Contents of Appendix

A Useful Lemmas

A.1 Remarks on Assumption 1

Here we show that the assumption on the functional class of f is satisfied if f is linear and injective, whenever the support of $\mathbb{P}_U$ has non-empty interior. Recall Assumption 1.

Assumption 1. *Let U be a p-dimensional random vector. Following [3], we assume that the interior of the support of $\mathbb{P}_U$ is a non-empty subset of $\mathbb{R}^p$, and that f is a full row rank polynomial.*[7]

Denote the support of $\mathbb{P}_U, \mathbb{P}_X$ as $\mathbb{U}, \mathbb{X}$ respectively. Let $\mathbb{U}^\circ$ be the interior of $\mathbb{U}$.

Lemma 2. *Suppose $\mathbb{U}^\circ$ is a non-empty subset of $\mathbb{R}^p$. If $f : \mathbb{U} \to \mathbb{X}$ is linear and injective, then it must be a full row rank polynomial.*

Proof. Since f is linear, it can be written as $f(U) = UH + h$ for some $H \in \mathbb{R}^{p\times n}$ and $h \in \mathbb{R}^p$. If H is not of full row rank, then there exists a non-zero vector $V \in \mathbb{R}^p$ such that $VH = 0$. Let $U \in \mathbb{U}^\circ$, then there exists $\epsilon > 0$ such that $U + \epsilon V \in \mathbb{U}$. We have $f(U + \epsilon V) = f(U)$, which violates f being injective. Therefore H must have full row rank. □

A.2 Proof of Lemma 1

The proof of Lemma 1 follows from [3]. For completeness, we present a concise proof here. Then we state a few remarks. Recall Lemma 1.

Lemma 1. *Under Assumption 1, we can identify the dimension p of U as well as its linear transformation $U\Lambda + b$ for some non-singular matrix Λ and vector b. In fact, with observational data, we can only identify U up to such linear transformations.*

Proof. We solve for the smallest integer $\hat{p}$ such that there exists a full row rank polynomial $\hat{f} : \mathbb{R}^{\hat{p}} \to \mathbb{R}^n$ where $\hat{U} := \hat{f}^{-1}(X)$ for $X \in \mathbb{X}$ has non-empty support $\hat{\mathbb{U}}^\circ \subseteq \mathbb{R}^{\hat{p}}$. In other words, denote all pairs of $\mathbb{P}_U, f$ that satisfy Assumption 1 as $\mathcal{F}_p$, we solve for

$$\min_{(\mathbb{P}_{\hat{U}}, \hat{f})\in\mathcal{F}_{\hat{p}}} \hat{p} \quad \text{subject to } \mathbb{P}_{\hat{f}(\hat{U})} = \mathbb{P}_X. \tag{4}$$

Note that $\hat{f}(\hat{U}) = X = f(U)$ for all $U \in \mathbb{U}$. Since $\hat{f}, f$ are full row rank polynomials, there exist full row rank matrices $\hat{H} \in \mathbb{R}^{(p+\ldots+p^{\hat{d}})\times n}, H \in \mathbb{R}^{(p+\ldots+p^d)\times n}$ and vectors $\hat{h}, h \in \mathbb{R}^n$ such that

$$(\hat{U}, \bar{\otimes}\hat{U}^2, ..., \bar{\otimes}\hat{U}^d)\hat{H} + \hat{h} = \hat{f}(\hat{U}) = X = f(U) = (U, \bar{\otimes}U^2, ..., \bar{\otimes}U^d)H + h. \tag{5}$$

Since $\hat{H}, H$ are of full rank, they have pseudo-inverses $\hat{H}^\dagger, H^\dagger$ such that $\hat{H}\hat{H}^\dagger = \mathbf{I}_{p+\ldots+p^{\hat{d}}}$ and $HH^\dagger = \mathbf{I}_{p+\ldots+p^d}$. Multiplying $\hat{H}^\dagger$ to Eq. (5), we have

$$(\hat{U}, \bar{\otimes}\hat{U}^2, ..., \bar{\otimes}\hat{U}^d) = (U, \bar{\otimes}U^2, ..., \bar{\otimes}U^d)H\hat{H}^\dagger + (h - \hat{h})\hat{H}^\dagger.$$

Therefore $\hat{U}$ can be written as a polynomial of U, i.e., $\hat{U} = poly_1(U)$. Similarly, we have $U = poly_2(\hat{U})$. Therefore $U = poly_2(poly_1(U))$ for all $U \in \mathbb{U}$. Since $\mathbb{U}^\circ$ is non-empty, we know that $U = poly_2(poly_1(U))$ on some open set. By the fundamental theorem of algebra [14], we know that $poly_1$ and $poly_2$ must have degree 1. Thus $\hat{U} = U\Lambda + b$ for some full row rank matrix Λ and vector b. Since $\Lambda \in \mathbb{R}^{p\times\hat{p}}$ is of full row rank, it indicates that $p \leq \hat{p}$. Since $\mathbb{P}_U, f \in \mathcal{F}_p$ satisfy $\mathbb{P}_{f(U)} = \mathbb{P}_X$, by Eq. (4), we must have $\hat{p} \leq p$. Thus $\hat{p} = p$ and $\hat{U} = U\Lambda + b$ for some non-singular matrix Λ and vector b.

This proof also shows that we can only identify U up to such linear transformations with observational data. Since for any non-singular matrix Λ and vector b, let $\hat{f}(\hat{U}) = f((\hat{U} - b)\Lambda^{-1})$. We have $\mathbb{P}_{\hat{U}}, \hat{f}$ satisfy Assumption 1 and they generate the same observational data. □

[7] There exists some integer d, a full row rank $H \in \mathbb{R}^{(p+\ldots+p^d)\times n}$ and a vector $h \in \mathbb{R}^n$ such that $f(U) = (U, \bar{\otimes}U^2, ..., \bar{\otimes}U^d)H + h$, where $\bar{\otimes}U^k$ denotes the size-p^k vector with degree-k polynomials of U as its entries.

Remark 1. *With observational data $X = f(U) \in \mathcal{D}$, we can identify $\hat{U} = \hat{g}(X)$ such that $\hat{U} = U\Lambda + b$ for non-singular Λ. Then for any interventional data $X = f(U) \in \mathcal{D}^I$, the analytic continuation of $\hat{g}$ to $\mathcal{D}^I$ satisfies $\hat{U} := \hat{g}(X) = U\Lambda + b$ for all $X \in \mathcal{D}^I$.*

Proof. The proof follows immediately by writing $\hat{g}, f^{-1}$ as polynomial functions. □

Next, we discuss identifiability of the underlying DAG $\mathcal{G}$. First, we give an example showing that any causal DAG can explain the observational data.

Example 4. *Suppose the ground-truth DAG is an empty graph $\mathcal{G} = \varnothing$. With observational data alone, any DAG can explain the data.*

Proof. Let $\hat{\mathcal{G}}$ be an arbitrary DAG with topological order $\tau(1), ..., \tau(p)$, i.e., $\tau(j) \in \mathrm{pa}_{\hat{\mathcal{G}}}(\tau(i))$ only if $j < i$. Let Λ be the permutation matrix such that $\hat{U} = U\Lambda$ satisfies $\hat{U}_{\tau(i)} = U_i$ for any $i \in [p]$. Then $\hat{U}$ factorizes as $\mathbb{P}(\hat{U}) = \mathbb{P}(U) = \prod_{i=1}^p \mathbb{P}(U_i) = \prod_{i=1}^p \mathbb{P}(\hat{U}_{\tau(i)})$. This implies $\hat{U}_{\tau(i)} \perp\!\!\!\perp \hat{U}_{\tau(j)}$ for $j \leq i-1$. Therefore $\mathbb{P}(\hat{U}_{\tau(i)}) = \mathbb{P}(\hat{U}_{\tau(i)} \mid \hat{U}_{\mathrm{pa}_{\hat{\mathcal{G}}(i)}})$ and $\mathbb{P}(\hat{U}) = \prod_{i=1}^p \mathbb{P}(\hat{U}_{\tau(i)} \mid \hat{U}_{\mathrm{pa}_{\hat{\mathcal{G}}(i)}})$ factorizes with respect to $\hat{\mathcal{G}}$. Thus $\hat{\mathcal{G}}$ can explain the data. □

Therefore with observational data alone, we cannot identify the underlying DAG $\mathcal{G}$ up to any nontrivial equivalence class. In [3], it was shown that with a *do* intervention[8] per latent node and assuming the interior of the support of the non-targeted variables is non-empty, one can identify U up to a finer class of linear transformations. Namely, one can identify U up to CD-equivalence (permutation and element-wise affine transformation); see Definition 1. Then, assuming for example faithfulness and influentiality [57], one can identify $\mathcal{G}$.

While several extensions beyond do-interventions are discussed in [3], they all involve manipulating the *support* of the intervention targets. In the case where the support of the intervention targets remains unchanged (e.g., additive Gaussian SCMs with shift interventions), a completely new approach and theory needs to be developed.

B Proof of Identifiability with Soft Interventions

In this section, we provide the proofs for the results in Section 4. While our main focus is on general types of soft interventions, our results also apply to hard interventions which include do-interventions as a special case.

Notation. We let e_i denote the indicator vector with the i-th entry equal to one and all other entries equal to zero. To be consistent with other notation in the paper, let $e_i \in \mathbb{R}^p$ be a row vector. We call $j \in \mathrm{ch}_{\mathcal{G}}(i)$ a *maximal child* of i if $\mathrm{pa}_{\mathcal{G}}(j) \cap \mathrm{de}_{\mathcal{G}}(i) = \varnothing$. Denote the set of all maximal children of i as $\mathrm{mch}_{\mathcal{G}}(i)$. For node i, define $\overline{\mathrm{de}}_{\mathcal{G}}(i) := \mathrm{de}_{\mathcal{G}}(i) \cup \{i\}$. Given a DAG $\mathcal{G}$, we denote the transitive closure of $\mathcal{G}$ by $\mathcal{TS}(\mathcal{G})$, i.e., $i \to j \in \mathcal{TS}(\mathcal{G})$ if and only if there is a directed path from i to j in $\mathcal{G}$.

B.1 Faithfulness Assumptions

We start by discussing previous interventional faithfulness assumptions. Prior interventional faithfulness assumptions [57, 66, 24] vary by a few technicalities; but they all assume that all causal variables are observed (causal sufficiency), and, more importantly, that intervening on a node will always change the marginal of its descendants. In particular, [57] (Definition 2, called "influentiality") only made this assumption and showed that the causal graph is identifiable up to its transitive closure by detecting marginal changes. [57] showed that their algorithm consistently identifies the full causal graph by assuming additionally that intervening on a node changes the conditional distribution of its direct children giving its neighbors (details can be found in Assumption 4.5 of [66]). A similar notion was also introduced in [24] where they made further assumptions regarding changes in the conditional distributions.

[8] Do interventions are a special type of hard interventions where the intervention target collapses to one specific value.

We now show our linear interventional faithfulness (Assumption 2) is satisfied by a large class of nonlinear SCMs and soft interventions. Recall Assumption 2.

Assumption 2. *Intervention I with target i satisfies* linear interventional faithfulness *if for every $j \in \{i\} \cup \mathrm{ch}_{\mathcal{G}}(i)$ such that $\mathrm{pa}_{\mathcal{G}}(j) \cap \mathrm{de}_{\mathcal{G}}(i) = \varnothing$, it holds that $\mathbb{P}(U_j + U_S C^\top) \neq \mathbb{P}^I(U_j + U_S C^\top)$ for all constant vectors $C \in \mathbb{R}^{|S|}$, where $S = [p] \setminus (\{j\} \cup \mathrm{de}_{\mathcal{G}}(i))$.*

In Example 2, we discussed a 2-node graph where Assumption 2 is satisfied. This example can be extended in the following way, which subsumes the case in Example 3.

Example 5. *Consider an SCM with additive noise, where each mechanism $\mathbb{P}(U_k \mid U_{\mathrm{pa}_{\mathcal{G}}(k)})$ is specified by $U_k = s_k(U_{\mathrm{pa}_{\mathcal{G}}(k)}) + \epsilon_k$, where ϵ_k for $k \in [p]$ are independent exogenous noise variables. Assumption 2 is satisfied if I only changes the variance of ϵ_i and s_j is a quadratic function with non-zero coefficient of U_i^2 for each $j \in \mathrm{mch}_{\mathcal{G}}(i)$.*

Proof. If $j = i$ in Assumption 2, then $S = [p] \setminus \overline{\mathrm{de}}_{\mathcal{G}}(i) \supset \mathrm{pa}_{\mathcal{G}}(i)$. Since $U_S \perp\!\!\!\perp \epsilon_i$, we have

$$\mathrm{Var}(U_i + U_S C^\top) = \mathrm{Var}(\epsilon_i) + \mathrm{Var}(s_i(U_{\mathrm{pa}_{\mathcal{G}}(i)}) + U_S C^\top).$$

Note that $\mathbb{P}^I$ does not change the joint distribution of U_S, and therefore

$$\mathrm{Var}_{\mathbb{P}}(s_i(U_{\mathrm{pa}_{\mathcal{G}}(i)}) + U_S C^\top) = \mathrm{Var}_{\mathbb{P}^I}(s_i(U_{\mathrm{pa}_{\mathcal{G}}(i)}) + U_S C^\top).$$

By $\mathrm{Var}_{\mathbb{P}}(\epsilon_i) \neq \mathrm{Var}_{\mathbb{P}^I}(\epsilon_i)$, we then know that$\mathrm{Var}_{\mathbb{P}}(U_i + U_S C^\top) \neq \mathrm{Var}_{\mathbb{P}^I}(U_i + U_S C^\top)$. Thus $\mathbb{P}(U_i + U_S C^\top) \neq \mathbb{P}^I(U_i + U_S C^\top)$.

If $j \neq i$ in Assumption 2, then by linearity of expectation $\mathbb{E}(U_j + U_S C^\top) = \mathbb{E}(U_j) + \mathbb{E}(U_S) C^\top$. Note that $S = [p] \setminus (\{j\} \cup \mathrm{de}_{\mathcal{G}}(i)) = [p] \setminus \mathrm{de}_{\mathcal{G}}(i)$, and therefore $\mathbb{E}_{\mathbb{P}}(U_S) = \mathbb{E}_{\mathbb{P}^I}(U_S)$. Next we show that $\mathbb{E}_{\mathbb{P}}(U_j) \neq \mathbb{E}_{\mathbb{P}^I}(U_j)$. Once this is proven, then we have that $\mathbb{E}_{\mathbb{P}}(U_j + U_S C^\top) \neq \mathbb{E}_{\mathbb{P}^I}(U_j + U_S C^\top)$, which concludes the proof for $\mathbb{P}(U_j + U_S C^\top) \neq \mathbb{P}^I(U_j + U_S C^\top)$.

Since s_j is a quadratic function of U_i, suppose the coefficient of U_i^2 in s_j is $\beta \neq 0$. Then

$$\begin{aligned}
\mathbb{E}(U_j) - \mathbb{E}(\epsilon_j) &= \mathbb{E}(U_j - \epsilon_j) \\
&= \mathbb{E}\big(s_{j,0}(U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) + s_{j,1}(U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) \cdot U_i + \beta U_i^2\big) \\
&= \mathbb{E}\big(s_{j,0}(U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) + s'_{j,1}(U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}, U_{\mathrm{pa}_{\mathcal{G}}(i)}) \cdot \epsilon_i + \beta \epsilon_i^2\big) \\
&= \mathbb{E}\big(s_{j,0}(U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}})\big) + \mathbb{E}\big(s'_{j,1}(U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}, U_{\mathrm{pa}_{\mathcal{G}}(i)}) \cdot \epsilon_i\big) + \beta \mathbb{E}(\epsilon_i^2),
\end{aligned} \tag{6}$$

for some functions $s_{j,0}$, $s_{j,1}$ and $s'_{j,1}$. Since $\mathrm{pa}_{\mathcal{G}}(j) \cap \mathrm{de}_{\mathcal{G}}(i) = \varnothing$, we know that $\mathbb{P}^I$ will not change the joint distribution of $U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}$ and that $\epsilon_i \perp\!\!\!\perp U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}, U_{\mathrm{pa}_{\mathcal{G}}(i)}$. Therefore we have

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\big(s_{j,0}(U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}})\big) &= \mathbb{E}_{\mathbb{P}^I}\big(s_{j,0}(U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}})\big), \\
\mathbb{E}_{\mathbb{P}}\big(s'_{j,1}(U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}, U_{\mathrm{pa}_{\mathcal{G}}(i)}) \cdot \epsilon_i\big) &= \mathbb{E}_{\mathbb{P}}\big(s'_{j,1}(U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}, U_{\mathrm{pa}_{\mathcal{G}}(i)})\big) \cdot \mathbb{E}_{\mathbb{P}}(\epsilon_i) \\
&= \mathbb{E}_{\mathbb{P}^I}\big(s'_{j,1}(U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}, U_{\mathrm{pa}_{\mathcal{G}}(i)})\big) \cdot \mathbb{E}_{\mathbb{P}^I}(\epsilon_i) \\
&= \mathbb{E}_{\mathbb{P}^I}\big(s'_{j,1}(U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}, U_{\mathrm{pa}_{\mathcal{G}}(i)}) \cdot \epsilon_i\big).
\end{aligned}$$

By $\mathbb{E}_{\mathbb{P}}(\epsilon_j) = \mathbb{E}_{\mathbb{P}^I}(\epsilon_j)$, $\mathbb{E}_{\mathbb{P}}(\epsilon_i^2) \neq \mathbb{E}_{\mathbb{P}^I}(\epsilon_i^2)$ and Eq. (6), we have $\mathbb{E}_{\mathbb{P}}(U_j) \neq \mathbb{E}_{\mathbb{P}^I}(U_j)$, which concludes the proof. □

This example shows how we may check $\mathbb{P}(U_j + U_S C^\top) \neq \mathbb{P}^I(U_j + U_S C^\top)$ by examining the mean and variance of $U_j + U_S C^\top$. In general, this can be extended to checking any finite moments of $U_j + U_S C^\top$ as stated in the following lemma.

Lemma 3. *Assumption 2 is satisfied if for each $i \in [p]$ one of the following conditions holds:*

(1) if $\mathbb{E}_{\mathbb{P}}(U_i \mid U_{\mathrm{pa}_{\mathcal{G}}(i)}) = \mathbb{E}_{\mathbb{P}^I}(U_i \mid U_{\mathrm{pa}_{\mathcal{G}}(i)})$, then there exits an integer $m > 1$ such that

$$\mathbb{E}_{\mathbb{P}}(U_i^m \mid U_{\mathrm{pa}_{\mathcal{G}}(i)}) \neq \mathbb{E}_{\mathbb{P}^I}(U_i^m \mid U_{\mathrm{pa}_{\mathcal{G}}(i)}),$$

and the smallest m that satisfies this also satisfies $\mathbb{E}_{\mathbb{P}}(U_i^m) \neq \mathbb{E}_{\mathbb{P}^I}(U_i^m)$. In addition, for all $j \in \mathrm{mch}_{\mathcal{G}}(i)$, it holds that $\mathbb{E}_{\mathbb{P}}(U_j) \neq \mathbb{E}_{\mathbb{P}^I}(U_j)$;

(2) if $\mathbb{E}_{\mathbb{P}}(U_i) \neq \mathbb{E}_{\mathbb{P}^I}(U_i)$*, then for all* $j \in \mathrm{mch}_{\mathcal{G}}(i)$*, there exists an integer* $m > 1$ *such that*

$$\mathbb{E}_{\mathbb{P}}((U_j + c_j U_i)^m \mid U_{S\setminus\{i\}}) \neq \mathbb{E}_{\mathbb{P}^I}((U_j + c_j U_i)^m \mid U_{S\setminus\{i\}}),$$

where S *is as defined in Assumption 2, and the smallest* m *that satisfies this also satisfies* $\mathbb{E}_{\mathbb{P}}((U_j + c_j U_i)^m) \neq \mathbb{E}_{\mathbb{P}^I}((U_j + c_j U_i)^m)$*, where*

$$c_j = -\frac{(\mathbb{E}_{\mathbb{P}}(U_j) - \mathbb{E}_{\mathbb{P}^I}(U_j))}{(\mathbb{E}_{\mathbb{P}}(U_i) - \mathbb{E}_{\mathbb{P}^I}(U_i))}.$$

Proof. Suppose (1) holds true. If $j = i$ in Assumption 2, then $\mathbb{P}(U_S) = \mathbb{P}^I(U_S)$ for $S = [p]\setminus\overline{\mathrm{de}}_{\mathcal{G}}(i)$, and

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}}\big((U_i + U_S C^\top)^m\big) \\
&= \mathbb{E}_{\mathbb{P}}(U_i^m) + \sum_{l=0}^{m-1} \binom{m}{l} \mathbb{E}_{\mathbb{P}}\big(U_i^l (U_S C^\top)^{m-l}\big) \\
&= \mathbb{E}_{\mathbb{P}}(U_i^m) + \sum_{l=0}^{m-1} \binom{m}{l} \mathbb{E}_{\mathbb{P}}\Big(\mathbb{E}_{\mathbb{P}}(U_i^l | U_S) \cdot (U_S C^\top)^{m-l}\Big) \quad (\textit{law of total expectation}) \\
&= \mathbb{E}_{\mathbb{P}}(U_i^m) + \sum_{l=0}^{m-1} \binom{m}{l} \mathbb{E}_{\mathbb{P}}\Big(\mathbb{E}_{\mathbb{P}}(U_i^l | U_{\mathrm{pa}_{\mathcal{G}}(i)}) \cdot (U_S C^\top)^{m-l}\Big) \quad (\text{since } U_i \perp\!\!\!\perp U_{S\setminus \mathrm{pa}_{\mathcal{G}}(i)} \mid U_{\mathrm{pa}_{\mathcal{G}}(i)}) \\
&\neq \mathbb{E}_{\mathbb{P}^I}(U_i^m) + \sum_{l=0}^{m-1} \binom{m}{l} \mathbb{E}_{\mathbb{P}}\Big(\mathbb{E}_{\mathbb{P}^I}(U_i^l | U_{\mathrm{pa}_{\mathcal{G}}(i)}) \cdot (U_S C^\top)^{m-l}\Big) \\
&= \mathbb{E}_{\mathbb{P}^I}(U_i^m) + \sum_{l=0}^{m-1} \binom{m}{l} \mathbb{E}_{\mathbb{P}^I}\Big(\mathbb{E}_{\mathbb{P}^I}(U_i^l | U_{\mathrm{pa}_{\mathcal{G}}(i)}) \cdot (U_S C^\top)^{m-l}\Big) = \mathbb{E}_{\mathbb{P}^I}\big((U_i + U_S C^\top)^m\big),
\end{aligned}$$

where the inequality is because of $\mathbb{E}_{\mathbb{P}}(U_i^m) \neq \mathbb{E}_{\mathbb{P}^I}(U_i^m)$ and $\mathbb{E}_{\mathbb{P}}(U_i^l | U_{\mathrm{pa}_{\mathcal{G}}(i)}) = \mathbb{E}_{\mathbb{P}^I}(U_i^l | U_{\mathrm{pa}_{\mathcal{G}}(i)})$ for any $l < m$. Therefore $\mathbb{P}(U_i + U_S C^\top) \neq \mathbb{P}^I(U_i + U_S C^\top)$.

If $j \neq i$ in Assumption 2, then $\mathbb{E}_{\mathbb{P}}(U_j) \neq \mathbb{E}_{\mathbb{P}^I}(U_j)$ implies $\mathbb{E}_{\mathbb{P}}(U_j + U_S C^\top) \neq \mathbb{E}_{\mathbb{P}^I}(U_j + U_S C^\top)$, which proves that $\mathbb{P}(U_j + U_S C^\top) \neq \mathbb{P}^I(U_j + U_S C^\top)$.

Suppose (2) holds true. If $j = i$ in Assumption 2, then $\mathbb{E}_{\mathbb{P}}(U_i) \neq \mathbb{E}_{\mathbb{P}^I}(U_i)$ implies $\mathbb{E}_{\mathbb{P}}(U_i + U_S C^\top) \neq \mathbb{E}_{\mathbb{P}^I}(U_i + U_S C^\top)$, which proves that $\mathbb{P}(U_i + U_S C^\top) \neq \mathbb{P}^I(U_i + U_S C^\top)$.

If $j \neq i$ in Assumption 2, then for $C \in \mathbb{R}^{|S|}$, if the coordinate for U_i is not c_j, then $\mathbb{E}_{\mathbb{P}}(U_i + U_S C^\top) = \mathbb{E}_{\mathbb{P}}(U_i) + \mathbb{E}_{\mathbb{P}}(U_S)C^\top \neq \mathbb{E}_{\mathbb{P}^I}(U_i) + \mathbb{E}_{\mathbb{P}^I}(U_S)C^\top = \mathbb{E}_{\mathbb{P}^I}(U_i + U_S C^\top)$, since $\mathbb{E}_{\mathbb{P}}(U_{S\setminus\{i\}}) = \mathbb{E}_{\mathbb{P}^I}(U_{S\setminus\{i\}})$. If the coordinate for U_i in C is c_j, denote $U_S C^\top = U_{S\setminus\{i\}} C_{-j}^\top + c_j U_i$, and then similar to above we obtain

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}}\big((U_j + U_S C^\top)^m\big) \\
&= \mathbb{E}_{\mathbb{P}}\big((U_j + c_j U_i + U_{S\setminus\{i\}} C_{-j}^\top)^m\big) \\
&= \mathbb{E}_{\mathbb{P}}\big((U_i + c_j U_i)^m\big) + \sum_{l=0}^{m-1} \binom{m}{l} \mathbb{E}_{\mathbb{P}}\big((U_i + c_j U_i)^l (U_{S\setminus\{i\}} C_{-j}^\top)^{m-l}\big) \\
&= \mathbb{E}_{\mathbb{P}}\big((U_i + c_j U_i)^m\big) + \sum_{l=0}^{m-1} \binom{m}{l} \mathbb{E}_{\mathbb{P}}\Big(\mathbb{E}_{\mathbb{P}}\big((U_i + c_j U_i)^l | U_{S\setminus\{i\}}\big) \cdot (U_{S\setminus\{i\}} C_{-j}^\top)^{m-l}\Big) \\
&\neq \mathbb{E}_{\mathbb{P}^I}\big((U_i + c_j U_i)^m\big) + \sum_{l=0}^{m-1} \binom{m}{l} \mathbb{E}_{\mathbb{P}}\Big(\mathbb{E}_{\mathbb{P}^I}\big((U_i + c_j U_i)^l | U_{S\setminus\{i\}}\big) \cdot (U_{S\setminus\{i\}} C_{-j}^\top)^{m-l}\Big) \\
&= \mathbb{E}_{\mathbb{P}^I}\big((U_i + c_j U_i)^m\big) + \sum_{l=0}^{m-1} \binom{m}{l} \mathbb{E}_{\mathbb{P}^I}\Big(\mathbb{E}_{\mathbb{P}^I}\big((U_i + c_j U_i)^l | U_{S\setminus\{i\}}\big) \cdot (U_{S\setminus\{i\}} C_{-j}^\top)^{m-l}\Big) \\
&= \mathbb{E}_{\mathbb{P}^I}\big((U_j + U_S C^\top)^m\big).
\end{aligned}$$

Thus $\mathbb{P}(U_j + U_S C^\top) \neq \mathbb{P}^I(U_j + U_S C^\top)$, which completes the proof. □

This lemma gives a sufficient condition for Assumption 2 to hold. Since it involves only finite moments of the variables, one can easily check if this is satisfied for a given SCM associated with soft interventions. Note that Example 5 satisfies the first condition of Lemma 3 for $m = 2$.

Next we show that Assumption 3 is satisfied on a tree graph if Assumption 2 holds, under mild regularity conditions such as that the interventional support lies within the observational support. Recall Assumption 3.

Assumption 3. *For every edge $i \to j \in \mathcal{G}$, there do* not *exist constants $c_j, c_k \in \mathbb{R}$ for $k \in S$ such that $U_i \perp\!\!\!\perp U_j + c_j U_i \mid \{U_l\}_{l \in \mathrm{pa}_{\mathcal{G}}(j)\setminus(S\cup\{i\})}, \{U_k + c_k U_i\}_{k\in S}$, where $S = \mathrm{pa}_{\mathcal{G}}(j) \cap \mathrm{de}_{\mathcal{G}}(i)$.*

Lemma 4. *Suppose $\mathcal{G}$ is a polytree and Assumption 2 holds for an intervention I targeting node i. Then for any edge $i \to j \in \mathcal{G}$, Assumption 3 holds if*[9]

$$\mathbb{P}(U_i = u \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) = 0 \quad \Rightarrow \quad \mathbb{P}^I(U_i = u \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) = 0, \tag{7}$$

for almost every u and all realizations of $U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}$.

Proof. Suppose $\mathcal{G}$ is a tree graph and Assumption 2 holds for I targeting i. For any edge $i \to j \in \mathcal{G}$, since there is only one undirected path between i and j, we have $S = \mathrm{pa}_{\mathcal{G}}(j) \cap \mathrm{de}_{\mathcal{G}}(i) = \varnothing$. Therefore we only need to show that U_i and $U_j + c_j U_i$ are not conditionally independent given $U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}$ for any $c_j \in \mathbb{R}$.

The regularity condition in Eq. (7) ensures that

$$\int_{\mathbb{P}(U_i = r \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}})\neq 0} \mathbb{P}^I(U_i = r \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) dr = 1, \tag{8}$$

for any realization of $U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}$.

Suppose U_i and $U_j + c_j U_i$ are conditionally independent given $U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}$. Then for any $l \in \mathbb{R}$ and realization of $U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}, U_i$ (denote the realization of U_i as r),

$$\begin{aligned}\mathbb{P}(U_j + c_j U_i = l \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) &= \mathbb{P}(U_j + c_j U_i = l \mid U_i = r, U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) \\ &= \mathbb{P}(U_j = l - c_j r \mid U_i = r, U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}).\end{aligned}$$

Since this is true for any r with $\mathbb{P}(U_i = r \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) \neq 0$, by Eq. (8), we have

$$\begin{aligned}&\mathbb{P}(U_j + c_j U_i = l \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) \\ =& \int_{\mathbb{P}(U_i = r \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}})\neq 0} \mathbb{P}(U_j + c_j U_i = l \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) \cdot \mathbb{P}^I(U_i = r \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) dr \\ =& \int_{\mathbb{P}(U_i = r \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}})\neq 0} \mathbb{P}(U_j = l - c_j r \mid U_i = r, U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) \cdot \mathbb{P}^I(U_i = r \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) dr.\end{aligned}$$

Note that $\mathbb{P}(U_j \mid U_i, U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) = \mathbb{P}^I(U_j \mid U_i, U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}})$ since I targets i, and we therefore have

$$\begin{aligned}&\mathbb{P}(U_j + c_j U_i = l \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) \\ =& \int_{\mathbb{P}(U_i = r \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}})\neq 0} \mathbb{P}^I(U_j = l - c_j r \mid U_i = r, U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) \cdot \mathbb{P}^I(U_i = r \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) dr \\ =& \int_{\mathbb{P}(U_i = r \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}})\neq 0} \mathbb{P}^I(U_j = l - c_j r, U_i = r \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) dr \\ =& \int_{\mathbb{P}^I(U_i = r \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}})\neq 0} \mathbb{P}^I(U_j = l - c_j r, U_i = r \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) dr \\ =& \mathbb{P}^I(U_j + c_j U_i = l \mid U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}),\end{aligned}$$

where the second-to-last equality uses the regularity condition in Eq. (7).

[9] For simplicity, we assume U is continuous and treat $\mathbb{P}$ as the density. For discrete U, the proofs extend by replacing $\int$ with $\sum$.

Since $\mathrm{pa}_{\mathcal{G}}(j) \cap \mathrm{de}_{\mathcal{G}}(i) = \varnothing$, it holds that $\mathbb{P}(U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}) = \mathbb{P}^I(U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}})$, and thus $\mathbb{P}(U_j + c_j U_i) = \mathbb{P}^I(U_j + c_j U_i)$, which is a contradiction to linear interventional faithfulness of I. Therefore, we must have that U_i and $U_j + c_j U_i$ are not conditionally independent given $U_{\mathrm{pa}_{\mathcal{G}}(j)\setminus\{i\}}$, which completes the proof. □

Essentially, Assumption 2 guarantees influentiality and Assumption 3 guarantees adjacency faithfulness. These assumptions differ from existing faithfulness conditions (c.f., [57, 58, 67]) due to the fact that we can only observe a linear mixing of the causal variables.

B.2 Summary of representations

In the remainder of this appendix, we will develop a series of representations which are increasingly related to the underlying representation U. These representations are summarized in Table 1.

Symbol	Definition	Section
U		Section 2
X	$X = U\Lambda + b, \quad \Lambda \in \mathbb{R}^{p\times p}, b \in \mathbb{R}^p$	Section 2
$\tilde{U}$	$\tilde{U} = U\tilde{\Gamma} + \tilde{c}, \quad \tilde{\Gamma} = \Lambda\Pi, \tilde{c} = b\Pi$ for $\Pi \in \mathbb{R}^{p\times p}$	Appendix B.3.1
$\hat{U}$	$\hat{U} = U\hat{\Gamma} + \hat{c}, \quad \hat{\Gamma} = \tilde{\Gamma}\hat{R}, \hat{c} = \tilde{c}\hat{R}$ for $\hat{R} \in \mathbb{R}^{p\times p}$ upp. tri.	Appendix B.3.2
$\bar{U}$	$\bar{U} = U\bar{\Gamma} + \bar{c}, \quad \bar{\Gamma} = \hat{\Gamma}\bar{R}, \bar{c} = \hat{c}\bar{R}$ for $\bar{R} \in \mathbb{R}^{p\times p}$ upp. tri.	Appendix B.4

Table 1: **Representations of U that are used in this appendix.** Note that, under Assumption 1, we can assume $X = U\Lambda + b$ without loss of generality, by Lemma 1 and Remark 1.

B.3 Proof of Theorem 1

In the main text (Section 4.2), we laid out an illustrative procedure to identify the transitive closure of $\mathcal{G}$ when we consider a simpler setting with $K = p$. This process relies on iteratively finding source nodes of $\mathcal{G}$. In the generalized setting with $K \geq p$, the proof works in the reversed way, where we iteratively identify the sink nodes[10] of $\mathcal{G}$.

In Section B.3.1, we introduce the concept of a *topological representation*: a representation $\tilde{U}$ of the data for which marginal distributions change in a way consistent with an assignment $\rho_1, \ldots, \rho_p$ of intervention targets. In Lemma 5, we show that under Assumptions 1 and 2, a topological representation is guaranteed to exist. In Lemma 6, we show that any topological representation is also topologically consistent in a natural way with the underlying representation U.

In Section B.3.2, we consider transforming a topological representation $\tilde{U}$ into a different topological representation $\hat{U}$. For any such representation $\hat{U} = \tilde{U}\hat{R}'$, we define an associated graph $\hat{\mathcal{G}}^{\hat{R}'}$. In Lemma 7, we show that picking $\hat{R}$ so that $\hat{\mathcal{G}}^{\hat{R}'}$ has the fewest edges will yield that $\hat{\mathcal{G}}^{\hat{R}} = \mathcal{TS}(\mathcal{G}_\tau)$.

Together, these results are used to prove Theorem 1: that we can identify $\mathcal{G}$ up to transitive closure.

B.3.1 Topologically ordered representations

We begin by introducing the concept of a topological representation.

Definition 2. *Suppose $X = U\Lambda + b$. Let $\Pi \in \mathbb{R}^{p\times p}$ be a non-singular matrix, let $\tilde{U} = X\Pi$, and let $\rho_1, \ldots, \rho_p \in [K]$. We call $\tilde{U}$ a* topological representation *of X with intervention targets $\rho_1, \ldots, \rho_p$ if the following two conditions are satisfied for all $j \in [p]$:*

(Condition 1) $\mathbb{P}(\tilde{U}_j) \neq \mathbb{P}^{I_{\rho_j}}(\tilde{U}_j)$.

(Condition 2) $\mathbb{P}(\tilde{U}_{1:j-1}C^\top) = \mathbb{P}^{I_{\rho_j}}(\tilde{U}_{1:j-1}C^\top)$ *and for any* $C \in \mathbb{R}^{j-1}$.

Here, $\mathbb{P}(\tilde{U}), \mathbb{P}^{I_k}(\tilde{U})$ are the induced distributions for $\tilde{U}$ when $X \sim \mathbb{P}_X$ and $X \sim \mathbb{P}_X^{I_k}$, respectively.

[10] A sink node is a node without children

The next result shows that a topological representation always exists. In particular, we show that a topological representation can be recovered simply be re-ordering the nodes of $\mathcal{G}$.

Lemma 5. *Suppose that Assumption 2 hold. Then, there exists a topological representation of X.*

Proof. Assume without loss of generality that $\mathcal{G}$ has topological order $\tau = (1, 2, ..., p)$, i.e., $i \to j \in \mathcal{G}$ only if $i < j$. Let $\Pi = \Lambda^{-1}$, then $\tilde{U} = U + \tilde{c}$ for constant vector $\tilde{c} = b\Pi$. Set $\rho_1, ..., \rho_p$ to be such that $T(I_{\rho_j}) = j$ for $j \in [p]$. Let $j \in [p]$ and $C \in \mathbb{R}^{j-1}$.

Condition 1. Since I_{ρ_j} targets U_j, by Assumption 2, we have $\mathbb{P}(U_j) \neq \mathbb{P}^{I_{\rho_j}}(U_j)$, and thus $\mathbb{P}(\tilde{U}_j) \neq \mathbb{P}^{I_{\rho_j}}(\tilde{U}_j)$.

Condition 2. Since I_{ρ_j} targets U_j and $U_{1:j-1} \subset U_{[p]\setminus\overline{\mathrm{de}}_{\mathcal{G}}(j)}$, we have $\mathbb{P}(U_{1:j-1}C^\top) \neq \mathbb{P}^{I_{\rho_j}}(U_{1:j-1}C^\top)$, and thus $\mathbb{P}(\tilde{U}_{1:j-1}C^\top) \neq \mathbb{P}^{I_{\rho_j}}(\tilde{U}_{1:j-1}C^\top)$. □

Now, we show that any topological representation is also consistent with the underlying representation U up to some linear transformation which respects the topological ordering.

Lemma 6. *Suppose that Assumptions 2 hold. Let $\tilde{U} = X\Pi$ be a topological representation of X and denote $\tilde{\Gamma} = \Lambda\Pi$. Then, there exists a topological ordering τ of $\mathcal{G}$ such that for any $j \in [p]$, we have that*

$$i < j \quad \Longrightarrow \quad \tilde{\Gamma}_{\tau(j),i} = 0 \quad \text{and} \quad \tilde{\Gamma}_{\tau(j),j} \neq 0. \tag{9}$$

Proof. We prove by induction. Let $\tilde{c} = b\Pi$. Note that $\tilde{U} = U\tilde{\Gamma} + \tilde{c}$.

Base case.
Consider I_{ρ_p}. Let $T(I_{\rho_p}) = i$. We will show that i must be a sink node in $\mathcal{G}$.

Suppose i is not a sink node, and let $j \in \mathrm{mch}_{\mathcal{G}}(i)$. Since $\tilde{\Gamma}$ is nonsingular, $\mathrm{rank}(\mathrm{span}(\tilde{\Gamma}_{:,1}, ..., \tilde{\Gamma}_{:,p-1})) = p-1$. Therefore, we must have $\mathrm{span}(e_i, e_j) \cap \mathrm{span}(\tilde{\Gamma}_{:,1}, ..., \tilde{\Gamma}_{:,p-1}) \neq \{0\}$. Thus,

$$\tilde{\Gamma}\gamma^\top = ae_i^\top + be_j^\top \quad \text{for some } a, b \in \mathbb{R}, \gamma \in \mathbb{R}^p \text{ such that } a^2 + b^2 \neq 0, \gamma_p = 0$$

By Condition 2, we have that $\mathbb{P}(\tilde{U}\gamma^\top) = \mathbb{P}_{I_{\rho_p}}(\tilde{U}\gamma^\top)$. Since $\tilde{c}\gamma^\top$ is a constant, this implies that $\mathbb{P}(aU_i + bU_j) = \mathbb{P}^{I_{\rho_p}}(aU_i + bU_j)$. However, this contradicts Assumption 2. Thus, i must be a sink node, which we denote by $\tau(p)$.

We also have that $\tilde{\Gamma}_{\tau(p),i} = 0$ for any $i < p$. Otherwise suppose $\tilde{\Gamma}_{\tau(p),i} \neq 0$, then $\tilde{U}_i = U\tilde{\Gamma}_{:,i} + h_i$ can be written as $\tilde{\Gamma}_{\tau(p),i} \cdot U_{\tau(p)} + U_S C^\top + \tilde{c}_i$ with $S = [p] \setminus (\{\tau(p)\}) = [p] \setminus \overline{\mathrm{de}}_{\mathcal{G}}(\tau(p))$. By Assumption 2, we have $\mathbb{P}(\tilde{U}_i) \neq \mathbb{P}^{I_{\rho_p}}(\tilde{U}_i)$, a contradiction to Condition 2.

Induction step.
Suppose that we have proven the statement for $q \leq p$. Denote the intervention targets of $I_{\rho_q}, \ldots, I_{\rho_p}$ as $\tau(q), \ldots, \tau(p)$, respectively. Let $K = [p] \setminus \{\tau(q), \ldots, \tau(p)\}$.

Consider $I_{\rho_{q-1}}$ with $T(I_{\rho_{q-1}}) = i$. Let $\mathcal{G}_q$ denote the graph $\mathcal{G}$ after removing the nodes $\tau(q), \ldots, \tau(p)$. We will show that i must be a sink node in $\mathcal{G}_q$.

Suppose that i is not a sink node $\mathcal{G}_q$ and let j be a maximal child of i in $\mathcal{G}_q$. Since $\tilde{\Gamma}_{[p]\setminus K,[q]} = 0$, $|K| = q$, and $\tilde{\Gamma}$ is nonsingular, we have that $\tilde{\Gamma}_{K,[q]}$ is nonsingular. Thus, as above,

$$\tilde{\Gamma}\gamma^\top = ae_{i(q)}^\top + be_{j(q)}^\top \quad \text{for some } a, b \in \mathbb{R}, \gamma \in \mathbb{R}^p \text{ such that } a^2 + b^2 \neq 0, \gamma_q = \gamma_{q+1} = \ldots = \gamma_p = 0$$

where $e_{i(q)}, e_{j(q)}$ are indicator vectors in $\mathbb{R}^q$ with ones at positions of i, j in $1, ..., p$ after removing $\tau(q+1), ..., \tau(p)$, respectively. Thus, by Condition 2, we have that $\mathbb{P}(\tilde{U}\gamma^\top) = \mathbb{P}^{I_{\rho_q}}(\tilde{U}\gamma^\top)$, which contradicts Assumption 2. Therefore I_{ρ_q} targets a sink node of $\mathcal{G}_q$.

To show that $\tilde{\Gamma}_{\tau(q),i} = 0$ for any $i < q$, use $\tilde{\Gamma}_{\tau(k),i} = 0$ for all $k \geq q$ and write $\tilde{U}_i = U\tilde{\Gamma}_{:,i} + \tilde{c}_i$ as $\tilde{\Gamma}_{\tau(q-1),i} \cdot U_{\tau(q-1)} + U_S C^\top + \tilde{c}_i$ with $S = [p] \setminus \{\tau(q-1), \tau(q), ..., \tau(p)\} \subset [p] \setminus \overline{\mathrm{de}}_{\mathcal{G}}(\tau(q))$. By Assumption 2, we have $\mathbb{P}(\tilde{U}_i) \neq \mathbb{P}^{I_{\rho_q}}(\tilde{U}_i)$ if $\tilde{\Gamma}_{\tau(q),i} \neq 0$, a contradiction to Condition 2.

By induction, we have thus proven that the solution to Condition 1 and Condition 2 satisfies $i < j \Rightarrow \tilde{\Gamma}_{\tau(j),i} = 0$. Therefore $\tilde{\Gamma}_{\tau,:}$ is upper triangular. Since it is also non-singular, it must hold that $\tilde{\Gamma}_{\tau(j),j} \neq 0$. Thus Eq. (9) holds for some unknown τ. Furthermore, the proof shows that $I_{\rho_1}, ..., I_{\rho_p}$ target $U_{\tau(1)}, ..., U_{\tau(p)}$ respectively. □

B.3.2 Sparsest topological representation

In the section, we will introduce a graph associated to any topological representation. We consider picking a topological representation such that the associated graph is as sparse is possible, and we show that this choice recovers the underlying graph $\mathcal{G}$ up to transitive closure.

We begin by establishing the following property of a topological representation $\tilde{U}$, which relates ancestral relationships in the underlying graph $\mathcal{G}$ to changes in marginals of $\tilde{U}$.

Proposition 1. *Suppose that Assumptions 2 hold. Let $\tilde{U}$ be a topological representation with intervention targets $\rho_1, \ldots, \rho_p$.*

Then, for any $i < j$ such that $\tau(j) \in \mathrm{de}_{\mathcal{G}}(\tau(i))$, we must have

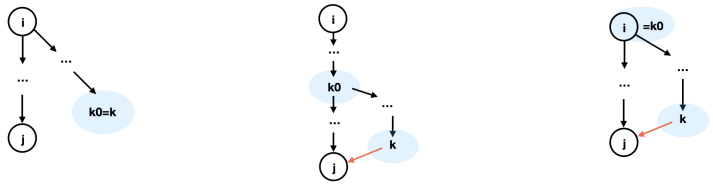
$$\mathbb{P}(\tilde{U}_j) \neq \mathbb{P}^{I_{\rho_k}}(\tilde{U}_j) \quad \textit{for some } i \leq k < j \quad \textit{such that } \tau(k) \in \overline{\mathrm{de}}_{\mathcal{G}}(\tau(i)).$$

Proof. By Lemma 6, Eq. (9), $\tilde{U}_j$ is a linear combination of $U_{\tau(1)}, ..., U_{\tau(j)}$ with nonzero coefficient of $U_{\tau(j)}$. Let k_0 be

(Case 1) the largest such that $i \leq k_0 < j$ where $\tau(k_0) \in \overline{\mathrm{de}}_{\mathcal{G}}(\tau(i))$ and the coefficient of $U_{\tau(k_0)}$ in $\tilde{U}_j$ is nonzero,

(Case 2) i, if no k_0 satisfies Case 1.

Then let $k = k_0$ if $\tau(j) \notin \mathrm{de}_{\mathcal{G}}(\tau(k_0))$; otherwise let k be such that $\tau(k) \in \overline{\mathrm{de}}_{\mathcal{G}}(\tau(k_0))$ and $\tau(j) \in \mathrm{mch}_{\mathcal{G}}(\tau(k))$ (such k exists by considering the parent of $\tau(j)$ on the longest directed path from $\tau(k_0)$ to $\tau(j)$ in $\mathcal{G}$). Figure 8 illustrates the different scenarios for k_0, k. Note that we always have $\tau(k) \in \overline{\mathrm{de}}_{\mathcal{G}}(\tau(i))$.



(A) Case 1-1 (B) Case 1-2 (C) Case 2

Figure 8: Illustration of k_0, k.

(Case 1): We first show that $\tilde{U}_j$ can be written as a linear combination of $U_{\tau(j)}, U_{\tau(k_0)}$ and U_S for $S \subset [p] \setminus \overline{\mathrm{de}}_{\mathcal{G}}(\tau(k_0))$ with nonzero coefficient for $U_{\tau(j)}, U_{\tau(k_0)}$. Consider an arbitrary $l \in [p]$. If the coefficient for $U_{\tau(l)}$ in $\hat{U}_j$ is nonzero, by Eq. (9), we have $l \leq j$. Also since k_0 is the largest, we have $l = k_0$ or $l = j$ or $l < k_0$ or $\tau(l) \notin \overline{\mathrm{de}}_{\mathcal{G}}(\tau(i))$. If $l < k_0$, then by the topological order, it holds that $\tau(l) \notin \mathrm{de}_{\mathcal{G}}(\tau(k_0))$. If $\tau(l) \notin \overline{\mathrm{de}}_{\mathcal{G}}(\tau(i))$, since $\tau(k_0) \in \overline{\mathrm{de}}_{\mathcal{G}}(\tau(i))$, it also holds that $\tau(l) \notin \overline{\mathrm{de}}_{\mathcal{G}}(\tau(k_0))$. Therefore $\tilde{U}_j$ can be written as a linear combination of $U_{\tau(j)}, U_{\tau(k_0)}$ and U_S with nonzero coefficient for $U_{\tau(j)}, U_{\tau(k_0)}$. Next, we show that $\mathbb{P}(\tilde{U}_j) \neq \mathbb{P}^{I_{\rho_k}}(\tilde{U}_j)$ by considering two subcases of Case 1.

If $\tau(j) \notin \mathrm{de}_{\mathcal{G}}(\tau(k_0))$, then $k = k_0$ (illustrated in Figure 8A). Then $S \cup \{\tau(j)\} \subset [p] \setminus \overline{\mathrm{de}}_{\mathcal{G}}(\tau(k))$. Therefore $\tilde{U}_j$ can be written as a linear combination of $U_{\tau(k)}$ and $U_{S'}$ for $S' \subset [p] \setminus \overline{\mathrm{de}}_{\mathcal{G}}(\tau(k))$ with nonzero coefficient for $U_{\tau(k)}$. By Assumption 2, we have $\mathbb{P}(\tilde{U}_j) \neq \mathbb{P}^{I_{\rho_k}}(\tilde{U}_j)$.

If $\tau(j) \in \mathrm{de}_{\mathcal{G}}(\tau(k_0))$, then since $\tau(k) \in \overline{\mathrm{de}}_{\mathcal{G}}(\tau(k_0))$, we have $\tau(k_0) \in [p] \setminus \mathrm{de}_{\mathcal{G}}(\tau(k))$ (illustrated in Figure 8B). Then we have $S \subset [p] \setminus \overline{\mathrm{de}}_{\mathcal{G}}(\tau(k_0)) \subset [p] \setminus \mathrm{de}_{\mathcal{G}}(\tau(k))$, and thus $S \cup \{\tau(k_0)\} \subset [p] \setminus \mathrm{de}_{\mathcal{G}}(\tau(k))$. In fact, $S \cup \{\tau(k_0)\}$ is a subset of $[p] \setminus (\mathrm{de}_{\mathcal{G}}(\tau(k)) \cup \{\tau(j)\})$, since $\tau(j) \in \mathrm{de}_{\mathcal{G}}(\tau(k))$ by $\tau(k) \in \mathrm{pa}_{\mathcal{G}}(\tau(i))$ (definition of k). Therefore $\tilde{U}_j$ can be written as a linear combination of $U_{\tau(j)}$ and $U_{S'}$ for $S' \subset [p] \setminus (\mathrm{de}_{\mathcal{G}}(\tau(k)) \cup \{\tau(j)\})$ with nonzero coefficient for $U_{\tau(j)}$. Since $\tau(j) \in \mathrm{mch}_{\mathcal{G}}(\tau(k))$ (definition of k), by Assumption 2, we have $\mathbb{P}(\tilde{U}_j) \neq \mathbb{P}^{I_{\rho_k}}(\tilde{U}_j)$, as I_{ρ_k} targets $U_{\tau(k)}$.

(Case 2): In this case $k_0 = i$ (illustrated in Figure 8C). Then for any $l < j$ such that $\tau(l) \in \mathrm{de}_{\mathcal{G}}(\tau(k))$, the coefficient of $U_{\tau(l)}$ in $\tilde{U}_j$ is zero. Otherwise since $\mathrm{de}_{\mathcal{G}}(\tau(k)) \subset \mathrm{de}_{\mathcal{G}}(\tau(k_0)) = \mathrm{de}_{\mathcal{G}}(\tau(i))$, it holds that $\tau(l) \in \mathrm{de}_{\mathcal{G}}(\tau(i))$, which by Eq. (11) implies $i < l < j$. Thus l satisfies Case 1, a contradiction. Therefore, also by Eq. (11), $\tilde{U}_i$ can be written as a linear combination of $U_{\tau(j)}$ and U_S with nonzero coefficient of $U_{\tau(j)}$, where $S \subset [p] \setminus \mathrm{de}_{\mathcal{G}}(\tau(k))$.

Since $\tau(j) \in \mathrm{de}_{\mathcal{G}}(\tau(i)) = \mathrm{de}_{\mathcal{G}}(\tau(k_0))$, by definition of k, we have $\tau(j) \in \mathrm{mch}_{\mathcal{G}}(\tau(k))$. Note that $\tilde{U}_i$ can be written as a linear combination of $U_{\tau(j)}$ and $U_{S'}$ with nonzero coefficient of $U_{\tau(j)}$, where $S' \subset [p] \setminus (\mathrm{de}_{\mathcal{G}}(k) \cup \{\tau(j)\})$. By Assumption 2, $\mathbb{P}(\tilde{U}_j) \neq \mathbb{P}^{I_{\rho_k}}(\tilde{U}_j)$, as I_{ρ_k} targets $U_{\tau(k)}$.

Therefore, in both cases it holds that $\mathbb{P}(\tilde{U}_j) \neq \mathbb{P}^{I_{\rho_k}}(\tilde{U}_j)$. Since $i \leq k < j$ and $\tau(k) \in \overline{\mathrm{de}}_{\mathcal{G}}(\tau(i))$, the claim is proven. □

Now, we use marginal changes to define a graph associated to any topologically-ordered representation. We use Proposition 1 to show that picking the the topologically-ordered representation which yields the sparsest graph will recover the transitive closure of $\mathcal{G}$.

Lemma 7. *Let $\tilde{U}$ be a topological representation of X with intervention targets $\rho_1, ..., \rho_p$. Let $\hat{R}' \in \mathbb{R}^{p \times p}$ be an invertible upper triangular and let $\hat{U} = \tilde{U}\hat{R}'$. Define the following:*

- *Let $\hat{\mathcal{G}}_0^{\hat{R}'}$ be the DAG such that $i \to j \in \hat{\mathcal{G}}_0$ if and only if $i < j \in [p]$ and $\mathbb{P}(\hat{U}_j) \neq \mathbb{P}^{I_{\rho_i}}(\hat{U}_j)$.*
- *Let $\hat{\mathcal{G}}^{\hat{R}'} = \mathcal{TS}(\hat{\mathcal{G}}_0^{R'})$.*

Let $\hat{R}$ be such that $\hat{\mathcal{G}}^{\hat{R}}$ has the fewest edges over any choice of $\hat{R}'$. Then $\hat{\mathcal{G}}^{\hat{R}} = \mathcal{TS}(\mathcal{G}_\tau)$. We call $\hat{U}$ a sparsest topological representation *of X.*

Proof.
Direction 1.
We first show that for any $\hat{R}'$,

$$\mathcal{TS}(\mathcal{G}_\tau) \subseteq \hat{\mathcal{G}}^{\hat{R}'}. \tag{10}$$

Let $i \to j \in \mathcal{TS}(\mathcal{G}_\tau)$, so $i < j$. By Proposition 1, we have k such that $i \leq k < j$ with $\tau(k) \in \overline{\mathrm{de}}_{\mathcal{G}}(\tau(i))$. By definition of $\hat{\mathcal{G}}_0^{\hat{R}'}$, we have $k \to j \in \hat{\mathcal{G}}_0^{\hat{R}'}$. Repeating this argument iteratively, we obtain a directed path from i to j in $\hat{\mathcal{G}}_0^{\hat{R}'}$. Thus, by definition of $\hat{\mathcal{G}}^{\hat{R}'}$, we have $\mathcal{TS}(\mathcal{G}_\tau) \subseteq \hat{\mathcal{G}}^{\hat{R}'}$.

Direction 2.
Now we give an example of $\hat{R}$ such that the constructed $\hat{\mathcal{G}}^{\hat{R}}$ satisfies

$$\hat{\mathcal{G}}^{\hat{R}} \subseteq \mathcal{TS}(\mathcal{G}_\tau).$$

Denote $\tilde{\Gamma}\hat{R} = \hat{\Gamma}$ and $\hat{c} = \tilde{c}\hat{R}$. Since $\hat{R}$ is upper-triangular and invertible, by Eq. (9), we have $\hat{U} = U\hat{\Gamma} + \hat{c}$, where

$$i < j \Rightarrow \hat{\Gamma}_{\tau(j),i} = 0 \quad \text{and} \quad \hat{\Gamma}_{\tau(j),j} \neq 0, \tag{11}$$

where τ is the topological order in Eq. (9). By Eq. (11), there exists an invertible upper-triangular matrix $R \in \mathbb{R}^{p \times p}$ such that $\hat{U} = (U\tilde{\Gamma} + \tilde{c})R = (U_{\tau(1)}, ..., U_{\tau(p)}) + c$ for some constant vector c. Now for $i < j \in [p]$, we have $i \to j \in \hat{\mathcal{G}}_0^R \Leftrightarrow \mathbb{P}(U_{\tau(j)}) \neq \mathbb{P}^{I_{\rho_i}}(U_{\tau(j)})$. Since I_{ρ_i} targets $U_{\tau(i)}$, this would only be true when $\tau(j) \in \mathrm{de}_{\mathcal{G}}(\tau(i))$. Therefore $i \to j \in \hat{\mathcal{G}}_0^{\hat{R}} \Rightarrow \tau(j) \in \mathrm{de}_{\mathcal{G}}(\tau(i))$. Thus $\hat{\mathcal{G}}_0^{\hat{R}} \subseteq \mathcal{TS}(\mathcal{G}_\tau)$. As $\mathcal{TS}(\mathcal{G}_\tau)$ is a transitive closure, this means $\hat{\mathcal{G}}^{\hat{R}} = \mathcal{TS}(\hat{\mathcal{G}}_0^{\hat{R}}) \subseteq \mathcal{TS}(\mathcal{G}_\tau)$. □

Analogously to Lemma 6, the following result shows that a sparsest topological representation is topologically consistent with U in a stronger sense than a topological representation.

Lemma 8. *Let Assumption 2 hold. For $\tilde{\Gamma} \in \mathbb{R}^{p\times p}$ and $\tilde{c} \in \mathbb{R}^p$, let $\tilde{U} = U\tilde{\Gamma} + \tilde{c}$ be a sparsest topological representation of U with intervention targets $\rho_1, \dots, \rho_p$. Let τ be a topological ordering of $\mathcal{G}$ such that*

$$i < j \Rightarrow \tilde{\Gamma}_{\tau(j),i} = 0 \quad \text{and} \quad \tilde{\Gamma}_{\tau(j),j} \neq 0. \tag{12}$$

Then $\tilde{\Gamma}_{\tau(j),l} = 0$ for $\tau(l) \notin \mathrm{de}_{\mathcal{G}}(\tau(j))$.

Proof. For sake of contradiction, let $l, j \in [p]$. Without loss of generality, let j be the largest value for which $\tilde{\Gamma}_{\tau(j),l} \neq 0$ and $\tau(l) \notin \mathrm{de}_{\mathcal{G}}(\tau(j))$. By transitivity of $\mathrm{de}_{\mathcal{G}}$ and the choice of j as the largest value, there is no j' such that $\tau(j') \in \mathrm{de}_{\mathcal{G}}(\tau(j))$ and $\tilde{\Gamma}_{\tau(j'),l} \neq 0$. Therefore $\tilde{U}_l$ can be written as a linear combination of $U_{\tau(j)}$ and U_S with nonzero coefficient of $U_{\tau(j)}$, where $S \subset [p] \setminus \overline{\mathrm{de}}_{\mathcal{G}}(\tau(j))$.

By Assumption 2, we have $\mathbb{P}(\tilde{U}_l) \neq \mathbb{P}^{I_{\rho_j}}(\tilde{U}_l)$. Since $\tau(l) \notin \mathrm{de}_{\mathcal{G}}(\tau(j))$ and $\hat{\mathcal{G}} = \mathcal{TS}(\mathcal{G}_\tau)$, we have $j \to l \notin \hat{\mathcal{G}}$, in which case $\mathbb{P}(\tilde{U}_l) \neq \mathbb{P}^{I_{\rho_{j'}}}(\tilde{U}_l)$ violates Condition 1, a contradiction. □

B.3.3 Proof of Theorem 1

Theorem 1. *Under Assumption 1 and Assumption 2 for $I_1, ..., I_K$, we can identify $\langle \hat{\mathcal{G}}, \hat{I}_1, ..., \hat{I}_K \rangle$, where $\hat{\mathcal{G}} = \mathcal{TS}(\mathcal{G}_\pi)$, and $\hat{I}_k = (I_k)_\pi$ for some permutation π.*

Here, we combine the results of the previous two sections to show that we can recover $\mathcal{G}$ up to transitive closure and permutation, and that we recover the intervention targets $I_1, \dots, I_K$ up to the same permutation.

Proof. By Lemma 1 and Remark 1, we can assume, without loss of generality, that p is known and that $X = f(U) = U\Lambda + b$ for some non-singular matrix Λ, as this can be identified from observational data $\mathcal{D}$.

By Lemma 7, we can identify a topological representation $\hat{U} = U\hat{\Gamma} + \hat{c}$ with intervention targets $\rho_1, ..., \rho_p \in [K]$, where $\hat{\Gamma} \in \mathbb{R}^{p\times p}$ and $\hat{c} \in \mathbb{R}^p$. Further, for some unknown topological ordering τ of $\mathcal{G}$, $\hat{\Gamma}$ satisfies Eq. (11), $T(I_{\rho_i}) = \tau(i)$ for $i \in [p]$, and we identify $\hat{\mathcal{G}} = \mathcal{TS}(\mathcal{G}_\tau)$.

Identifying additional intervention targets.
So far, we only guarantee that we identify the intervention targets for $I_{\rho_1}, \dots, I_{\rho_p}$. Now, consider any $k \in [K] \setminus \{\rho_1, ..., \rho_p\}$. Let l be such that $T_{\mathcal{G}}(I_k) = \tau(l)$. We now argue that l can be identified as the smallest l' in $[p]$ such that $\mathbb{P}(\hat{U}_{l'}) \neq \mathbb{P}^{I_k}(\hat{U}_{l'})$.

By Assumption 2, we have $\mathbb{P}(\hat{U}_l) \neq \mathbb{P}^{I_k}(\hat{U}_l)$, since I_k targets $U_{\tau(l)}$ and $\hat{U}_l$ can be written as a linear combination of $U_{\tau(1)}, ..., U_{\tau(l)}$ with nonzero coefficient $U_{\tau(l)}$ (note that $U_{\tau(1)}, ..., U_{\tau(l-1)} \in [p] \subset \overline{\mathrm{de}}_{\mathcal{G}}(\tau(l))$).

On the other hand, for $l' < l$, we have $\mathbb{P}(\hat{U}_l) = \mathbb{P}^{I_k}(\hat{U}_l)$, since $\hat{U}_{l'}$ can be written as a linear combination of $U_{\tau(1)}, ..., U_{\tau(l')}$ and τ is the topological order. □

B.4 Proof of Theorem 2

In this section, we show that by introducing Assumption 3, we can go beyond recovering the transitive closure of $\mathcal{G}$, and we instead recover $\mathcal{G}$. We begin by establishing a basic fact about conditional independences in our setup.

Claim 1. *Under Assumption 1, let $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ denote (potentially linear combinations of) components of U, and assume that $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}, \mathbf{D}$ and $\mathbf{A} \perp\!\!\!\perp \mathbf{C} \mid \mathbf{B}, \mathbf{D}$. Then $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{D}$.*

Proof. By Assumption 1, $\mathbb{P}_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}}$ has positive measure on some full-dimensional set. By Proposition 2.1 of [56], $\mathbb{P}_{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}}$ is a graphoid, i.e., it obeys the intersection property. Invoking this property, we obtain $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{D}$, as desired. □

With this, we are ready to prove Theorem 2, which we recall here.

Theorem 2. *Under Assumptions 1,2,3, $\langle \mathcal{G}, I_1, ..., I_K \rangle$ is identifiable up to its CD-equivalence class.*

Note that, from Theorem 1, we have already identified the interventions $I_1, \ldots, I_K$ up to CD-equivalence for a permutation τ. Thus, the only remaining result to show is that we identify $\mathcal{G}$ up to the same permutation.

In particular, we can again characterize the solution in terms of the sparsest solution.

Theorem 2, Constructive. *Let $\hat{U}$ be a sparsest topological representation of X with intervention targets $\rho_1, ..., \rho_p$. Let $\bar{R}' \in \mathbb{R}^{p \times p}$ be an invertible upper triangular matrix, and let $\bar{U} = \hat{U}\bar{R}'$. Define the following:*

- *Let $\bar{\mathcal{G}}^{\bar{R}'}$ be the DAG such that $i \to j$ for $i < j \in [p]$, if and only if $\bar{U}_i \not\perp\!\!\!\perp \bar{U}_j \mid \bar{U}_1, \ldots, \bar{U}_{i-1}, \bar{U}_{i+1}, \ldots, \bar{U}_{j-1}$*

Let $\bar{R}$ be such that $\bar{\mathcal{G}}^{\bar{R}}$ has the fewest edges over any choice of $\bar{R}'$. Then $\bar{\mathcal{G}}^{\bar{R}} = \mathcal{G}_\tau$ for τ satisfying Eq. (12).

Proof. By Lemma 7, we have $\hat{U} = U\hat{\Gamma}$ for some matrix $\hat{\Gamma} \in \mathbb{R}^{p \times p}$ satisfying Eq. (9) under some topological order τ of $\mathcal{G}$. Further, we identify $\hat{\mathcal{G}} = \mathcal{TS}(\mathcal{G}_\tau)$.

Denoting $\hat{\Gamma}\bar{R} = \bar{\Gamma}$ and $\bar{c} = \hat{c}\bar{R}$, by Lemma 8, we have $\bar{U} = U\bar{\Gamma} + \bar{c}$ with

$$\begin{aligned} &i < j \Rightarrow \bar{\Gamma}_{\tau(j),i} = 0 \quad \text{and} \quad \bar{\Gamma}_{\tau(j),j} \neq 0, \\ &\tau(l) \notin \mathrm{de}_{\mathcal{G}}(\tau(j)) \Rightarrow \bar{\Gamma}_{\tau(j),l} = 0. \end{aligned} \tag{13}$$

Direction 1.
First, we show that

$$\mathcal{G}_\tau \subseteq \bar{\mathcal{G}}^{\bar{R}}.$$

Assume on the contrary that there exists $\tau(i) \to \tau(j) \in \mathcal{G}$ such that $i \to j \notin \bar{\mathcal{G}}^{\bar{R}}$. By definition, we have $\bar{U}_i \not\perp\!\!\!\perp \bar{U}_j \mid \bar{U}_1, \ldots, \bar{U}_{i-1}, \bar{U}_{i+1}, \ldots, \bar{U}_{j-1}$. By Eq. (13), we know that we can retrieve $U_{\tau(1)}, ..., U_{\tau(i-1)}$ by linearly transforming $\bar{U}_1, ..., \bar{U}_{i-1}$; this implies $U_{\tau(i)} \not\perp\!\!\!\perp \bar{U}_j \mid U_{\tau(1)}, ..., U_{\tau(i-1)}, \bar{U}_{i+1}, ..., \bar{U}_{j-1}$. By subtracting terms in $U_{\tau(1)}, ..., U_{\tau(i-1)}$ from $\bar{U}_{i+1}, ..., \bar{U}_j$ and then subtracting terms $\bar{U}_l$ from $\bar{U}_{l+1}, ..., \bar{U}_j$ for $l = i+1, ..., j-1$, we have that

$$\begin{aligned} U_{\tau(i)} \not\perp\!\!\!\perp U_{\tau(j)} + c_j U_{\tau(i)} \mid U_{\tau(1)}, ..., U_{\tau(i-1)}, \\ U_{\tau(i+1)} + c_{i+1}U_{\tau(i)}, ..., U_{\tau(j-1)} + c_{j-1}U_{\tau(i)}, . \end{aligned} \tag{14}$$

for some $c_{i+1}, ..., c_j \in \mathbb{R}$. Since by Eq. (13) there is $\bar{\Gamma}_{\tau(i),l} = 0$ for any $\tau(l) \notin \mathrm{de}_{\mathcal{G}}(\tau(i))$, this subtraction gives us $c_l = 0$ if $\tau(l) \notin \mathrm{de}_{\mathcal{G}}(\tau(i))$.

Therefore let

$$\begin{aligned} &\mathbf{A} = U_{\tau(j)} + c_j U_{\tau(i)}, && \mathbf{B} = U_{\tau(i)} \\ &\mathbf{C} = \{U_{\tau(l)} + c_l U_{\tau(i)}\}_{l \leq j-1, \tau(l) \notin \mathrm{pa}_{\mathcal{G}}(\tau(j))}, \text{and} && \mathbf{D} = \{U_{\tau(l)} + c_l U_{\tau(i)}\}_{\tau(l) \in \mathrm{pa}_{\mathcal{G}}(\tau(j)) \setminus \{\tau(i)\}}, \end{aligned}$$

where $c_1 = ... = c_{i-1} = 0$. There is $\mathbf{A} \not\perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}, \mathbf{D}$.

On the other hand, since $\tau(i) \to \tau(j) \in \mathcal{G}$, i.e., $\tau(i) \in \mathrm{pa}_{\mathcal{G}}(\tau(j))$. We will now show that this implies $\mathbf{A} \not\perp\!\!\!\perp \mathbf{C} \mid \mathbf{B}, \mathbf{D}$. Starting with the local Markov property, we have for any $c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_j$ that

$$\begin{aligned} & U_{\tau(j)} \not\perp\!\!\!\perp \{U_{\tau(l)}\}_{l \leq j-1, \tau(l) \notin \mathrm{pa}_{\mathcal{G}}(\tau(j))} \mid \{U_{\tau(l)}\}_{\tau(l) \in \mathrm{pa}_{\mathcal{G}}(\tau(j))} \\ \Longrightarrow \quad & U_{\tau(j)} + c_j U_{\tau(i)} \not\perp\!\!\!\perp \{U_{\tau(l)} + c_l U_{\tau(i)}\}_{l \leq j-1, \tau(l) \notin \mathrm{pa}_{\mathcal{G}}(\tau(j))} \mid \{U_{\tau(l)}\}_{\tau(l) \in \mathrm{pa}_{\mathcal{G}}(\tau(j))} \\ \Longrightarrow \quad & U_{\tau(j)} + c_j U_{\tau(i)} \not\perp\!\!\!\perp \{U_{\tau(l)} + c_l U_{\tau(i)}\}_{l \leq j-1, \tau(l) \notin \mathrm{pa}_{\mathcal{G}}(\tau(j))} \\ & \qquad \mid U_{\tau(i)}, \{U_{\tau(l)} + c_l U_{\tau(i)}\}_{\tau(l) \in \mathrm{pa}_{\mathcal{G}}(\tau(j)) \setminus \{\tau(i)\}} \end{aligned} \tag{15}$$

where the first implication follows from the definition of conditional independence, and the second implication follows since $\{U_{\tau(l)}\}_{\tau(l)\in \mathrm{pa}_{\mathcal{G}}(\tau(j))}$ is a deterministic function of $U_{\tau(i)}, \{U_{\tau(l)} + c_l U_{\tau(i)}\}_{\tau(l)\in \mathrm{pa}_{\mathcal{G}}(\tau(j))\setminus\{\tau(i)\}}$.

Thus, by Claim 1, if $i \to j \notin \bar{\mathcal{G}}^{\bar{R}}$ and $\tau(i) \to \tau(j) \in \mathcal{G}$, then $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{D}$, i.e.,

$$U_{\tau(i)} \perp\!\!\!\perp U_{\tau(j)} + c_j U_{\tau(i)} \mid \{U_{\tau(l)} + c_l U_{\tau(i)}\}_{l\in \mathrm{pa}_{\mathcal{G}}(\tau(j))\setminus\{\tau(i)\}}.$$

Since $c_l = 0$ for any $\tau(l) \notin \mathrm{de}_{\mathcal{G}}(\tau(i))$ and τ is the topological order, this can be further written as

$$U_{\tau(i)} \perp\!\!\!\perp U_{\tau(j)} + c_j U_{\tau(i)} \mid \{U_{\tau(l)}\}_{l\in \mathrm{pa}_{\mathcal{G}}(\tau(j))\setminus(S\cup\{\tau(i)\})}, \{U_{\tau(l)} + c_l U_{\tau(i)}\}_{l\in S},$$

where $S = \mathrm{pa}_{\mathcal{G}}(\tau(j)) \cap \mathrm{de}_{\mathcal{G}}(\tau(i))$, which violates Assumption 3. Therefore we must have $\mathcal{G}_\tau \subseteq \bar{\mathcal{G}}^{\bar{R}}$.

Direction 2.

There exists an invertible upper-triangular matrix $\bar{R} \in \mathbb{R}^{p\times p}$ such that $\bar{U} = \hat{U}\bar{R} = (U_{\tau(1)}, ..., U_{\tau(p)}) + \bar{c}$ for some constant vector $\bar{c}$. Note that clearly $\bar{U}$ satisfies Condition 1. Also for $i < j \in [p]$ such that $\tau(i) \to \tau(j) \notin \mathcal{G}$, by the Markov property and τ being the topological order, we have $\bar{U}_i \perp\!\!\!\perp \bar{U}_j \mid \bar{U}_1, ..., \bar{U}_{i-1}, \bar{U}_{i+1}, ..., \bar{U}_{j-1}$. Thus $\tau(i) \to \tau(j) \notin \mathcal{G} \Rightarrow i \to j \notin \bar{\mathcal{G}}$, and hence $\bar{\mathcal{G}} \subseteq \mathcal{G}_\tau$, which completes the proof. □

Remark 2. *These proofs (Lemma 1, Theorem 1,2) together indicate that under Assumptions 1,2,3, we can identify $\langle \mathcal{G}, I_1, ..., I_K\rangle$ up to its CD-equivalence class by solving for the smallest $\hat{p}$, an encoder $\hat{g} : \mathbb{R}^n \to \mathbb{R}^{\hat{p}}$, $\hat{\mathcal{G}}$ and $\hat{I}_1, ..., \hat{I}_{\hat{K}}$ that satisfy*

(1) there exists a full row rank polynomial decoder $\hat{f}(\cdot)$ such that $\hat{f} \circ \hat{g}(X) = X$ for all $X \in \mathcal{D} \cup \mathcal{D}^{I_1} \cup ... \cup \mathcal{D}^{I_K}$;

(2) the induced distribution on $\hat{U} := \hat{g}(X)$ by $X \in \mathcal{D}$ factorizes with respect to $\hat{\mathcal{G}}$;

(3) the induced distribution on $\hat{U}$ by $X \in \mathcal{D}^{I_k}$ where $k \in [K]$ changes the distribution of $\hat{U}_{T_{\hat{\mathcal{G}}}(\hat{I}_k)}$ but does not change the joint distribution of non-descendants of $\hat{U}_{T_{\hat{\mathcal{G}}}(\hat{I}_k)}$ in $\hat{\mathcal{G}}$;

(4) $[\hat{p}] \subseteq T_{\hat{\mathcal{G}}}(\hat{I}_1) \cup ... \cup T_{\hat{\mathcal{G}}}(\hat{I}_{\hat{K}})$;

(5) $\hat{\mathcal{G}}$ has topological order $1, ..., \hat{p}$;

(6) the transitive closure $\mathcal{TS}(\hat{\mathcal{G}})$ of the DAG $\hat{\mathcal{G}}$ is the sparsest amongst all solutions that satisfy (1)-(5);

(7) the DAG $\hat{\mathcal{G}}$ is the sparsest amongst all solutions that satisfy (1)-(6);

We will use these observations in Appendix E to develop a discrepancy-based VAE and show that it is consistent in the limit of infinite data.

Proof. We first show that there is a solution to (1)-(7). For this, it suffices to show that there is a solution to (1)-(5). Then since $\hat{p}$ and $\hat{\mathcal{G}}$ are discrete, one can find the solution to (1)-(7) by searching amongst all solutions to (1)-(5) such that $\hat{p}$ is the smallest and (6)-(7) are satisfied. Assume without loss of generality that $\mathcal{G}$ has topological order $1, ..., p$. Then $\hat{p} = p$, $\hat{g} = f^{-1}$, $\hat{\mathcal{G}} = \mathcal{G}$, and $\hat{I}_k = I_k$ for $k \in [K]$ satisfy (1)-(5).

Next we show that any solution must recover $\hat{p} = p$ and $\langle \hat{\mathcal{G}}, \hat{I}_1, \ldots, \hat{I}_K\rangle$ that is in the same CD equivalence class as $\langle \mathcal{G}, I_1, \ldots, I_K\rangle$. Since we solve for the smallest $\hat{p}$, the former paragraph also implies that $\hat{p} \leq p$. By the proof of Lemma 1, (1) guarantees that $\hat{p} \geq p$. Therefore it must hold that $\hat{p} = p$.

Since we solve for the sparsest transitive closure, the first paragraph implies that $\mathcal{TS}(\hat{\mathcal{G}}) \subset \mathcal{TS}(\mathcal{G})$. Also by the proof of Lemma 1, $\hat{U}$ can be written as an invertible linear mixing of U. Then (3)-(5) guarantee that Condition 1 and Condition 2 in Step 1 in the proof of Theorem 1 are satisfied. Then by the proof of Step 2 in the proof of Theorem 1, we have $\mathcal{TS}(\mathcal{G}) \subset \mathcal{TS}(\hat{\mathcal{G}})$. Therefore, it must hold that $\mathcal{TS}(\hat{\mathcal{G}}) = \mathcal{TS}(\mathcal{G})$.

Lastly, by (2) and (5), we obtain that $\hat{\mathcal{G}}$ satisfies Condition 1 and Condition 2 in Theorem 2. Therefore by the proof of Theorem 2, we obtain $\mathcal{G} \subset \hat{\mathcal{G}}$. Again, by the first paragraph and the fact that the sparsest transitive closure satisfies $\mathcal{TS}(\hat{\mathcal{G}}) = \mathcal{TS}(\mathcal{G})$, we obtain that the sparsest $\hat{\mathcal{G}}$ with this transitive closure must satisfy $\hat{\mathcal{G}} \subset \mathcal{G}$, and thus $\hat{\mathcal{G}} = \mathcal{G}$. With this result, it is easy to see that $\hat{I}_k = I_k$ for all $k \in [K]$, as I_k changes the distribution of $\hat{U}_{T_{\mathcal{G}}(I_k)}$ but does not change the joint distribution of $\hat{U}_{[p]\setminus\overline{\mathrm{de}}_{\mathcal{G}}(T_{\mathcal{G}}(I_k))}$.

Therefore we can recover p and the CD equivalence class of $\langle \mathcal{G}, I_1, ..., I_K \rangle$ by solving (1)-(7). Note that this proof assumes the topological order of $\mathcal{G}$ is $1, ..., p$, and therefore it does not violate the fact that $\mathcal{G}, I_1, ..., I_K$ cannot be recovered exactly. □

B.5 Proof of Theorem 3

Now, we will show that recovering $\langle U, \mathcal{G}, I_1, \ldots, I_K \rangle$ up to Theorem 1 is sufficient for predicting the effect of combinatorial interventions..

Theorem 3. *Letting $\langle \hat{U}, \hat{\mathcal{G}}, \hat{I}_1, ..., \hat{I}_K \rangle$ be the solution identified in the proof of Theorem 1. Then the interventional distribution $\mathbb{P}^{\mathcal{I}}$ for any combinatorial intervention $\mathcal{I} \subset \{I_1, ..., I_K\}$ is given by Eq. (2), i.e., we can generate samples X from the distribution $X = f(U), U \sim \mathbb{P}^{\mathcal{I}}$.*

Proof. Since $\mathcal{I}$ contains interventions with different intervention targets, for each $i \in [p]$, we can define $\mathbb{P}^{\hat{\mathcal{I}}}(\hat{U}_i \mid \hat{U}_{\mathrm{pa}_{\hat{\mathcal{G}}}(i)})$ as $\mathbb{P}^{\hat{I}_k}(\hat{U}_i \mid \hat{U}_{\mathrm{pa}_{\hat{\mathcal{G}}}(i)})$ if $i = T_{\hat{\mathcal{G}}}(\hat{I}_k)$ for some $I_k \in \mathcal{I}$ and otherwise $\mathbb{P}(\hat{U}_i \mid \hat{U}_{\mathrm{pa}_{\hat{\mathcal{G}}}(i)})$. Using this definition, we define the joint distribution of $\hat{U}$ as $\mathbb{P}^{\hat{\mathcal{I}}}(\hat{U}) = \prod_{i=1}^{p} \mathbb{P}^{\hat{\mathcal{I}}}(\hat{U}_i \mid \hat{U}_{\mathrm{pa}_{\hat{\mathcal{G}}}(i)})$. In the following we show that $\mathbb{P}^{\hat{\mathcal{I}}}(\hat{U}) = \mathbb{P}^{\mathcal{I}}(U)$ in the sense that $\mathbb{P}^{\hat{\mathcal{I}}}\big(\hat{U} = \hat{f}^{-1}(x)\big) = \mathbb{P}^{\mathcal{I}}(U = f^{-1}(x))$ for all $x \in \mathbb{R}^n$.

Our proof combines the following equalities. For any $i \in [p]$, we have

$$\textit{Equality 1: } \mathbb{P}^{\hat{\mathcal{I}}}(\hat{U}_i \mid \hat{U}_{\mathrm{pa}_{\hat{\mathcal{G}}}(i)}) = \mathbb{P}^{\hat{\mathcal{I}}}(\hat{U}_i \mid \hat{U}_{\mathrm{an}_{\hat{\mathcal{G}}}(i)}),$$

$$\textit{Equality 2: } \mathbb{P}^{\hat{\mathcal{I}}}(\hat{U}_i \mid \hat{U}_{\mathrm{an}_{\hat{\mathcal{G}}}(i)}) = \mathbb{P}^{\mathcal{I}}(U_i \mid U_{\mathrm{an}_{\mathcal{G}}(i)}),$$

$$\textit{Equality 3: } \mathbb{P}^{\mathcal{I}}(U_i \mid U_{\mathrm{an}_{\mathcal{G}}(i)}) = \mathbb{P}^{\mathcal{I}}(U_i \mid U_{\mathrm{pa}_{\mathcal{G}}(i)}).$$

Proof of Equality 1. This follows by definition of $\hat{\mathcal{G}}$, since it is transitively closed, we have $\mathrm{pa}_{\hat{\mathcal{G}}}(i) = \mathrm{an}_{\hat{\mathcal{G}}}(i)$.

Proof of Equality 2. By similar arguments below Eq. (13), we have $\hat{U} = U\hat{\Gamma} + \hat{c}$ for an invertible matrix $\hat{\Gamma}$, where $\hat{\Gamma}_{\tau(j),l} = 0$ for any $\tau(l) \notin \mathrm{de}_{\mathcal{G}}(\tau(j))$. Therefore we can recover $U_{\mathrm{an}_{\mathcal{G}}(i)}$ by linear transforming $U\hat{\Gamma}_{:,\mathrm{an}_{\mathcal{G}}(i)}$ and vise versa. We can also recover U_i by subtracting linear terms of $U_{\mathrm{an}_{\mathcal{G}}(i)}$ from $U\hat{\Gamma}_{:,i}$.

Note also, since $\mathcal{TS}(\mathcal{G}) = \mathcal{TS}(\hat{\mathcal{G}})$, there must be $\mathrm{an}_{\hat{\mathcal{G}}}(i) = \mathrm{an}_{\mathcal{G}}(i)$. Thus

$$\begin{aligned}
\mathbb{P}^{\hat{\mathcal{I}}}(\hat{U}_i \mid \hat{U}_{\mathrm{an}_{\hat{\mathcal{G}}}(i)}) &= \mathbb{P}^{\mathcal{I}}(U\hat{\Gamma}_{:,i} \mid U\hat{\Gamma}_{:,\mathrm{an}_{\hat{\mathcal{G}}}(i)}) \\
&= \mathbb{P}^{\mathcal{I}}(U\hat{\Gamma}_{:,i} \mid U\hat{\Gamma}_{:,\mathrm{an}_{\mathcal{G}}(i)}) \\
&= \mathbb{P}^{\mathcal{I}}(U\hat{\Gamma}_{:,i} \mid U_{\mathrm{an}_{\mathcal{G}}(i)}) = \mathbb{P}^{\mathcal{I}}(U_i \mid U_{\mathrm{an}_{\mathcal{G}}}(i)).
\end{aligned}$$

Proof of Equality 3. Follows from the Markov property on U.

Combining these equalities, we have $\mathbb{P}^{\hat{\mathcal{I}}}(\hat{U}_i \mid \hat{U}_{\mathrm{pa}_{\hat{\mathcal{G}}}(i)}) = \mathbb{P}^{\mathcal{I}}(U_i \mid U_{\mathrm{pa}_{\mathcal{G}}(i)})$ for all $i \in [p]$. Thus $\mathbb{P}^{\hat{\mathcal{I}}}(\hat{U}) = \prod_{i=1}^{p} \mathbb{P}^{\hat{\mathcal{I}}}(\hat{U}_i \mid \hat{U}_{\mathrm{pa}_{\hat{\mathcal{G}}}(i)}) = \prod_{i=1}^{p} \mathbb{P}^{\mathcal{I}}(U_i \mid U_{\mathrm{pa}_{\mathcal{G}}(i)}) = \mathbb{P}^{\mathcal{I}}(U)$. Therefore the procedure in Section 4.4 generates X from the same distribution as $X = f(U), U \sim \mathbb{P}^{\mathcal{I}}$. □

C Details on Discrepancy-based VAE

In previous sections, we have shown that the data-generating process in Section 2 is identifiable up to equivalence classes. However, the proofs (Appendix A, B) do not lend themselves to an algorithmically efficient approach to learning the latent causal variables from data. Therefore, we propose a discrepancy-based VAE in Section 5, which inherits scalable tools of VAEs that can in principle learn flexible deep latent-variable models. In this framework, Eq. (3) can be computed and optimized efficiently using the reparametrization trick [29] and gradient-based optimizers.

C.1 Maximum Mean Discrepancy

We recall the definition of the maximum mean discrepancy measure between two distributions, and its empirical counterpart.

Definition 3. *Let k be a positive definite kernel function and let $\mathcal{H}$ be the reproducing kernel Hilbert space defined by this kernel. Given distributions $\mathbb{P}$ and $\mathbb{P}'$, we define*

$$\mathrm{MMD}(\mathbb{P}, \mathbb{P}') := \sup_{f \in \mathcal{H}} \left(\mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{P}'}[f(X)] \right)$$

The following empirical counterpart is an unbiased estimate of the squared MMD, see Lemma 6 of [18].

Definition 4. *Let k be a positive definite kernel. Let $\{X_{(i)}\}_{i=1}^m$ be samples from $\mathbb{P}$ and $\{X'_{(i)}\}_{i=1}^m$ be samples from $\mathbb{P}'$. We define*

$$\widehat{\mathrm{MMD}}^2(\{X_{(i)}\}_{i=1}^m, \{X'_{(i)}\}_{i=1}^m) = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i} k(X_i, X_j) + \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i} k(X'_i, X'_j) - \frac{2}{m^2} \sum_{i=1}^m \sum_{j=1}^m k(X_i, X'_j)$$

C.2 Discrepancy VAE Details

We walk through the details of this model in this section, where we illustrate it using two types of interventions, namely do interventions and shift interventions.

Noiseless vs. Noisy Measurement Model with General SCMs. Recall that each latent causal variable U_i is a function of its parents in $\mathcal{G}$ and an exogenous noise term Z_i. All the Z_i's are mutually independent. The overall model can be defined (recursively) as

$$\begin{aligned} U_j &= s_j(U_{\mathrm{pa}_{\mathcal{G}}(j)}, Z_j), \\ X &= f(U_1, ..., U_p). \end{aligned} \tag{16}$$

In particular, there exists a function $s_{\varnothing}^{\text{full}}$ such that $U = s_{\varnothing}^{\text{full}}(Z)$. We model each intervention I as a set of intervention targets $T(I)$ and a vector a^I. Under I, the observations X are generated by

$$\begin{aligned} U_j^I &= \begin{cases} s_j(U^I_{\mathrm{pa}_{\mathcal{G}}(j)}, Z_j)\mathbb{1}_{j \notin T(I)} + a_j^I \mathbb{1}_{j \in T(I)}, & \text{for do intervention,} \\ s_j(U^I_{\mathrm{pa}_{\mathcal{G}}(j)}, Z_j) + a_j^I \mathbb{1}_{j \in T(I)}, & \text{for shift intervention,} \end{cases} \\ X^I &= f(U_1^I, ..., U_p^I). \end{aligned} \tag{17}$$

As above, there exists a function s_I^{full} such that $U^I = s_I^{\text{full}}(Z)$. Note that here we assume that the measurements (sometimes called "observations" in the literature[11]) X are *noiseless*. Our theoretical results are built upon noiseless measurements. In practice, however, one can consider the *noisy* measurement model in which $X = f(U) + \epsilon$ (resp. $X^I = f(U^I) + \epsilon$), where ϵ is some measurement noise independent of U.

We leave as future work to prove consistency under the noisy measurement model. [28] established identifiability results of the noisy measurement model, when the latent variables conditioned on

[11] We use "measurements" to distinguish from the observational distribution defined for U.

additionally observed variables follow a factorized distribution in an exponential family. Their techniques can be potentially used to generalize our results to the noisy measurement model; however, further assumptions on the mechanisms s_i's will be needed.

Discrepancy-based VAE. We use one *decoder*

$$p_\theta(X|U)$$

parameterized by θ to approximate both $X = f(U)$ and $X^I = f(U^I)$ in the noiseless measurement model (or $X = f(U) + \epsilon$ and $X^I = f(U^I) + \epsilon$ in the noisy measurement model). As for the encoder, we do not directly learn the posteriors $\mathbb{P}(U|X)$ and $\mathbb{P}(U^I|X^I)$. Instead, we approximate one posterior $\mathbb{P}(Z|X)$ and then use Eq. (16), (17) to transform Z into U, U^I respectively. This is done by two *encoders* for Z and $(T(I), a^I)$ parameterized by ϕ and denoted as

$$q_\phi(Z|X), (T_\phi(I), a_\phi(I))$$

The dimension p of Z is set as a hyperparameter. Note that the procedure of learning a posterior $\mathbb{P}(Z|X)$ in the observational distribution and then mapping to U^I using Eq. (17) can be regarded as learning the counterfactual posterior of $\mathbb{P}(U^I|X)$.

In the following, to better distinguish data from observational and interventional distributions, we use $X^\varnothing, U^\varnothing$ instead of X, U to denote samples generated by Eq. (16). After encoding $X^\varnothing$ and I into Z and $(T(I), a^I)$ respectively, we parameterize the causal mechanisms s_j's in Eq. (16), (17) as neural networks (e.g., multi-layer perceptrons or linear layers). We absorb the paramterizations of s_j's into θ and denote

$$p_{\theta,\varnothing}(X^\varnothing|Z) = p_\theta\big(X^\varnothing \mid U^\varnothing = s_\varnothing^{\text{full}}(Z)\big),$$
$$p_{\theta,I}(X^I|Z) = p_\theta\big(X^I \mid U^I = s_I^{\text{full}}(Z)\big).$$

Note that in implementation, to make sure U_j only depends on its parents $U_{\text{pa}_\mathcal{G}(j)}$, one can train an adjacency matrix A that is upper-triangular up to permutations and then apply any layers after individual rows of matrix $U \otimes A$[12]. Since identifiability can be only up to permutations of latent nodes, one can simply use an upper-triangular adjacency matrix A.

D Lower Bound to Paired Log-Likelihood

In this section, we consider the *paired* setting, in which we have access to samples from the joint distribution $\mathbb{P}(X^\varnothing, X^I)$. To discuss counterfactual pairs, we must introduce structure beyond the structure described in Section 2. In particular, in the observational setting, assume that the latent variables $U^\varnothing$ are generated from a structural causal model with exogenous noise terms Z. This implies that there is a function $g_\varnothing$ such that $U^\varnothing = g_\varnothing(Z)$. Similarly, under intervention I, assume there is a function g_I such that $U^I = g_I(Z)$. Then, given a distribution $\mathbb{P}(Z)$, the joint distribution $\mathbb{P}(X^\varnothing, X^I)$ is simply the induced distribution under the maps $X^\varnothing = f(U^\varnothing)$ and $X^I = f(U^I)$.

Since $X^\varnothing$ and X^I are independent conditioned on Z, we have

$$\log \mathbb{P}(X^\varnothing, X^I) \geq \mathbb{E}_{\mathbb{P}(X^\varnothing, X^I)}\big[\mathbb{E}_{q_\phi(Z|X^\varnothing)} \log p_{\theta,\varnothing}(X^\varnothing \mid Z) + \mathbb{E}_{q_\phi(Z|X^\varnothing)} \log p_{\theta,I}(X^I|Z) - D_{\text{KL}}\left(q_\phi(Z|X^\varnothing) \| p(Z)\right)\big] \tag{18}$$

We have the following result on the loss function in Eq. (3).

Proposition 2. *Let k be a Gaussian kernel with width ϵ, i.e., $k(x, y) = \exp\left(-\frac{\|x-y\|_2^2}{2\epsilon^2}\right)$. Let $p_{\theta,I}(X^I \mid U)$ be Gaussian with mean $\mu_\theta^I(U)$ and a fixed variance σ^2. Then, for ϵ sufficiently large, for α given in the proof, and for some constant c depending only on σ and data dimension d,*

$$\mathbb{E}_{\mathbb{P}(X^\varnothing, X^I)}\left[\mathbb{E}_{q_\phi(Z|X^\varnothing)} \log p_{\theta,I}(X^I|Z)\right] \geq -\alpha \cdot \text{MMD}\left(p_{\theta,I}(X^I), \mathbb{P}^I(X^I)\right) + c.$$

Thus, up to an additive constant, $\mathcal{L}_{\theta,\phi}^{\alpha,1,0}$ lower bounds the paired-data ELBO in Eq.(18) and by extension the paired-data log-likelihood $\log \mathbb{P}(X^\varnothing, X^I)$.

[12]Here $\otimes$ denotes the Kronecker product.

Proof. By the choice of a Gaussian distribution for $p_{\theta,I}(X^I \mid U)$, we have

$$\log p_{\theta,I}(X^I \mid Z) = \log p_\theta(X^I \mid U^I = s_I^{\text{full}}(Z)) = c - \frac{1}{2\sigma^2}\|X^I - \mu_\theta^I(U)\|_2^2, \tag{19}$$

where c is a constant depending only on σ and data dimension d. Let $\{(X_{(i)}^{\varnothing}, X_{(i)}^I)\}_{i=1}^m$ be independent and identically distributed according to $\mathbb{P}(X^{\varnothing}, X^I)$. Then

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}(X^{\varnothing},X^I)}\left[\mathbb{E}_{q_\phi(Z|x^{(0)})}[\log p_{\theta,I}(X^I|Z)]\right] \\
&= \mathbb{E}_{\mathbb{P}(X^{\varnothing},X^I)}\left[\mathbb{E}_{q_\phi(Z_i|X_i^{\varnothing})}\left[\frac{1}{m}\sum_{i=1}^m \log p_{\theta,I}(X_i^I|Z_i)\right]\right] \\
&= c - \frac{1}{2\sigma^2}\mathbb{E}_{\mathbb{P}(X^{\varnothing},X^I)}\left[\mathbb{E}_{q_\phi(Z_{(i)}|X_{(i)}^{\varnothing})}\left[\frac{1}{m}\sum_{i=1}^m \|X_{(i)}^I - \mu_\theta^I(U_{(i)})\|_2^2\right]\right]
\end{aligned}$$

Now, for the empirical MMD, we have

$$\begin{aligned}
&\widehat{\text{MMD}}^2\left(\{X_{(i)}^I\}_{i=1}^m, \{\hat{X}_{(i)}^I\}_{i=1}^m\right) \\
&= \frac{1}{m(m-1)}\sum_{i=1}^m\sum_{j\neq i}\exp\left(-\frac{\|X_{(i)}^I - X_{(j)}^I\|_2^2}{2\epsilon^2}\right) + \frac{1}{m(m-1)}\sum_{i=1}^m\sum_{j\neq i}\exp\left(-\frac{\|\hat{X}_{(i)}^I - \hat{X}_{(j)}^I\|_2^2}{2\epsilon^2}\right) \\
&\qquad - \frac{2}{m^2}\sum_{i=1}^m\sum_{j=1}^m\exp\left(-\frac{\|X_{(i)}^I - \hat{X}_{(j)}^I\|_2^2}{2\epsilon^2}\right) \\
&\geq -\frac{2}{m^2}\sum_{i=1}^m\sum_{j=1}^m\exp\left(-\frac{\|X_{(i)}^I - \hat{X}_{(j)}^I\|_2^2}{2\epsilon^2}\right) \\
&\geq -2 + \frac{1}{2m^2\epsilon^2}\sum_{i=1}^m\sum_{j=1}^m\|X_{(i)}^I - \hat{X}_{(j)}^I\|_2^2 \\
&\geq -2 + \frac{1}{2m^2\epsilon^2}\sum_{i=1}^m\|X_{(i)}^I - \hat{X}_{(i)}^I\|_2^2,
\end{aligned}$$

where we have used the positivity of the exponential function and for the penultimate inequality used the fact that ϵ is large enough and that $e^{-x} \leq 1 - x/2$ for x sufficiently small. Substituting into (20) yields the theorem, with $\alpha = \frac{1}{2m\sigma^2\epsilon^2}$. □

E Consistency of Discrepancy-based VAE

We consider Discrepancy-based VAE described in the last section. Suppose the conditions in Theorem 2 is satisfied by the ground-truth model, i.e., it is possible to identify CD-equivalence class in theory.

E.1 CD-Equivalence Class

Theorem 4. *Let $X^{\varnothing}$, X^{I_1}, …, X^{I_K} be generated as in Section 2. Suppose that Assumptions 1, 2, and 3 hold. Define*

$$\begin{aligned}
M_1 &= \textit{argmin}_{\theta,\phi}\mathcal{L}_{\theta,\phi} \\
M_2 &= \textit{argmin}_{\theta,\phi\in M_1}|\mathcal{TS}(\mathcal{G}_\theta)| \\
\hat{\theta}, \hat{\phi} &\in \textit{argmin}_{\theta,\phi\in M_2}|\mathcal{G}_\theta|
\end{aligned}$$

for $\mathcal{L}_{\theta,\phi}$ defined in Equation 3. Further, suppose that the VAE prior $p(Z)$ is equal to the true distribution over Z, that $p_\theta(X \mid U)$ and $q_\phi(Z \mid X)$ are Dirac distributions. Let $\langle \hat{U}, \hat{\mathcal{G}}, \hat{I}_1, \ldots, \hat{I}_K\rangle$ be the solution induced by $\hat{\theta}, \hat{\phi}$.

Then $\langle \hat{U}, \hat{\mathcal{G}}, \hat{I}_1, ..., \hat{I}_K \rangle$ *is CD-equivalent to* $\langle U, \mathcal{G}, I_1, ..., I_K \rangle$.

Proof. Note that the parameterization of s_j, $\mathcal{G}$, and the induced distributions of U through prior $p(Z)$ using Eq. (16), (17) satisfy (2),(3) and (5) in Remark 2.

The first two terms in combined in Eq. (18) satisfy

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}(X^{\varnothing})}\left[\mathbb{E}_{q_\phi(Z|X^{\varnothing})} \log p_{\theta,\varnothing}(X^{\varnothing}|Z) - D_{KL}\big(q_\phi(Z|X^{\varnothing}) \| p(Z)\big)\right] \\
=\ &\mathbb{E}_{\mathbb{P}(X^{\varnothing})}\left[\log p_{\theta,\varnothing}(X^{\varnothing}) - D_{KL}\big(q_\phi(Z|X^{\varnothing}) \| p_{\theta,\varnothing}(Z|X^{\varnothing})\big)\right] \\
\leq\ &\mathbb{E}_{\mathbb{P}(X^{\varnothing})} \log p_{\theta,\varnothing}(X^{\varnothing}) \\
=\ &\mathbb{E}_{\mathbb{P}(X^{\varnothing})} \log \mathbb{P}(X^{\varnothing}) - D_{KL}\big(p_{\theta,\varnothing}(X^{\varnothing}) \| \mathbb{P}(X^{\varnothing})\big) \\
\leq\ &\mathbb{E}_{\mathbb{P}(X^{\varnothing})} \log \mathbb{P}(X^{\varnothing}),
\end{aligned}$$

where the equality holds if and only if $q_\phi(Z|X^{\varnothing}) = p_{\theta,\varnothing}(Z|X^{\varnothing})$ and $p_{\theta,\varnothing}(X^{\varnothing}) = \mathbb{P}(X^{\varnothing})$. On the other hand, since $\text{MMD}(\cdot,\cdot)$ is a valid measure between distributions, we have

$$-\text{MMD}\left(\mathbb{P}_{\theta,\phi}\left(\hat{X}^{\hat{I}_k}\right), \mathbb{P}(X^{I_k})\right) \leq 0,$$

where the inequality is satisfied with equality if and only if $\hat{X}^{\hat{I}_k}$ and X^{I_k} are equal in distribution.

Therefore if the learned intervention targets of $I_1, ..., I_K$ cover $[\hat{p}]$ and the minimum loss function is not larger that for $\hat{p} = K$, we have the solution satisfy (1)-(5) in Remark 2. Since $\mathcal{G}$ has the sparsest transitive closure and $\mathcal{G}$ is the sparsest with this transitive closure, (6)-(7) in Remark 2 are also satisfied. Therefore Remark 2 guarantees the smallest $\hat{p} \leq K$ satisfying the conditions recovers the CD-equivalence class. □

Note that in practice, it can be hard to ensure that the gradient-based approach returns a DAG $\mathcal{G}$ that has the sparsest transitive closure and is simultaneously the sparsest DAG with this transitive closure. We instead search for sparser DAGs $\mathcal{G}$ by penalizing its corresponding adjacency in Eq. (3).

E.2 Consistency for Multi-Node Interventions

Theorem 3 guarantees that in an SCM with additive noises where interventions modify the exogenous noises, if the CD equivalence can be identified, we can extrapolate to unseen combinations of interventions with different intervention targets. In fact, for certain types of interventions, extrapolation to unseen combinations of any interventions is possible. We illustrate this for shift interventions in an SCM with additive Gaussian noises, where an intervention changes the mean of the exogenous noise variable.

For single-node intervention I, let a^I denote the corresponding changes in the mean of the exogenous noise variables, i.e.,

$$a_i^I = \begin{cases} \mathbb{E}(\epsilon_i^I) - \mathbb{E}(\epsilon_i), & i \in T(I), \\ 0, & i \notin T(I). \end{cases}$$

We encode it as $\hat{I}$ with $T(\hat{I})$ containing one element and $\hat{\mathbf{a}}^{\hat{I}}$ being a one-hot vector, where

$$\hat{a}_i^{\hat{I}} = \begin{cases} \hat{a}_i, & i \in T(\hat{I}), \\ 0, & i \notin T(\hat{I}). \end{cases}$$

We extend this notation for I with potentially multiple intervention targets (i.e., sets $I, \hat{I}$ that contain multiple elements) where $\mathbf{a}^I, \hat{\mathbf{a}}^{\hat{I}}$ can be a multi-hot vector.

In the shift intervention case, from Theorem 3, we know that the encoded $\hat{\mathbf{a}}^{\hat{I}_1}, ..., \hat{\mathbf{a}}^{\hat{I}_K}$ satisfy $\hat{\mathbf{a}}^{\hat{I}_k} = M(\mathbf{a}^{I_k})$ in the limit of infinite data, where M is a linear operation with $M(\mathbf{a})_i = \Upsilon_{\tau(i),i} a_{\tau(i)}$. Thus for single-node interventions $I_{t(1)}, ..., I_{t(k)}$ amongst $I_1, ..., I_K$, the multi-node intervention $\mathcal{I} = I_{t(1)} \cup ... \cup I_{t(k)}$[13] corresponds multi-hot vector $\mathbf{a}^{\mathcal{I}}$ that satisfies $M(\mathbf{a}^{\mathcal{I}}) = M(\mathbf{a}^{I_{t(1)}} + ... +$

[13]Note that we allow overlapping intervention targets among $I_{t(1)}, ..., I_{t(k)}$, where $I_{t(1)} \cup ... \cup I_{t(k)}$ adds up all the shift values for intervention target i.

$\mathbf{a}^{I_{t(k)}}) = \hat{\mathbf{a}}^{\hat{I}_{t(1)}} + ... + \hat{\mathbf{a}}^{\hat{I}_{t(k)}}$. Thus if we encode $\mathcal{I}$ as $\hat{\mathbf{a}}^{\hat{\mathcal{I}}} := \hat{\mathbf{a}}^{\hat{I}_{t(1)}} + ... + \hat{\mathbf{a}}^{\hat{I}_{t(k)}}$, we can also generate $\hat{X}^{\hat{\mathcal{I}}}$ from the ground-truth distribution of $X = f(U)$ where $U \sim \mathbb{P}_U^{\mathcal{I}}(U)$ following the encoding-decoding process of Fig. 4.

F Discrepancy-based VAE Implementation Details

We summarize our hyperparameters in Table 2. Below, we describe where they are used in more detail. We use a linear structural equation with shift interventions. In practice, due to the nonlinear encoding from the latent U to observed X, not much expressive power is lost. Code for our method is at https://github.com/uhlerlab/discrepancy_vae.

Loss function	
Kernel width (MMD)	200
Number of kernels (MMD)	10
λ	0.1
β_{max}	1
α_{max}	1
Training	
t_{max}	100
Learning rate	0.001
Batch size	32

Table 2: Hyper-Parameters

VAE Parameterization. As is standard with VAEs, our encoder and decoder are parameterized as neural networks, and the exogenous variables are described via the reparameterization trick. We use a standard isotropic normal prior for $p(Z)$. To encode interventions, the function $T_\phi(\cdot)$ is parameterized as a fully connected neural network, where for differentiable training $T_\phi(C)$ is encoded as a one-hot vector via a softmax function, i.e., $T_\phi(C)_i = \exp(tT'_\phi(C)_i) / \sum_{j=1}^{p} \exp(tT'_\phi(C)_j)$ for some fully connected T'_ϕ and temperature $t > 0$. During training, we adopt an annealing temperature for t. In particular, $t = 1$ until half of the epochs elapse, and t is linearly increased to t_{max} over the remaining epochs. At test time, the temperature of the softmax is set to a large value, recovering a close-to-true one-hot encoding.

Loss Functions. We use a mixture of MMD discrepancies, each with a Gaussian kernel with widths that are dyadically spaced [18]. This helps prevent numerical issues and vanishing gradient issues in training. The coefficient α of the discrepancy loss term $\mathcal{L}_{\theta,\phi}^{\text{discrep}}$ is given the following schedule: $\alpha = 0$ for the first 5 epochs, then α is linearly increased to α_{max} until half of the epochs elapse, at which point it remains at α_{max} for the rest of training. Similarly, the coefficient β of the KL regularization term is given the following schedule: $\beta = 0$ for the first 10 epochs, then β is linearly increased to β_{max} until half of the epochs elapse, at which point it remains at β_{max} for the rest of training.

Optimization. We train using the Adam optimizer, with the default parameters from PyTorch and a learning rate of 0.001.

Biological Data. For the experiments described in Section 6, the encoder q_ϕ was implemented as a 2-layer fully connected network with leaky ReLU activations and 128 hidden units. The intervention encoder T_ϕ uses 128 hidden units. To account for interventions with less samples, we use a batch size of 32. We train for 100 epochs in total, which takes less than 45 minutes on a single GPU.

G Extended Results on Biological Dataset

In this section, we provide additional evaluations of the experiments on the Perturb-seq dataset. The computation of RMSE are computed for individual interventional distributions. The computation of R^2 (we capped the minimum by 0 to avoid overflow) records the coefficient of determination by regressing the mean of the generated samples on the ground-truth distribution mean.

G.1 Single-node interventions

Figure 9 shows the same visualization as Figure 5 in the main text for the remaining $11 = 14 - 3$ single target-gene interventions with more than 800 cells. Figure 10 presents this side-by-side for the training samples. For the entire 105 single interventions, we visualize for each individual intervention the empirical MMD between the generated populations and ground-truth populations in Figure 11, where the bars record the MMD in different batches.

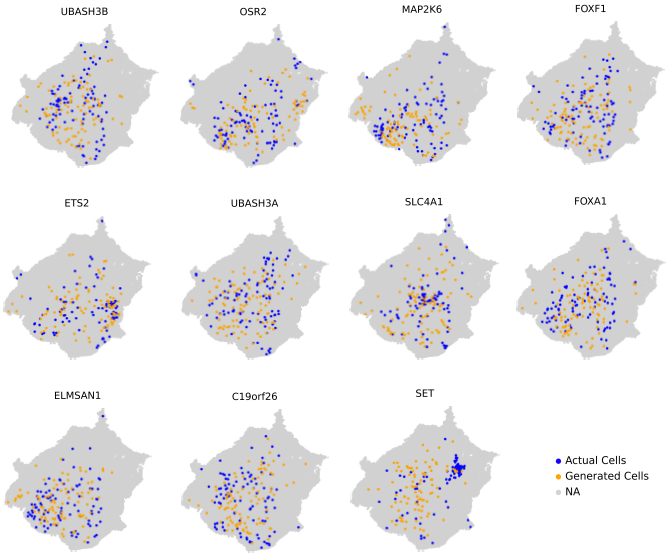


Figure 9: **For single-node interventions, the distribution of generated test samples visually mirrors the distribution of the actual samples.** A UMAP visualization of 11 single target interventions shows that the generated and the actual distributions closely match.

Figure 10: **For single-node interventions, the distribution of generated training samples visually mirrors the distribution of the actual samples.** As with the test samples, the distributions of the generated training samples closely match the actual distributions.

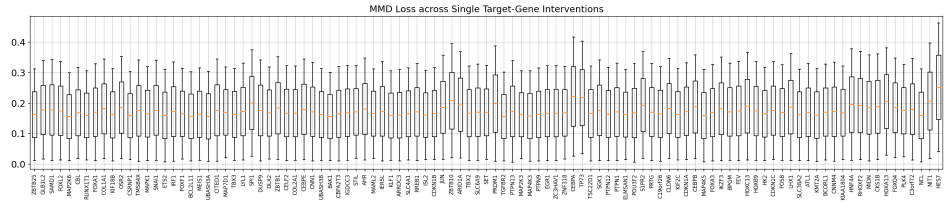


Figure 11: **For single-node interventions, the distribution of generated training samples quantitatively mirrors the distribution of the actual samples.** The figure shows the empirical MMD, defined in Appendix C.1, between the generated populations and ground-truth populations for 105 single target-node interventions.

G.2 Double-node interventions

We plot the generated samples for 11 random double target-gene interventions in Figure 12. In Figure 13, we highlight two interventions for which the generated samples differ from the actual samples. The plots for all 112 interventions are provided at https://github.com/uhlerlab/discrepancy_vae.

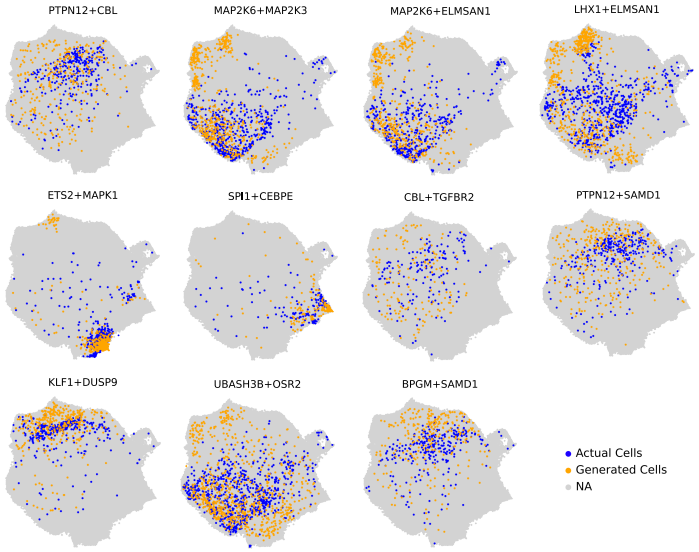


Figure 12: **UMAP visualization for a random sampling of double-node interventions.** Compared to single-node interventions, the generated samples of the double-node interventions match only for certain pairs.

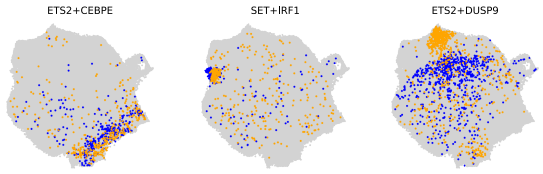


Figure 13: **For some double-node interventions, the generated samples match the actual samples, and for some combinations they do not.** The model accurately predicts the effect of the combinations ETS2+CEBPE and SET+IRF1, but does not accurately predict the effect of ETS2+DUSP9.

The MMD losses for all 112 interventions are summarized in Figure 14. Similar to Figure 6 in the main text, Figure 15 shows the distribution of RMSE and R^2 of the 112 interventions.

We remark here that this task has also been studied in previous works (e.g., [39, 8, 68, 49]) with different setups. Formally benchmarking the empirical results under a unified setting would be of interest in future works.

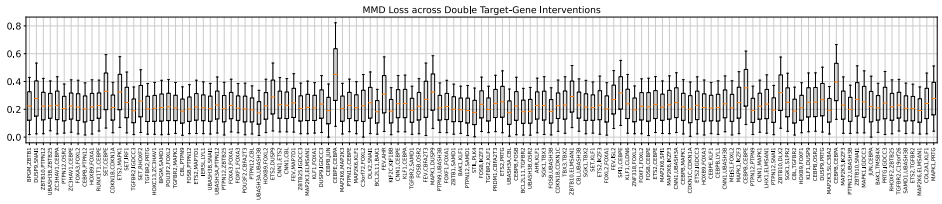


Figure 14: **For some double-node interventions, the distribution of generated samples quantitatively mirrors the distribution of the actual samples.** The figure shows the empirical MMD, defined in Appendix C.1, between the generated populations and ground-truth populations for 105 single target-node interventions.

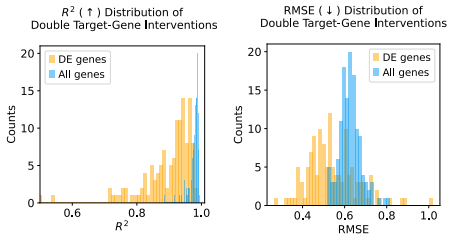
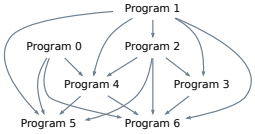


Figure 15: **Our model accurately predicts the effect of many double-node interventions.** 'All genes' indicates measurements using the entire 5000-dimensional vectors; 'DE genes' indicates measurements using the 20-dimensional vectors for the top 20 most differentially expressed genes.

G.3 Structure Learning

In Figure 16, we show the learned latent structure between gene programs, along with descriptions of each gene program.



Program 0: OSR2
Program 1: SLC38A2, MAP2K3, ELMSAN1, TBX2, MAML2, MAP2K6
Program 2: TGFBR2, ZBTB1, PLK4, HNF4A, ARID1A, ATL1, IKZF3, BPGM, BCORL1, LYL1, BCL2L11, HK2, RUNX1T1, CLDN6, CDKN1A, NCL, CNN1, MAP4K3, TSC22D1, HOXA13, CDKN1B, KIF2C, AHR, SGK1, DUSP9, ZC3HAV1, S1PR2, TP73, CKS1B, C19orf26, POU3F2, ZBTB25, CNNM4, FOXA3, SAMD1, STIL, PTPN12, IGDCC3, KMT2A, MAP7D1, PRDM1, SLC6A9, PTPN9, ZBTB10, CDKN1C, BAK1, MEIS1, CSRNP1, RREB1, MAP4K5, PTPN13, KIF18B, IER5L, CITED1, FOXF1, FOXA1, EGR1, KIAA1804, ZNF318, UBASH3B, PTPN1, PRTG, UBASH3A, RHOXF2, FOXO4, ARRDC3, GLB1L2, JUN, HOXC13, CBL
Program 3: MAPK1, CEBPE, TMSB4X, CELF2, DLX2, SPI1, HES7, NIT1, ETS2, LHX1, CBFA2T3, HOXB9, MIDN, CEBPA, SNAI1, FOSB, COL1A1, C3orf72, FOXL2, CEBPB, ISL2
Program 4: FEV, COL2A1, IRF1
Program 5: SLC4A1, SET, TBX3
Program 6: KLF1

Figure 16: Regulatory relationships between programs learned in $\mathcal{G}$ and full list of genes in each program.

H Extended Experiments

In this section, we provide additional experimental results. First, we perform ablation studies of different components of the proposed architecture on biological data. Then, we provide a simple simulation study to examine the performance of the framework on different tasks.

H.1 Ablation Studies

For the ablation studies of different components, we compared the performance of our final model (depicted in Figure 4) against three alternative versions. All models are trained with the same setting (data split, schedule, learning rate, etc). In particular, we compared against

- **Models without the discrepancy loss.** These models learn the distributions similar to conditional VAE [51], where both an interventional sample and its interventional label are fed in to learn the exogenous Z. Then inside the latent space, we use the same causal layer as our model to generate a virtual sample. During inference, we can generate interventional samples via two approaches. One is sampling the exogenous Z from $p(Z)$ and decoding. The other is sampling an observational sample, obtaining its exogenous Z using the encoder

then decoding. These two approaches correspond to the second and third rows of Table 3 respectively.

- **A model without the causal layer.** This model uses a similar workflow as our final model in Figure 4, where we do not use a causal-based decoder but a simple MLP decoder. This corresponds to the fourth row of Table 3.

We note that the encoder, decoder, DSCM, and intervention encoder are needed to learn distributions and the latent causal graph from this setting where observational and interventional data are present.

For the metrics, we report both MMD and R^2 in Table 3. However, MMD is more meaningful as we are assessing the quality of generating a distribution. We observe that models without discrepancy perform much worse due to mode collapses, whereas the sampling approach using observational data performs slightly better. Our final model works the best in general; however on the MMD for double-node interventions, the version without a causal layer seems to work slightly better. This is potentially because some double-node interventions that act non-additively can be captured better without imposing the structure.

Method	MMD (single)	R^2 (single)	MMD (double)	R^2 (double)
ours	**0.324±0.007**	**0.986±0.001**	0.432±0.006	**0.978±0.001**
ours w/o discrepancy	2.966±0.054	0.984±0.003	3.358±0.031	0.972±0.002
ours w/o discrepancy (obs)	2.965±0.054	0.984±0.002	3.355±0.030	0.972±0.002
ours w/o causal layer	0.348±0.009	0.982±0.002	**0.427±0.006**	**0.978±0.002**

Table 3: **Ablation studies.** We report testing metrics and their standard error on the biological datasets. The results on single-node interventions are computed over 14 interventions. The results on double-node interventions are computed over all 112 interventions.

H.2 Simulation

For the simulation study, as a proof-of-concept, we tested on a simple 5-node graph, where we generate 2048 samples in each of the 5 interventional datasets. We map this to a 10-dimensional observation space, where we pad zeros to the additional dimensions. This ensures clear visualization of the generated samples in Figure 17, where we compare the zero-shot learned double-node interventional samples against ground truth. In Table 4, we report the quantitative metrics. In addition to the MMD on left-out single and double-node interventions, we also report the training MMD and Structural Hamming Distance (SHD) of the learned graph.

Due to the combinatorial nature of learning a DAG and the small sample sizes in this setting, we observe that the learned intervention targets can be quite sensitive to initializations. Therefore during evaluation, we report the metrics while fixing the intervention targets to be of different transposition distances to the true targets. For single-node generations, different transposition distances return similar results, meaning that the model is expressive enough to learn these distributions, although we observe that the result with zero transposition distance is marginally better. This also holds during training, which can potentially be used as model selection to overcome the initialization issue. For double-node extrapolation, the result with zero transposition distance shows a larger benefit, as expected from our theory.

Transposition Distance	MMD (training)	MMD (single)	MMD (double)	SHD
0	0.030±0.007	0.047±0.008	0.041±0.004	2
1	0.057±0.028	0.058±0.030	0.181±0.048	6
10	0.042±0.007	0.041±0.009	0.119±0.023	11

Table 4: **A simple simulation study.** On a 5-node DAG, we test the model performance with varying transposition distances of the identified intervention targets. For sample generations, we report MMD and its standard error. The training metric is evaluated on all single-node interventions, where the third and forth rows are evaluated based on held-out samples of single and double-node interventions.

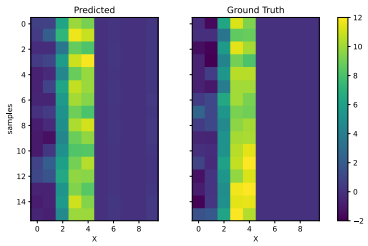


Figure 17: **An illustration of double-node intervention extrapolation in simulation.** We visualize 16 samples of the double-node intervention on nodes $2, 3$. The generated samples are shown on the left, where the ground-truth samples are shown on the right.

I Extended Discussion

I.1 Limitations and Future Work

This paper opens up several direction for future theoretical and empirical work, which we now discuss.

Theoretical Perspective. We have focused on the setting where a single-node intervention on each latent node is available, similar to prior works on causal disentanglement [3, 53]. However, we highlight three issues in this setup and discuss potential remedies. First, by assuming access to data from intervening on every single latent node, we inherently possess partial knowledge of all the latent variables, even though we are unaware of their specific values or whether multiple interventions act on the same variable. The setups that do not assume interventions but the existence of anchored observed variables (i.e., variables with only one latent parent) [19, 9, 64, 65] face the same issue. This assumption can be unsatisfying in the context of causal representation learning, where the causal variables are assumed to be entirely unknown. Second, it may be impossible to intervene on all latent causal variables, especially in scenarios involving latent confounding. For instance, in climate research, it might be impossible to intervene on a variable like the precipitation level in a particular region. Finally, the assumption of single-node interventions can be overly optimistic in many applications. For example, in the case of chemical perturbations on cells, it is known that drugs often target multiple variables.

Nevertheless, the results obtained in the current setup can serve as a foundation and stepping stone towards the ultimate goal of general causal representation learning. On one hand, our analysis showed what can be learned from each intervention. This is helpful when considering cases where only a subset of the latent causal variables can be intervened on. On the other hand, the key techniques employed in our proofs can be extended to the multi-node setting. Specifically, in the latent space, one should expect only the marginals of variables downstream of a multi-node intervention to change.

Moreover, we have primarily focused on the infinite data regime for analyzing identifiability. Considering the expensive nature of obtaining interventional samples in practice, there is ample room for further investigation concerning sample complexity. Aside from the feasibility of identifiability, many applications are concerned with specific downstream tasks. Full identification of the underlying causal representations provides a comprehensive understanding of the system and would be beneficial for multiple downstream tasks. However, in certain cases, full identification may be unnecessary or inefficient for a particular task. Therefore, it is of interest to develop task-specific identifiability criteria for causal representation learning.

Empirical Perspective. We make two remarks on the VAE framework proposed in this work. First, as shown in our experiments in Section 6, our proposed framework can still be applied in settings with multi-node interventions and fewer single-node interventions. For instance, one can model multi-node interventions by reducing the temperature in the softmax layer. Second, due to the permutation symmetry of CD-equivalence, we impose an upper-triangular structure on the adjacency matrix in the deep SCM and learn the intervention targets. Alternatively, when there is exactly one intervention available for each latent node, one can instead prefix the intervention targets and learn the adjacency matrix. Specifically, we can set the intervention targets of $I_1, ..., I_p$ to be

a random permutation of $[p]$. Subsequently, the adjacency matrix can be learned for example via the nontears penalty [70] to enforce acyclicity. However, both methods inherit the combinatorial nature of learning a DAG, and therefore their performance may require large sample sizes and can be sensitive to initialization [27]. Consequently, endeavors to improve the optimization process and robustness of such models would be valuable.

I.2 Discussion of Contemporaneous Works.

This work is concurrent with a number of other works in interventional causal representation learning. Unless otherwise noted, all of these works consider single-node interventions, as we do in this paper. Most similar to our setting is [59], which studies identifiability of nonparametric latent SCMs under linear mixing. They consider the case where exactly one intervention per latent node is available, which is an easier setting as we discussed in Section 2. In that setting, they provide a characterization of the learned causal variables. On the other hand, [7] studies identifiability of a linear latent SCM under nonparametric mixing. They also consider both hard and soft interventions, but in the form of linear SCM with additive Gaussian noises. Three concurrent works [25, 60, 35] consider *both* nonparametric SCMs and nonparametric mixing functions: [60] prove identifiability for the case of $p = 2$ latent variables when there is one intervention per latent variable. They provide an extension to arbitrary p for settings where there are paired interventions on each latent variable. Meanwhile, [25] consider arbitrary p, without paired interventions. However, they use only conditional independence statements over the observed variables X to recover the latent causal graph. As a result, their identifiability guarantees place restrictions on the latent causal graph, unlike the other works discussed here. The third work [35] studies the *Causal Component Analysis* problem, where the latent causal graph is assumed to be known. Finally, we note that other concurrent works study causal representation learning without interventional data [40, 32] or with vector-valued contexts instead of interventions [31].