# Appendices

## Table of Contents



## A Graph terminology

A directed graph $G = (V, T)$ is an object consisting of a set of vertices $V$ and a set of ordered pairs of vertices $T \subset V \times V$ corresponding to directed edges in $G$. A *path* is a sequence of vertices $(V_{i_1}, \ldots, V_{i_n})$ with $n \geq 2$ such that $V_{i_k} \to V_{i_{k+1}}$ or $V_{i_k} \leftarrow V_{i_{k+1}}$ and in a *directed path* $V_{i_k} \to V_{i_{k+1}}$ for all $k$. In graph $G$: the *children* $\mathbf{CH}_j^G$ of $V_j$ are all $V_m$ such that $V_j \to V_m$, the *parents* $\mathbf{PA}_j^G$ of $V_j$ are all $V_m$ such that $V_m \to V_j$, the *ancestors* $\mathbf{AN}_j^G$ of $V_j$ are all $V_m$ such that there exists a directed path $(V_m, \ldots, V_j)$, and the *descendants* $\mathbf{DE}_j^G$ of $V_j$ are $V_j$ and all $V_m$ such that there exists a directed path $(V_j, \ldots, V_m)$. The graph superscript will be omitted unless needed. A *cycle* is a path such that $V_{i_1} = V_{i_n}$ and a *directed acyclic graph (DAG)* is a directed graph with no directed cycles.

On a path $(V_{i_1}, \ldots, V_{i_k}, \ldots, V_{i_n})$, we say variable $V_{i_k}$ is a *collider* if $V_{i_{k-1}} \to V_{i_k}$ and $V_{i_k} \leftarrow V_{i_{k+1}}$. A subset $\mathbf{Z} \in \mathbf{V} \setminus \{V_{i_1}, V_{i_n}\}$ *blocks* the path if either (i) $\mathbf{Z}$ contains at least one non-collider vertex on the path or (ii) the path contains a collider with no descendants in $\mathbf{Z}$ (this includes the collider itself by the descendant definition). With this terminology, we say that on the disjoint variable sets $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{Z}$, $\mathbf{A}$ is *d-separated* from $\mathbf{B}$ by $\mathbf{Z}$ iff every path between $\mathbf{A}$ and $\mathbf{B}$ is blocked by $\mathbf{Z}$ [45, Def. 6.1]. This is denoted as $\mathbf{A} \perp\!\!\!\perp_G \mathbf{B} \mid \mathbf{Z}$. If $\mathbf{A}$ and $\mathbf{B}$ are not d-separated, and hence there exists an unblocked path, we say that they are *d-connected*.

## B Full proofs

### B.1 Characterization of the learned equivalence class

Our first result is not focused on in the main text but is nonetheless interesting in relation to related works and creating a coherent multi-environment causal discovery framework. Specifically, it shows that the MSS solution set is an equivalence class of DAGs. Our main results demonstrate when, how, and under what conditions this equivalence class shrinks to only the true DAG.

**Proposition B.1.** *$\mathcal{G}^{\min}_{\text{MEC}}$ is the $\Psi$-MEC corresponding to (i.e., containing) the true (but unknown) graph $G^*$ and (unobserved) intervention targets $\{\mathcal{I}^e\}_{e=1}^{n_e}$.*

*Proof.* First we must introduce the $\Psi$-MEC concept as introduced by Jaber et al. [29], which relies on what we will call the *pairwise augmented graph*, in order to distinguish it from the augmented CGM defined in Defn. 2.5. The following definition is rephrased from Jaber et al. [29, Def. 4] to match our existing notation, and without latent variables as in the context of our setting.

**Definition B.2** (Pairwise augmented graph)**.** Let $\{(G, \mathbb{P}^e_{\mathbf{X}})\}_{e\in\mathcal{E}}$ be a collection of CGMs over the DAG $G = (\mathbf{X}, T)$ from environments $\mathcal{E}$. For a pair of environments $e, e' \in \mathcal{E}$, $e \neq e'$, construct an auxiliary vertex $E_{e,e'}$ and an auxiliary set of edges

$$T_{e,e'} := \{E_{e,e'} \to X_j : \mathbb{P}^e_{\mathbf{X}}(X_j \mid \mathbf{PA}^G_j) \neq \mathbb{P}^{e'}_{\mathbf{X}}(X_j \mid \mathbf{PA}^G_j)\}.^{7}$$

from $E_{e,e'}$ to each variable with a change in mechanism. The *pairwise augmented graph* is

$$Aug(G; \{\mathbb{P}^e_{\mathbf{X}}\}_{e\in\mathcal{E}}) := (\mathbf{X} \cup \{E_{e,e'}\}_{e,e'\in\mathcal{E}, e\neq e'}, T \cup \{T_{e,e'}\}_{e,e'\in\mathcal{E}, e\neq e'})$$

For the remainder of the proof, we will refer to the augmented graph only with respect to this pairwise augmented graph, not the augmented GCM. Note that the pairwise augmented graph differs from the augmented CGM in that it adds vertices for all pairs of environments, but these vertices are treated as parameters rather than as random variables; thus, in contrast to the augmented CGM whose environmental variable is a (discrete) random variable with support over the environments, there is not an explicit distribution over the entire pairwise augmented graph, rather only pairs of environments. Using the pairwise augmented graph, Jaber et al. [29] provide the following result

**Corollary B.3** (Jaber et al. [29])**.** *Let $G_1$ and $G_2$ be two DAGs on $\mathbf{X}$; let $\{\mathcal{I}^e_1\}_{e\in\mathcal{E}}$ and $\{\mathcal{I}^e_2\}_{e\in\mathcal{E}}$ be two sets of (unobservable) intervention targets, which by definition respectively induce two sets of (observable) interventional distributions $\{\mathbb{P}^e_{1,\mathbf{X}}\}_{e\in\mathcal{E}}$ and $\{\mathbb{P}^e_{2,\mathbf{X}}\}_{e\in\mathcal{E}}$.*[8] *The pairs of graphs and intervention targets $(G_1, \{\mathcal{I}^e_1\}_{e\in\mathcal{E}})$ and $(G_2, \{\mathcal{I}^e_2\}_{e\in\mathcal{E}})$ are $\Psi$-Markov equivalent iff $Aug(G_1; \{\mathbb{P}^e_{1,\mathbf{X}}\}_{e\in\mathcal{E}})$ and $Aug(G_2; \{\mathbb{P}^e_{2,\mathbf{X}}\}_{e\in\mathcal{E}})$ are in the same MEC, i.e., have the same skeleton and v-structures.*

Now recall that by definition

$$\mathcal{G}^{\min}_{\text{MEC}} := \arg\min_{G\in\mathcal{G}_{\text{MEC}}} \text{MSS}(G; \mathcal{P})$$

To show that $\mathcal{G}^{\min}_{\text{MEC}}$ is the $\Psi$-MEC corresponding to (i.e., containing) the true (but unknown) graph $G^*$ and (unobserved) intervention targets $\{\mathcal{I}^e\}_{e=1}^{n_e}$, we need to verify the two conditions of Cor. B.3 for the true DAG, which is contained in $\mathcal{G}^{\min}_{\text{MEC}}$, and any other graph in this set. Specifically, under the distribution shifts implied by $\{\mathcal{I}^e\}_{e=1}^{n_e}$, we show that ($\Rightarrow$) the pairwise augmented graphs of the true DAG $G^*$ and any graph $G \in \mathcal{G}^{\min}_{\text{MEC}}$ share the same skeleton and v-structure, and ($\Leftarrow$) the pairwise augmented graphs of the true DAG $G^*$ and any graph $G \notin \mathcal{G}^{\min}_{\text{MEC}}$ differ either in skeleton or v-structures.

($\Rightarrow$) To verify the forward direction, we verify that the skeleton and v-structures are shared.

[7] Note that in the graph definition in Jaber et al. [29], they start with the (unobservable) intervention targets, which induce (observable) distribution shifts that define the pairwise augmented graph. We start from the distribution shifts directly since we do not explicitly model the sets of possible intervention targets which could have given rise to the observed distribution shifts.

[8] As with the pairwise augmented graph, we prefer to think about the observed distributions rather than interventions. The fact that the $\Psi$-MEC is defined over both graphs and interventions is because without knowledge of a baseline environment, multiple (unobservable) interventions can induce the same observed distributions.

- **Skeleton**: Since $\mathcal{G}^{\min}_{\text{MEC}} \subseteq \mathcal{G}_{\text{MEC}}$, by definition of the MEC all *unaugmented* DAGs in the set share the same skeleton. For sake of contradiction, assume that some graph $G$ contains an augmented edge, which the true augmented causal DAG does not. Then, by definition of the augmented graph (Defn. B.2), $G$ has a variable with changing mechanism which does not change under the true causal DAG, so $G$ cannot be in the minimal set $\mathcal{G}^{\min}_{\text{MEC}}$. Hence, all augmented DAGs in $\mathcal{G}^{\min}_{\text{MEC}}$ share the same skeleton.
- **v-structures**: Since $\mathcal{G}^{\min}_{\text{MEC}} \subseteq \mathcal{G}_{\text{MEC}}$, by definition of the MEC all *unaugmented* DAGs in the set share the same v-structures. *Augmented* DAGs form additional v-structures through edges from the augmented variables $E^{e,e'}$. In the case of a v-structure between two augmented variables, $E^{e^1,e^2} \to X_i \leftarrow E^{e^3,e^4}$, it is necessarily shared because of the shared skeleton condition and orientation of edges out of augmented variables, by definition.Otherwise, under the true DAG $G^*$, let the augmented graph $Aug(G^*, \{\mathbb{P}^e_{\mathbf{X}}\}_{e\in\mathcal{E}})$ contain the v-structure $E^{e,e'} \to X_i \leftarrow X_i$ and $E_{e,e'} \not\to X_j$. Thus the mechanism of $X_j$ is invariant across environments $e$ and $e'$. If we assume that some other graph $G$ with augmented graph $Aug(G, \{\mathbb{P}^e_{\mathbf{X}}\}_{e\in\mathcal{E}})$ shares the same skeleton as the augmented graph of $G^*$ but does not contain this v-structure, then $Aug(G, \{\mathbb{P}^e_{\mathbf{X}}\}_{e\in\mathcal{E}})$ must contain the structure $E^{e,e'} \to X_i \to X_i$. But then $G's$ mechanism $\mathbb{P}^e_{\mathbf{X}}(X_j \mid \mathbf{PA}^G_j) \neq \mathbb{P}^{e'}_{\mathbf{X}}(X_j \mid \mathbf{PA}^G_j)$ would differ since under $G$ we would condition on the true collider $X_i$ (as specified in $G^*$) and unblock the path from $E$ to $X_j$. By definition, $Aug(G, \{\mathbb{P}^e_{\mathbf{X}}\}_{e\in\mathcal{E}})$ would then necessarily contain the edge $E_{e,e'} \to X_j$ which is not contained in $Aug(G^*, \{\mathbb{P}^e_{\mathbf{X}}\}_{e\in\mathcal{E}})$, contradicting the shared skeleton assumption.

($\Leftarrow$) For the sake of contradiction, assume there exists some DAG $G \notin \mathcal{G}^{\min}_{\text{MEC}}$ but which satisfies the $\Psi$-MEC equivalence conditions with $G^*$, i.e., shares the skeleton and v-structures with the true augmented graph. The graph $G$ must be in the same MEC as $G^*$ or else by definition differ in skeleton or v-structures in the normal graph and hence also in the augmented graph. Since $G$ is in the MEC but $G \notin \mathcal{G}^{\min}_{\text{MEC}}$, we know that $\text{MSS}(G) > \text{MSS}(G^*)$ and thus there must be some variable $X_j$ with changing mechanism across some pair of environments $e, e' \in \mathcal{E}$ in graph $G$ but not $G^*$. But this would immediately imply the existence of the edge $E^{e,e'} \to X_j$ in the augmented graph of $G$ but not in $G^*$. This is a contradiction with the shared skeleton condition of the $\Psi$-MEC and so no graphs outside of $\mathcal{G}^{\min}_{\text{MEC}}$ can be in the $\Psi$-MEC.

□

### B.2 Identifiability rate of the true DAG

### B.3 Proof of Lemma 4.3

**Lemma 4.3.** *For any $X_j \in \mathbf{X}$ and set $\mathbf{Z} \subseteq \mathbf{X}\setminus\{X_j\}$, the conditional distribution $\mathbb{P}(X_j \mid \mathbf{Z})$ changes if and only if the following d-connectedness relationship holds:*

$$X_j \not\perp\!\!\!\perp_{G_{\mathbf{X}\cup E}} E \mid \mathbf{Z}\,.$$

*Proof.* ($\Rightarrow$) If $\mathbb{P}(X_j \mid \mathbf{Z})$ changes across environments $E$, then $X_j \not\perp\!\!\!\perp E \mid \mathbf{Z}$. The global Markov property of the CGM states that d-separation implies conditional independence, and thus by the contra-positive the d-connectedness relationship follows.

($\Leftarrow$) d-connectedness implies conditional dependence by faithfulness, and thus a change across environments. □

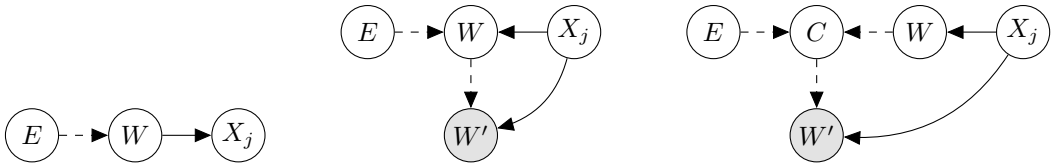
### B.4 Proof of Corollary 4.4

**Corollary 4.4.** *For any variable $X_j \in \mathbf{X}$ and set $\mathbf{Z} \subseteq (\boldsymbol{PA}^G_j \cup \boldsymbol{CH}^G_j)$ in the augmented graph, the conditional distribution $\mathbb{P}(X_j \mid \mathbf{Z})$ changes if and only if at least one of the following holds:*

*(i) $E \to X_j$ [a direct cause].*

*(ii) $\exists W_{\boldsymbol{PA}} \in \boldsymbol{PA}^G_j \setminus \mathbf{Z}$ such that $W_{\boldsymbol{PA}} \not\perp\!\!\!\perp_{G_{\mathbf{X}\cup E}} E \mid \mathbf{Z}$ [unblocked path to unconditioned parent].*

*(iii) $\exists W_{\boldsymbol{CH}} \in \boldsymbol{CH}^G_j \cap \mathbf{Z}$ such that $W_{\boldsymbol{CH}} \not\perp\!\!\!\perp_{G_{\mathbf{X}\cup E}} E \mid \mathbf{Z} \setminus W_{\boldsymbol{CH}}$ [unblocked path to conditioned child].*




(a) $W$ is a parent. (b) $W$ is a child and collider. (b) $W$ is a child but not a collider.

Figure 6: Cases from the proof of Cor. 4.4. Case (a) iff case (ii). Case (b) induces two subcases either of which occur iff case (iii).

*Proof.* By Lemma 4.3, $\mathbb{P}(X_j \mid \mathbf{Z})$ changes iff $X_j \not\perp\!\!\!\perp_{G_{\mathbf{X}\cup E}} E \mid \mathbf{Z}$ (d-connection) and equivalently, iff there is an unblocked path $(E, \dots, X_j)$ in $G_{\mathbf{X}\cup E}$. We assume a generic path and through casework establish that it is unblocked if and only if one of cases (i), (ii), or (iii) holds. The casework is visualized in Figure 7.

Either $(E, \dots, X_j)$ is just $E \to X_j$ [case (i)] or there must exist $W$ such that $(E, \dots, W, X_j)$ and $W$ is either (a) a collider or (b) not a collider.

(a) If $W$ is a collider, it necessarily a child of $X_j$ The collided path in unblocked iff $W \not\perp\!\!\!\perp_{G_{\mathbf{X}\cup E}} E \mid \mathbf{Z} \setminus W$ and some descendant $W' \in \mathbf{DE}_W^G$ of $W$ is conditioned on. Thus $W' \in \mathbf{Z} \subset \mathbf{PA}_j^G \cup \mathbf{CH}_j^G$. Without loss of generality, assume $W'$ is the closest descendant to $W$ and hence the path $(W, \dots, W')$ is unblocked by $\mathbf{Z}$. $W'$ cannot be a parent of $X_j$, else induce the cycle $(X_j, W, \dots, W', X_j)$, and so must be a child and case (iii) holds. Specifically, there exists $W'$ such that $W' \not\perp\!\!\!\perp_{G_{\mathbf{X}\cup E}} E \mid \mathbf{Z} \setminus W'$ and $W'$ is a child in $\mathbf{Z}$.

(b) If $W$ is not a collider, by definition the path is unblocked iff $W \not\perp\!\!\!\perp_{G_{\mathbf{X}\cup E}} E \mid \mathbf{Z}$ and $W \notin \mathbf{Z}$. If $W$ is a parent of $X_j$ (since they are adjacent), case (ii) holds. If $W$ is a child of $X_j$, because $E$ has an outgoing edge there must exist some collider $C$ on the path such that $(E, \dots, C, \dots, W, X_j)$ and the subpath from $W$ to $C$ is directed into $C$. The condition $W \not\perp\!\!\!\perp_{G_{\mathbf{X}\cup E}} E \mid \mathbf{Z}$ holds iff some descendant $W'$ of $C$ is in $\mathbf{Z}$. As before, $W'$ cannot be a parent of $X_j$ or else induce a cycle, and so it must be a child and case (iii) holds.

□

### B.5 Proof of Lemma 5.2

**Lemma 5.2** (Identifiability of causal parents)**.** *Let $G^*$ be the true DAG in the MEC $\mathcal{G}_{\text{MEC}}$ and $\rho_i$ the probability that the causal mechanism of $X_i$ is different across any two environments. Under Asms. 2.2 to 2.4 and 2.6, for any $j \in \{1, \dots, d\}$, graph $G \in \mathcal{G}_{\text{MEC}}$ such that $\mathbf{PA}_j^{G^*} \neq \mathbf{PA}_j^G$, and lower and upper bounds on the shift probabilities $\rho_i^{\text{LB}} \leq \rho_i \leq \rho_i^{\text{UB}}$ for all $i$, we have that*

$$\Pr[\text{MSS}_j(G^*; \mathcal{P}) < \text{MSS}_j(G; \mathcal{P})] \geq 1 - \left(1 - (1 - \rho_j^{\text{UB}}) \min_i \rho_i^{\text{LB}}\right)^{\lfloor n_\mathcal{E}/2 \rfloor}.$$

*Proof.* By Assumption 2.3, the distribution $\mathbb{P}_{\mathbf{X}}^e$ in each environment $e \in \{1, \dots, n_\mathcal{E}\}$ is the result of changing mechanisms from some underlying yet unknown distribution $\mathbb{P}_{\mathbf{X}}$. Let $\Delta^{e,e'}(X_j)$ denote the event $\mathbb{I}[\mathbb{P}_{\mathbf{X}}^e(X_j \mid \mathbf{PA}_j^{G^*}) \neq \mathbb{P}_{\mathbf{X}}^{e'}(X_j \mid \mathbf{PA}_j^{G^*})]$ that the mechanism of variable $X_j$, with respect to the true graph $G^*$, changes across environments $e$ and $e'$. Abbreviate $\rho_j^{e,e'} := \Pr[\Delta^{e,e'}(X_j) = 1]$.

Since $\mathbf{PA}_j^{G^*} \neq \mathbf{PA}_j^G$ and $G$ shares the same skeleton as $G^*$, at least one edge must be oriented incorrectly in $G$. In the conditioning set $\mathbf{PA}_j^G$ according to the incorrect graph $G$, there thus exists either an unconditioned true parent $Z \in \mathbf{PA}_j^{G^*} \setminus \mathbf{PA}_j^G$ or a conditioned-upon true child $Z \in \mathbf{CH}_j^{G^*} \cap \mathbf{PA}_j^G$. By Cor. 4.4, we know that if $Z$ is not d-separated from $E$ in the augmented graph, then the conditional $\mathbb{P}(X_j \mid \mathbf{PA}_j^G)$ changes across $E$. This occurs at least if the mechanism of $Z$ directly changes, e.g. there is the edge $E \to Z$ in the augmented graph.

Consider first the case of two environments. We know from Prop. 5.1 that $\text{MSS}_j(G^*; \mathcal{P})$ cannot be greater than $\text{MSS}_j(G; \mathcal{P})$, and will be less if the mechanism of $X_j$ remains invariant while the

mechanism of $Z$ changes. By the assumption of independent changing mechanisms,

$$\begin{aligned}
&\Pr[\text{MSS}_j(G^*; \{\mathcal{D}^1, \mathcal{D}^2\}) = \text{MSS}_j(G; \{\mathcal{D}^1, \mathcal{D}^2\})] \\
&= 1 - \Pr[\text{MSS}_j(G^*; \{\mathcal{D}^1, \mathcal{D}^2\}) < \text{MSS}_j(G; \{\mathcal{D}^1, \mathcal{D}^2\})] \\
&\leq 1 - \Pr[\Delta^{1,2}(X_j) = 0, \Delta^{1,2}(Z) = 1] \\
&= 1 - \Pr[\Delta^{1,2}(X_j) = 0] \Pr[\Delta^{1,2}(Z) = 1] \\
&= 1 - (1 - \rho_j^{1,2})\rho_Z^{1,2}
\end{aligned}$$

Given $n_{\mathcal{E}} > 2$ environments, it follows that

$$\begin{aligned}
&\Pr[\text{MSS}_j(G^*; \mathcal{P}) = \text{MSS}_j(G; \mathcal{P})] \\
&= \Pr\left[\bigcap_{e,e'>e} \text{MSS}_j(G^*, \{\mathcal{D}^e, \mathcal{D}^{e'}\}) = \text{MSS}_j(G, \{\mathcal{D}^e, \mathcal{D}^{e'}\})\right] \\
&\leq \Pr\left[\bigcap_{e\in\{1,\dots,\lfloor\mathcal{E}/2\rfloor\}} \text{MSS}_j(G^*, \{\mathcal{D}^{2e-1}, \mathcal{D}^{2e}\}) = \text{MSS}_j(G, \{\mathcal{D}^{2e-1}, \mathcal{D}^{2e}\})\right] \\
&= \prod_{e\in\{1,\dots,\lfloor\mathcal{E}/2\rfloor\}} \Pr\left[\text{MSS}_j(G^*, \{\mathcal{D}^{2e-1}, \mathcal{D}^{2e}\}) = \text{MSS}_j(G, \{\mathcal{D}^{2e-1}, \mathcal{D}^{2e}\})\right] \\
&\leq \prod_{e\in\{1,\dots,\lfloor\mathcal{E}/2\rfloor\}} \left(1 - (1 - \rho_j^{2e-1,2e})\rho_Z^{2e-1,2e}\right) \\
&\leq \left(1 - \min_{e\in\{1,\dots,\lfloor\mathcal{E}/2\rfloor\}} (1 - \rho_j^{2e-1,2e})\rho_Z^{2e-1,2e}\right)^{\lfloor n_{\mathcal{E}}/2\rfloor}.
\end{aligned}$$

Since $Z$ is arbitrary, we construct an upper bound using the worst case, in which a variable frequently or rarely changes.

$$\begin{aligned}
1 - \min_{e\in\{1,\dots,\lfloor\mathcal{E}/2\rfloor\}} (1 - \rho_j^{2e-1,2e})\rho_Z^{2e-1,2e} &\leq 1 - (1 - \max_{e\in\{1,\dots,\lfloor\mathcal{E}/2\rfloor\}} \rho_j^{2e-1,2e}) \min_{e\in\{1,\dots,\lfloor\mathcal{E}/2\rfloor\}} \rho_Z^{2e-1,2e} \\
&\leq 1 - (1 - \max_{e,e'\neq e} \rho_j^{e,e'}) \min_{e,e'\neq e} \rho_Z^{e,e'}
\end{aligned}$$

and so to acquire the final bound with simplified notation, for any variable $X_i$ denote the minima and maxima of $\rho_i^{e,e'}$ across any two environments with $\rho_i^{\text{LB}}$ and $\rho_i^{\text{UB}}$, respectively. □

### B.6 Proof of Theorem 5.3

**Theorem 5.3** (Identifiability of the graph). *Let $G^*$ be the true DAG in the MEC $\mathcal{G}_{\text{MEC}}$ and $\rho_j$ the probability that the causal mechanism of $X_j$ is different across any two environments. Under assumptions 2.2, 2.3, 2.4, and 2.6, and bounds $\rho_i^{\text{LB}} \leq \rho_i \leq \rho_i^{\text{UB}}$ for all $i$, we have that*

$$\Pr[\mathcal{G}_{\text{MEC}}^{\min} = \{G^*\}] \geq 1 - |\mathcal{G}_{\text{MEC}}| \left(1 - (1 - \min_i \rho_i^{\text{UB}}) \min_i \rho_i^{\text{LB}}\right)^{\lfloor n_{\mathcal{E}}/2\rfloor}.$$

*Proof.* Since $\Pr[\mathcal{G}^{\min}_{\text{MEC}} = \{G^*\}] = 1 - \Pr[\mathcal{G}^{\min}_{\text{MEC}} \neq \{G^*\}]$ and by Lemma 5.2,

$$
\begin{aligned}
\Pr[\mathcal{G}^{\min}_{\text{MEC}} \neq \{G^*\}] &= \Pr\left[\bigcup_{G \in \mathcal{G}_{\text{MEC}} \setminus \{G^*\}} \text{MSS}(G^*; \mathcal{P}) = \text{MSS}(G; \mathcal{P})\right] \\
&\leq \sum_{G \in \mathcal{G}_{\text{MEC}}} \Pr\left[\text{MSS}(G^*; \mathcal{P}) = \text{MSS}(G, \mathcal{D})\right] \\
&\leq \sum_{G \in \mathcal{G}_{\text{MEC}}} \Pr\left[\sum_j \text{MSS}_j(G^*; \mathcal{P}) = \sum_j \text{MSS}_j(G, \mathcal{D})\right] \\
&= \sum_{G \in \mathcal{G}_{\text{MEC}}} \Pr\left[\sum_j \text{MSS}_j(G^*; \mathcal{P}) = \sum_j \text{MSS}_j(G, \mathcal{D})\right] \\
&= \sum_{G \in \mathcal{G}_{\text{MEC}}} \Pr\left[\bigcap_j \text{MSS}_j(G^*; \mathcal{P}) = \text{MSS}_j(G, \mathcal{D})\right] \\
&\leq \sum_{G \in \mathcal{G}_{\text{MEC}}} \min_j \Pr\left[\text{MSS}_j(G^*; \mathcal{P}) = \text{MSS}_j(G, \mathcal{D})\right] \\
&\leq \sum_{G \in \mathcal{G}_{\text{MEC}}} \min_j \left(1 - (1 - \rho_j^{\text{UB}}) \min_i \rho_i^{\text{LB}}\right)^{\lfloor n_{\mathcal{E}}/2 \rfloor} \\
&\leq \sum_{G \in \mathcal{G}_{\text{MEC}}} \left(1 - (1 - \min_j \rho_j^{\text{UB}}) \min_i \rho_i^{\text{LB}}\right)^{\lfloor n_{\mathcal{E}}/2 \rfloor} \\
&= |\mathcal{G}_{\text{MEC}}| \left(1 - (1 - \min_j \rho_j^{\text{UB}}) \min_i \rho_i^{\text{LB}}\right)^{\lfloor n_{\mathcal{E}}/2 \rfloor}.
\end{aligned}
$$

□

# C Details of assumptions and methods

## C.1 Pseudo causal sufficiency and the Independent Causal Mechanisms (ICM) assumption

Huang et al. [28] introduced the idea of pseudo-causal sufficiency (Asm. 2.6) and provide a useful discussion on its relation to results on soft interventions by Eberhardt and Scheines [10]. Guo et al. [18] provide a useful formalization of multi-environment data, specifically through a plate-notation representation. An environment $e$ specifies parameters of the causal mechanisms in the CGM over $\mathbf{X}$; we can think of environments as encapsulating specific experimental settings, or broad contexts such as climate or time [38]. Under the context of $e$, there is some distribution $\mathbb{P}^e_{\mathbf{X}}$ and we observe a dataset sampled i.i.d. The ICM assumption tells us that the parameters for each causal mechanism in an environment are chosen or sampled independently, and thus in the augmented CGM the edges from $E$ appear independently.

Within each environment, i.e., when we condition on $E$, the environmental parameters are fixed; thus we are in the typical i.i.d. setting and causal sufficiency is implied by the CGM. However, without conditioning on $E$, the environmental parameters are not fixed and across two samples either all remain the same (if the samples are in the same environment) or some change. This dependence between samples through the parameters defined by $E$ is the result of $E$ being a confounder; thus causal sufficiency cannot hold over $\mathbf{X}$ without conditioning on $E$. Because $E$ is not necessarily a true causal variable but rather an environment encoding a fixed set of unmeasured variables, Huang et al. [28] call it a *pseudo-confounder*. It is worth noting that the second stage of the approach of Huang et al. [28] relies on a novel kernel-based test, which computes a measure of mechanism dependence across all samples. They correctly compare the test statistics rather than examine $p$-values because the dependence between samples without controlling for environment would lead to a samll $p$-value even if the mechanisms were independent.

### C.2 The *p*-values "soft" score

We provide further details on the *p*-value "soft" score. Recall the modified score definition to be

$$\widehat{\text{MSS}}_j(G;\mathcal{D}) = \sum_{e=1,e'>e}^{n_\mathcal{E}} \left[1 - p\text{-value}\left(\mathbb{P}^e(X_j \mid \mathbf{PA}_j^G) \neq \mathbb{P}^{e'}(X_j \mid \mathbf{PA}_j^G)\right)\right].$$

Using a test of equality of distribution, we calculate a test statistic; at a pre-specified level $\alpha$, if the test is well specified [53], the one-sided *p*-value is valid and corresponds to the probability under the null hypothesis $H_0 : \mathbb{P}^e(X_j \mid \mathbf{PA}_j^G) = \mathbb{P}^{e'}(X_j \mid \mathbf{PA}_j^G)$ of a test statistic as large or larger than the observed test statistic.

If a mechanism changes, a powerful test should yield a small *p*-value and thus a term close to 1 in the summation, similar to the "hard" score. If a mechanism doesn't change, since *p*-values are uniformly distributed in $[0, 1]$ under the null hypothesis, the term in the sum would be similarly uniformly distributed. With enough variables and environments, the variance of the sum of random uniform variables will decrease and the behavior of the score will be dominated by the *p*-values under the alternatives. It must be noted that the *p*-values are not independent, as some will use data from the same environments.

### C.3 Comparison to other augmented graph methods

The reviewing process and conference brought to our attention multiple other relevant works that merit more in-depth discussion. The idea of the augmented graph or CGM is not new: Mooij et al. [38] and Huang et al. [28] both pool all data by incorporating a *single* environmental variable. As we have demonstrated, however, with more environments this variable becomes more densely connected and thus the size of the learned equivalence class *increases*.

In addition to the present work, two additional works have identified the utility in pooling only pairs of environments. Jaber et al. [29] explicitly construct the pairwise augmented graph, adding a node for each pair of environments, and then apply the constraint-based PC algorithm. As mentioned, this is similar to what was done by Huang et al. [28] who only added one augmented variable. Similarly, Squires et al. [57] use an existing score-based method on the pairwise augmented graph, which penalizes the number of edges (i) between variables and (ii) from augmented nodes. These penalties restrict the solution set to the (i) MEC and (ii) minimum shift set therein, respectively.

The estimands of both Jaber et al. [29] and Squires et al. [57] are equivalent to ours under an oracle test. In practice, on finite samples, there are two points of disagreement. First, we require the MEC as an input while both other methods learn it as part of their algorithm. Although this may appear to be a downside of our approach and does require specifying a separate procedure to first learn the MEC, it actually permits additional flexibility in that we may use any of the many existing methods to learn the MEC. Notably, this allows us to pool *all* environments as in Huang et al. [28], Mooij et al. [38] to learn the MEC, decreasing errors by increasing sample sizes of the conditional independence tests over the samples sizes attained by only pairwise pooling. Second, we efficiently enumerate over graphs in the MEC, while Jaber et al. [29] use PC orientation rules and Squires et al. [57] use a consistent greedy search. Even when done efficiently, enumeration may be slower. Yet, it is possible in finite samples for the other two approaches to rule out graphs in their procedures which would otherwise have the fewest mechanism shifts. For example, a falsely detected mechanism shift may incorrectly orient an edge, removing some DAGs from consideration which otherwise could have fewer total shifts. Overall, ours and their approaches have demonstrated the benefits of pooling only subsets of environments; in particular, we show the connection to the sparse mechanism shift hypothesis and prove asymptotic identifiability.

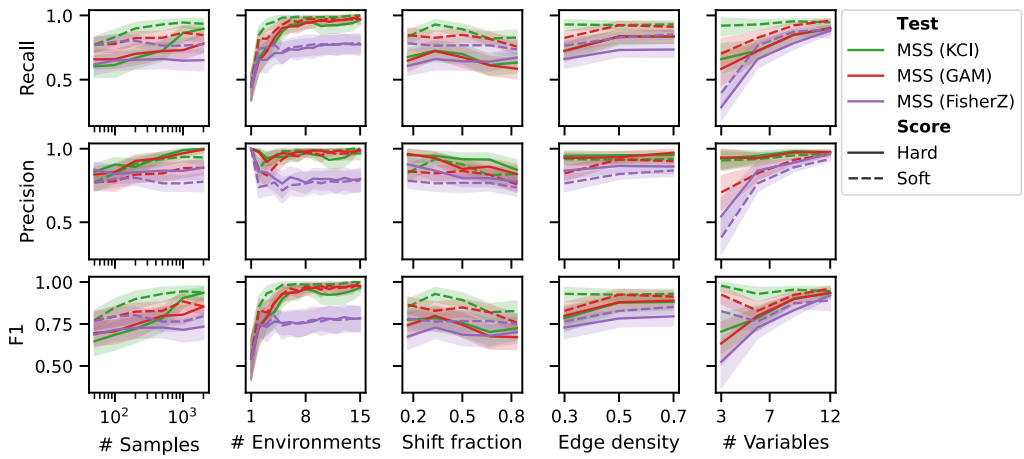



Figure 7: Nonparametric hypothesis tests perform well in nonlinear simulations, and soft scores succeed. Notably, recall converges with increasing environments. KCI appears to best balance high recall and precision.

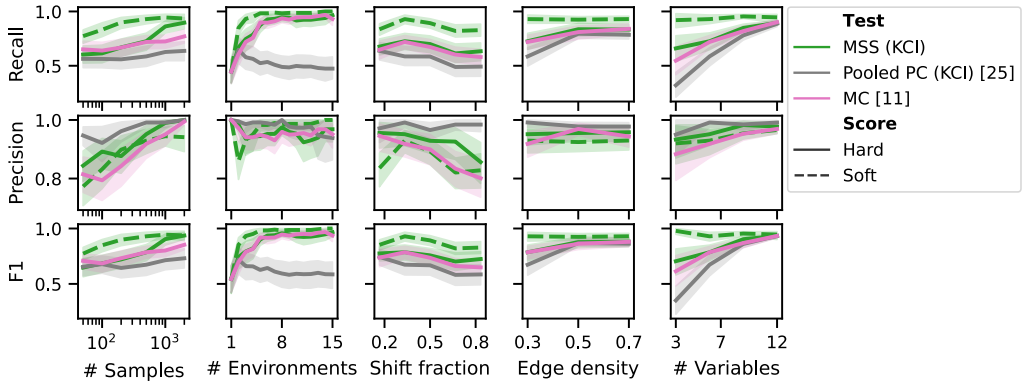



Figure 8: Pairwise approaches improve with more environments, unlike Pooled PC. Although the parametric MC works surprisingly well across settings, the nonparametric MSS with KCI has superior precision and recall.

# D Supporting experiments

## D.1 F1 Scores

In the main text, we present the precision and recall separately as they are important metrics to consider. Here, we also present the F1 score which equals their harmonic mean and conveniently provides a single numeric summary. As before, Fig. 7 provides a simulated comparison of the MSS estimator using various equality of distribution tests, while Fig. 8 provides a simulated comparison of MSS to other approaches in the literature which we discuss heavily in the main body.

## D.2 Additional simulations

### D.2.1 MSS improves upon pooled PC across random graph models.

Previously in Fig. 4, we demonstrated that the MSS improves upon pooled PC under an oracle test in simulation settings where DAGs were sampled according to the Erdős-Rènyi (ER) random DAG model; in the ER model, each edge is sampled i.i.d. with a fixed probability. Here, we expand upon that simulation by further comparing rates under the Barabasi-Alberts (Hub) scale-free random DAG model; in the Hub model, vertices are sequentially added to the DAG and edges are connected to previous vertices with probability proportional to their existing number of edges.

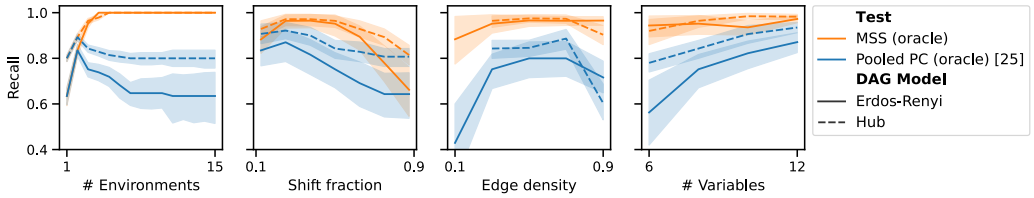



Figure 9: Oracle MSS improves upon pooled PC across both random and hub random graph models.

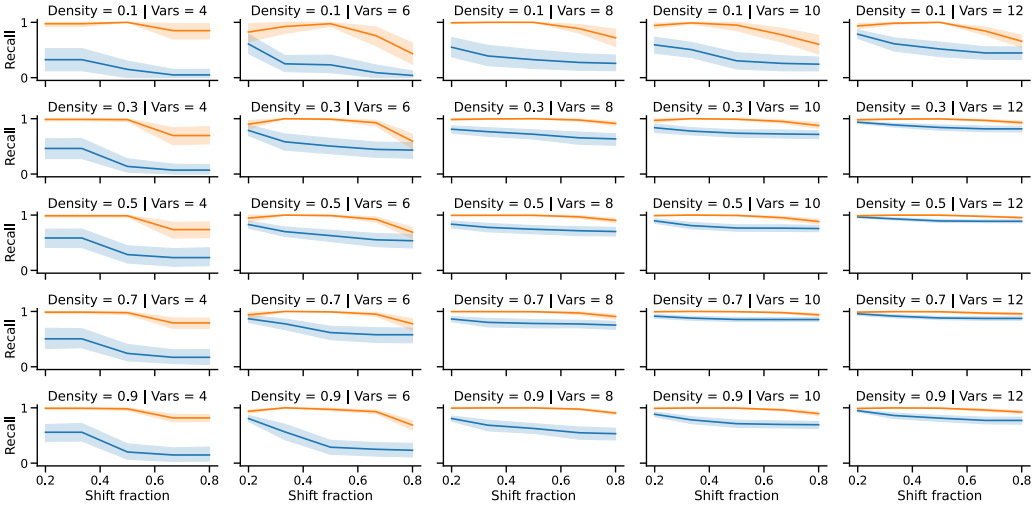



Figure 10: Differences between oracle MSS and pooled PC on 5 environments are most pronounced on smaller and sparse DAGs. For readability, the legend is omitted but we refer back to the same legend in Fig. 4. Specifically, the orange line corresponds to the MSS while the blue line corresponds to pooled PC. Only five environments are sampled, but differences would be exacerbated with additional environments.

As seen in Fig. 9, we vary the same parameters as before but compare the two oracle methods across both DAG models this time. First, note that at 1 environment the Hub model exhibits greater recall, indicating that the observational MEC of the Hub graph has fewer unoriented edges than that of the ER graph. Thus, the gap between the methods is lessened in a Hub model as compared to an ER model. Otherwise, the qualitative trends between the two methods are almost identical across the two random graph models. The MSS appears at least mildly robust to the graph structure.

### D.2.2 Differences between oracle MSS and pooled PC are most pronounced on sparser and smaller DAGs.

Although Fig. 4 highlighted the most important trends of oracle methods in certain fixed settings, for completeness we examine rates of recall across additional fixed settings. As before, we sample DAGs from an Erdős-Rènyi distribution and in five environments vary the DAG density, shift fraction, and number of variables. The set of experimental results shown in Fig. 10 convey broader trends in oracle recall rates as multiple variables change across row, column, and the x-axis. We do not vary the number of environments as we can only visualize three variables through our plot and the trend across environments is best understood from the theory. Differences in oracle recall rates are less pronounced on graphs with more variables and when the density of edges is large. Note that we only compare five environments here and that with more environments, differences will again be more pronounced; with enough environments, pooled PC cannot learn more than the MEC.

### D.2.3 KCI-based approaches perform the best on bivariate CGMs.

We previously examined the empirical rates of recall and precision across various simulated settings, highlighting when methods succeed and fail. Due to the size and complexity of those studied DAGs, not all results are fully interpretable. We seek to further understand empirical performance through the simple bivariate DAG, which contains no indirect effects and few possible interventions to analyze. Specifically, on the DAG $X_1 \to X_2$, shifts can occur to either $\mathbb{P}(X_1)$, $\mathbb{P}(X_2|X_1)$, neither mechanism, or both mechanisms; the first two shifts are sparse and provide oracle identifiability of the true DAG. Following the simulation setup described by eq. (6.1) on the DAG $X_1 \to X_2$, we simulate data from one base environment and from one interventional environment subject to one of the four possible shifts. Each environment has 500 samples. We compare the four different MSS methods using parametric and nonlinear equality of distribution tests and conditional independence tests. We also compare the pooled PC and MC approaches. Since only two environments are compared, we conjecture MSS and pooled PC to be equivalent under an oracle test.

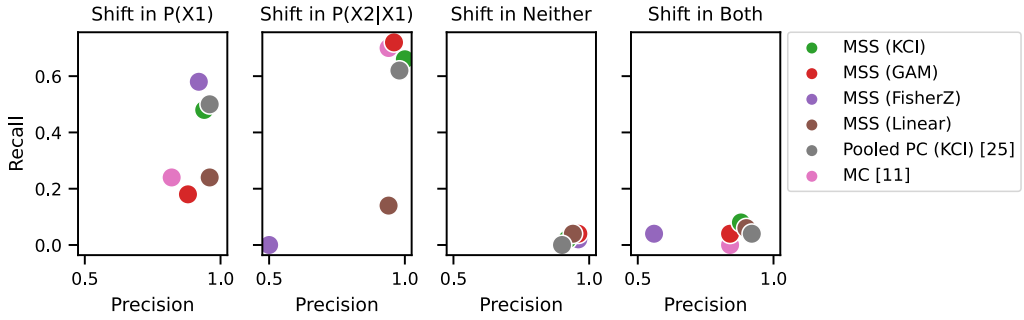



Figure 11: KCI-based tests perform well for causal identification in a bivariate CGM. 500 samples are drawn from a base environment, and a second environment subject to one of four shifts given by the columns; the first two columns are sparse shift settings where we have identifiability. The precision and recall are plotted for each of the methods. It appears that the two KCI-based methods (MSS and pooled PC) achieve the best balance of high power in both sparse shift settings while maintaining high precision in both non-sparse settings. Other methods either have drastically lower recall or precision close to 0.5, indicating random guessing.

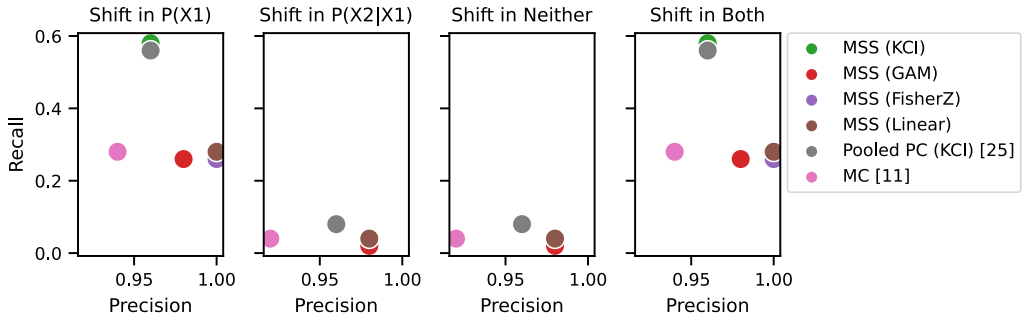



Figure 12: Test performance is dependent on the data generating process. In the same experimental setup as in Fig. 11, we modify the simulation such that the noise is multiplicative. As we see, although the marginal change in $\mathbb{P}(X_1)$ can still be detected, the conditional distribution tests are not powerful enough for shifts in $\mathbb{P}(X_2|X_1)$. However, notice that this means that the true graph is still identifiable when both mechanisms shift. This is contrary to theory about the oracle, *exactly because the finite sample tests are not powerful enough.*

Results are shown in Fig. 11. For reference, an oracle method would have recall 1 in the first two (sparse shift) columns and 0 in the other two columns. Although Fisher-Z has high precision when $\mathbb{P}(X_1)$ shifts, it has chance precision when $\mathbb{P}(X_2|X_1)$ shifts. The KCI methods maintain high precision while the precision of other methods is comparable or noticeably lower. With respect to recall, when neither or both mechanisms change and thus the DAG is not identifiable, all methods correctly have low recall. However, when just the marginal $\mathbb{P}(X_1)$ changes, the KCI

methods dominate in recall whereas the linear MSS MS, and GAM approaches have lower recall, implying they are less often able to detect a change in the reverse conditional $\mathbb{P}(X_1 \mid X_2)$. When the mechanism $\mathbb{P}(X_2 \mid X_1)$ shifts, all methods have high recall. Notably, the linear MSS performs much worse than MC. The only difference between them, however, is that MC explicitly counts how many parameters change while the linear MSS simply tests if there is a change; this does come at a slight cost in precision for MS though. In a small extension, we additionally run this experiment under a multiplicative noise data generating process. Those results are seen in Fig. 12 and highlight both that it is necessary to have access to a powerful hypothesis test and yet failing to reject the null can promote sparsity and lead to identifiability under dense changes..

### D.3 Application to real-world cytometry data

Although simulations with known ground truth provide useful reference points for comparing methods and evaluating empirical performance, in practice we are interested in studying real data with no known truth and additional challenges such as violated assumptions. To illustrate how one may apply our method in practice, and to analyze empirical performance on real data, we conduct a case-study application of MSS for causal discovery on a well-studied cytometry dataset [49].

#### D.3.1 Background

Sachs et al. [49] present a detailed study of the application of Bayesian discovery approaches to learning a causal DAG among protein concentration levels in human immune system cells. In each of 9 experimental environments subject to different perturbations, approximately 700-900 sample measurements were collected; each sample is the concentration levels of 11 proteins from a cell. The learned *Sachs network* is a proposed DAG among the variables, which the authors discuss and contrast with a domain-expert network from the "biologist's view". This cytometry data has subsequently been studied in further detail [8, 38, 47]. As is often pointed out, various assumptions may be violated, including the acyclicity assumption, since protein networks contain strong feedback loops [38]. As such, it is not necessarily useful to treat the Sachs network as a ground truth and there are numerous relationships and orientations which should rightfully be questioned [38]. Results must be considered in the context of domain-knowledge and various existing studies in the literature.

#### D.3.2 Experimental setup

In order to focus on learning edge orientations of undirected edges in the MEC, rather than learning the MEC, we start from the Sachs network despite the potential caveats. The Sachs network from Sachs et al. [49] is a DAG on the 11 variables with 17 edges. We compute the *Sachs MEC* which contains all DAGs which are Markov equivalent to the Sachs network. The Sachs MEC has no directed edges, and thus is simply the undirected skeleton of the graph in Fig. 5. Starting from the Sachs MEC makes our results more interpretable in light of previous works and saves costly computation of the MEC. In practice, we would advise starting from the pooled PC MEC; based on the number of environments and observed density of changes, we would not expect this to orient any edges beyond the observational MEC.

Starting from the Sachs MEC, we apply the MSS using the KCI test, which appears to perform the best among plug-in estimators for MSS in our simulations. Since the feature distributions are heavily skewed, we preprocess them by taking their natural logarithm [47]. Among all DAGs in the Sachs MEC, the DAG with the uniquely minimal number of shifts exhibits approximately 8.9 shifts per pair of environments; this is relatively high but satisfies the assumption of sparse shifts. Violations to assumptions may lead to more shifts than expected.

#### D.3.3 Results and comparison to related works

The DAG which minimizes the MSS is the unique minimizer and is visualized in Fig. 5. An edge in black is oriented in the same direction as in the Sachs network, while an edge in red is oriented in the opposite direction. Overall, the majority of edges match the Sachs network. The edges which do not match, however, reflect ambiguities and flawed assumptions. We list each edge which does not match the Sachs network and discuss why this might be the case in light of existing work.

- PIP2 → PIP3: As illustrated in Sachs et al. [49], these two proteins are actually cyclically related through bi-directed edges in the accepted "biologist's view". Indeed, PIP2 → PIP3 was similarly recovered by an analysis of Ramsey and Andrews [47], detailed in their Figure 11.
- Mek → Raf: that this edge does not match the Sachs network is discussed heavily by Mooij et al. [38] who point to it as a fundamental flaw of the Sachs network. The Mek → Raf edge is indeed found by many other methods [8, 38, 47].
- The PKA, PKC, P38 triangle: although there is not a detailed discussion of these variables in other studies, there is strong ambiguity in the edge directions among approaches. Notably, Mooij et al. [38] similarly find strong evidence in their approach for the edge P38 → PKC while Ramsey and Andrews [47] and Eaton and Murphy [8] find evidence for PKA → PKC. However, all other approaches agree that the edge P38 → PKA is incorrect. Although we do not explore further, it is worth noting that the 3rd minimal MSS DAG (not shown) is the same as the one shown, except it contains the presumed correct edge PKA → P38.

As an additional note, we see in Fig. 5 that the edge Mek → Erk is correctly recovered. Sachs et al. [49] similarly recover this well-known connection and point to it as strong evidence of success. In contrast, neither Eaton and Murphy [8], Ramsey and Andrews [47], nor Mooij et al. [38] recover the edge with their methods. Indeed, Ramsey and Andrews [47] specifically discuss how their approach incorrectly missed this edge, potentially the result of signal being lost when all the data is pooled. Pooled PC would face a similar issue, exacerbated by additional environments, while the pairwise comparisons of the MSS help to avoid this issue.