DECOUPLED CAUSAL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal discovery from observational data is a fundamental yet challenging task in scientific research. While existing approaches are primarily based on conditional independence tests, structure scores, or restrictive functional assumptions, we propose Decoupled Causal Discovery (DCD), a novel decoupling-based perspective that does not rely on these methodologies. DCD directly identifies the Markov boundary (MB) by decoupling non-target variables via weighting functions, such that only variables within the MB preserve dependence with the target under the decoupled distribution. Building on this, DCD iteratively constructs the Completed Partially Directed Acyclic Graph (CPDAG) by exploiting structural asymmetries within the MBs. We establish the theoretical identifiability, soundness, and completeness of DCD. Empirical evaluations demonstrate that DCD achieves strong performance, particularly excelling in challenging noise regimes.

1 INTRODUCTION

Causal discovery aims to uncover underlying causal structures from observational data and serves as a fundamental task in scientific research and decision-making (Spirtes & Zhang, 2016; Glymour et al., 2019). Following decades of development, existing approaches have driven remarkable progress in the field and primarily fall into three paradigms: constraint-based methods (Spirtes & Glymour, 1991), which rely on sequential conditional independence tests (CITs) (Colombo & Maathuis, 2014; Peters et al., 2017); score-based methods (Chickering, 2002), which search for a graph maximizing a goodness-of-fit score (Schwarz, 1978) and have been extended through continuous optimization to handle larger-scale graphs (Zheng et al., 2018); and functional causal models (FCMs) (Shimizu et al., 2006), which exploit specific data-generating assumptions to achieve stronger identifiability (Chen et al., 2026). Nevertheless, causal discovery from complex observational data remains inherently challenging, motivating the exploration of perspectives beyond these established methodologies.

In this paper, we propose Decoupled Causal Discovery (DCD), a novel decoupling-based perspective that does not rely on CITs, structure scores, or restrictive functional assumptions. The core mechanism of DCD is to decouple the distribution via weighting functions that render non-target variables mutually independent. We theoretically show that under this decoupled distribution, a variable preserves dependence on the target if and only if it belongs to the MB of the target, thereby enabling the direct identification of MBs. Building upon this identifiability, DCD iteratively constructs the Completed Partially Directed Acyclic Graph (CPDAG) by exploiting the structural asymmetries within the MBs.

The main contributions of this paper are summarized as follows:

- We provide a decoupling-based perspective for causal discovery and propose DCD, a novel framework that is orthogonal to CITs, structure scores, and FCM assumptions.
- We establish the theoretical foundation for this framework, proving the identifiability of MBs under the decoupled distribution, and the soundness and completeness of the construction algorithm.
- Empirical evaluations show that DCD achieves strong performance, particularly under challenging noise regimes where several representative baselines become unreliable.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 Related Work

We outline several representative and recent causal discovery methods across three classical paradigms, while referring to Nogueira et al. (2022) and Vowels et al. (2022) for a comprehensive overview. As a modeling paradigm with strong identifiability, FCMs exploit asymmetries in structural data-generating assumptions to identify causal order. Pioneered by linear non-Gaussian models (Shimizu et al., 2006; 2011; Hoyer et al., 2008b), research has broadened the scope of such identifiability (Hoyer et al., 2008a; Bühlmann et al., 2014; Zhang & Hyvärinen, 2009; Tagasovska et al., 2020), with recent work moving toward a more unified framework (Xi et al., 2025; Chen et al., 2026; Dai et al., 2026). Score-based methods maximize a goodness-of-fit score (Schwarz, 1978; Heckerman et al., 1995; Huang et al., 2018) over the space of graph structures. Classical approaches perform direct search over structures (Cooper & Herskovits, 1992; Chickering, 2002), while continuous optimization (Zheng et al., 2018; Lachapelle et al., 2020; Ng et al., 2020) has enabled this paradigm to scale to larger graphs.

Constraint-based methods (Spirtes & Glymour, 1991; Spirtes et al., 1995), representing one of the earliest paradigms (Spirtes et al., 1989; Pearl & Verma, 1991), rely on CITs (Colombo & Maathuis, 2014; Li & Fan, 2020) to infer graphical separation (e.g., d-separation (Pearl, 1988)) in a causal graph (Spirtes et al., 2000; Peters et al., 2017). To perform CITs more efficiently, a natural strategy is to exploit local structure. Some early approaches first identified MBs and then assembled the global structure (Margaritis & Thrun, 1999; Pellet & Elisseeff, 2008). More recently, a line of research (Akbari et al., 2021; Mokhtarian et al., 2021; 2022; 2023; 2025) has further exploited MBs to locally identify removable variables, thereby progressively reducing the problem size during algorithm execution. Despite leveraging local structural information from MBs, these methods are ultimately driven by CITs.

3 Decoupled Causal Discovery

We first introduce the procedure for MB identification (Section 3.1), followed by the construction of the global CPDAG by directly identifying MBs (Section 3.2), and present the practical implementation (Section 3.3). For notational readability, we slightly abuse the notation by treating sets of random variables and random vectors interchangeably. Thus, calligraphic and boldface symbols may be subjected to both set-theoretic and probabilistic operations.

3.1 Identifiability of MB via Decoupling

In this section, we elucidate the direct identification of the MB via distribution decoupling. Notably, this process operates independently of the graph structure, relying purely on statistical relationships. This implies that the proposed procedure serves as a general statistical identification framework with broad applicability in scenarios beyond causal discovery. We first provide the formal definition of the Markov blanket and Markov boundary (Pearl, 1988).

Definition 1 (Markov Blanket and Markov Boundary)**.** *Given a set of random variables $\mathcal{O}$ and a target variable $X_t \in \mathcal{O}$, a **Markov blanket** of X_t relative to $\mathcal{O}$ is a subset $\mathbf{M} \subseteq \mathcal{O} \setminus \{X_t\}$ such that*

$$X_t \perp\!\!\!\perp \big(\mathcal{O} \setminus (\{X_t\} \cup \mathbf{M})\big) \mid \mathbf{M}.$$

*The **Markov boundary (MB)** of X_t relative to $\mathcal{O}$, denoted by $\mathrm{MB}(X_t \mid \mathcal{O})$, is defined as an inclusion-minimal Markov blanket. Formally, it is a Markov blanket $\mathbf{M}$ that satisfies:*

$$\forall \mathbf{M}' \subsetneq \mathbf{M}, \quad X_t \not\perp\!\!\!\perp \big(\mathcal{O} \setminus (\{X_t\} \cup \mathbf{M}')\big) \mid \mathbf{M}'.$$

While prior works occasionally conflate the Markov blanket with the Markov boundary (the minimal conditioning set), we explicitly distinguish between the two to ensure theoretical rigor.

We next introduce the core principle of DCD, which leverages a weighting scheme to decouple the joint distribution of non-target variables, thereby enabling direct identification of the MB for a target variable. This framework draws inspiration from the line of work on stable prediction (Kuang et al., 2018; 2023; Shen et al., 2018; 2020; Xu et al., 2022; Cui & Athey, 2022), and our theoretical results can be viewed as a generalization of that paradigm. We begin with the definitions of weighting and distribution decoupling.

Definition 2 (Weighting Function and Weighted Distribution)**.** *Given a target variable X_t and a random vector $\mathbf{Z}$ with domain Val($\mathbf{Z}$), the weighting function set $\mathcal{W}(\mathbf{Z})$ is defined as:*

$$\mathcal{W}(\mathbf{Z}) = \{w : \mathit{Val}(\mathbf{Z}) \to \mathbb{R}^+ \mid \mathbb{E}[w(\mathbf{Z})] = 1\}.$$

For any $w \in \mathcal{W}(\mathbf{Z})$, the weighted distribution $\tilde{P}_w(X_t, \mathbf{Z})$ is defined as:

$$\tilde{P}_w(X_t, \mathbf{Z}) = w(\mathbf{Z})P(X_t, \mathbf{Z}).$$

Intuitively, Definition 2 characterizes the space of valid weighting functions that alter the distribution of non-target variables. Unlike Xu et al. (2022) (for linear models), our definition does not impose first-moment balance conditions like $\mathbb{E}[w(\mathbf{Z})\mathbf{Z}] = \mathbf{0}$. We then define the subset of weights that eliminates the statistical entanglement among $\mathbf{Z}$.

Definition 3 (Decoupled Distribution)**.** *Given the target variable X_t and the random vector $\mathbf{Z}$, the set of decoupling weights is defined as:*

$$\mathcal{W}_\perp(\mathbf{Z}) = \{w \in \mathcal{W}(\mathbf{Z}) \mid \mathbf{Z} \textit{ are mutually independent in } \tilde{P}_w\}. \tag{1}$$

For any $w \in \mathcal{W}_\perp(\mathbf{Z})$, we refer to $\tilde{P}_w(X_t, \mathbf{Z})$ as a decoupled distribution of $\mathbf{Z}$.

In a decoupled distribution $\tilde{P}_w$, the joint distribution factorizes as $\tilde{P}_w(\mathbf{Z}) = \prod_i \tilde{P}_w(Z_i)$, meaning that $\tilde{P}_w(\mathbf{Z}_j \mid \mathbf{Z}_k) = \tilde{P}_w(\mathbf{Z}_j)$ for any disjoint $\mathbf{Z}_j, \mathbf{Z}_k$, which is essential for isolating non-MB variables. In general, the set of such decoupling weights $\mathcal{W}_\perp(\mathbf{Z})$ is non-trivial (see Appendix C).

We next introduce two assumptions necessary for MB identifiability under the decoupled distribution.

Assumption 1 (Strict Positivity)**.** *The joint distribution (and consequently all marginal and conditional distributions) of random variables is strictly positive.*

Assumption 1 is a standard regularity condition widely adopted in stable prediction and probabilistic graphical models (Xu et al., 2022; Pearl, 2009). Combined with the strictly positive weighting functions defined in Definition 2, Assumption 1 ensures that all probability ratios and conditioning operations in our subsequent analysis are well-defined. Crucially, this assumption guarantees that the conditional independence relation satisfies the *intersection axiom* (Pearl, 1988), which ensures the uniqueness of the MB (Appendix A, Lemmas A.4 and A.5).

While the strict positivity guarantees the fundamental properties, we further introduce a non-cancellation assumption to rule out pathological cases where the influence of a relevant variable vanishes merely due to numerical coincidences.

Assumption 2 (Weighting Faithfulness)**.** *Let $\mathcal{O}$ be a set of variables, $X_t \in \mathcal{O}$ be the target variable, and $\mathbf{X}_{-t} = \mathcal{O} \setminus \{X_t\}$. Given a decoupling weight $w \in \mathcal{W}_\perp(\mathbf{X}_{-t})$ (and its corresponding $\tilde{P}_w(X_t, \mathbf{X}_{-t})$), we assume that for any $X_e \in \mathbf{X}_{-t}$ and $\mathbf{X}_c \subseteq \mathbf{X}_{-t} \setminus \{X_e\}$:*

$$\textit{If } P(X_t \mid X_e, \mathbf{X}_c) \textit{ depends on } X_e, \textit{ then so does } \mathbb{E}_{\tilde{P}_w(\mathbf{X}_c)}[P(X_t \mid X_e, \mathbf{X}_c)].$$

Remark 1. This assumption states that integrating out $\mathbf{X}_c$ under $\tilde{P}_w$ does not perfectly cancel the original dependence of X_t on X_e given $\mathbf{X}_c$. This generally holds when the expectation is taken over the original distribution. Furthermore, since the weight $w \in \mathcal{W}_\perp(\mathbf{X}_{-t})$ is derived solely from $P(\mathbf{X}_{-t})$, it is strictly agnostic to the mechanism $P(X_t \mid \cdot)$. Consequently, it is extremely unlikely for the decoupled distribution $\tilde{P}_w$ to adversarially align with $P(X_t \mid \cdot)$ to induce an exact cancellation. We provide a formal proof showing that this assumption holds almost everywhere (Proposition 2).

Remark 2. This assumption mirrors the standard faithfulness (Assumption 5, although our current MB discussion is purely statistical), serving as an intuitive and mild condition in practice. Both assumptions ensure that true statistical dependencies are not masked by exact numerical cancellations, thereby guaranteeing that the underlying relationships can be properly identified.

Remark 3. This assumption parallels several standard concepts in causal inference:

- *Covariate Balancing:* The term $\mathbb{E}_{\tilde{P}_w(\mathbf{X}_c)}[P(X_t \mid X_e, \mathbf{X}_c)]$ matches the standard form of estimators using covariate balancing (Rosenbaum & Rubin, 1983). Assuming it is non-degenerate is a standard implicit setting for causal inference, ensuring that heterogeneous effects do not perfectly cancel out, which is necessary for identification.

- *Backdoor Adjustment:* If the expectation is taken over the original distribution, it recovers the formula of backdoor adjustment (Pearl, 1995; 2009). In this context, this condition acts as a natural guarantee that the adjusted effect does not coincidentally vanish if a genuine effect exists.

Based on the definitions and the two assumptions above, we establish the following property to achieve the identifiability of the MB (see Appendix A.1 for the complete proof).

Proposition 1 (MB Conditional Distribution Invariance)**.** *Let $\mathcal{O}$ be a set of variables, $X_t \in \mathcal{O}$ be the target variable, and $\mathbf{X}_{-t} = \mathcal{O} \setminus \{X_t\}$. Under Assumption 1, for any weighting function $w \in \mathcal{W}(\mathbf{X}_{-t})$ (and its corresponding $\tilde{P}_w(X_t, \mathbf{X}_{-t})$), the following holds for any $\mathbf{X}^+_{\mathrm{MB}}$ satisfying $\mathrm{MB}(X_t \mid \mathcal{O}) \subseteq \mathbf{X}^+_{\mathrm{MB}} \subseteq \mathbf{X}_{-t}$:*

$$\tilde{P}_w(X_t \mid \mathbf{X}^+_{\mathrm{MB}}) = P(X_t \mid \mathbf{X}^+_{\mathrm{MB}}) = P(X_t \mid \mathrm{MB}(X_t \mid \mathcal{O})) = P(X_t \mid \mathbf{X}_{-t}).$$

Note that this proposition holds for any weighted distribution (Definition 2, not necessarily decoupled), requiring only the strict positivity assumption. It extends the conditional distribution invariance of unweighted variables given all weighted variables (Shimodaira, 2000; Xu et al., 2022) to the context of MBs. Specifically, as long as the conditioning set of the target variable in the weighted distribution contains the complete MB, the resulting conditional distribution remains identical to that under the original distribution, both coinciding with $P(X_t \mid \mathbf{X}_{-t})$ and $P(X_t \mid \mathrm{MB}(X_t \mid \mathcal{O}))$. This also implies that the original MB remains a Markov blanket under weighted distributions.

Building on this proposition, we further establish the main theorem regarding MB identifiability.

Theorem 1 (MB Identifiability)**.** *Let $\mathcal{O}$ be a set of variables, $X_t \in \mathcal{O}$ be the target variable, and $\mathbf{X}_{-t} = \mathcal{O} \setminus \{X_t\}$. Under Assumption 1, for any decoupling weight $w \in \mathcal{W}_\perp(\mathbf{X}_{-t})$ (and its corresponding $\tilde{P}_w(X_t, \mathbf{X}_{-t})$) satisfying Assumption 2, we have for any variable $X_i \in \mathbf{X}_{-t}$:*

$$X_i \in \mathrm{MB}(X_t \mid \mathcal{O}) \iff X_i \not\perp\!\!\!\perp_{\tilde{P}_w} X_t.$$

According to Theorem 1, MB identification can be a direct process, achieved by learning decoupling weights over the non-target variables and evaluating which variables remain dependent on the target variable under the weighted distribution.

Importantly, our conclusion relies exclusively on the statistical definition of the MB along with two basic assumptions, without requiring any specific model specifications or graph structures. In this context, our result can be viewed as a generalization of the findings in Xu et al. (2022). Their theoretical objective focuses on the mean prediction problem and therefore concerns the *minimal stable variable set*, which is the subset of the MB that has an effect on the first moment. Additionally, their analysis is grounded in an additive noise model and a linear estimation framework. In contrast, beyond the standard strict positivity, our framework requires only a single non-coincidental assumption. Furthermore, we show that Assumption 2 holds almost everywhere under a natural measure on the weight space, confirming its genericity.

Condition 1 (Non-degeneracy of Decoupling Space)**.** *The space of decoupled distributions of $\mathbf{Z}$ (generated by $w \in \mathcal{W}_\perp(\mathbf{Z})$) is sufficiently rich to avoid degenerating onto any hyperplane.*

Formally, let $P_{\mathbf{Z}}$ be the original density of $\mathbf{Z}$, and equip $\mathcal{W}_\perp(\mathbf{Z})$ with a Borel probability measure $\mu^{\mathbf{Z}}_{\mathcal{W}}$. Then, for any nonzero continuous linear functional $\ell : L^1(\lambda_{\mathbf{Z}}) \to \mathbb{R}$ (where $\lambda_{\mathbf{Z}}$ is the dominating measure on $\mathbf{Z}$), the following holds:

$$\mu^{\mathbf{Z}}_{\mathcal{W}}\big(\{w \in \mathcal{W}_\perp(\mathbf{Z}) \;\big|\; \ell(w \cdot P_{\mathbf{Z}}) = 0\}\big) = 0$$

Condition 1 formalizes that the functional space of valid weights is sufficiently rich (discussed in Appendix C) and characterizes a practical estimation process through the probability measure $\mu^{\mathbf{Z}}_{\mathcal{W}}$. Under this condition, we establish that Assumption 2 holds almost everywhere with respect to $\mu^{\mathbf{Z}}_{\mathcal{W}}$.

Proposition 2 (Genericity of Weighting Faithfulness)**.** *Let $\mathcal{O}$ be a set of variables, $X_t \in \mathcal{O}$ be the target variable, and $\mathbf{X}_{-t} = \mathcal{O} \setminus \{X_t\}$. Under Assumption 1 and Condition 1, for $\mu^{\mathbf{X}_{-t}}_{\mathcal{W}}$-almost every weight $w \in \mathcal{W}_\perp(\mathbf{X}_{-t})$, Assumption 2 (Weighting Faithfulness) holds.*

Consequently, Proposition 2 confirms that Assumption 2 is a generic property in practical scenarios, providing a theoretical foundation for our framework.

Algorithm 1 DCD-C: Construct CPDAG from Local MBs

Require: Set of variables $\mathcal{V} = \{X_1, \ldots, X_d\}$, Markov boundary identifier $\text{MB}(\cdot \mid \cdot)$
Ensure: Completed Partially Directed Acyclic Graph (CPDAG) $\hat{\mathcal{G}} = (\mathcal{V}, \hat{\mathcal{E}})$

Initialize graph $\hat{\mathcal{G}} = (\mathcal{V}, \hat{\mathcal{E}})$ with empty edge set $\hat{\mathcal{E}} \leftarrow \emptyset$
Initialize v-structure set $\mathcal{K} \leftarrow \emptyset$
for each node $X_t \in \mathcal{V}$ **do**
  Obtain initial Markov boundary: $\mathcal{M}_t \leftarrow \text{MB}(X_t \mid \mathcal{V})$
  Initialize neighbor set $\mathcal{N}_t \leftarrow \mathcal{M}_t$
  {Traverse by increasing subset size to find exclusion sets}
  for $k = 1$ **to** $|\mathcal{M}_t| - 1$ **do**
    for each subset $\mathcal{Z} \subset \mathcal{M}_t$ with $|\mathcal{Z}| = k$ **do**
      $\mathcal{M}_{\setminus \mathcal{Z}} \leftarrow \text{MB}(X_t \mid (\mathcal{M}_t \cup \{X_t\}) \setminus \mathcal{Z})$ {Local MB within the restricted subset}
      $\mathcal{D}_{\text{drop}} \leftarrow (\mathcal{M}_t \setminus \mathcal{Z}) \setminus \mathcal{M}_{\setminus \mathcal{Z}}$
      if $\mathcal{D}_{\text{drop}} \neq \emptyset$ **then**
        for each $Y \in \mathcal{D}_{\text{drop}}$ **do**
          for each $Z \in \mathcal{Z}$ **do**
            $\mathcal{K} \leftarrow \mathcal{K} \cup \{(X_t, Z, Y)\}$ {Identify potential v-structures}
          end for
          $\mathcal{N}_t \leftarrow \mathcal{N}_t \setminus \{Y\}$ {Remove spouse from candidate neighbors}
        end for
      end if
    end for
  end for {Remaining nodes in $\mathcal{N}_t$ are neighbors of X_t, form the undirected skeleton}
  for each $W \in \mathcal{N}_t$ **do**
    Add undirected edge to $\hat{\mathcal{E}}$: $X_t - W$
  end for
end for
for each $\{(X, Z, Y)\} \in \mathcal{K}$ **do**
  if Z is adjacent to both X and Y in $\hat{\mathcal{G}}$ **then**
    if X and Y are not adjacent in $\hat{\mathcal{G}}$ **then**
      Orient edges in $\hat{\mathcal{E}}$ as: $X \to Z$ and $Y \to Z$
    end if
  end if
end for
Apply Meek rules (Meek, 1995) to $\hat{\mathcal{G}}$ to exhaustively orient remaining undirected edges
return $\hat{\mathcal{G}}$

3.2 Global Graph Construction

Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a random vector following a joint probability distribution P. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed acyclic graph (DAG) characterizing the underlying causal structure of $\mathbf{X}$, where the node set $\mathcal{V} = \{X_1, \ldots, X_d\}$ corresponds to the random variables. A directed edge from X_i to X_j, denoted by $(X_i, X_j) \in \mathcal{E}$, indicates that X_i is a direct cause of X_j. For any node $X_i \in \mathcal{V}$, we denote its parents, children, and spouses in $\mathcal{G}$ as $\text{Pa}(X_i)$, $\text{Ch}(X_i)$, and $\text{Sp}(X_i)$, respectively. Specifically, $\text{Pa}(X_i) = \{X_p \mid (X_p, X_i) \in \mathcal{E}\}$, $\text{Ch}(X_i) = \{X_c \mid (X_i, X_c) \in \mathcal{E}\}$ and $\text{Sp}(X_i) = \{X_s \in \mathcal{V} \setminus \{X_i\} \mid \text{Ch}(X_i) \cap \text{Ch}(X_s) \neq \emptyset\}$. We adopt standard assumptions (Spirtes et al., 2000) to bridge the distribution P with the graph $\mathcal{G}$.

Assumption 3 (Causal Sufficiency)**.** *There are no unobserved confounders. That is, for any $X_i, X_j \in \mathcal{V}$, if a variable X_u is a common cause of X_i and X_j, then $X_u \in \mathcal{V}$.*

Assumption 4 (Markov Condition)**.** *The joint distribution P satisfies the Markov condition with respect to $\mathcal{G}$. That is, every variable $X_i \in \mathcal{V}$ is conditionally independent of its non-descendants given its parents* $\text{Pa}(X_i)$.

Assumption 5 (Faithfulness)**.** *The joint distribution P is faithful to the DAG $\mathcal{G}$. This implies that all conditional independencies present in P are exactly those entailed by the d-separation in $\mathcal{G}$. Formally, for any disjoint subsets $\mathbf{U}, \mathbf{V}, \mathbf{W} \subset \mathcal{V}$:*

$$\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W} \text{ in } P \iff \mathbf{U} \text{ and } \mathbf{V} \text{ are d-separated by } \mathbf{W} \text{ in } \mathcal{G}.$$

Under the three assumptions above, the MB of any variable in $\mathcal{G}$ consists exactly of its parents, children, and spouses (Lemma A.6). Building upon this property, we propose DCD-C (Algorithm 1), an algorithm designed to recover the CPDAG solely via MB identification.

The core idea of DCD-C relies on the structural asymmetry between direct neighbors (parents and children) and spouses. If a variable is a true neighbor of a target, it will always remain in the target's MB across any subset of variables containing it. Conversely, this is not necessarily true for a spouse.

DCD-C (Algorithm 1) exploits this property to isolate true neighbors. For each variable, the algorithm first identifies its MB over the full set of variables to form an initial set of candidate neighbors. We then eliminate non-neighbors by iteratively examining the MBs of the target variable over subsets of the full MB. Specifically, in each iteration, we exclude a chosen subset from the candidate neighbors and re-evaluate the MB over the remaining variables. If any additional variable drops out of the MB, it is guaranteed to be a non-adjacent node. We establish theoretical guarantees for DCD-C (see Appendix A.2 for the complete proof).

Theorem 2 (Soundness and Completeness of DCD-C)**.** *Under Assumptions 3, 4, and 5, with an oracle MB identifier* $\mathrm{MB}(\cdot \mid \cdot)$*, Algorithm 1 (DCD-C) returns the true CPDAG of* $\mathcal{G}$*.*

This theorem guarantees the strict correctness of DCD-C in recovering the structure up to its Markov equivalence class. While DCD-C requires iterating over subsets of the MB, its time complexity is bounded by $O(d \cdot 2^m)$ MB identifications, where m denotes the maximum MB size. Consequently, the algorithm is highly efficient for sparse graphs where m remains small. Importantly, since the operations are localized to individual variables, DCD-C inherently offers potential for parallel execution. We provide a detailed discussion on the computational complexity in Appendix B.

3.3 Practical Estimation

Building upon the theoretical identifiability, we propose a practical framework for estimating MBs. We learn the sample weights as an optimization problem via backpropagation on a decoupling metric, and subsequently perform weighted independence tests to infer relations under the decoupled distribution. Equipped with this, the overall discovery process follows Algorithm 1 to construct the global graph.

Weight Parameterization. To ensure the sample weights are non-negative and satisfy the expectation constraints in Definition 2, we parameterize them using a softmax function. Given n samples and learnable parameters $\theta \in \mathbb{R}^n$, the weight for the k-th sample is defined as $w_k = n \cdot \exp(\theta_k) / \sum_{l=1}^{n} \exp(\theta_l)$. This parameterization guarantees that $\sum_{k=1}^{n} w_k = n$, thereby empirically satisfying $\mathbb{E}[w] = 1$.

Decoupling via Weighted HSIC. According to the definition, the decoupled distribution requires mutual independence among the non-target variables $\mathbf{X}_{-t}$. To empirically enforce this, we minimize the pairwise weighted Hilbert-Schmidt Independence Criterion (HSIC) using a Radial Basis Function (RBF) kernel with median-heuristic bandwidth. The empirical objective function is formulated as:

$$\mathcal{L}(\theta) = \sum_{i \neq j, X_i, X_j \in \mathbf{X}_{-t}} \widehat{\mathrm{HSIC}}_w(X_i, X_j)$$

where $\widehat{\mathrm{HSIC}}_w$ denotes the empirical weighted HSIC estimator. To prevent weight degeneration during optimization, we monitor Kish's Effective Sample Size, defined as $\mathrm{ESS} = (\sum_{k=1}^{n} w_k)^2 / \sum_{k=1}^{n} w_k^2$, and employ early stopping if it drops below a predefined threshold.

Weighted Correlation Test. Once the optimal weights w^* are learned via gradient descent, we proceed to the identification step dictated by Theorem 1. For each non-target variable $X_i \in \mathbf{X}_{-t}$, we compute its weighted Pearson correlation coefficient r_w with the target variable X_t. We then conduct a two-tailed Student's t-test with degrees of freedom adjusted by the ESS, using a default significance level of $\alpha = 0.05$. Ultimately, we include X_i in the estimated MB if the null hypothesis is rejected.

4 Experiments

We empirically evaluate the effectiveness of our DCD under different mechanisms (Section 4.1), noise conditions (Section 4.2), graph topologies (Section 4.3), and sample sizes (Section 4.4), comparing against representative baselines: PC (Spirtes & Glymour, 1991), GES (Chickering, 2002), ICALiNGAM (Shimizu et al., 2006), DirectLiNGAM (Shimizu et al., 2011), Notears (Zheng et al., 2018), GraNDAG (Lachapelle et al., 2020), and GOLEM (Ng et al., 2020). Notably, for the CIT in the PC algorithm, we implement the Kernel-based CIT (KCIT) (Zhang et al., 2011), the Fisher-z test

Table 1: Performance under different mechanisms across graph sizes $d \in \{5, 10, 20\}$.

Method	Func				MLP			
	$\text{SHD}_c \downarrow$	Precision ↑	Recall ↑	F1 ↑	$\text{SHD}_c \downarrow$	Precision ↑	Recall ↑	F1 ↑
$d = 5$								
DCD	**2.4**$_{\pm 1.1}$	**0.86**$_{\pm 0.31}$	**0.43**$_{\pm 0.24}$	**0.55**$_{\pm 0.26}$	**0.9**$_{\pm 1.1}$	**0.97**$_{\pm 0.15}$	**0.79**$_{\pm 0.26}$	**0.85**$_{\pm 0.22}$
PC (FisherZ)	3.5$_{\pm 1.2}$	0.41$_{\pm 0.34}$	0.30$_{\pm 0.25}$	0.34$_{\pm 0.26}$	3.7$_{\pm 1.7}$	0.32$_{\pm 0.36}$	0.28$_{\pm 0.29}$	0.29$_{\pm 0.30}$
PC (E-FisherZ)	3.4$_{\pm 1.3}$	0.41$_{\pm 0.38}$	0.28$_{\pm 0.24}$	0.32$_{\pm 0.28}$	3.4$_{\pm 1.5}$	0.32$_{\pm 0.36}$	0.28$_{\pm 0.30}$	0.29$_{\pm 0.31}$
PC (KCIT)	3.2$_{\pm 1.2}$	0.38$_{\pm 0.41}$	0.26$_{\pm 0.28}$	0.30$_{\pm 0.31}$	2.1$_{\pm 1.7}$	0.52$_{\pm 0.42}$	0.52$_{\pm 0.42}$	0.52$_{\pm 0.41}$
PC (E-KCIT)	3.2$_{\pm 1.1}$	0.37$_{\pm 0.41}$	0.25$_{\pm 0.27}$	0.28$_{\pm 0.30}$	2.7$_{\pm 1.6}$	0.39$_{\pm 0.40}$	0.36$_{\pm 0.37}$	0.37$_{\pm 0.38}$
GES	3.1$_{\pm 1.6}$	0.53$_{\pm 0.41}$	0.38$_{\pm 0.30}$	0.43$_{\pm 0.33}$	2.4$_{\pm 1.7}$	0.60$_{\pm 0.39}$	0.54$_{\pm 0.35}$	0.56$_{\pm 0.35}$
ICALiNGAM	3.0$_{\pm 0.9}$	0.68$_{\pm 0.47}$	0.26$_{\pm 0.21}$	0.37$_{\pm 0.28}$	2.9$_{\pm 1.0}$	0.66$_{\pm 0.48}$	0.27$_{\pm 0.25}$	0.36$_{\pm 0.30}$
DirectLiNGAM	2.9$_{\pm 0.9}$	0.70$_{\pm 0.46}$	0.28$_{\pm 0.22}$	0.38$_{\pm 0.28}$	2.9$_{\pm 1.0}$	0.66$_{\pm 0.48}$	0.27$_{\pm 0.26}$	0.36$_{\pm 0.31}$
Notears	3.3$_{\pm 0.8}$	0.52$_{\pm 0.51}$	0.18$_{\pm 0.20}$	0.25$_{\pm 0.27}$	3.6$_{\pm 0.6}$	0.30$_{\pm 0.46}$	0.10$_{\pm 0.16}$	0.14$_{\pm 0.23}$
GraNDAG	3.9$_{\pm 2.0}$	0.45$_{\pm 0.43}$	0.32$_{\pm 0.28}$	0.34$_{\pm 0.31}$	3.2$_{\pm 2.1}$	0.55$_{\pm 0.38}$	0.50$_{\pm 0.33}$	0.49$_{\pm 0.32}$
GOLEM	3.0$_{\pm 0.9}$	0.64$_{\pm 0.48}$	0.25$_{\pm 0.22}$	0.35$_{\pm 0.29}$	3.2$_{\pm 0.8}$	0.60$_{\pm 0.49}$	0.21$_{\pm 0.21}$	0.30$_{\pm 0.28}$
$d = 10$								
DCD	**5.5**$_{\pm 1.7}$	0.87$_{\pm 0.16}$	**0.45**$_{\pm 0.16}$	**0.58**$_{\pm 0.16}$	**3.3**$_{\pm 2.0}$	0.88$_{\pm 0.16}$	**0.72**$_{\pm 0.18}$	**0.78**$_{\pm 0.17}$
PC (FisherZ)	9.4$_{\pm 2.9}$	0.28$_{\pm 0.22}$	0.24$_{\pm 0.17}$	0.26$_{\pm 0.19}$	10.4$_{\pm 2.5}$	0.15$_{\pm 0.15}$	0.17$_{\pm 0.17}$	0.16$_{\pm 0.16}$
PC (E-FisherZ)	8.6$_{\pm 2.4}$	0.29$_{\pm 0.25}$	0.22$_{\pm 0.18}$	0.25$_{\pm 0.20}$	9.1$_{\pm 2.6}$	0.20$_{\pm 0.19}$	0.21$_{\pm 0.19}$	0.20$_{\pm 0.19}$
PC (KCIT)	8.2$_{\pm 1.5}$	0.19$_{\pm 0.18}$	0.15$_{\pm 0.13}$	0.16$_{\pm 0.14}$	6.3$_{\pm 2.9}$	0.37$_{\pm 0.30}$	0.37$_{\pm 0.30}$	0.37$_{\pm 0.30}$
PC (E-KCIT)	7.9$_{\pm 1.3}$	0.21$_{\pm 0.22}$	0.14$_{\pm 0.14}$	0.16$_{\pm 0.15}$	6.4$_{\pm 2.3}$	0.33$_{\pm 0.27}$	0.32$_{\pm 0.25}$	0.32$_{\pm 0.26}$
GES	8.5$_{\pm 2.7}$	0.34$_{\pm 0.29}$	0.25$_{\pm 0.20}$	0.29$_{\pm 0.23}$	9.3$_{\pm 3.5}$	0.21$_{\pm 0.28}$	0.22$_{\pm 0.27}$	0.21$_{\pm 0.27}$
ICALiNGAM	7.0$_{\pm 1.3}$	0.87$_{\pm 0.31}$	0.23$_{\pm 0.14}$	0.34$_{\pm 0.18}$	6.7$_{\pm 1.4}$	0.89$_{\pm 0.31}$	0.25$_{\pm 0.16}$	0.38$_{\pm 0.21}$
DirectLiNGAM	6.5$_{\pm 1.4}$	0.93$_{\pm 0.24}$	0.28$_{\pm 0.15}$	0.41$_{\pm 0.19}$	6.7$_{\pm 1.5}$	0.89$_{\pm 0.31}$	0.26$_{\pm 0.17}$	0.38$_{\pm 0.21}$
Notears	7.8$_{\pm 1.0}$	0.76$_{\pm 0.43}$	0.13$_{\pm 0.11}$	0.22$_{\pm 0.16}$	7.9$_{\pm 0.9}$	0.72$_{\pm 0.45}$	0.12$_{\pm 0.10}$	0.21$_{\pm 0.16}$
GraNDAG	9.7$_{\pm 4.9}$	0.45$_{\pm 0.38}$	0.26$_{\pm 0.18}$	0.31$_{\pm 0.24}$	10.8$_{\pm 4.8}$	0.33$_{\pm 0.26}$	0.26$_{\pm 0.17}$	0.28$_{\pm 0.19}$
GOLEM	6.7$_{\pm 1.3}$	**0.94**$_{\pm 0.24}$	0.26$_{\pm 0.14}$	0.39$_{\pm 0.18}$	6.8$_{\pm 1.2}$	**0.90**$_{\pm 0.30}$	0.25$_{\pm 0.14}$	0.38$_{\pm 0.18}$
$d = 20$								
DCD	**14.5**$_{\pm 3.4}$	0.66$_{\pm 0.13}$	**0.45**$_{\pm 0.13}$	**0.53**$_{\pm 0.13}$	15.1$_{\pm 6.4}$	0.46$_{\pm 0.31}$	**0.46**$_{\pm 0.31}$	**0.46**$_{\pm 0.31}$
PC (FisherZ)	26.2$_{\pm 3.8}$	0.11$_{\pm 0.08}$	0.11$_{\pm 0.09}$	0.11$_{\pm 0.08}$	29.2$_{\pm 6.8}$	0.11$_{\pm 0.07}$	0.15$_{\pm 0.08}$	0.13$_{\pm 0.08}$
PC (E-FisherZ)	21.0$_{\pm 3.8}$	0.19$_{\pm 0.14}$	0.15$_{\pm 0.10}$	0.17$_{\pm 0.12}$	23.5$_{\pm 6.7}$	0.17$_{\pm 0.17}$	0.19$_{\pm 0.16}$	0.18$_{\pm 0.16}$
PC (KCIT)	20.0$_{\pm 2.8}$	0.15$_{\pm 0.13}$	0.11$_{\pm 0.09}$	0.13$_{\pm 0.10}$	15.6$_{\pm 4.5}$	0.30$_{\pm 0.23}$	0.30$_{\pm 0.22}$	0.30$_{\pm 0.22}$
PC (E-KCIT)	17.6$_{\pm 1.9}$	0.19$_{\pm 0.12}$	0.13$_{\pm 0.07}$	0.15$_{\pm 0.09}$	**14.2**$_{\pm 3.6}$	0.33$_{\pm 0.22}$	0.29$_{\pm 0.19}$	0.31$_{\pm 0.20}$
GES	24.3$_{\pm 4.0}$	0.12$_{\pm 0.16}$	0.10$_{\pm 0.11}$	0.11$_{\pm 0.13}$	28.9$_{\pm 7.0}$	0.08$_{\pm 0.08}$	0.11$_{\pm 0.09}$	0.09$_{\pm 0.08}$
ICALiNGAM	15.7$_{\pm 2.0}$	0.82$_{\pm 0.33}$	0.18$_{\pm 0.09}$	0.29$_{\pm 0.14}$	15.2$_{\pm 1.9}$	0.89$_{\pm 0.24}$	0.20$_{\pm 0.10}$	0.32$_{\pm 0.14}$
DirectLiNGAM	14.9$_{\pm 1.7}$	0.93$_{\pm 0.22}$	0.23$_{\pm 0.09}$	0.36$_{\pm 0.12}$	14.5$_{\pm 2.8}$	0.91$_{\pm 0.25}$	0.24$_{\pm 0.15}$	0.36$_{\pm 0.18}$
Notears	16.8$_{\pm 1.4}$	0.91$_{\pm 0.28}$	0.12$_{\pm 0.08}$	0.20$_{\pm 0.12}$	17.4$_{\pm 1.4}$	0.72$_{\pm 0.46}$	0.08$_{\pm 0.07}$	0.14$_{\pm 0.12}$
GraNDAG	33.4$_{\pm 13.1}$	0.23$_{\pm 0.23}$	0.18$_{\pm 0.09}$	0.18$_{\pm 0.10}$	32.7$_{\pm 9.4}$	0.18$_{\pm 0.16}$	0.16$_{\pm 0.08}$	0.15$_{\pm 0.09}$
GOLEM	14.8$_{\pm 1.5}$	**0.95**$_{\pm 0.21}$	0.23$_{\pm 0.08}$	0.36$_{\pm 0.11}$	14.5$_{\pm 2.6}$	**0.96**$_{\pm 0.20}$	0.24$_{\pm 0.14}$	0.36$_{\pm 0.18}$

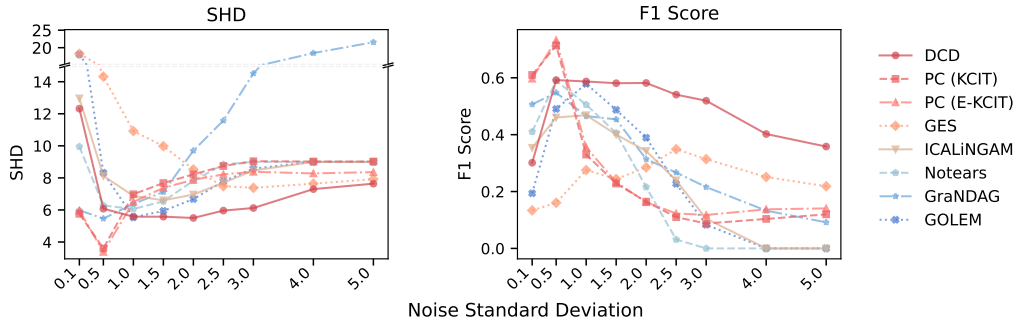
(FisherZ) (Fisher, 1921), and their respective ensemble variants (E-KCIT/E-FisherZ) (Guan & Kuang, 2026), to achieve optimal empirical performance. Since the evaluated methods output either DAG or CPDAG, we uniformly convert all estimated structures into their corresponding CPDAG to ensure a fair and consistent comparison. We evaluate the performance using standard metrics: Structural Hamming Distance adapted for CPDAGs (SHD_c), Precision, Recall, and F1-score.

4.1 Performance Across Mechanisms

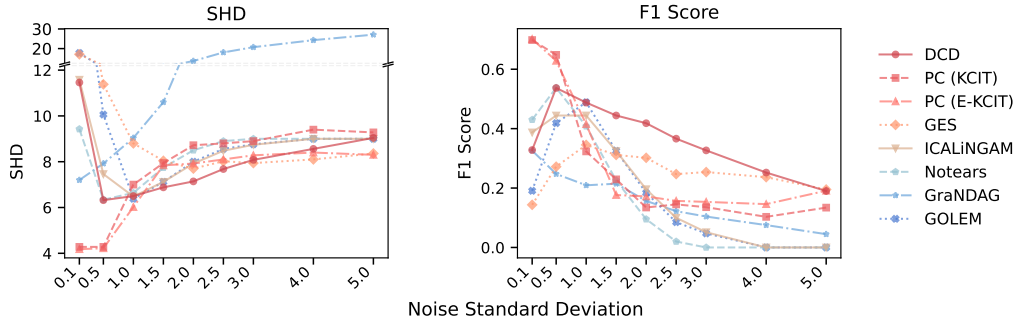
In this section, we primarily evaluate the effectiveness of our method under different data generation mechanisms. We randomly generated DAGs by the Barabási-Albert model with an attachment parameter of 1 to create scale-free graphs (approximately degree 2 per node, i.e., in-degree or out-degree 1). Given a DAG, the data are generated according to

$$X_i = f_i(\text{Pa}(X_i)) + \epsilon_i,$$

where ϵ_i is an exogenous noise randomly sampled from one of the following distributions: Gaussian, Laplace, Student's t, and Uniform, with a standard deviation equal to 2. We consider two structural mechanisms $f_i(\cdot)$: (i) *Functional mechanisms (Func)*, where $f_i(\text{Pa}(X_i)) = g_i(W_i^\top \text{Pa}(X_i))$, with each entry of $W_i^\top$ independently sampled from $U(0.5, 2.0)$ with random signs and divided by



(a) Functional mechanism



(b) Noise-in-mechanism

Figure 1: Performance under increasing noise levels.

$\sqrt{\dim(\mathrm{Pa}(X_i))}$, and $g_i(\cdot)$ randomly selected from $\{x, x^2, x^3, \tanh(x), \exp(x), \log(|x|)\}$. (ii) ***MLP mechanisms (MLP)***, where $f_i(\mathrm{Pa}(X_i))$ is implemented as a 5-layer MLP with hidden dimension 20 and $\tanh$ activations, with weights sampled as in Func and normalized to preserve layer-wise output variance. For both mechanisms, the resulting structural signal $f_i(\mathrm{Pa}(X_i))$ is standardized to zero mean and unit variance before adding noise, to prevent numerical explosion and ensure comparability with noise magnitudes.

Table 1 presents the performance evaluation across two mechanisms and varying graph sizes, with sample size $n = 800$, reporting the mean and standard deviation over 50 independent trials (25 trials for $d = 20$). Across both functional and MLP mechanisms and all graph sizes, DCD achieves superior overall performance, yielding the lowest SHD_c in almost all settings with consistently best Recall and F1-scores. While some baselines (e.g., GOLEM) attain high Precision at $d = 20$, these overly conservative edge predictions compromise Recall, leading to suboptimal overall accuracy.

4.2 Noise Robustness

We evaluate noise robustness using the same setting as in Section 4.1 (Func, $d = 10$), while varying the standard deviation of noise from 0.1 to 5.0. Additionally, we consider a more stringent noise-in-mechanism variant, where noise is injected before the nonlinear transformation ($X_i = g_i(W_i^\top \mathrm{Pa}(X_i) + \epsilon_i)$).

Figure 1 illustrates the trajectories of SHD_c (left) and F1-score (right) under two noise mechanisms, based on 50 independent trials. For visual clarity, we omit the variance bars and lines for PC with FisherZ/E-FisherZ (which lag behind PC with KCIT/E-KCIT) and DirectLiNGAM (which is slightly inferior to ICALiNGAM) from the figures. Under low-noise regimes, PC with KCIT/E-KCIT stands out. However, as noise increases, the performance of all baselines degrades rapidly, except GES which consistently underperforms. In contrast, DCD demonstrates superior robustness, remaining stable even under high-noise conditions. For instance, in Figure 1a at noise standard deviation near 5.0, DCD maintains an F1-score around 0.4, whereas some baselines drop close to zero.

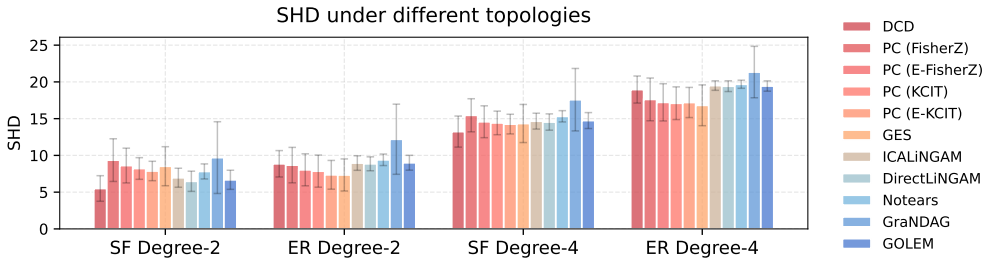


Figure 2: Performance under different graph topologies and densities.

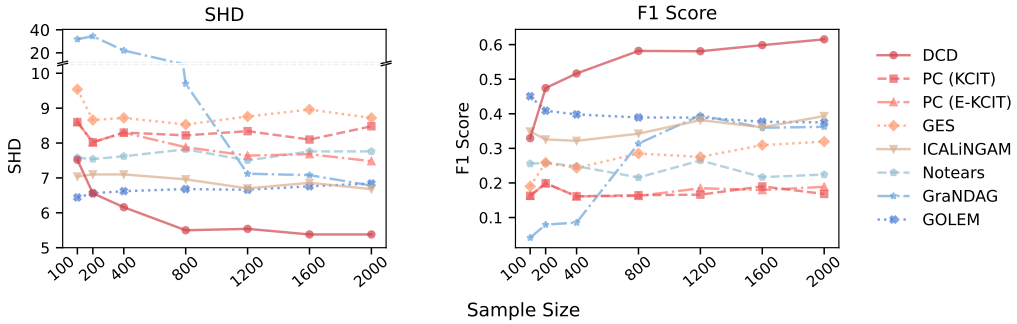


Figure 3: Performance under increasing sample sizes.

4.3 Performance Across Topologies

We evaluate graph topology robustness using the same setting as in Section 4.1 (Func, $d = 10$), while varying the topological structures and densities via Scale-Free (SF) and Erdős-Rényi (ER) graphs across average degree 2 and 4 (i.e., 10 and 20 edges per graph). To guarantee connectivity, SF DAGs are generated via the Barabási-Albert model with attachment parameters 1 or 2, while ER DAGs are approximately generated via Prüfer sequences and augmented with additional edges accordingly.

Figure 2 presents the SHD_c under four different graph settings, averaged over 50 trials. Error bars represent the standard deviation across the independent trials. DCD exhibits solid robustness with a substantial advantage across SF graphs of both densities. Although PC and GES show superior performance on ER graphs, DCD remains highly competitive and outperforms all other baselines.

4.4 Sample Efficiency

We evaluate sample efficiency using the same setting as in Section 4.1 (Func, $d = 10$), while varying the sample size from 100 to 2000. Figure 3 illustrates the trajectories over sample sizes (with the same plotting conventions as Figure 1). DCD outperforms all baselines across almost all sample sizes, with its performance steadily improving as the sample size increases.

5 Discussion

In this paper, we have introduced Decoupled Causal Discovery (DCD) and established its theoretical foundation. By leveraging the properties of decoupled distributions, DCD directly identifies the MB and iteratively constructs the global structure. Although DCD has demonstrated empirical effectiveness, there are several limitations and future directions, including the practical challenges of learning decoupling weights (e.g., the requirement of a certain amount of noise, high dimensionality, and the risk of information loss) and the potential of the purely MB-driven Algorithm 1 (especially combining decoupling with CIT). Due to space limitations, detailed discussions are provided in Appendix C. Overall, as a decoupling-based perspective orthogonal to CITs, score functions, or restrictive functional assumptions, DCD broadens the methodological scope of causal discovery.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation and assisting in the writing of proofs. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, or any other tasks with required disclosure. Additionally, we used generative AI tools for editing the paper to improve readability. We have reviewed all AI-assisted work. All revised text was manually checked for clarity and technical accuracy. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide a complete and anonymous source code package containing all algorithm implementations and experiment scripts in the Supplementary Material. Additionally, the main text includes the detailed pseudo-code for our algorithm (Section 3.2), implementation details (Section 3.3), and comprehensive experimental settings (Section 4). Complete proofs for all theoretical results in the main text are provided in Appendix A.

REFERENCES

Sina Akbari, Ehsan Mokhtarian, AmirEmad Ghassami, and Negar Kiyavash. Recursive causal structure learning in the presence of latent variables and selection bias. *Advances in Neural Information Processing Systems*, 34:10119–10130, 2021.

Peter Bühlmann, Jonas Peters, and Jan Ernest. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Yikang Chen, Xingzhe Sun, and Dehui du. Causal discovery via quantile partial effect. In *International Conference on Learning Representations*, 2026.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Diego Colombo and Marloes H Maathuis. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

Gregory F Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347, 1992.

Peng Cui and Susan Athey. Stable learning establishes some common ground between causal inference and machine learning. *Nature Machine Intelligence*, 4(2):110–115, February 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00445-z.

Haoyue Dai, Immanuel Albrecht, Peter Spirtes, and Kun Zhang. Distributional equivalence in linear non-gaussian latent-variable cyclic causal models: Characterization and learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ronald A Fisher. On the" probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Zhengkang Guan and Kun Kuang. Efficient ensemble conditional independence test framework for causal discovery. In *International Conference on Learning Representations*, 2026.

David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008a.

Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008b.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1551–1560, 2018.

Ilmun Kim, Matey Neykov, Sivaraman Balakrishnan, and Larry A. Wasserman. Local permutation tests for conditional independence. *The Annals of Statistics*, 50(6):3388–3414, 2022. doi: 10.1214/22-AOS2233.

Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. Stable prediction across unknown environments. In *KDD*, pp. 1617–1626, 2018.

Kun Kuang, Haotian Wang, Yue Liu, Ruoxuan Xiong, Runze Wu, Weiming Lu, Yueting Zhuang, Fei Wu, Peng Cui, and Bo Li. Stable prediction with leveraging seed variable. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):6392–6404, 2023. doi: 10.1109/TKDE.2022.3169333.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. In *International Conference on Learning Representations*, 2020.

Chun Li and Xiaodan Fan. On nonparametric conditional independence tests for continuous variables. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(3):e1489, 2020.

Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

Christopher Meek. Causal inference and causal explanation with background knowledge. In Philippe Besnard and Steve Hanks (eds.), *UAI '95: Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, August 18-20, 1995*, pp. 403–410. Morgan Kaufmann, 1995.

Ehsan Mokhtarian, Sina Akbari, AmirEmad Ghassami, and Negar Kiyavash. A recursive Markov boundary-based approach to causal structure learning. In *The KDD'21 Workshop on Causal Discovery*, pp. 26–54. PMLR, 2021.

Ehsan Mokhtarian, Sina Akbari, Fateme Jamshidi, Jalal Etesami, and Negar Kiyavash. Learning bayesian networks in the presence of structural side information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7814–7822, 2022.

Ehsan Mokhtarian, Mohmmadsadegh Khorasani, Jalal Etesami, and Negar Kiyavash. Novel ordering-based approaches for causal structure learning in the presence of unobserved variables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 12260–12268, 2023.

Ehsan Mokhtarian, Sepehr Elahi, Sina Akbari, and Negar Kiyavash. Recursive causal discovery. *Journal of Machine Learning Research*, 26(61):1–65, 2025.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. Methods and tools for causal discovery and causal inference. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 12(2):e1449, 2022.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge university press, 2nd edition, 2009.

Judea Pearl and Thomas Verma. A theory of inferred causation. In *International Conference on Principles of Knowledge Representation and Reasoning*, 1991.

Jean-Philippe Pellet and André Elisseeff. Using Markov blankets for causal structure learning. *Journal of Machine Learning Research*, 9(43):1295–1342, 2008.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning series. MIT Press, Cambridge, MA, 2017. ISBN 9780262037310.

Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30 (4):962–1030, 2002. ISSN 00905364.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pp. 461–464, 1978.

Rajen Dinesh Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020. doi: 10.1214/19-AOS1857.

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li, and Peixuan Chen. Causally regularized learning with agnostic data selection bias. In *ACM Multimedia*, pp. 411–419, 2018.

Zheyan Shen, Peng Cui, Tong Zhang, and Kun Kuang. Stable learning via sample reweighting. In *AAAI*, pp. 5692–5699, 2020.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. DirectLiNGAM: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr):1225–1248, 2011.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. doi: 10.1016/S0378-3758(00)00115-4.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9:62–72, 1991.

Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pp. 1–28. SpringerOpen, 2016.

Peter Spirtes, Clark Glymour, and Richard Scheines. Causality from probability. 1989.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 499–506. Morgan Kaufmann, 1995.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

Eric V Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1):20180017, 2019.

Natasa Tagasovska, Valérie Chavez-Demoulin, and Thibault Vatter. Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9311–9323. PMLR, 2020.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In Piero P. Bonissone, Max Henrion, Laveen N. Kanal, and John F. Lemmer (eds.), *UAI '90: Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, MIT, Cambridge, MA, USA, July 27-29, 1990*, pp. 255–270. Elsevier, 1990.

Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.

Johnny Xi, Hugh Dance, Peter Orbanz, and Benjamin Bloem-Reddy. Distinguishing cause from effect with causal velocity models. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Renzhe Xu, Xingxuan Zhang, Zheyan Shen, Tong Zhang, and Peng Cui. A theoretical analysis on independence-driven importance weighting for covariate-shift generalization. In *ICML*, pp. 24803–24829, 2022.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Conference on Uncertainty in Artificial Intelligence*, pp. 647–655, 2009.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In Fábio Gagliardi Cozman and Avi Pfeffer (eds.), *UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, July 14-17, 2011*, pp. 804–813. AUAI Press, 2011.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Neural Information Processing Systems*, 2018.

A OMITTED PROOFS

To facilitate the subsequent derivations, we first recall several fundamental properties of conditional independence (i.e., the graphoid axioms) and the Markov boundary (MB) (Pearl, 1988).

Lemma A.1 (Decomposition Axiom)**.** *For any disjoint sets of random variables* $\mathbf{U}$*,* $\mathbf{V}$*,* $\mathbf{W}$*, and* $\mathbf{T}$*:*

$$\mathbf{U} \perp\!\!\!\perp (\mathbf{V} \cup \mathbf{T}) \mid \mathbf{W} \implies \mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W} \text{ and } \mathbf{U} \perp\!\!\!\perp \mathbf{T} \mid \mathbf{W}.$$

Lemma A.2 (Weak Union Axiom)**.** *For any disjoint sets of random variables* $\mathbf{U}$*,* $\mathbf{V}$*,* $\mathbf{W}$*, and* $\mathbf{T}$*:*

$$\mathbf{U} \perp\!\!\!\perp (\mathbf{V} \cup \mathbf{T}) \mid \mathbf{W} \implies \mathbf{U} \perp\!\!\!\perp \mathbf{T} \mid (\mathbf{W} \cup \mathbf{V}).$$

Lemma A.3 (Contraction Axiom)**.** *For any disjoint sets of random variables* $\mathbf{U}$*,* $\mathbf{V}$*,* $\mathbf{W}$*, and* $\mathbf{T}$*:*

$$\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid \mathbf{W} \text{ and } \mathbf{U} \perp\!\!\!\perp \mathbf{T} \mid (\mathbf{W} \cup \mathbf{V}) \implies \mathbf{U} \perp\!\!\!\perp (\mathbf{V} \cup \mathbf{T}) \mid \mathbf{W}.$$

Lemma A.4 (Intersection Axiom)**.** *Under Assumption 1, for any disjoint sets of random variables* $\mathbf{U}$*,* $\mathbf{V}$*,* $\mathbf{W}$*, and* $\mathbf{T}$*:*

$$\mathbf{U} \perp\!\!\!\perp \mathbf{V} \mid (\mathbf{W} \cup \mathbf{T}) \text{ and } \mathbf{U} \perp\!\!\!\perp \mathbf{T} \mid (\mathbf{W} \cup \mathbf{V}) \implies \mathbf{U} \perp\!\!\!\perp (\mathbf{V} \cup \mathbf{T}) \mid \mathbf{W}.$$

Lemma A.5 (Uniqueness of MB)**.** *Let* $\mathcal{O}$ *be a set of random variables. Under Assumption 1, for any variable* $X_i \in \mathcal{O}$*,* $\mathrm{MB}(X_i \mid \mathcal{O})$ *is unique.*

Proof. We prove this by contradiction. Suppose there exist two distinct MBs for X_i, denoted by $\mathbf{M}_1$ and $\mathbf{M}_2$, with $\mathbf{M}_1 \neq \mathbf{M}_2$. By Definition 1, we have:

$$X_i \perp\!\!\!\perp \big(\mathcal{O} \setminus (\{X_i\} \cup \mathbf{M}_1)\big) \mid \mathbf{M}_1 \quad \text{and} \quad X_i \perp\!\!\!\perp \big(\mathcal{O} \setminus (\{X_i\} \cup \mathbf{M}_2)\big) \mid \mathbf{M}_2.$$

Let $\mathbf{D} = \mathbf{M}_1 \cap \mathbf{M}_2$, $\mathbf{V}_1 = \mathbf{M}_1 \setminus \mathbf{D}$, $\mathbf{V}_2 = \mathbf{M}_2 \setminus \mathbf{D}$, and $\mathbf{R} = \mathcal{O} \setminus (\{X_i\} \cup \mathbf{M}_1 \cup \mathbf{M}_2)$. The two conditional independence statements can be rewritten respectively as:

$$X_i \perp\!\!\!\perp (\mathbf{V}_2 \cup \mathbf{R}) \mid (\mathbf{D} \cup \mathbf{V}_1) \quad \text{and} \quad X_i \perp\!\!\!\perp (\mathbf{V}_1 \cup \mathbf{R}) \mid (\mathbf{D} \cup \mathbf{V}_2).$$

By the decomposition axiom (Lemma A.1), this implies:

$$X_i \perp\!\!\!\perp \mathbf{V}_2 \mid (\mathbf{D} \cup \mathbf{V}_1) \quad \text{and} \quad X_i \perp\!\!\!\perp \mathbf{V}_1 \mid (\mathbf{D} \cup \mathbf{V}_2).$$

Applying the intersection axiom (Lemma A.4) yields:

$$X_i \perp\!\!\!\perp (\mathbf{V}_1 \cup \mathbf{V}_2) \mid \mathbf{D}. \tag{2}$$

Meanwhile, by applying the weak union axiom (Lemma A.2) to $X_i \perp\!\!\!\perp (\mathbf{V}_2 \cup \mathbf{R}) \mid (\mathbf{D} \cup \mathbf{V}_1)$, we obtain:

$$X_i \perp\!\!\!\perp \mathbf{R} \mid (\mathbf{D} \cup \mathbf{V}_1 \cup \mathbf{V}_2). \tag{3}$$

By the contraction axiom (Lemma A.3), combining equations (2) and (3) yields:

$$X_i \perp\!\!\!\perp (\mathbf{V}_1 \cup \mathbf{V}_2 \cup \mathbf{R}) \mid \mathbf{D}.$$

Since $\mathbf{V}_1 \cup \mathbf{V}_2 \cup \mathbf{R} = \mathcal{O} \setminus (\{X_i\} \cup \mathbf{D})$, this is equivalent to $X_i \perp\!\!\!\perp \big(\mathcal{O} \setminus (\{X_i\} \cup \mathbf{D})\big) \mid \mathbf{D}$. Thus, by Definition 1, $\mathbf{D} = \mathbf{M}_1 \cap \mathbf{M}_2$ is a Markov blanket of X_i. Since $\mathbf{M}_1 \neq \mathbf{M}_2$, we have $\mathbf{D} \subsetneq \mathbf{M}_1$, which contradicts the minimality of $\mathbf{M}_1$ (Definition 1). Therefore, the MB is unique. □

While Lemmas A.1–A.4 are standard axioms (Pearl, 1988), we provide the brief proof for Lemma A.5 as a direct corollary, since the strict uniqueness of the MB is a foundational prerequisite for our theoretical framework. Beyond these purely probabilistic properties, incorporating a graph structure yields the following characterization (Pearl, 1988):

Lemma A.6 (MB for DAGs)**.** *Under Assumptions 3, 4, and 5, for the MB of* X_i *in* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$*, we have:*

$$\mathrm{MB}(X_i \mid \mathcal{V}) = \mathrm{Pa}(X_i) \cup \mathrm{Ch}(X_i) \cup \mathrm{Sp}(X_i).$$

In the following subsections, we provide detailed proofs for Theorems 1 and 2, respectively.

A.1 MB IDENTIFIABILITY

We first prove Proposition 1, a general property of the weighted distribution.

Proposition 1 (MB Conditional Distribution Invariance). *Let $\mathcal{O}$ be a set of variables, $X_t \in \mathcal{O}$ be the target variable, and $\mathbf{X}_{-t} = \mathcal{O} \setminus \{X_t\}$. Under Assumption 1, for any weighting function $w \in \mathcal{W}(\mathbf{X}_{-t})$ (and its corresponding $\tilde{P}_w(X_t, \mathbf{X}_{-t})$), the following holds for any $\mathbf{X}^+_{\text{MB}}$ satisfying $\text{MB}(X_t \mid \mathcal{O}) \subseteq \mathbf{X}^+_{\text{MB}} \subseteq \mathbf{X}_{-t}$:*

$$\tilde{P}_w(X_t \mid \mathbf{X}^+_{\text{MB}}) = P(X_t \mid \mathbf{X}^+_{\text{MB}}) = P(X_t \mid \text{MB}(X_t \mid \mathcal{O})) = P(X_t \mid \mathbf{X}_{-t}).$$

Proof. By the definition of MB (Definition 1), it is straightforward to obtain that

$$P(X_t \mid \mathbf{X}^+_{\text{MB}}) = P(X_t \mid \text{MB}(X_t \mid \mathcal{O})) = P(X_t \mid \mathbf{X}_{-t}).$$

Thus, we need only establish that $\tilde{P}_w(X_t \mid \mathbf{X}^+_{\text{MB}}) = P(X_t \mid \mathbf{X}^+_{\text{MB}})$. To simplify the integration, we denote $\mathbf{X}_{\text{MB}} = \text{MB}(X_t \mid \mathcal{O})$, $\mathbf{R}_t = \mathcal{O} \setminus \mathbf{X}^+_{\text{MB}}$ and $\mathbf{R}_{-t} = \mathbf{X}_{-t} \setminus \mathbf{X}^+_{\text{MB}}$. Then we have:

$$\begin{aligned}
\tilde{P}_w(X_t \mid \mathbf{X}^+_{\text{MB}}) &= \frac{\tilde{P}_w(X_t, \mathbf{X}^+_{\text{MB}})}{\tilde{P}_w(\mathbf{X}^+_{\text{MB}})} \\
&= \frac{\int \tilde{P}_w(X_t, \mathbf{X}_{-t})\, d\mathbf{R}_{-t}}{\int \tilde{P}_w(X_t, \mathbf{X}_{-t})\, d\mathbf{R}_t} \\
&\overset{\text{(a)}}{=} \frac{\int w(\mathbf{X}_{-t}) P(X_t, \mathbf{X}_{-t})\, d\mathbf{R}_{-t}}{\int w(\mathbf{X}_{-t}) P(X_t, \mathbf{X}_{-t})\, d\mathbf{R}_t} \\
&= \frac{\int w(\mathbf{X}_{-t}) P(X_t \mid \mathbf{X}_{-t}) P(\mathbf{X}_{-t})\, d\mathbf{R}_{-t}}{\int w(\mathbf{X}_{-t}) P(\mathbf{X}_{-t})\, d\mathbf{R}_{-t}} \\
&\overset{\text{(b)}}{=} \frac{\int w(\mathbf{X}_{-t}) P(X_t \mid \mathbf{X}_{\text{MB}}) P(\mathbf{X}_{-t})\, d\mathbf{R}_{-t}}{\int w(\mathbf{X}_{-t}) P(\mathbf{X}_{-t})\, d\mathbf{R}_{-t}} \\
&\overset{\text{(c)}}{=} \frac{P(X_t \mid \mathbf{X}_{\text{MB}}) \int w(\mathbf{X}_{-t}) P(\mathbf{X}_{-t})\, d\mathbf{R}_{-t}}{\int w(\mathbf{X}_{-t}) P(\mathbf{X}_{-t})\, d\mathbf{R}_{-t}} \\
&= P(X_t \mid \mathbf{X}_{\text{MB}})
\end{aligned}$$

where equation (a) follows from the definition of $\tilde{P}_w$ in Definition 2, equation (b) follows from the definition of MB (Definition 1), and (c) follows from the fact that $P(X_t \mid \mathbf{X}_{\text{MB}})$ does not depend on $\mathbf{R}_{-t}$. □

Building on this property, together with the definition of the decoupled distribution (Definition 3) and Weighting Faithfulness (Assumption 2), we prove Theorem 1.

Theorem 1 (MB Identifiability). *Let $\mathcal{O}$ be a set of variables, $X_t \in \mathcal{O}$ be the target variable, and $\mathbf{X}_{-t} = \mathcal{O} \setminus \{X_t\}$. Under Assumption 1, for any decoupling weight $w \in \mathcal{W}_\perp(\mathbf{X}_{-t})$ (and its corresponding $\tilde{P}_w(X_t, \mathbf{X}_{-t})$) satisfying Assumption 2, we have for any variable $X_i \in \mathbf{X}_{-t}$:*

$$X_i \in \text{MB}(X_t \mid \mathcal{O}) \iff X_i \not\perp\!\!\!\perp_{\tilde{P}_w} X_t.$$

Proof. To prove the equivalence $X_i \in \text{MB}(X_t \mid \mathcal{O}) \iff X_i \not\perp\!\!\!\perp_{\tilde{P}_w} X_t$, we show that both the necessity and sufficiency hold. For notational convenience, let $\mathbf{X}_{\text{MB}} = \text{MB}(X_t \mid \mathcal{O})$ and $\mathbf{X}_{\text{MB}\setminus i} = \text{MB}(X_t \mid \mathcal{O}) \setminus \{X_i\}$.

Necessity. We show that for every $X_i \in \mathrm{MB}(X_t \mid \mathcal{O})$, it holds that $X_i \not\perp\!\!\!\perp_{\tilde{P}_w} X_t$, which is equivalent to showing that the conditional distribution $\tilde{P}_w(X_t \mid X_i)$ is a nontrivial function of X_i. We can reformulate this conditional distribution:

$$\begin{aligned}
\tilde{P}_w(X_t \mid X_i) &= \int \tilde{P}_w(X_t, \mathbf{X}_{\mathrm{MB}\setminus i} \mid X_i)\, d\mathbf{X}_{\mathrm{MB}\setminus i} \\
&= \int \tilde{P}_w(X_t \mid \mathbf{X}_{\mathrm{MB}\setminus i}, X_i)\tilde{P}_w(\mathbf{X}_{\mathrm{MB}\setminus i} \mid X_i)\, d\mathbf{X}_{\mathrm{MB}\setminus i} \\
&\overset{\text{(a)}}{=} \int \tilde{P}_w(X_t \mid \mathbf{X}_{\mathrm{MB}})\tilde{P}_w(\mathbf{X}_{\mathrm{MB}\setminus i})\, d\mathbf{X}_{\mathrm{MB}\setminus i} \\
&\overset{\text{(b)}}{=} \int P(X_t \mid \mathbf{X}_{\mathrm{MB}})\tilde{P}_w(\mathbf{X}_{\mathrm{MB}\setminus i})\, d\mathbf{X}_{\mathrm{MB}\setminus i}
\end{aligned}$$

where equation (a) follows from $X_i \perp\!\!\!\perp_{\tilde{P}_w} \mathbf{X}_{\mathrm{MB}\setminus i}$ by Definition 3, and equation (b) follows from Proposition 1.

We next prove $X_t \not\perp\!\!\!\perp X_i \mid \mathbf{X}_{\mathrm{MB}\setminus i}$ by contradiction. Assume the contrary:

$$X_t \perp\!\!\!\perp X_i \mid \mathbf{X}_{\mathrm{MB}\setminus i}. \tag{4}$$

Meanwhile, by the definition of MB (Definition 1), we have

$$X_t \perp\!\!\!\perp \mathbf{R} \mid (\mathbf{X}_{\mathrm{MB}\setminus i} \cup \{X_i\}), \tag{5}$$

where $\mathbf{R} = \mathcal{O} \setminus (\mathbf{X}_{\mathrm{MB}} \cup \{X_t\})$. Applying the contraction axiom (Lemma A.3) to combine equations (4) and (5) yields:

$$X_t \perp\!\!\!\perp (\mathbf{R} \cup \{X_i\}) \mid \mathbf{X}_{\mathrm{MB}\setminus i},$$

which is a contradiction to the definition of MB (inclusion-minimal, Definition 1). Therefore, we conclude that $X_t \not\perp\!\!\!\perp X_i \mid \mathbf{X}_{\mathrm{MB}\setminus i}$, which implies that $P(X_t \mid X_i, \mathbf{X}_{\mathrm{MB}\setminus i})$ depends on X_i.

Note that

$$\begin{aligned}
\tilde{P}_w(X_t \mid X_i) &= \int P(X_t \mid \mathbf{X}_{\mathrm{MB}})\tilde{P}_w(\mathbf{X}_{\mathrm{MB}\setminus i})\, d\mathbf{X}_{\mathrm{MB}\setminus i} \\
&= \mathbb{E}_{\tilde{P}_w(\mathbf{X}_{\mathrm{MB}\setminus i})}[P(X_t \mid X_i, \mathbf{X}_{\mathrm{MB}\setminus i})].
\end{aligned}$$

Therefore, by Assumption 2, $\tilde{P}_w(X_t \mid X_i)$ also depends on X_i. Thus, the proof of necessity is complete.

Sufficiency. We prove this by contraposition, i.e., showing $X_i \notin \mathrm{MB}(X_t \mid \mathcal{O}) \implies X_i \perp\!\!\!\perp_{\tilde{P}_w} X_t$. Equivalently, we show $\tilde{P}_w(X_t \mid X_i) = \tilde{P}_w(X_t)$:

$$\begin{aligned}
\tilde{P}_w(X_t \mid X_i) &= \int \tilde{P}_w(X_t, \mathbf{X}_{\mathrm{MB}} \mid X_i)\, d\mathbf{X}_{\mathrm{MB}} \\
&= \int \tilde{P}_w(X_t \mid \mathbf{X}_{\mathrm{MB}}, X_i)\tilde{P}_w(\mathbf{X}_{\mathrm{MB}} \mid X_i)\, d\mathbf{X}_{\mathrm{MB}} \\
&\overset{\text{(a)}}{=} \int P(X_t \mid \mathbf{X}_{\mathrm{MB}}, X_i)\tilde{P}_w(\mathbf{X}_{\mathrm{MB}})\, d\mathbf{X}_{\mathrm{MB}} \\
&\overset{\text{(b)}}{=} \int P(X_t \mid \mathbf{X}_{\mathrm{MB}})\tilde{P}_w(\mathbf{X}_{\mathrm{MB}})\, d\mathbf{X}_{\mathrm{MB}} \\
&\overset{\text{(c)}}{=} \int \tilde{P}_w(X_t \mid \mathbf{X}_{\mathrm{MB}})\tilde{P}_w(\mathbf{X}_{\mathrm{MB}})\, d\mathbf{X}_{\mathrm{MB}} \\
&= \tilde{P}_w(X_t)
\end{aligned}$$

where equation (a) follows from Proposition 1 and $X_i \perp\!\!\!\perp_{\tilde{P}_w} \mathbf{X}_{\mathrm{MB}}$ by Definition 3, equation (b) follows from the definition of MB (Definition 1), and (c) follows from Proposition 1. Thus, the proof of sufficiency is complete.

With both sufficiency and necessity proved, the equivalence follows.

□

Finally, we show that Weighting Faithfulness (Assumption 2) required by Theorem 1 holds almost everywhere under Condition 1.

Proposition 2 (Genericity of Weighting Faithfulness)**.** *Let $\mathcal{O}$ be a set of variables, $X_t \in \mathcal{O}$ be the target variable, and $\mathbf{X}_{-t} = \mathcal{O} \setminus \{X_t\}$. Under Assumption 1 and Condition 1, for $\mu_{\mathcal{W}}^{\mathbf{X}_{-t}}$-almost every weight $w \in \mathcal{W}_{\perp}(\mathbf{X}_{-t})$, Assumption 2 (Weighting Faithfulness) holds.*

Proof. Consider any $X_e \in \mathbf{X}_{-t}$ and $\mathbf{X}_c \subseteq \mathbf{X}_{-t} \setminus \{X_e\}$. We prove that if $P(X_t \mid X_e, \mathbf{X}_c)$ depends on X_e, then

$$g(X_t, X_e) = \mathbb{E}_{\tilde{P}_w(\mathbf{X}_c)}[P(X_t \mid X_e, \mathbf{X}_c)]$$

is not a constant function of X_e, for $\mu_{\mathcal{W}}^{\mathbf{X}_{-t}}$-almost every weight $w \in \mathcal{W}_{\perp}(\mathbf{X}_{-t})$.

Let $\nu = P_{X_e}$. For measurable events A and B in the state spaces of X_t and X_e, respectively, define

$$h_A(x_e, \mathbf{x}_c) = P(X_t \in A \mid X_e = x_e, \mathbf{X}_c = \mathbf{x}_c),$$

and

$$r_B(x_e) = \mathbf{1}_B(x_e) - \nu(B).$$

Since $P(X_t \mid X_e, \mathbf{X}_c)$ depends on X_e, there exist such events A and B for which

$$\Delta(\mathbf{x}_c) := \int r_B(x_e) h_A(x_e, \mathbf{x}_c)\, \nu(dx_e)$$

is nonzero on a set of positive $\lambda_{\mathbf{X}_c}$-measure.

We prove this by contradiction. Consider countable determining classes of events for X_t and X_e (for real-valued variables, rational half-lines suffice to determine the distribution). If the above integral were zero almost everywhere for every pair A, B from these classes, then outside a common null set of $\mathbf{x}_c$ values, we would have

$$\int_B h_A(x_e, \mathbf{x}_c)\, \nu(dx_e) = \nu(B) \int h_A(x_e, \mathbf{x}_c)\, \nu(dx_e)$$

for every such pair. By uniqueness of finite measures, $h_A(\cdot, \mathbf{x}_c)$ would be ν-almost everywhere constant for every A in the determining class. Under Assumption 1, the joint distribution of $(X_e, \mathbf{X}_c)$ is equivalent to $\nu \otimes \lambda_{\mathbf{X}_c}$. Consequently, $P(X_t \mid X_e, \mathbf{X}_c)$ would depend only on $\mathbf{X}_c$, which contradicts the fact that $P(X_t \mid X_e, \mathbf{X}_c)$ depends on X_e.

Fix such events A and B. Since $0 \le h_A \le 1$ and $|r_B| \le 1$, $\Delta(\mathbf{x}_c)$ is bounded and measurable, with $|\Delta(\mathbf{x}_c)| \le 1$. Moreover,

$$\int r_B(x_e)\, \nu(dx_e) = 0.$$

If some $w' \in \mathcal{W}_{\perp}(\mathbf{X}_{-t})$ renders $g(X_t, X_e)$ constant as a function of X_e, then

$$g_A(x_e) := \int h_A(x_e, \mathbf{x}_c)\, \tilde{P}_{w'}(\mathbf{x}_c)\, \lambda_{\mathbf{X}_c}(d\mathbf{x}_c)$$

is also constant ν-almost everywhere. Therefore, by Fubini's theorem,

$$\begin{aligned} 0 &= \int r_B(x_e) g_A(x_e)\, \nu(dx_e) \\ &= \int \left(\int r_B(x_e) h_A(x_e, \mathbf{x}_c)\, \nu(dx_e) \right) \tilde{P}_{w'}(\mathbf{x}_c)\, \lambda_{\mathbf{X}_c}(d\mathbf{x}_c) \\ &= \int \Delta(\mathbf{x}_c)\, \tilde{P}_{w'}(\mathbf{x}_c)\, \lambda_{\mathbf{X}_c}(d\mathbf{x}_c). \end{aligned}$$

Using the relation

$$\tilde{P}_{w'}(\mathbf{x}_c) = \int \tilde{P}_{w'}(\mathbf{x}_{-t})\, \lambda_{\mathbf{R}_c}(d\mathbf{r}_c),$$

where $\mathbf{r}_c$ denotes a value of $\mathbf{R}_c = \mathbf{X}_{-t} \setminus \mathbf{X}_c$, we obtain

$$\int \Delta(\mathbf{x}_c)\, \tilde{P}_{w'}(\mathbf{x}_c)\, \lambda_{\mathbf{X}_c}(d\mathbf{x}_c) = \int \Delta(\mathbf{x}_c)\, \tilde{P}_{w'}(\mathbf{x}_{-t})\, \lambda_{\mathbf{X}_{-t}}(d\mathbf{x}_{-t}) = 0.$$

Define the linear functional $G : L^1(\lambda_{\mathbf{X}_{-t}}) \to \mathbb{R}$ by

$$G(f) = \int \Delta(\mathbf{x}_c) f(\mathbf{x}_{-t})\, \lambda_{\mathbf{X}_{-t}}(d\mathbf{x}_{-t}).$$

Thus, the set of $w \in \mathcal{W}_\perp(\mathbf{X}_{-t})$ for which $g(X_t, X_e)$ is constant as a function of X_e is contained in the set of weights satisfying

$$G(w \cdot P_{\mathbf{X}_{-t}}) = \int \Delta(\mathbf{x}_c)\, w(\mathbf{x}_{-t}) P(\mathbf{x}_{-t})\, \lambda_{\mathbf{X}_{-t}}(d\mathbf{x}_{-t}) = 0.$$

Since $|\Delta(\mathbf{x}_c)| \leq 1$, we have $|G(f)| \leq \|f\|_1$, so G is continuous. Moreover, because $\Delta(\mathbf{x}_c)$ is nonzero on a set of positive $\lambda_{\mathbf{X}_c}$-measure and $\lambda_{\mathbf{X}_{-t}}$ is the corresponding product measure, G is not the zero functional. Hence, by Condition 1,

$$\mu_{\mathcal{W}}^{\mathbf{X}_{-t}}\big(\{w \in \mathcal{W}_\perp(\mathbf{X}_{-t}) \mid G(w \cdot P_{\mathbf{X}_{-t}}) = 0\}\big) = 0.$$

Therefore,

$$\mu_{\mathcal{W}}^{\mathbf{X}_{-t}}\big(\{w \in \mathcal{W}_\perp(\mathbf{X}_{-t}) \mid g(X_t, X_e) \text{ is constant as a function of } X_e\}\big) = 0.$$

Since there are finitely many choices of X_e and $\mathbf{X}_c$, the union of the corresponding exceptional sets also has measure zero.

□

A.2 Soundness and Completeness of DCD-C

We prove Theorem 2 in several steps. Lemmas A.7 and A.8 establish basic properties of MBs in DAGs. Based on these, Lemma A.9 gives a sufficient condition for v-structure exclusion, and Corollary A.1 further demonstrates the existence of such a condition. We then prove the soundness and completeness of the skeleton and v-structures (Lemmas A.10 and A.11), thereby establishing Theorem 2.

Since our setting assumes causal sufficiency (Assumption 3), we adopt a direct and self-contained proof strategy based on the statistical definition of MB (Definition 1) for intuitive simplicity, without introducing Maximal Ancestral Graphs (MAGs) or latent projections (Richardson & Spirtes, 2002). All graphical relations we discussed (e.g., $\mathrm{Pa}(\cdot)$, $\mathrm{Ch}(\cdot)$, and $\mathrm{Sp}(\cdot)$) are defined with respect to $\mathcal{G}$. If the MAG framework were introduced, the overall structure would be similar, while some of the arguments could instead be formulated based on corresponding graphical tools such as inducing paths.

Lemma A.7 (Neighbor Persistence)**.** *Let $\mathcal{O} \subseteq \mathcal{V}$ be a set of variables containing X_i and a neighbor $A \in \mathrm{Pa}(X_i) \cup \mathrm{Ch}(X_i)$. Then $A \in \mathrm{MB}(X_i \mid \mathcal{O})$.*

Proof. We prove this by contradiction. Assuming $A \notin \mathrm{MB}(X_i \mid \mathcal{O})$, by the definition of the MB (Definition 1) and the decomposition axiom (Lemma A.1), we have

$$X_i \perp\!\!\!\perp A \mid \mathrm{MB}(X_i \mid \mathcal{O}),$$

with $\mathrm{MB}(X_i \mid \mathcal{O}) \subseteq \mathcal{O} \subseteq \mathcal{V}$.

Since X_i and A are adjacent in $\mathcal{G}$, they cannot be d-separated. Therefore, by Assumption 5 (Faithfulness), X_i and A cannot be conditionally independent under any conditioning set. This contradicts $X_i \perp\!\!\!\perp A \mid \mathrm{MB}(X_i \mid \mathcal{O})$. Hence, $A \in \mathrm{MB}(X_i \mid \mathcal{O})$.

□

Lemma A.8 (Spouse Persistence). *Let $\mathcal{O} \subseteq \mathcal{V}$ be a set of variables containing X_i, $B \in \mathrm{Sp}(X_i)$ and $A \in \mathrm{Ch}(X_i) \cap \mathrm{Ch}(B)$. Then $B \in \mathrm{MB}(X_i \mid \mathcal{O})$.*

Proof. If B and X_i are adjacent, the statement reduces to Lemma A.7 and holds trivially. We therefore consider the case where B and X_i are non-adjacent.

Assume, for the sake of contradiction, that $B \notin \mathrm{MB}(X_i \mid \mathcal{O})$. There exists a path $X_i \to A \leftarrow B$ in $\mathcal{G}$, and by Lemma A.7, $A \in \mathrm{MB}(X_i \mid \mathcal{O})$. Thus, the set $\mathrm{MB}(X_i \mid \mathcal{O})$ does not block the path $X_i \to A \leftarrow B$, which implies that X_i and B are not d-separated by $\mathrm{MB}(X_i \mid \mathcal{O})$ (in $\mathcal{G}$).

Consequently, by Assumption 5 (Faithfulness), we have

$$X_i \not\perp\!\!\!\perp B \mid \mathrm{MB}(X_i \mid \mathcal{O}).$$

This contradicts the definition of MB (Definition 1). Hence, $B \in \mathrm{MB}(X_i \mid \mathcal{O})$.

□

Lemma A.9 (Sufficient Condition for v-structure Exclusion). *Suppose X_i and Y_j in $\mathcal{G}$ form exactly K v-structures $X_i \to Z_k \leftarrow Y_j$ for $k = 1, \ldots, K$. Let $\mathcal{Z}^* = \{Z_1, \ldots, Z_K\} \cup \bigcup_{k=1}^{K} \mathrm{De}(Z_k)$, where $\mathrm{De}(Z_k)$ denotes the set of all descendants of Z_k ($X_i, Y_j \notin \mathcal{Z}^*$ by acyclicity of $\mathcal{G}$). Then $Y_j \notin \mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*})$, where $\mathcal{O}_{i\setminus\mathcal{Z}^*} = (\mathrm{MB}(X_i \mid \mathcal{V}) \cup \{X_i\}) \setminus \mathcal{Z}^*$.*

Proof. Since $Y_j \in \mathrm{Sp}(X_i)$, Lemma A.6 yields $Y_j \in \mathrm{MB}(X_i \mid \mathcal{V})$. Combined with $Y_j \notin \mathcal{Z}^*$, it follows that $Y_j \in (\mathrm{MB}(X_i \mid \mathcal{V}) \cup \{X_i\}) \setminus \mathcal{Z}^* = \mathcal{O}_{i\setminus\mathcal{Z}^*}$. Consequently, $Y_j \notin \mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*})$ does not trivially hold. We now prove it holds by contradiction.

Assuming $Y_j \in \mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*})$, i.e., $Y_j \notin \mathbf{R}^*$, we first show X_i and Y_j are d-separated (in $\mathcal{G}$) by $\mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*}) \setminus \{Y_j\}$. We proceed by considering all possible paths (in $\mathcal{G}$) between X_i and Y_j, and showing each is blocked.

- ***First-order paths.*** By the definition of v-structures, X_i and Y_j are non-adjacent, so no path of length 1 exists.
- ***Second-order paths.*** Any length-2 path between X_i and Y_j takes one of two forms:
  - v-structure path $X_i \to Z_k \leftarrow Y_j$: The collider node Z_k and all its descendants belong to $\mathcal{Z}^*$ and are therefore not in $\mathcal{O}_{i\setminus\mathcal{Z}^*}$. Consequently, this path will always remain blocked given $\mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*}) \setminus \{Y_j\}$, as Z_k and all its descendants are absent from this set.
  - Non-v-structure paths (i.e., $X_i \to A \to Y_j$, $X_i \leftarrow A \leftarrow Y_j$, or $X_i \leftarrow A \to Y_j$):
    - If $A \in \mathcal{Z}^*$: There must exist directed paths from both X_i and Y_j to A. Since the graph $\mathcal{G}$ is acyclic, these directed paths preclude the existence of the reverse edges $A \to X_i$ and $A \to Y_j$ in $\mathcal{E}$. Therefore, this case is impossible.
    - If $A \notin \mathcal{Z}^*$: Combined with $A \in \mathrm{MB}(X_i \mid \mathcal{V})$ (Lemma A.6, as A and X_i are adjacent), it yields $A \in \mathcal{O}_{i\setminus\mathcal{Z}^*}$. By Lemma A.7 we get $A \in \mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*})$, which means this path is blocked by $\mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*}) \setminus \{Y_j\}$.
- ***Higher-order paths.*** Every path of length ≥ 3 begins with a 2-edge prefix adjacent to X_i. We classify by the structure of this prefix:
  - Prefix $X_i \to A \leftarrow B$:
    - If $A \in \mathcal{Z}^*$: We have $\mathrm{De}(A) \subseteq \bigcup_{k=1}^{K} \mathrm{De}(Z_k)$, thus A and all its descendants belong to $\mathcal{Z}^*$ and are therefore not in $\mathcal{O}_{i\setminus\mathcal{Z}^*}$. Therefore, this path will always remain blocked given $\mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*}) \setminus \{Y_j\}$, as A and all its descendants are absent from this set.
    - If $A \notin \mathcal{Z}^*$:
      - If $B \in \mathcal{Z}^*$: We have $\mathrm{De}(B) \subseteq \bigcup_{k=1}^{K} \mathrm{De}(Z_k)$, and since $A \in \mathrm{De}(B)$, we get that $A \in \bigcup_{k=1}^{K} \mathrm{De}(Z_k)$. Therefore, this case is impossible.
      - If $B \notin \mathcal{Z}^*$: We have $A, B \notin \mathcal{Z}^*$. Combined with $A, B \in \mathrm{MB}(X_i \mid \mathcal{V})$ (Lemma A.6, as $A \in \mathrm{Ch}(X_i)$ and $B \in \mathrm{Sp}(X_i)$), it follows that $A, B \in \mathcal{O}_{i\setminus\mathcal{Z}^*}$. By Lemma A.8, this yields $B \in \mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*})$. Moreover, since the edge $B \to A$ exists, B is not a collider on this path, so the path is blocked by $\mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*}) \setminus \{Y_j\}$.

- Prefix $X_i \to A \to B$:
  - If $A \in \mathcal{Z}^*$:
    - If $B \in \mathcal{Z}^*$: Because $A \in \mathcal{Z}^*$, there must exist directed paths from Y_j to A. Since the graph $\mathcal{G}$ is acyclic, these directed paths preclude the reverse directed paths from A to Y_j, which means there must exist a collider node C on this path such that $C \in \mathrm{De}(A) \subseteq \bigcup_{k=1}^{K} \mathrm{De}(Z_k)$. Thus, C and all its descendants belong to $\mathcal{Z}^*$ and are therefore not in $\mathcal{O}_{i\setminus\mathcal{Z}^*}$. Consequently, this path will always remain blocked given $\mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*})\setminus\{Y_j\}$, as C and all its descendants are absent from this set.
    - If $B \notin \mathcal{Z}^*$: We have $\mathrm{De}(A) \subseteq \bigcup_{k=1}^{K} \mathrm{De}(Z_k)$, and since $B \in \mathrm{De}(A)$, we get that $B \in \bigcup_{k=1}^{K} \mathrm{De}(Z_k)$. Therefore, this case is impossible.
  - If $A \notin \mathcal{Z}^*$: Combined with $A \in \mathrm{MB}(X_i \mid \mathcal{V})$ (Lemma A.6, as A and X_i are adjacent), it yields $A \in \mathcal{O}_{i\setminus\mathcal{Z}^*}$. By Lemma A.7 we get $A \in \mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*})$, which means this path is blocked by $\mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*}) \setminus \{Y_j\}$.
- Prefix $X_i \leftarrow A \leftarrow B$ or $X_i \leftarrow A \to B$:
  - If $A \in \mathcal{Z}^*$: There must exist directed paths from X_i to A. Since the graph $\mathcal{G}$ is acyclic, these directed paths preclude the existence of the reverse edges $A \to X_i$ in $\mathcal{E}$. Therefore, this case is impossible.
  - If $A \notin \mathcal{Z}^*$: Combined with $A \in \mathrm{MB}(X_i \mid \mathcal{V})$ (Lemma A.6, as A and X_i are adjacent), it yields $A \in \mathcal{O}_{i\setminus\mathcal{Z}^*}$. By Lemma A.7 we get $A \in \mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*})$, which means this path is blocked by $\mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*}) \setminus \{Y_j\}$.

Since every path between X_i and Y_j is blocked given $\mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*}) \setminus \{Y_j\}$, the d-separation holds. Consequently, under Assumption 5 (Faithfulness), we have

$$X_i \perp\!\!\!\perp Y_j \mid \big(\mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*}) \setminus \{Y_j\}\big) \tag{6}$$

Meanwhile, by the definition of MB (Definition 1), we have

$$X_i \perp\!\!\!\perp \mathbf{R}^* \mid \mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*}), \tag{7}$$

where $\mathbf{R}^* = \mathcal{O}_{i\setminus\mathcal{Z}^*} \setminus (\mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*}) \cup \{X_i\})$. By the contraction axiom (Lemma A.3), combining equations (6) and (7) yields:

$$X_i \perp\!\!\!\perp (\mathbf{R}^* \cup \{Y_j\}) \mid \big(\mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*}) \setminus \{Y_j\}\big),$$

which is a contradiction to the minimality of MB (Definition 1). Thus, our assumption is false, i.e., $Y_j \notin \mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*})$.

□

Corollary A.1 (Existence of v-structure Exclusion)**.** *For any v-structure $X_i \to Z' \leftarrow Y_j$ in $\mathcal{G}$, there exists $\mathcal{Z}' \subseteq \mathrm{MB}(X_i \mid \mathcal{V}) \setminus \{Y_j\}$ such that $Y_j \notin \mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}'})$ and $Z' \in \mathcal{Z}'$, where $\mathcal{O}_{i\setminus\mathcal{Z}'} = (\mathrm{MB}(X_i \mid \mathcal{V}) \cup \{X_i\}) \setminus \mathcal{Z}'$.*

Proof. We have that X_i and Y_j form one or more v-structures in $\mathcal{G}$. Denoting all of them as $X_i \to Z_k \leftarrow Y_j$ for $k = 1, \ldots, K$, we have that $Z' \in \{Z_1, \ldots, Z_K\}$. Let $\mathcal{Z}^* = \{Z_1, \ldots, Z_K\} \cup \bigcup_{k=1}^{K} \mathrm{De}(Z_k)$, where $\mathrm{De}(Z_k)$ denotes the set of all descendants of Z_k. Then, by Lemma A.9, we immediately have $Y_j \notin \mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*})$.

Let $\mathcal{Z}' = \mathcal{Z}^* \cap \mathrm{MB}(X_i \mid \mathcal{V})$. By Lemma A.6, $Z', Y_j \in \mathrm{MB}(X_i \mid \mathcal{V})$ since $Z' \in \mathrm{Ch}(X_i)$ and $Y_j \in \mathrm{Sp}(X_i)$. Thus, we have $Z' \in \mathcal{Z}'$. Since $\mathcal{G}$ is acyclic, X_i and Y_j cannot be descendants of any Z_k. We have $X_i, Y_j \notin \bigcup_{k=1}^{K} \mathrm{De}(Z_k)$, which implies $X_i, Y_j \notin \mathcal{Z}^*$. Thus, by $\mathcal{Z}' = \mathcal{Z}^* \cap \mathrm{MB}(X_i \mid \mathcal{V})$, we get $Y_j \notin \mathcal{Z}'$ and $\mathcal{Z}' \subseteq \mathrm{MB}(X_i \mid \mathcal{V}) \setminus \{Y_j\}$.

Meanwhile, by definition,

$$\begin{aligned}
\mathcal{O}_{i\setminus\mathcal{Z}'} &= (\mathrm{MB}(X_i \mid \mathcal{V}) \cup \{X_i\}) \setminus \mathcal{Z}' \\
&= (\mathrm{MB}(X_i \mid \mathcal{V}) \cup \{X_i\}) \setminus (\mathcal{Z}^* \cap \mathrm{MB}(X_i \mid \mathcal{V})) \\
&= (\mathrm{MB}(X_i \mid \mathcal{V}) \cup \{X_i\}) \setminus \mathcal{Z}^* \\
&= \mathcal{O}_{i\setminus\mathcal{Z}^*}.
\end{aligned}$$

Therefore, it follows that $\mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}^*}) = \mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}'})$, which means $Y_j \notin \mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}'})$. Consequently, for any such v-structure $X_i \to Z' \leftarrow Y_j$ in $\mathcal{G}$, there exists $\mathcal{Z}' \subseteq \mathrm{MB}(X_i \mid \mathcal{V}) \setminus \{Y_j\}$ such that $Y_j \notin \mathrm{MB}(X_i \mid \mathcal{O}_{i\setminus\mathcal{Z}'})$ and $Z' \in \mathcal{Z}'$. □

Lemma A.10 (Soundness and Completeness of Skeleton)**.** *Algorithm 1 (DCD-C) adds the undirected edge $X_t - W$ to $\hat{\mathcal{E}}$ if and only if X_t and W are adjacent in $\mathcal{G}$.*

Proof. We prove the soundness and completeness separately given any $X_t \in \mathcal{V}$. Since we assume the MB identifier in Algorithm 1 operates as an oracle, we directly use the ground-truth set operator $\mathrm{MB}(\cdot \mid \cdot)$ throughout the proof.

According to Algorithm 1, the undirected edge $X_t - W$ is added to $\hat{\mathcal{E}}$ if and only if $W \in \mathcal{M}_t$ and there does not exist a subset $\mathcal{Z} \subseteq \mathcal{M}_t \setminus \{W\}$ such that $W \notin \mathrm{MB}(X_t \mid \mathcal{O}_{t\setminus\mathcal{Z}})$, where $\mathcal{O}_{t\setminus\mathcal{Z}} = (\mathcal{M}_t \cup \{X_t\}) \setminus \mathcal{Z}$ and $\mathcal{M}_t = \mathrm{MB}(X_t \mid \mathcal{V})$.

Soundness. To establish soundness, we show that if $W \in \mathcal{M}_t$ and there does not exist any subset $\mathcal{Z} \subseteq \mathcal{M}_t \setminus \{W\}$ such that $W \notin \mathrm{MB}(X_t \mid \mathcal{O}_{t\setminus\mathcal{Z}})$, then X_t and W must be adjacent in the true graph $\mathcal{G}$.

Equivalently, we prove the contrapositive: if $W \in \mathcal{M}_t$ but is not adjacent to X_t in $\mathcal{G}$, then there exists a subset $\mathcal{Z} \subseteq \mathcal{M}_t \setminus \{W\}$ such that $W \notin \mathrm{MB}(X_t \mid \mathcal{O}_{t\setminus\mathcal{Z}})$.

Since X_t and W are non-adjacent in $\mathcal{G}$ (i.e., $W \notin \mathrm{Pa}(X_t) \cup \mathrm{Ch}(X_t)$) and $W \in \mathrm{MB}(X_t \mid \mathcal{V}) = \mathcal{M}_t$, by Lemma A.6, it follows that $W \in \mathrm{Sp}(X_t)$. Hence, X_t and W must form one or more v-structures in $\mathcal{G}$. Therefore, by Corollary A.1, there exists a set $\mathcal{Z} \subseteq \mathcal{M}_t \setminus \{W\}$ such that $W \notin \mathrm{MB}(X_t \mid \mathcal{O}_{t\setminus\mathcal{Z}})$, thereby establishing soundness.

Completeness. To establish completeness, we show that if X_t and W are adjacent in the true graph $\mathcal{G}$, then $W \in \mathcal{M}_t$ and there does not exist any subset $\mathcal{Z} \subseteq \mathcal{M}_t \setminus \{W\}$ such that $W \notin \mathrm{MB}(X_t \mid \mathcal{O}_{t\setminus\mathcal{Z}})$.

Since X_t and W are adjacent in $\mathcal{G}$ (i.e., $W \in \mathrm{Pa}(X_t) \cup \mathrm{Ch}(X_t)$), by Lemma A.6, it follows that $W \in \mathrm{MB}(X_t \mid \mathcal{V}) = \mathcal{M}_t$. Hence, for any $\mathcal{Z} \subseteq \mathcal{M}_t \setminus \{W\}$, we have $W \in \mathcal{O}_{t\setminus\mathcal{Z}} = (\mathcal{M}_t \cup \{X_t\}) \setminus \mathcal{Z}$. By Lemma A.7, it follows that $W \in \mathrm{MB}(X_t \mid \mathcal{O}_{t\setminus\mathcal{Z}})$ for all such $\mathcal{Z}$. Consequently, there does not exist any $\mathcal{Z} \subseteq \mathcal{M}_t \setminus \{W\}$ such that $W \notin \mathrm{MB}(X_t \mid \mathcal{O}_{t\setminus\mathcal{Z}})$, thereby establishing completeness.

□

Lemma A.11 (Soundness and Completeness of v-structure)**.** *Algorithm 1 (DCD-C) orients the v-structure $X_t \to Z \leftarrow Y$ in $\hat{\mathcal{E}}$ if and only if the v-structure $X_t \to Z \leftarrow Y$ exists in $\mathcal{G}$.*

Proof. We prove the soundness and completeness separately given any $X_t \in \mathcal{V}$. Since we assume the MB identifier in Algorithm 1 operates as an oracle, we directly use the ground-truth set operator $\mathrm{MB}(\cdot \mid \cdot)$ throughout the proof.

According to Algorithm 1, the orientation of v-structures occurs after all undirected edges have been added. At this stage, by Lemma A.10, the skeleton of $\hat{\mathcal{G}} = (\mathcal{V}, \hat{\mathcal{E}})$ is identical to the true graph $\mathcal{G}$. The v-structure $X_t \to Z \leftarrow Y$ is oriented in $\hat{\mathcal{E}}$ if and only if all of the following conditions are satisfied:

- $Y \in \mathcal{M}_t$.
- There exists $\mathcal{Z} \subseteq \mathcal{M}_t \setminus \{Y\}$ such that $Y \notin \mathrm{MB}(X_t \mid \mathcal{O}_{t\setminus\mathcal{Z}})$, where $\mathcal{O}_{t\setminus\mathcal{Z}} = (\mathcal{M}_t \cup \{X_t\}) \setminus \mathcal{Z}$.
- $Z \in \mathcal{Z}$.
- Z is adjacent to both X_t and Y in $\hat{\mathcal{G}}$.
- X_t and Y are not adjacent in $\hat{\mathcal{G}}$.

Soundness. To establish soundness, we show that if all the conditions hold, there is a v-structure $X_t \to Z \leftarrow Y$ in $\mathcal{G}$.

We prove this by contradiction. Suppose there is no v-structure $X_t \to Z \leftarrow Y$ in $\mathcal{G}$. Because the skeleton of $\hat{\mathcal{G}}$ is correctly identified, Z is adjacent to both X_t and Y, while X_t and Y are not adjacent in $\hat{\mathcal{G}}$. Therefore, $\mathcal{G}$ must contain $X_t \leftarrow Z \to Y$ (or $X_t \to Z \to Y$, or $X_t \leftarrow Z \leftarrow Y$).

Since $Z \in \mathcal{Z}$, we have $Z \notin \mathcal{O}_{t\setminus\mathcal{Z}}$, and thus $Z \notin \mathrm{MB}(X_t \mid \mathcal{O}_{t\setminus\mathcal{Z}})$. Meanwhile, we already have $X_t, Y \notin \mathrm{MB}(X_t \mid \mathcal{O}_{t\setminus\mathcal{Z}})$. Therefore, the path $X_t \leftarrow Z \to Y$ (or $X_t \to Z \to Y$, or $X_t \leftarrow Z \leftarrow Y$) is not blocked by $\mathrm{MB}(X_t \mid \mathcal{O}_{t\setminus\mathcal{Z}})$, which means X_t and Y are not d-separated by $\mathrm{MB}(X_t \mid \mathcal{O}_{t\setminus\mathcal{Z}})$ (in $\mathcal{G}$). By Assumption 5 (Faithfulness), we have

$$X_t \not\perp\!\!\!\perp Y \mid \mathrm{MB}(X_t \mid \mathcal{O}_{t\setminus\mathcal{Z}}),$$

which contradicts the definition of MB (Definition 1), given that $Y \in \mathcal{O}_{t\setminus\mathcal{Z}} = (\mathcal{M}_t \cup \{X_t\}) \setminus \mathcal{Z}$. Thus, our assumption is false, i.e., $\mathcal{G}$ must contain the v-structure $X_t \to Z \leftarrow Y$, thereby establishing soundness.

Completeness. To establish completeness, we show that if there is a v-structure $X_t \to Z \leftarrow Y$ in $\mathcal{G}$, then all conditions are satisfied.

By Lemma A.6, we have $Y \in \mathrm{MB}(X_t \mid \mathcal{V}) = \mathcal{M}_t$, as $Y \in \mathrm{Sp}(X_t)$. Meanwhile, since the skeleton of $\hat{\mathcal{G}}$ is correctly identified, Z is adjacent to both X_t and Y, while X_t and Y are not adjacent in $\hat{\mathcal{G}}$.

Meanwhile, by Corollary A.1, there exists a subset $\mathcal{Z} \subseteq \mathcal{M}_t \setminus \{Y\}$ such that $Y \notin \mathrm{MB}(X_t \mid \mathcal{O}_{t\setminus\mathcal{Z}})$ and $Z \in \mathcal{Z}$.

Consequently, all conditions are satisfied, thereby establishing completeness.

□

Theorem 2 (Soundness and Completeness of DCD-C)**.** *Under Assumptions 3, 4, and 5, with an oracle MB identifier* $\mathrm{MB}(\cdot \mid \cdot)$*, Algorithm 1 (DCD-C) returns the true CPDAG of* $\mathcal{G}$*.*

Proof. By Lemmas A.10 and A.11, Algorithm 1 correctly identifies both the exact skeleton and v-structures. According to Verma & Pearl (1990), the skeleton and v-structures uniquely determine a Markov equivalence class, which is represented by a CPDAG. Based on the soundness and completeness of Meek rules (Meek, 1995), Algorithm 1 returns the true CPDAG. This completes the proof.

□

B Computational Complexity of DCD-C

We analyze the computational complexity of DCD-C (Algorithm 1) in terms of the number of MB identifications. Given $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{X_1, \ldots, X_d\}$ and $d = |\mathcal{V}|$, denote the maximum MB size in $\mathcal{G}$ by $m = \max_{X \in \mathcal{V}} |\mathrm{MB}(X \mid \mathcal{V})|$ and let $m_i = |\mathrm{MB}(X_i \mid \mathcal{V})|$ ($m_i \le m$). Assume that $\mathcal{E} \neq \emptyset$, so that $m > 0$.

As shown in Algorithm 1, DCD-C identifies the MB of each variable once, and then iterates over all non-empty proper subsets of the identified MB. Therefore, the total number of MB identifications is

$$N_{\mathrm{MB}} = d + \sum_{i:m_i>0} (2^{m_i} - 2) \le d + d(2^m - 2).$$

Hence, we have

$$N_{\mathrm{MB}} = O(d \cdot 2^m).$$

Consequently, the number of MB identifications required by DCD-C is controlled by both the number of variables and the maximum local structural complexity of the graph.

Compared with the complexity of constraint-based methods in terms of the number of CITs, DCD-C has a more favorable complexity of MB identifications. Under the standard assumptions (Spirtes et al., 2000) of causal sufficiency, the Markov condition, and faithfulness (Assumptions 3, 4 and 5),

without imposing additional structural assumptions (Mokhtarian et al., 2022), Mokhtarian et al. (2025) establish a theoretical lower bound for the number of CITs required by constraint-based methods:

$$\Omega(d^2 + d\Delta_{\text{in}}(\mathcal{G})2^{\Delta_{\text{in}}(\mathcal{G})}).$$

For example, the classical PC algorithm (Spirtes & Glymour, 1991) has a worst-case complexity that scales exponentially with the maximum degree of the graph, requiring $O(d^{\Delta(\mathcal{G})+2})$ CITs, where $\Delta(\mathcal{G})$ denotes the maximum degree (both in-degree and out-degree) of any node in $\mathcal{G}$.

Moreover, in DCD-C, the results for different nodes do not affect each other and can therefore be computed in parallel. In the ideal parallel case, the empirical runtime can be close to that required for $(2^m - 1)$ MB identifications. Meanwhile, the MB estimation procedure we use (Section 3.3) may be more efficient than some nonparametric CIT methods. As a simple optimization problem, it can be solved on a single ordinary CPU core, and empirically the optimization speed is insensitive to the sample size n. In contrast, the commonly used KCIT (Zhang et al., 2011) scales cubically with n, although some methods can approximate it in linear time (Strobl et al., 2019; Guan & Kuang, 2026).

C Limitations and Future Directions

Decoupling Weights. Theoretically, while no single CIT is uniformly effective across all dependence structures (Shah & Peters, 2020; Kim et al., 2022), satisfying Equation (1) at the population level to guarantee MB identifiability is straightforward in our decoupling-based approach.

According to Definition 3, for any $\tilde{P}_w(\mathbf{Z}) = \prod_i \tilde{P}_w(Z_i) = w(\mathbf{Z})P(\mathbf{Z})$ under Assumption 1 (strict positivity), we can explicitly construct the weight as:

$$w(\mathbf{Z}) = \frac{\prod_i \tilde{P}_w(Z_i)}{P(\mathbf{Z})}.$$

That is, for any desired collection of valid marginal distributions $\{\tilde{P}_w(Z_i)\}$, the corresponding density ratio directly defines a weighting function under which the joint distribution becomes a product of marginals. For continuous $\mathbf{Z}$, there are infinitely many admissible choices of the marginal distributions $\{\tilde{P}_w(Z_i)\}$. Each such choice induces a corresponding decoupling weight through the density ratio construction above. Therefore, the set of decoupling weights $\mathcal{W}_\perp(\mathbf{Z})$ can contain infinitely many distinct weights.

However, in practical estimation with finite samples (Section 3.3), we still face challenges:

Noise. Empirically, an extremely low noise level degrades the performance of DCD (Figure 1), likely because learning effective decoupling weights becomes more challenging under such conditions. Overly deterministic dependencies may limit the variation available in observational samples, making it difficult to decouple these relationships through sample weighting alone. Future work could improve the current framework for learning weights and systematically investigate the interplay between the learning of decoupling weights and noise levels to further explore the potential of decoupling-based causal discovery.

High dimensionality. Although DCD is highly effective on small-scale graphs, its performance degrades as the graph size increases (Table 1), a limitation shared by most existing methods. This is because causal discovery inherently requires measuring the relationships among multiple variables. Similar to how large conditioning sets render CITs difficult, increasing the number of variables makes it challenging to achieve mutual independence via sample weighting (which is why we employ pairwise HSIC as an approximation in Section 3.3). Fundamentally, the challenge of high dimensionality in causal discovery arises from the rapidly increasing sample complexity required to characterize distributional relationships as dimensionality grows. Mitigating this issue represents a crucial direction for future exploration in learning decoupling weights, as well as in the broader field of causal discovery.

Sample selection. Practically, the weighting mechanism implicitly performs sample selection. This provides an intuitive explanation for why, within the context of DAGs, non-adjacent spouse nodes maintain their dependence on the target after decoupling. However, this selection process may assign substantially lower weights to certain sample points in the support, thereby potentially weakening the signal of dependencies and increasing the risk of false negatives. Learning decoupling weights that better preserve information is a promising direction for future work.

Graph Construction via MB Identification As a novel framework, Algorithm 1 remains open to refinement and offers several directions for future development.

Large MB. As analyzed in Section B, the number of MB identifications required by Algorithm 1 is governed by the size of the MBs. When the underlying graph is dense, both error accumulation and computational costs will increase.

Graph Topology. As illustrated in Figure 2, DCD exhibits a pronounced advantage on SF graphs, whereas it slightly underperforms the constraint-based algorithm on more homogeneous ER graphs. On the one hand, the decoupling mechanism is less sensitive to variations in node degrees, maintaining robust performance across graph topologies. On the other hand, constraint-based methods can naturally benefit from homogeneous graphs due to lower-dimensional conditioning sets, leading to better performance on ER graphs than on SF graphs.

Combining Decoupling with Existing Strategies. Algorithm 1 currently relies entirely on MB identification, and integrating CITs to complement this process represents a highly promising direction. In fact, similar ideas rooted in CITs have been explored (Margaritis & Thrun, 1999; Pellet & Elisseeff, 2008). A line of research (Akbari et al., 2021; Mokhtarian et al., 2021; 2022; 2023; 2025) performs MB identification before locally identifying removable variables. A straightforward design would be to directly replace their CIT-based MB identification with our decoupling method. Interestingly, despite originating from different motivations, our strategy exhibits complementarity with the idea of removable variables. While they focus on preserving the stability of the projected MAG after variable removal, we focus on the structural disruption of the MB upon removal. Overall, decoupling offers a direct MB perspective, providing a natural basis for integration with existing causal discovery paradigms.