variable formula applied to (C.16), the marginal density $q_1(z_1)$ is given by

$$q_1(z_1) = p_1(\psi_1(z_1))\left|\frac{\partial \psi_1}{\partial z_1}(z_1)\right|\,. \tag{C.17}$$

Further, since $\boldsymbol{J}_\psi$ is triangular due to (C.15), its determinant is given by

$$\left|\det \boldsymbol{J}_\psi(\boldsymbol{z})\right| = \left|\frac{\partial \psi_1}{\partial z_1}(z_1)\,\frac{\partial \psi_2}{\partial z_2}(z_1, z_2)\right|\,. \tag{C.18}$$

Substituting (C.17) and (C.18) into (C.12) yields (after cancellation of equal terms):

$$\tilde{q}_2(z_2) = \tilde{p}_2\left(\psi_2(z_1, z_2)\right)\left|\frac{\partial \psi_2}{\partial z_2}(z_1, z_2)\right|\,. \tag{C.19}$$

The expression in (C.19) implies that, for all $z_1$, the mapping $\psi_2(z_1, \cdot) : \mathbb{R} \to \mathbb{R}$ is measure preserving for the differentiable $\tilde{q}_2$ and $\tilde{p}_2$. By Lemma C.1 (Lemma 2 of Brehmer et al. [18, § A.2]), it then follows that $\psi_2$ must, in fact, be constant in $z_1$, that is

$$\forall \boldsymbol{z}: \qquad \frac{\partial \psi_2}{\partial z_1}(\boldsymbol{z}) = 0\,. \tag{C.20}$$

Note that this last step is where the assumption of perfect interventions (Asm. 2.9) is leveraged: the conclusion would not hold for arbitrary imperfect interventions for which (3.8) would involve $\tilde{q}_2(z_2 \mid z_1)$ and $p_2\left(\psi_2(z_1, z_2) \mid \psi_1(z_1)\right)$.

Hence, we have shown that $\psi$ is an element-wise function:

$$\boldsymbol{V} = (V_1, V_2) = \psi(\boldsymbol{Z}) = (\psi_1(Z_1), \psi_2(Z_2))\,. \tag{C.21}$$

Finally, since $\psi$ is a diffeomorphism, (C.21) implies that

$$V_1 \perp\!\!\!\perp V_2 \iff Z_1 \perp\!\!\!\perp Z_2\,. \tag{C.22}$$

It then follows from the faithfulness assumption (Asm. 2.2) that we also must have $G = G'$.

This concludes the proof of Case 1, as we have shown that $(h, G') \sim_{\text{CRL}} (f^{-1}, G)$ with $G' = \pi(G) = G$ ($\pi$ being the identity permutation) and $h \circ f = \psi^{-1} =: \phi$ an element-wise diffeomorphism.

**Case 2: Misaligned Intervention Targets.** Writing down the system of equations similar to (C.10)–(C.12), but for the case with misaligned intervention targets across $p$ and $q$ yields:

$$q_1(z_1)q_2(z_2 \mid z_{\text{pa}(2;G')}) = p_1\left(\psi_1(\boldsymbol{z})\right)p_2\left(\psi_2(\boldsymbol{z}) \mid \psi_{\text{pa}(2)}(\boldsymbol{z})\right)\left|\det \boldsymbol{J}_\psi(\boldsymbol{z})\right| \tag{C.23}$$

$$\tilde{q}_1(z_1)q_2(z_2 \mid z_{\text{pa}(2;G')}) = p_1\left(\psi_1(\boldsymbol{z})\right)\tilde{p}_2\left(\psi_2(\boldsymbol{z})\right)\left|\det \boldsymbol{J}_\psi(\boldsymbol{z})\right| \tag{C.24}$$

$$q_1(z_1)\tilde{q}_2(z_2) = \tilde{p}_1\left(\psi_1(\boldsymbol{z})\right)p_2\left(\psi_2(\boldsymbol{z}) \mid \psi_{\text{pa}(2)}(\boldsymbol{z})\right)\left|\det \boldsymbol{J}_\psi(\boldsymbol{z})\right|\,. \tag{C.25}$$

Taking ratios of $e_1$ and $e_2$ with $e_0$ yields

$$\frac{\tilde{q}_1}{q_1}(z_1) = \frac{\tilde{p}_2\left(\psi_2(\boldsymbol{z})\right)}{p_2\left(\psi_2(\boldsymbol{z}) \mid \psi_{\mathrm{pa}(2)}(\boldsymbol{z})\right)} \tag{C.26}$$

$$\frac{\tilde{q}_2(z_2)}{q_2(z_2 \mid z_{\mathrm{pa}(2;G')})} = \frac{\tilde{p}_1}{p_1}\left(\psi_1(\boldsymbol{z})\right) . \tag{C.27}$$

We separate the remainder of the proof of Case 2 into different subcases depending on $G$ and $G'$: as we will see, we can use a similar reasoning as in Case 1, except for the case where both $G$ and $G'$ are missing no edge.

*Case 2a: $V_1 \not\to V_2$ in $G$, that is* $\mathrm{pa}(2) = \varnothing$. Then we can proceed similarly to Case 1. First, we take the partial derivative of (C.26) w.r.t. $z_2$ to arrive at:

$$0 = \left(\frac{\tilde{p}_2}{p_2}\right)'\left(\psi_2(\boldsymbol{z})\right)\frac{\partial \psi_2}{\partial z_2}(\boldsymbol{z}) . \tag{C.28}$$

Using (A3), this implies that $\psi_2$ does not depend on $Z_2$, that is, $V_2 = \psi_2(Z_1)$.

Next, we again write $q(z_1)$ in terms of $p_2(\psi_2(z_1))$ using the univariate change of variable formula, substitute into (C.25), cancel the corresponding terms, and arrive at:

$$\tilde{q}_2(z_2) = \tilde{p}_1\left(\psi_1(z_1, z_2)\right)\left|\frac{\partial \psi_1}{\partial z_2}(z_1, z_2)\right| \tag{C.29}$$

Lemma C.1 applied to $\psi_1(z_1, \cdot)$ which preserves $\tilde{q}_2$ and $\tilde{p}_1$ for all $z_1$ shows that $\psi_1$ is constant in $Z_1$, that is

$$\boldsymbol{V} = (V_1, V_2) = \psi(\boldsymbol{Z}) = (\psi_1(Z_2), \psi_2(Z_1)) . \tag{C.30}$$

Since $V_1 \perp\!\!\!\perp V_2$ by the assumption of Case 2a, it follows from the invertible element-wise reparametrisation above that $Z_1 \perp\!\!\!\perp Z_2$ and hence, by faithfulness, $Z_1 \not\to Z_2$ in $G'$.

Finally, note that there is no partial order on the empty graph and so $G' = \pi(G) = G$ and $\boldsymbol{Z} = \boldsymbol{P}_{\pi^{-1}} \cdot \psi^{-1}(\boldsymbol{V})$ where $\pi$ is the nontrivial permutation of $\{1, 2\}$.

*Case 2b: $V_1 \to V_2$ in $G$, that is* $\mathrm{pa}(2) = \{1\}$. If $G' \neq G$, that is $Z_1 \not\to Z_2$ in $G'$, then the same argument as in Case 2a, this time starting by taking the partial derivative of (C.27) w.r.t. $z_1$, can be used to reach the same conclusion in (C.30). However, this contradicts faithfulness since $V_1 \not\perp\!\!\!\perp V_2$ in $G$.

Hence, we must have $G' = G$, and the following two equations must hold for all $\boldsymbol{z}$:

$$\frac{\tilde{q}_1(z_1)}{q_1(z_1)} = \frac{\tilde{p}_2\left(\psi_2(\boldsymbol{z})\right)}{p_2\left(\psi_2(\boldsymbol{z}) \mid \psi_1(\boldsymbol{z})\right)} \tag{C.31}$$

$$\frac{\tilde{q}_2(z_2)}{q_2(z_2 \mid z_1)} = \frac{\tilde{p}_1\left(\psi_1(\boldsymbol{z})\right)}{p_1\left(\psi_1(\boldsymbol{z})\right)} \tag{C.32}$$

The remainder of the proof consists of exploring the implications of (C.32) and (C.31), ultimately resulting in a violation of the genericity condition (A4).

To ease notation, define the following auxiliary functions:

$$a(z_1) := \frac{\tilde{q}_1(z_1)}{q_1(z_1)} , \tag{C.33}$$

$$b(\boldsymbol{v}) := \frac{\tilde{p}_2(v_2)}{p_2(v_2 \mid v_1)} , \tag{C.34}$$

$$c(\boldsymbol{z}) := \frac{\tilde{q}_2(z_2)}{q_{2|1}(z_2 \mid z_1)} , \tag{C.35}$$

$$d(v_1) := \frac{\tilde{p}_1\left(v_1\right)}{p_1\left(v_1\right)} . \tag{C.36}$$

With this, (C.31) and (C.32) take the following form:

$$a(z_1) = b(\psi(\boldsymbol{z}))\,. \tag{C.37}$$
$$c(\boldsymbol{z}) = d(\psi_1(\boldsymbol{z}))\,, \tag{C.38}$$

Next, define the following maps:

$$\kappa : \boldsymbol{z} \mapsto \begin{bmatrix} a(z_1) \\ c(\boldsymbol{z}) \end{bmatrix} \tag{C.39}$$
$$\rho : \boldsymbol{v} \mapsto \begin{bmatrix} b(\boldsymbol{v}) \\ d(v_1) \end{bmatrix} \tag{C.40}$$

Then, (C.37) and (C.38) together imply that

$$\kappa = \rho \circ \psi\,. \tag{C.41}$$

Recalling that by (A1) all densities are continuously differentiable, the Jacobians of $\kappa$ and $\rho$ are given by:

$$\boldsymbol{J}_\kappa(\boldsymbol{z}) = \begin{bmatrix} a'(z_1) & 0 \\ \frac{\partial c}{\partial z_1}(\boldsymbol{z}) & \frac{\partial c}{\partial z_2}(\boldsymbol{z}) \end{bmatrix}\,, \tag{C.42}$$
$$\boldsymbol{J}_\rho(\boldsymbol{v}) = \begin{bmatrix} \frac{\partial b}{\partial v_1}(\boldsymbol{v}) & \frac{\partial b}{\partial v_2}(\boldsymbol{v}) \\ d'(v_1) & 0 \end{bmatrix}\,, \tag{C.43}$$

and the corresponding determinants are given by

$$\left|\det \boldsymbol{J}_\kappa(\boldsymbol{z})\right| = \left|a'(z_1)\frac{\partial c}{\partial z_2}(\boldsymbol{z})\right| \neq 0 \tag{C.44}$$
$$\left|\det \boldsymbol{J}_\rho(\boldsymbol{v})\right| = \left|d'(v_1)\frac{\partial b}{\partial v_2}(\boldsymbol{v})\right| \neq 0 \tag{C.45}$$

where the inequalities for all $\boldsymbol{z}$ follow since, by assumption (A3), the derivatives of ratios of intervened and original mechanisms are non-vanishing everywhere:

$$a'(z_1) \neq 0 \neq \frac{\partial c}{\partial z_2}(\boldsymbol{z}) \qquad \text{and} \qquad d'(v_1) \neq 0 \neq \frac{\partial b}{\partial v_2}(\boldsymbol{v})\,, \tag{C.46}$$

This implies that the following families of maps are continuously differentiable, monotonic, and invertible,

$$a : z_1 \mapsto a(z_1)\,, \tag{C.47}$$
$$b_{v_1} : v_2 \mapsto b(v_1, v_2)\,, \tag{C.48}$$
$$c_{z_1} : z_2 \mapsto c(z_1, z_2)\,, \tag{C.49}$$
$$d : v_1 \mapsto d(v_1)\,, \tag{C.50}$$

with continuously differentiable inverses

$$a^{-1} : w_1 \mapsto a^{-1}(w_1)\,, \tag{C.51}$$
$$b_{v_1}^{-1} : w_1 \mapsto b_{v_1}^{-1}(w_1)\,, \tag{C.52}$$
$$c_{z_1}^{-1} : w_2 \mapsto c_{z_1}^{-1}(w_2)\,, \tag{C.53}$$
$$d^{-1} : w_2 \mapsto d^{-1}(w_2)\,. \tag{C.54}$$

This implies that $\rho$ and $\kappa$ are valid diffeomorphisms onto their image and their inverses are given by:

$$\kappa^{-1} : \boldsymbol{w} \mapsto \begin{bmatrix} a^{-1}(w_1) \\ c^{-1}_{a^{-1}(w_1)}(w_2) \end{bmatrix}\,, \tag{C.55}$$
$$\rho^{-1} : \boldsymbol{w} \mapsto \begin{bmatrix} d^{-1}(w_2) \\ b^{-1}_{d^{-1}(w_2)}(w_1) \end{bmatrix}\,. \tag{C.56}$$

Since $\boldsymbol{V} = \psi(\boldsymbol{Z})$, by (C.41) we have

$$\boldsymbol{W} := \rho(\boldsymbol{V}) = \rho \circ \psi(\boldsymbol{Z}) = \kappa(\boldsymbol{Z})\,. \tag{C.57}$$

Denote the distributions of $\boldsymbol{W}$ by $R_{\boldsymbol{W}}$ and its density by $r(\boldsymbol{w})$. Since for all $e$, we have

$$P^e_{\boldsymbol{V}} = \psi_*(Q^e_{\boldsymbol{Z}}) \tag{C.58}$$

it follows from (C.57) that

$$R^e_{\boldsymbol{W}} = \rho_*(P^e_{\boldsymbol{V}}) = \kappa_*(Q^e_{\boldsymbol{Z}})\,. \tag{C.59}$$

This provides two different ways of applying the change of variable formula to compute $r(\boldsymbol{w})$.

First, we consider the pushforward of $Q^{e_0}_{\boldsymbol{Z}}$ by $\kappa$:

$$r(\boldsymbol{w}) = q\left(\kappa^{-1}(\boldsymbol{w})\right)\left|\det \boldsymbol{J}_{\kappa^{-1}}(\boldsymbol{w})\right| \tag{C.60}$$

$$= q_1\left(a^{-1}(w_1)\right) q_2\left(c^{-1}_{a^{-1}(w_1)}(w_2) \;\middle|\; a^{-1}(w_1)\right)\left|\frac{\mathrm{d}}{\mathrm{d}w_1}a^{-1}(w_1)\frac{\mathrm{d}}{\mathrm{d}w_2}c^{-1}_{a^{-1}(w_1)}(w_2)\right| \tag{C.61}$$

By integrating this joint density with respect to $w_2$, we obtain the following expression for the marginal $r_1(w_1)$:

$$r_1(w_1) = \left|\frac{\mathrm{d}}{\mathrm{d}w_1}a^{-1}(w_1)\right| q_1\left(a^{-1}(w_1)\right)\int q_2\left(c^{-1}_{a^{-1}(w_1)}(w_2) \;\middle|\; a^{-1}(w_1)\right)\left|\frac{\mathrm{d}}{\mathrm{d}w_2}c^{-1}_{a^{-1}(w_1)}(w_2)\right|\mathrm{d}w_2\,. \tag{C.62}$$

By the diffeomorphic change of variable $z_2 = c^{-1}_{a^{-1}(w_1)}(w_2)$, [13] this can be written as

$$r_1(w_1) = \left|\frac{d}{dw_1}a^{-1}(w_1)\right| q_1\left(a^{-1}(w_1)\right)\int q_2\left(z_2 \;\middle|\; a^{-1}(w_1)\right)dz_2 \tag{C.63}$$

$$= \left|\frac{d}{dw_1}a^{-1}(w_1)\right| q_1\left(a^{-1}(w_1)\right) \tag{C.64}$$

Next, we carry out the same calculation for the pushforward of $P^{e_0}_{\boldsymbol{V}}$ by $\rho$:

$$r(\boldsymbol{w}) = p\left(\rho^{-1}(\boldsymbol{w})\right)\left|\det \boldsymbol{J}_{\rho^{-1}}(\boldsymbol{w})\right| \tag{C.65}$$

$$= p_1\left(d^{-1}(w_2)\right) p_2\left(b^{-1}_{d^{-1}(w_2)}(w_1) \;\middle|\; d^{-1}(w_2)\right)\left|\frac{\mathrm{d}}{\mathrm{d}w_2}d^{-1}(w_2)\frac{\mathrm{d}}{\mathrm{d}w_1}b^{-1}_{d^{-1}(w_2)}(w_1)\right|\,, \tag{C.66}$$

leading to the marginal

$$r_1(w_1) = \int p_1\left(d^{-1}(w_2)\right) p_2\left(b^{-1}_{d^{-1}(w_2)}(w_1) \;\middle|\; d^{-1}(w_2)\right)\left|\frac{\mathrm{d}}{\mathrm{d}w_1}b^{-1}_{d^{-1}(w_2)}(w_1)\right|\left|\frac{\mathrm{d}}{\mathrm{d}w_2}d^{-1}(w_2)\right|\mathrm{d}w_2 \tag{C.67}$$

$$= \int p_1(v_1)p_2\left(b^{-1}_{v_1}(w_1) \;\middle|\; v_1\right)\left|\frac{\mathrm{d}}{\mathrm{d}w_1}b^{-1}_{v_1}(w_1)\right|\mathrm{d}v_1\,, \tag{C.68}$$

where the second line is obtained by the diffeomorphic change of variable $v_1 = d^{-1}(w_2)$.

Equating the two expressions for $r(w_1)$ in $e_0$ in (C.68) and (C.64), we obtain for all $w_1$:

$$\left|\frac{\mathrm{d}}{\mathrm{d}w_1}a^{-1}(w_1)\right| q_1\left(a^{-1}(w_1)\right) = \int p_1(v_1)p_2\left(b^{-1}_{v_1}(w_1) \;\middle|\; v_1\right)\left|\frac{\mathrm{d}}{\mathrm{d}w_1}b^{-1}_{v_1}(w_1)\right|\mathrm{d}v_1\,. \tag{C.69}$$

Applying the same approach to the environment in which $V_1$ is intervened upon changing $p_1$ to $\tilde{p}_1$ while $Z_2$ is intervened upon leaving $q_1$ invariant, yields for all $w_1$:

$$\left|\frac{\mathrm{d}}{\mathrm{d}w_1}a^{-1}(w_1)\right| q_1\left(a^{-1}(w_1)\right) = \int \tilde{p}_1(v_1)p_2\left(b^{-1}_{v_1}(w_1) \;\middle|\; v_1\right)\left|\frac{\mathrm{d}}{\mathrm{d}w_1}b^{-1}_{v_1}(w_1)\right|\mathrm{d}v_1\,. \tag{C.70}$$

[13] Note that: $\int q_2(z_2(w_2))\left|\frac{\mathrm{d}z_2}{\mathrm{d}w_2}\right|\mathrm{d}w_2 = \int q_2(z_2)\,\mathrm{d}z_2\,.$

Finally, by equating (C.69) and (C.70), we arrive at the following expression which must hold for all $w_1$:

$$\int p_1(v_1) p_2\left(b_{v_1}^{-1}(w_1) \mid v_1\right)\left|\frac{\mathrm{d}}{\mathrm{d}w_1} b_{v_1}^{-1}(w_1)\right| \mathrm{d}v_1 = \int \tilde{p}_1(v_1) p_2\left(b_{v_1}^{-1}(w_1) \mid v_1\right)\left|\frac{\mathrm{d}}{\mathrm{d}w_1} b_{v_1}^{-1}(w_1)\right| \mathrm{d}v_1 \tag{C.71}$$

which we can rewrite as

$$\int \left(\tilde{p}_1(v_1) - p_1(v_1)\right) p_2\left(b_{v_1}^{-1}(w_1) \mid v_1\right)\left|\frac{\mathrm{d}}{\mathrm{d}w_1} b_{v_1}^{-1}(w_1)\right| \mathrm{d}v_1 = 0\,. \tag{C.72}$$

Multiplying by any continuous function $\varphi(w_1)$, integrating w.r.t. $w_1$ and applying the diffeomorphic change of variable $v_2 = b_{v_1}^{-1}(w_1)$, this can be expressed as:

$$0 = \int \varphi(w_1) \int \left(\tilde{p}_1(v_1) - p_1(v_1)\right) p_2\left(b_{v_1}^{-1}(w_1) \mid v_1\right)\left|\frac{\mathrm{d}}{\mathrm{d}w_1} b_{v_1}^{-1}(w_1)\right| \mathrm{d}v_1 \,\mathrm{d}w_1 \tag{C.73}$$

$$= \int\int \varphi\left(b_{v_1}(v_2)\right)\left(\tilde{p}_1(v_1) - p_1(v_1)\right) p_2(v_2 \mid v_1)\,\mathrm{d}v_2\,\mathrm{d}v_1 \tag{C.74}$$

$$= \int\int \varphi\left(\frac{\tilde{p}_2(v_2)}{p_2(v_2 \mid v_1)}\right)\left(\tilde{p}_1(v_1) - p_1(v_1)\right) p_2(v_2 \mid v_1)\,\mathrm{d}v_2\,\mathrm{d}v_1 \tag{C.75}$$

where we have resubstituted the expression for $b_{v_1}(v_2)$ in the last line.

Equivalently, this can be written as: *for any continuous function* $\varphi$,

$$\mathbb{E}_{\boldsymbol{v}\sim P_{\boldsymbol{V}}^{e_0}}\left[\varphi\left(\frac{\tilde{p}_2(v_2)}{p_2(v_2 \mid v_1)}\right)\right] = \mathbb{E}_{\boldsymbol{v}\sim P_{\boldsymbol{V}}^{e_1}}\left[\varphi\left(\frac{\tilde{p}_2(v_2)}{p_2(v_2 \mid v_1)}\right)\right]. \tag{C.76}$$

However, the genericity condition (A4) precisely rules this out, since the above equality must be violated for at least one $\varphi$, concluding this last case.

To sum up, all cases either lead to a contradiction, or imply the conclusion that $(f^{-1}, G) \sim_{\text{CRL}} (h, G')$, concluding the proof. □

### C.3 Proof of Thm. 3.4

**Theorem 3.4** (Identifiability up to $\sim_{\text{CRL}}$ from two paired perfect stochastic interventions per node)**.** *Suppose that we have access to multiple environments* $\{P_{\boldsymbol{X}}^e\}_{e\in\mathcal{E}}$ *generated as described in § 2 under Asms. 2.2, 2.3, 2.5, 2.8 and 2.9. Let* $(h, G')$ *be any candidate solution such that the inferred latent distributions* $Q_{\boldsymbol{Z}}^e = h_*(P_{\boldsymbol{X}}^e)$ *of* $\boldsymbol{Z} = h(\boldsymbol{X})$ *and the inferred mixing function* $h^{-1}$ *satisfy the above assumptions w.r.t. the candidate causal graph* $G'$*. Assume additionally that*

*(A1) all densities* $p^e$ *and* $q^e$ *are continuously differentiable and fully supported on* $\mathbb{R}^n$*;*

*(A2') we have access to at least one* pair *of single-node perfect interventions per node, with unknown targets: there exist* $m \geq n$ *known pairs of environments* $\mathcal{E} = \{(e_j, e_j')\}_{j=1}^m$ *such that for each* $i \in [n]$ *there exists some* unknown $j \in [m]$ *for which* $\mathcal{I}^{e_j} = \mathcal{I}^{e_j'} = \{i\}$*;*

*(A3') for all* $i \in [n]$*, the intervened mechanisms* $\tilde{p}_i(v_i)$ *and* $\tilde{\tilde{p}}_i(v_i)$ *differ everywhere, in the sense that*

$$\forall v_i: \qquad \left(\frac{\tilde{\tilde{p}}_i}{\tilde{p}_i}\right)'(v_i) \neq 0\,; \tag{3.10}$$

*Then the ground truth is identified in the sense of Defn. 2.6, that is,* $(f^{-1}, G) \sim_{\text{CRL}} (h, G')$.

*Proof.* First, we show that we can extract from the $m \geq n$ available pairs of environments a suitable subset $\mathcal{E}_n$ of exactly $n$ pairs, containing one pair of interventional environments for each node.

Let $\mathcal{E}_n \subseteq \mathcal{E}$ be a subset of $n$ pairs of environments which are assumed to correspond to distinct targets in the model $q$, and suppose for a contradiction that this is not actually the case for the ground truth $p$ (i.e., there are duplicate and missing interventions w.r.t. $p$). Then there must be two pairs of environments $(e_a, e_a'), (e_b, e_b') \in \mathcal{E}_n$, both corresponding to interventions on some $V_i$ in $p$, but which

are modelled as interventions on distinct nodes $Z_j$ and $Z_k$ with $j \neq k$ in $q$. We show that this implies that $V_i$ must simultaneously be a deterministic function of only $Z_j$ and only $Z_k$. Similar to the proof of Thm. 3.2, we obtain the following equations,

$$\frac{\tilde{\hat{q}}_j}{\tilde{q}_j}\left(z_j\right) = \frac{\tilde{\hat{p}}_i}{\tilde{p}_i}\left(\psi_i(\boldsymbol{z})\right) , \tag{C.77}$$

$$\frac{\tilde{\hat{q}}_k}{\tilde{q}_k}\left(z_k\right) = \frac{\hat{\tilde{p}}_i}{\hat{p}_i}\left(\psi_i(\boldsymbol{z})\right) . \tag{C.78}$$

By taking partial derivatives w.r.t. $z_l$ and applying assumption (A3'), we find that

$$\frac{\partial \psi_i}{\partial z_l} = 0 \qquad \forall l \neq j \,, \tag{C.79}$$

$$\frac{\partial \psi_i}{\partial z_l} = 0 \qquad \forall l \neq k \,. \tag{C.80}$$

Since $j \neq k$, this implies that $\partial \psi_i / \partial z_l = 0$ for all $l$ which contradicts invertibility of $\psi$. Thus, by contradiction, we find that $\mathcal{E}_n$ must contain exactly one pair of intervention per node also w.r.t. $p$. For the remainder of the proof, we only consider $\mathcal{E}_n$.

W.l.o.g., for any $(e_i, e_i') \in \mathcal{E}_n$ we now fix the intervention targets in $p$ to $\mathcal{I}^{e_i} = \mathcal{I}^{e_i'} = \{i\}$ and let $\pi$ be a permutation of $[n]$ such that $\pi(i)$ denotes the inferred intervention target in $q$ that by (A2') is shared across $(e_i, e_i')$. (We will show later that not all permutations are admissible, but only ones that preserve the partial order of $G$.)

The first part of the proof is similar to Case 1 in the proof of Thm. 3.2. Consider the densities in environments $e_i$ and $e_i'$, which are related through the change of variable formula by:

$$\tilde{q}_{\pi(i)}\left(z_{\pi(i)}\right) \prod_{j \in [n] \setminus \{\pi(i)\}} q_j\left(z_j \mid \boldsymbol{z}_{\mathrm{pa}(j;G')}\right) = \tilde{p}_i\left(\psi_i(\boldsymbol{z})\right) \prod_{j \in [n] \setminus \{i\}} p_j\left(\psi_j(\boldsymbol{z}) \mid \psi_{\mathrm{pa}(j)}(\boldsymbol{z})\right) \left|\det \boldsymbol{J}_\psi(\boldsymbol{z})\right| , \tag{C.81}$$

$$\tilde{\hat{q}}_{\pi(i)}\left(z_{\pi(i)}\right) \prod_{j \in [n] \setminus \{\pi(i)\}} q_j\left(z_j \mid \boldsymbol{z}_{\mathrm{pa}(j;G')}\right) = \tilde{\hat{p}}_i\left(\psi_i(\boldsymbol{z})\right) \prod_{j \in [n] \setminus \{i\}} p_j\left(\psi_j(\boldsymbol{z}) \mid \psi_{\mathrm{pa}(j)}(\boldsymbol{z})\right) \left|\det \boldsymbol{J}_\psi(\boldsymbol{z})\right| , \tag{C.82}$$

where $\boldsymbol{Z}_{\mathrm{pa}(j;G')} \subseteq \boldsymbol{Z} \setminus \{Z_j\}$ denotes the parents of $Z_j$ in $G'$.

Taking the quotient of the two equations yields

$$\frac{\tilde{\hat{q}}_{\pi(i)}}{\tilde{q}_{\pi(i)}}\left(z_{\pi(i)}\right) = \frac{\tilde{\hat{p}}_i}{\tilde{p}_i}\left(\psi_i(\boldsymbol{z})\right) . \tag{C.83}$$

Next, for any $j \neq \pi(i)$, taking partial derivatives w.r.t. $z_j$ on both sides yields

$$0 = \left(\frac{\tilde{\hat{p}}_i}{\tilde{p}_i}\right)'\left(\psi_i(\boldsymbol{z})\right) \frac{\partial \psi_i}{\partial z_j}(\boldsymbol{z}) \,. \tag{C.84}$$

By assumption (A3'), the first term on the RHS is non-zero everywhere. Hence, (C.84) implies

$$\forall j \neq \pi(i), \forall \boldsymbol{z} : \quad \frac{\partial \psi_i}{\partial z_j}(\boldsymbol{z}) = 0 \tag{C.85}$$

from which we can conclude that

$$V_i = \psi_i\left(Z_{\pi(i)}\right) \tag{C.86}$$

for all $i \in [n]$. That is, $\psi$ is the composition of the permutation $\pi$ with an element-wise reparametrisation.

It remains to show that $\pi$ must, in fact, be a graph isomorphism, which is equivalent to the statement

$$V_i \to V_j \quad \text{in} \quad G \qquad \Longleftrightarrow \qquad Z_{\pi(i)} \to Z_{\pi(j)} \quad \text{in} \quad G' . \tag{C.87}$$

($\implies$) Suppose for a contradiction that there exist $(i, j)$ such that $V_i \to V_j$ in $G$, but $Z_{\pi(i)} \not\to Z_{\pi(j)}$ in $G'$.

The main idea is to demonstrate that the lack of such direct arrow implies a certain conditional independence which, by faithfulness, would contradict the unconditional dependence of $V_i$ and $V_j$.

Consider environment $e_i$ in which there are perfect interventions on $Z_{\pi(i)}$ and $V_i$, which has the effect of removing all incoming arrows to $Z_{\pi(i)}$ and $V_i$ in the respective post-intervention graphs $G'_{\overline{Z}_{\pi(i)}}$ and $G_{\overline{V}_i}$.

As a result of this and the lack of direct arrow by assumption, any d-connecting path between $Z_{\pi(i)}$ and $Z_{\pi(j)}$ must enter the latter via $\boldsymbol{Z}_{\mathrm{pa}(\pi(j);G')}$ [95].

It then follows from Markovianity of $q$ w.r.t. $G'$ that the following holds in $Q^{e_i}_{\boldsymbol{Z}}$:

$$Z_{\pi(i)} \perp\!\!\!\perp Z_{\pi(j)} \mid \boldsymbol{Z}_{\mathrm{pa}(\pi(j);G')} \,. \tag{C.88}$$

We now consider the corresponding implication for $P^{e_i}_{\boldsymbol{V}}$. Define

$$\tilde{\boldsymbol{V}} = \left\{ V_k = \psi_k\left(Z_{\pi(k)}\right) : Z_{\pi(k)} \in \boldsymbol{Z}_{\mathrm{pa}(\pi(j);G')} \right\} \subseteq \boldsymbol{V} \setminus \{V_i, V_j\} \,, \tag{C.89}$$

and note that by assumption, $Z_{\pi(i)} \notin \boldsymbol{Z}_{\mathrm{pa}(\pi(j);G')}$ and hence $V_i \notin \tilde{\boldsymbol{V}}$.

By applying the corresponding diffeomorphic functions $\psi_i$ from (C.86) to (C.88), it follows from Lemma C.2 that

$$V_i \perp\!\!\!\perp V_j \mid \tilde{\boldsymbol{V}} \tag{C.90}$$

in $P^{e_i}_{\boldsymbol{V}}$. However, this violates faithfulness (Asm. 2.2) of $P_{\boldsymbol{V}}$ to $G$ since $V_i$ and $V_j$ are d-connected in $G_{\overline{V}_i}$.

Thus, by contradiction, we must have $Z_{\pi(i)} \to Z_{\pi(j)}$ in $G'$.

($\impliedby$) Now, suppose for a contradiction that there exist $(i, j)$ such that $Z_{\pi(i)} \to Z_{\pi(j)}$ in $G'$, but $V_i \not\to V_j$ in $G$.

By the same argument as before, we find that

$$V_i \perp\!\!\!\perp V_j \mid \boldsymbol{V}_{\mathrm{pa}(j)} \tag{C.91}$$

in $P^{e_i}_{\boldsymbol{V}}$, and thus by Lemma C.2

$$Z_{\pi(i)} \perp\!\!\!\perp Z_{\pi(j)} \mid \tilde{\boldsymbol{Z}} \tag{C.92}$$

in $Q^{e_i}_{\boldsymbol{Z}}$ where

$$\tilde{\boldsymbol{Z}} = \left\{ Z_{\pi(k)} : V_k \in \boldsymbol{V}_{\mathrm{pa}(j)} \right\} \subseteq \boldsymbol{Z} \setminus \{Z_{\pi(i)}, Z_{\pi(j)}\} \,.$$

However, this contradicts faithfulness of $Q_{\boldsymbol{Z}}$ to $G'$. Hence, we must have that $V_i \to V_j$ in $G$.

This shows that $\pi$ must be a graph isomorphism, thus concluding the proof. □

### C.4 Proof of Thm. 4.2

**Theorem 4.2** (Preservation of causal influences under $\sim_{\text{CRL}}$)**.** *Let $P_{\boldsymbol{V}}$ be Markovian w.r.t. $G$, let $\pi$ be a graph isomorphism of $G$, and let $\phi$ be an element-wise diffeomorphism. Let $\boldsymbol{Z} = \boldsymbol{P}_{\pi^{-1}} \circ \phi(\boldsymbol{V})$ and denote its induced distribution by $Q_{\boldsymbol{Z}}$. Then for any $V_i \to V_j$ in $G$ we have $\mathfrak{C}^{P_{\boldsymbol{V}}}_{i \to j} = \mathfrak{C}^{Q_{\boldsymbol{Z}}}_{\pi(i) \to \pi(j)}$.*

*Proof.* First, recall that according to Defn. 4.1,

$$\mathfrak{C}^{P_{\boldsymbol{V}}}_{i \to j} := D_{\mathrm{KL}}\big(P_{\boldsymbol{V}} \,\big\|\, P^{i \to j}_{\boldsymbol{V}}\big), \tag{C.93}$$

where $P^{i \to j}_{\boldsymbol{V}}$ denotes the interventional distribution obtained by replacing $p_j\left(v_j \mid \boldsymbol{v}_{\mathrm{pa}(j)}\right)$ with

$$p_j^{i \to j}\big(v_j \mid \boldsymbol{v}_{\mathrm{pa}(j) \setminus \{i\}}\big) = \int_{\mathcal{V}_i} p_j\big(v_j \mid \boldsymbol{v}_{\mathrm{pa}(j)}\big) p_i(v_i) \,\mathrm{d}v_i \,. \tag{C.94}$$

Writing out the KL divergence and noting that all terms except the interved mechanism $j$ cancel inside the $\log$, we obtain

$$\mathfrak{C}^{P_{\boldsymbol{V}}}_{i\to j} = \int_{\mathcal{V}} \log\left(\frac{p_j\left(v_j \mid \boldsymbol{v}_{\mathrm{pa}(j)}\right)}{\int_{\mathcal{V}_i} p_j\big(v_j \mid \boldsymbol{v}_{\mathrm{pa}(j)}\big) p_i(v_i)\,\mathrm{d}v_i}\right) p(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v}\,. \tag{C.95}$$

and similarly

$$\mathfrak{C}^{Q_{\boldsymbol{Z}}}_{\pi(i)\to\pi(j)} = \int_{\mathcal{Z}} \log\left(\frac{q_{\pi(j)}\left(z_{\pi(j)} \mid \boldsymbol{z}_{\mathrm{pa}(\pi(j);G')}\right)}{\int_{\mathcal{Z}_{\pi(i)}} q_{\pi(j)}\left(z_{\pi(j)} \mid \boldsymbol{z}_{\mathrm{pa}(\pi(j);G')}\right) q_{\pi(i)}(z_{\pi(i)})\,\mathrm{d}z_{\pi(i)}}\right) q(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}\,. \tag{C.96}$$

Since $\boldsymbol{Z} = \boldsymbol{P}_{\pi^{-1}} \circ \phi(\boldsymbol{V})$, we have $V_i = \psi_i(Z_{\pi(i)})$ for all $i \in [n]$ where $\psi = \phi^{-1}$.

Thus, by the change of variable formula, and using the fact that $\pi(\mathrm{pa}(i)) = \mathrm{pa}(\pi(i); G')$ since $\pi : G \mapsto G'$ is a graph isomorphism, we have for all $i \in [n]$:

$$q_{\pi(i)}\left(z_{\pi(i)} \mid \boldsymbol{z}_{\mathrm{pa}(\pi(i);G')}\right) = p_i\left(\psi_i\left(z_{\pi(i)}\right) \mid \psi_{\mathrm{pa}(i)}\left(\boldsymbol{z}_{\mathrm{pa}(\pi(i);G')}\right)\right)\left|\frac{\mathrm{d}\psi_i}{\mathrm{d}z_{\pi(i)}}\left(z_{\pi(i)}\right)\right|\,, \tag{C.97}$$

as well as for the marginal density

$$q_{\pi(i)}\left(z_{\pi(i)}\right) = p_i\left(\psi_i\left(z_{\pi(i)}\right)\right)\left|\frac{\mathrm{d}\psi_i}{\mathrm{d}z_{\pi(i)}}\left(z_{\pi(i)}\right)\right|\,, \tag{C.98}$$

and

$$q(\boldsymbol{z}) = p(\psi \circ \boldsymbol{P}_\pi(\boldsymbol{z}))\left|\det \boldsymbol{J}_\psi(\boldsymbol{z})\right|\,. \tag{C.99}$$

Substitution into the expression for $\mathfrak{C}^{Q_{\boldsymbol{Z}}}_{\pi(i)\to\pi(j)}$ yields:

$$\mathfrak{C}^{Q_{\boldsymbol{Z}}}_{\pi(i)\to\pi(j)} = \int_{\mathcal{Z}} \log\left(\frac{p_j\left(\psi_j(z_{\pi(j)}) \mid \psi_{\mathrm{pa}(j)}(\boldsymbol{z}_{\mathrm{pa}(\pi(j);G')})\right)}{\int_{\mathcal{Z}_{\pi(i)}} p_j\left(\psi_j(z_{\pi(j)}) \mid \psi_{\mathrm{pa}(j)}(\boldsymbol{z}_{\mathrm{pa}(\pi(j);G')})\right) p_i(\psi_i(z_{\pi(i)}))\left|\frac{\mathrm{d}\psi_i}{\mathrm{d}z_{\pi(i)}}(z_{\pi(i)})\right|\mathrm{d}z_{\pi(i)}}\right) \tag{C.100}$$

$$p(\psi \circ \boldsymbol{P}_\pi(\boldsymbol{z}))\left|\det \boldsymbol{J}_\psi(\boldsymbol{z})\right|\mathrm{d}\boldsymbol{z}\,. \tag{C.101}$$

$$= \int_{\mathcal{V}} \log\left(\frac{p_j\left(v_j \mid \boldsymbol{v}_{\mathrm{pa}(j)}\right)}{\int_{\mathcal{V}_i} p_j\big(v_j \mid \boldsymbol{v}_{\mathrm{pa}(j)}\big) p_i(v_i)\,\mathrm{d}v_i}\right) p(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} \tag{C.102}$$

$$= \mathfrak{C}^{P_{\boldsymbol{V}}}_{i\to j}\,. \tag{C.103}$$

where the second to last line follows by integration by substitution, applied to both integrals. □

# D Experimental Details and Additional Results

In this appemndix, we describe the experiments presented in § 6 in more details (Appx. D.1), and present additional results (Appx. D.2).

## D.1 Experimental Details for § 6

**Synthetic Data Generating Process.** We consider linear Gaussian latent SCMs of the form

$$V_1 := U_1, \qquad V_2 := \alpha V_1 + U_2, \tag{D.1}$$

with standard normal $U_1$ and $U_2$. As a mixing function, we use a three-layer multilayer perceptron (MLP),

$$f = \sigma \circ \boldsymbol{W}_3 \circ \sigma \circ \boldsymbol{W}_2 \circ \sigma \circ \boldsymbol{W}_1 \tag{D.2}$$

where $\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3 \in \mathbb{R}^{2\times 2}$ are invertible weight matrices, and $\sigma$ is an element-wise invertible nonlinear leaky-tanh activation function used in [41]:

$$\sigma(x) = \tanh(x) + 0.1x\,. \tag{D.3}$$

To compute averages of our results over multiple runs, we construct different ground truth data generating processes as follows. We generate different latent SCMs by drawing $\alpha$ uniformly from $[-10, -2] \cup [2, 10]$. (We exclude $(-2, 2)$ to avoid sampling near unfaithful models.) We generate the corresponding mixing functions by uniformly sampling each element of the weight matrices, $(\boldsymbol{W}_k)_{ij} \sim U(0, 1)$. (To avoid the sampled weight matrices being too close to singular, we reject and resample if $|\det \boldsymbol{W}_k| < 0.1$.)

**Interventional Environments.** In line with Thm. 3.2, for each choice of latent SCM and mixing function, we generate three environments: one observational environment and one interventional environment for each perfect single-node intervention. For $i = 1, 2$, we model a perfect intervention on $V_i$ by removing the influence of the parent variables and changing the exogenous noise by shifting its mean up or down. Specifically, we replace the corresponding assignment in (D.1) by

$$V_i := \tilde{U}_i\,, \quad \text{where} \quad \tilde{U}_i \sim \mathcal{N}(m_i, 1) \tag{D.4}$$

where the mean $m_i$ of the shifted Gaussian noise is fixed per environment and sampled uniformly from $\{\pm 2\}$.

We label the observational environment as $e = 0$ and the environment arising from intervention on $V_i$ by $e = i$ for $i = 1, 2$. Samples from $p^e$ are then generated by sampling latents $\boldsymbol{v}$ from the respective (un)intervened SCM and then applying the mixing function.

**Model Architecture.** We use normalizing flows [93] to model observations $\boldsymbol{x}$ as the result of an invertible, differentiable transformation $g$ of some latent (noise) variable $\boldsymbol{z}$,

$$\boldsymbol{x} = g(\boldsymbol{z})\,. \tag{D.5}$$

We apply a series of $L$ such transformations $g^l : \mathbb{R}^2 \to \mathbb{R}^2$ such that $g = g^L \circ \ldots \circ g^1$ which we refer to as *flow layers*. We use Neural Spline Flows [30] for the invertible transformation, with a 3-layer feedforward neural network with hidden dimension 128 and permutation in each flow layer and $L = 12$ layers. The transformations $g, g^1, \ldots, g^L$ have learnable parameters (the weights and biases of the neural networks), which we omit to simplify notation.

Typically, simple distributions such as a uniform or isotropic Gaussian are used as base distribution $q(\boldsymbol{z})$ in normalizing flows. Here, we instead choose a base distribution that encodes information about the latent SCM. Specifically, we model the base mechanism as

$$q_1(z_1) = \mathcal{N}\left(\mu_1, \sigma_1^2\right)\,, \qquad q_2(z_2 \mid z_1) = \mathcal{N}\left(\hat{\alpha} z_1, \sigma_2^2\right)\,, \qquad q_2(z_2) = \mathcal{N}\left(\mu_2, \hat{\sigma}_2^2\right) \tag{D.6}$$

and the intervened mechanism as

$$\tilde{q}_1(z_1) = \mathcal{N}(\tilde{\mu}_1, \tilde{\sigma}_1^2)\,, \qquad \tilde{q}_2(z_2) = \mathcal{N}(\tilde{\mu}_2, \tilde{\sigma}_2^2)\,. \tag{D.7}$$

**Candidate Graphs and Intervention Targets.** We train a separate normalizing-flow based model for each choice of candidate graph $G'$ and inferred intervention targets. For the bivariate case with $n = 2$, this gives rise to four models, depending on whether $G'$ matches $G$ or not, and whether the intervention targets are aligned or misaligned w.r.t. the ground truth intervention targets. To model the setting $G' \neq G$ in which $Z_1$ and $Z_2$ are assumed independent, we use $q_2(z_2)$ in place of $q_2(z_2 \mid z_1)$ in (D.6). If the intervention targets are aligned, we use $\tilde{q}_i$ instead of $q_i$ in $e = i$ for $i = 1, 2$. Else, if they are misaligned, we use $\tilde{q}_2$ instead of $q_2$ in $e = 1$ and $\tilde{q}_1$ instead of $q_1$ in $e = 2$. By multiplying the respective mechanisms, we thus obtain three environment-specific joint base distributions $q^e(\boldsymbol{z})$ for $e = 0, 1, 2$.

**Learning Objective.** Given multi-environment data, the parameters $\mu_1$, $\sigma_1$, $\hat{\alpha}$, $\sigma_2$, $\mu_2$, $\hat{\sigma}_2$, $\tilde{\mu}_2$, $\tilde{\sigma}_2$, $\tilde{\mu}_1$ and $\tilde{\sigma}_1$ are jointly learned with the parameters of the invertible transformations $g^l$ by maximising the log-likelihood of the data under our model, which is given by:

$$\sum_{e \in \mathcal{E}} \mathbb{E}_{\boldsymbol{x} \sim p^e(\boldsymbol{x})} \left[ \log p^e_{\text{model}}(\boldsymbol{x}) \right] = \sum_{e \in \mathcal{E}} \mathbb{E}_{\boldsymbol{x} \sim p^e(\boldsymbol{x})} \left[ \log q^e(h(\boldsymbol{x})) + \log \left| \det \boldsymbol{J}_h(\boldsymbol{x}) \right| \right] \tag{D.8}$$

where the encoder $h := g^{-1}$ is the inverse of the normalizing flow which is readily available by construction; and where the expectations are empirical averages over the respective datasets in practice.

**Training and Model Selection Details.** Each environment comprises a total of 200k data points. We use the ADAM optimizer [67] with cosine annealing learning rate scheduling, starting with a learning rate of $5 \times 10^{-3}$ and ending with $1 \times 10^{-7}$. We train the model for 200 epochs with a batch size of 4096. We split the dataset into 70% for training, and 15% for validation and held-out test data, each sampled randomly across all environments. For each drawn data generating process, we train three versions of each model with different random initializations and select the one with the highest validation log likelihood at the end of training for evaluation.

**Evaluation Metrics.** We evaluate the trained models w.r.t. *mean correlation coefficient* (MCC) on held-out data and *log-likelihood* on validation data (for model selection).

- The MCC measures the extent to which there is a one-to-one correspondence between the ground truth latents $V_i$ and (a permuted version of) the inferred latents $Z_i = h_i(\boldsymbol{X})$. Its maximum value of one indicates a perfect correlation between the two. MCC is thus a proxy measure for the level of identifiability up to permutation and invertible reparametrisation. We report MCC based on Pearson (linear) correlation, though we found the results based on Spearman (nonlinear monotonic) correlation to be almost identical.
- The log-likelihood, on the other hand, measures how well a model explains or fits the data. Since the ground truth is typically unknown, a reasonable procedure when training multiple models is to select the one that attains the highest likelihood. For this reason, we report the difference in log-likelihood between misspecified models (ones assuming a wrong graph or intervention targets) to the correctly specified model. Whenever this difference is larger than zero, the correct model fits the data better and would thus be selected.

### D.2 Additional Results: Learning Nonlinear Latent SCMs from Partial Causal Order

In this subsection, we present an additional experiment, in which we extend the setting investigated in § 6 and Appx. D.1 along the following axes.

- We fit generative models over three instead of two variables, corresponding to the setting of Thm. 3.4.
- The ground-truth SCM is now given by nonlinear mechanisms with non-additive, non-Gaussian noise.
- The generative model, including the learnt mechanisms, is now fully nonlinear.
- Despite Thm. 3.4 formally requiring two environments per single-node intervention, we only provide one interventional environment per node.
- Rather than searching over candidate graphs, we only fix the causal order and fit the reduced form of the SCM (see § 2.1) with a second normalizing flow.

Below, we describe these differences in more detail.

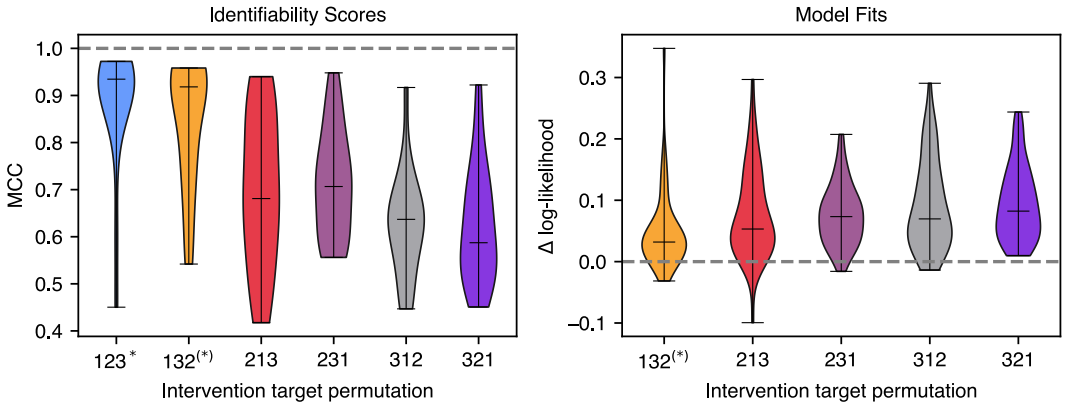


Figure 4: **Comparison of Correctly and Incorrectly Specified Models for $V_2 \leftarrow V_1 \rightarrow V_3$ with Fixed Causal Order and Nonlinear SCM.** Each violinplot corresponds to one setting where the intervention target labels are permuted. The blue plot ($123^*$) is the setting with correct intervention target labels. The yellow plot ($132^{(*)}$) has the targets for the two children $V_2$ and $V_3$ permuted, which also corresponds to a correct causal ordering and should thus be considered equivalent. We show mean correlation coefficients (MCCs) between the learned and ground truth latents *(Left)* and the difference in validation model log-likelihood between the well-specified (blue) and misspecified models *(Right)*. Each violin plot is based on 20 different ground truth data generating processes; the horizontal lines indicate the minimum, median and maximum values.

**Three-Variable Graph.** The *unknown* ground truth graph is given by

$$V_2 \leftarrow V_1 \rightarrow V_3 \,. \tag{D.9}$$

This is consistent with the partial ordering $V_1 \preceq V_2 \preceq V_3$, which is assumed for all models a priori w.l.o.g., see § 2.2. Note that, due to the encoding of causal structure in the nonparametric model explained below, we only iterate over different permutations of the intervention targets and not over latent graph configurations. Due to the causal order implied by the graph (D.9), the permutations $(1, 2, 3)$ (no permutaion) and $(1, 3, 2)$ (permutation of the two effects) are equivalent since the latter also implies the correct causal ordering.

**Nonlinear, Non-Gaussian SCM.** The mechanisms in the ground-truth SCM are now given by

$$V_i := \beta f_i^{\text{loc}}(\boldsymbol{V}_{\text{pa}(i)}) + f_i^{\text{scale}}(\boldsymbol{V}_{\text{pa}(i)}) U_i \tag{D.10}$$

for all $i$, where the location and scale functions $f_i^{\text{loc}}$, $f_i^{\text{scale}} : \mathbb{R}^{|\text{pa}(i)|} \to \mathbb{R}$ are parameterized by random 3-layer neural networks (sampled as the random mixing function in (D.2)) and the noise variables are Gaussian, $U_i \sim \mathcal{N}(0, 1)$. The factor $\beta$ controls the influence of the parent variables relative to the exogenous noise. As $\beta$ increases, variables tend to become more dependent, as also the mean shifts as a function of the parent variables. We set $\beta = 10$ for the experiments shown in Fig. 4.

**Nonparametric Latent SCM.** We use a second normalizing flow to learn a *reduced form of the latent SCM* via the transformation $g^{\text{SCM}} : \mathbb{R}^3 \to \mathbb{R}^3$ mapping an exogenous noise variable $\boldsymbol{\epsilon}$ to the latent variable $\boldsymbol{z}$,

$$\boldsymbol{z} = g^{\text{SCM}}(\boldsymbol{\epsilon}) \,. \tag{D.11}$$

The distribution of the exogenous noise variable $\boldsymbol{\epsilon}$ as well as the distribution of the intervened mechanisms $\tilde{q}_i(z_i)$ for $i = 1, 2, 3$ is fixed and standard (isotropic) Gaussian. The flow layers in $g^{\text{SCM}}$ have an *upper triangular Jacobian* and thus allow us to encode assumptions about the causal graph: by passing the variables in topological order, which we can assume w.l.o.g., we ensure that an exogenous noise variable $\epsilon_i$ can only influence endogenous variables in $\boldsymbol{z}$ that are descendants of $z_i$. The learned weights of the flow layers then implicitly encode which endogenous variables are connected. Therefore, only different choices of the permutations of the intervention targets need to be considered as candidate models. We use a similar architecture based on Neural Spline Flows. However, we omit permutation layers, which would violate the topological order of the variables.

**Results.** In Fig. 4, we present identifiability scores and model fits for both well-specified and misspecified models (corresponding to different intervention target choices). Notably, we observe that the well-specified model (in blue) or its equivalent (in yellow) yield the highest log-likelihood in the majority of cases, as depicted in Fig. 4 (*Right*). This demonstrates that, even in this nonparametric setting without fully specified graph, the log-likelihood remains a reliable criterion for selecting the correct intervention targets. Fig. 4 (*Left*) shows that the selected models (blue or yellow) approximately identify the ground-truth latent variables up to element-wise rescaling, whereas other choices lead to much lower MCCs.

It is worth noting that, compared to the parametric setting investigated in § 6 and Fig. 3, the nonparametric setting appears to be more challenging (as expected), as there is a less pronounced distinction between well-specified and misspecified models, both in terms of identifiability scores and model fits. Moreover, future work is needed to parse the implicitly learned causal relationships in the transformation $g^{\mathrm{SCM}}$ in (D.11): since only the (pre-imposed) causal order is specified, in practice, $g^{\mathrm{SCM}}$ may learn to use additional or fewer edges than in the true graph $G$.

# E Discussion of the Role of Our Assumptions

Below, we summarize the rationale and intuition behind each assumption:

- Asm. 2.2 helps rule out degenerate cases (cancellation along different paths) in which variables are (conditionally) independent despite being causally related. It is a standard assumption in classical causal discovery from observational data, and therefore also helps in CRL to recover the true causal graph.
- Asm. 2.3 is required to know how many latent variables we are looking for. It is a standard assumption in identifiable representation learning (that is often made implicitly). However, it may be dropped when suitable techniques for estimating the intrinsic dimensionality of $\mathcal{X}$ can be employed.
- Asm. 2.5 is needed for the mapping between latents and observations to be invertible in the first place. Without it, full recovery of the causal variables (up to CRL equivalence) is infeasible. This assumption is also standard for the simpler problem of nonlinear ICA.
- Asm. 2.8 is a characterisation of our generative setup. Sharing of some mechanisms and the mixing function is needed for the multi-environment setting to provide useful additional information: if everything may change across environments, the datasets can only be analysed in isolation, running into the non-identifiability of CRL from iid data.
- Asm. 2.9 and (A2) / (A2') are needed since with imperfect interventions or interventions not on all nodes, identifiability is not achievable even in the linear setting as shown by Squires et al. [117].
- Asm. (A1) is a technical assumption needed for our analysis. It is not strictly necessary (it can also be relaxed to fully supported on a Cartesian product of intervals) but substantially eases the readability and accessibility of the proof, without a major impact on the main causal aspects of the problem setup.
- Asm. (A3) / (A3') is needed to avoid spurious solutions based on applying a measure preserving transformation on a part of the domain unaffected by the intervention.
- Asm. (A4) is needed to rule out a fine-tuning of the ground-truth generating process that are possible due to fully non-parametric nature of the setup, see also Remark 4.2 and the following paragraph.