Unified Causal Discovery and Missing Data Imputation

Osman Mian
Institute for AI in Medicine,
University Medicine Essen
Lamarr Institute,
TU Dortmund

Jens Kleesiek
Institute for AI in Medicine,
University Medicine Essen

Michael Kamp
Lamarr Institute,
TU Dortmund
Institute for AI in Medicine,
University Medicine Essen

Abstract

Causal discovery and data imputation are often treated separately, yet both face challenges when data are missing. Existing causal discovery methods discard incomplete samples, losing valuable information, while standard imputation relies on spurious correlations that obscure the causal signal. We propose LoGIC, a framework that performs causal discovery and causally consistent imputation jointly. In contrast to prior work that assumes all source variables are observed, we derive a verifiable criterion for this assumption under MCAR and MAR missingness, grounded in the Algorithmic Markov Condition. LoGIC then proceeds layer by layer: identifying sources, recovering downstream relations, and imputing missing values, while explicitly declaring unknowns when imputation is unsupported. This design preserves causal reasoning even in challenging missingness regimes. Experiments on synthetic and real-world data show that LoGIC outperforms state-of-the-art baselines in both structure recovery and imputation accuracy.

1 INTRODUCTION

Causal discovery aims to uncover the underlying data-generating mechanisms from observational data. This task fundamentally relies on the structure of conditional dependencies between variables (Pearl, 2009). In practical domains, however, values may be missing because of measurement costs, data entry errors, or deliberate omission by participants. In medicine, for example, some patients may be asked to undergo extensive laboratory tests while others are not, and some attributes may be systematically unrecorded for subpopulations (Pezzullo, 2022). Missing data, however, disrupts the structure of conditional dependencies we rely on: When values are absent, new spurious dependencies can emerge (Tu et al., 2019). As a result, standard causal discovery algorithms, which assume fully observed data, become unreliable or even fail entirely when applied to incomplete datasets. This biases predictive modeling by training on a distribution that no longer reflects the underlying generating process, leading to overfit predictions and poor generalization.

A common remedy is to first impute the missing values and then apply causal discovery to the completed data. Classical approaches such as mean imputation (Little and Rubin, 2019) or matrix factorization (Koren et al., 2009) use only low-order statistical correlations. More advanced methods, including deep generative models (Yoon et al., 2018; Mattei and Frellsen, 2019), learn flexible conditional distributions but still treat the data as exchangeable and disregard causal directionality. While these methods can achieve low reconstruction error, they can also distort the signals that causal discovery relies on by introducing spurious dependencies (Tu et al., 2019).

Recent work has highlighted the potential of causal structure for imputation. The Miracle framework (Kyono et al., 2021) takes an important step in this direction, demonstrating that structural constraints can improve reconstruction accuracy and robustness. Miracle's reliance on acyclicity as the sole structural constraint, however, is problematic: Acyclicity is satisfied by any edge consistent with the underlying causal topological order, and does not identify causal relations on its own. This tends to produce unnecessarily dense causal structures as we observe in our evaluations. Moreover, the framework depends on

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

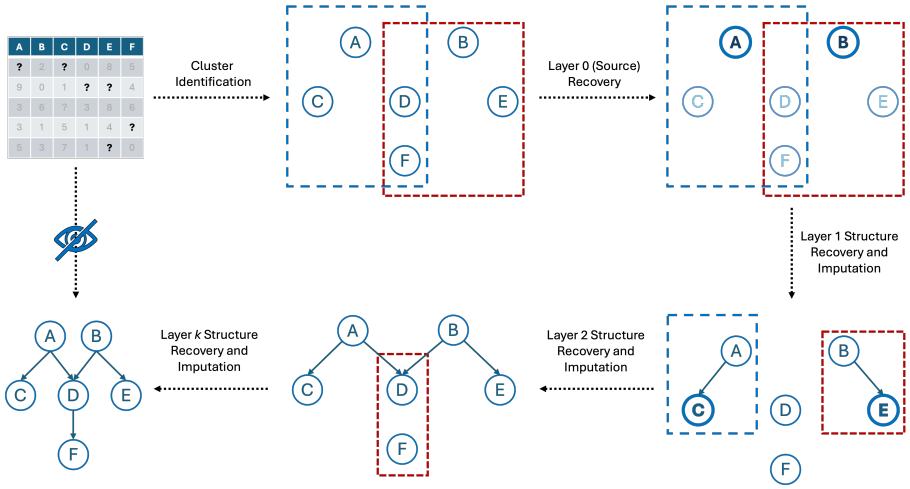

Figure 1: **Overview of LoGIC.** Given a dataset with missing values, LoGIC first discovers clusters, defined as maximal sets of variables that are pairwise dependent (Sec. 4.1). It then identifies a source variable within each cluster (Sec. 4.2) and uses the recovered source variables to validate the source-safe assumption. Building on this, LoGIC performs layer-wise causal structure discovery and leverages the inferred parent variables from preceding layers to impute missing values in subsequent layers (Sec. 5). If the parents of a variable cannot be reliably inferred due to insufficient causal information, LoGIC explicitly returns *cannot impute*.

the generally unverifiable source-safe missingness assumption, i.e., that source variables are never missing. Without theoretical guarantees, MIRACLE risks converging to behavior similar to standard non-causal imputation methods.

Treating imputation and causal discovery as separate steps is fundamentally problematic. If one first imputes without knowledge of the causal graph, the imputer may introduce dependencies that never existed or erase those that do, biasing any subsequent discovery. Conversely, if one attempts to discover structure on incomplete data, missingness can mask true independencies and produce spurious ones, leading to an incorrect graph and thus incorrect guidance for imputation. These two problems are coupled: imputations must respect causal structure, and structure learning must account for missingness. Solving them jointly allows each task to inform the other, enabling imputations that preserve the generative process and graphs that remain valid under incomplete observations.

We study the joint problem of imputing missing values and recovering causal structure from incomplete observational data. The input is a dataset with missing entries, assumed to be generated by a structural causal model with additive noise where samples are missing (completely) at random (MCAR/MAR). The output is an imputed dataset together with a causal graph consistent with data's independencies, with the ability to abstain when imputation is unsupported. This abstention occurs precisely when the causal parents required for reconstruction are themselves missing or not identifiable, making the imputation problem ill-posed under the assumed model. We therefore distinguish between causally identifiable and non-identifiable imputations, and only perform the former.

We propose LoGIC (**L**ayer **O**rdered **G**raph structure discovery and data **I**mputation via **C**ausality), which enforces the source-safe assumption, imputes values following the causal ordering of variables, and learns the causal structure hierarchically. As shown in Fig. 1, LoGIC identifies independent sources via pairwise tests, extracts source-safe rows, and incrementally constructs a layer-wise ordering of variables. Each variable is imputed from its causal parents, while noise is estimated using its most informative child, producing both an imputed dataset and a candidate causal graph. By grounding imputation in causal order, LoGIC preserves structural consistency and abstains when causally valid imputation is impossible. Specifically, our contributions are as follows:

1. We propose a theoretically sound algorithm, LoGIC, that hierarchically constructs the causal structure while imputing missing values, minimizing row deletion by ensuring at most two columns have missing values at any time.
2. We introduce a systematic approach, based on the algorithmic model of causality, to verify the source-safe assumption in incomplete datasets.
3. We demonstrate through extensive experiments

that LoGIC outperforms existing approaches on both synthetic and real-world datasets.

2 RELATED WORK

Classical causal discovery methods rely on conditional independence testing or score-based search to recover the structure of a causal graph from observational data. Structure learning approaches such as constraint-based PC algorithm (Spirtes et al., 2000), score-based methods (Chickering, 2002; Ramsey et al., 2017; Huang et al., 2018; Squires et al., 2020; Mian et al., 2023), and additive noise model-based methods (Shimizu et al., 2006; Peters et al., 2014) typically assume causal sufficiency and faithfulness and perform well with complete data. However, missing values can introduce spurious dependencies (Tu et al., 2019), limiting the reliability of standard discovery procedures. To address this, recent work has modified conditional independence tests and recovery criteria to account for missingness (Gao et al., 2022; Strobl et al., 2018; Tu et al., 2019), while graphical analyses have established theoretical conditions for identifiability under data Missing (Completely) At Random and Missing Not At Random mechanisms (Mohan et al., 2013; Shpitser et al., 2015). Although these work study the limits of learning under incomplete data, they do not directly provide imputation procedures, leaving causal discovery methods constrained under missingness.

Imputation has been studied extensively in statistics and machine learning. Classical approaches include multiple imputation by chained equations (MICE) (Van Buuren and Groothuis-Oudshoorn, 2011) and matrix factorization (Rubin, 1987). More recent methods employ machine learning models such as random forests, autoencoders, and generative models (Stekhoven and Bühlmann, 2012; Gondara and Wang, 2018; Yoon et al., 2018; Morales-Alvarez et al., 2021). While effective at reducing reconstruction error, these methods are not built to preserve causal dependencies crucial for robust inference. Causally aware imputation methods, such as Miracle (Kyono et al., 2021), introduce regularization schemes to align imputations with a learned missingness graph. However, Miracle offers no causal consistency guarantees, and cannot abstain from imputing unsupported values.

In contrast to prior work, we address causal discovery and imputation simultaneously in a unified framework. By combining layer-wise causal graph construction with verification of source-safe assumptions, our method provides both structural consistency and imputations that respect the underlying causal dependencies. This positions LoGIC as a first step toward bridging the gap between theoretically grounded discovery and practical, causally consistent imputation.

3 BACKGROUND

We consider data over joint distribution of M continuous random variables $\mathcal{X} = \{X_1, ..., X_M\}$ with their corresponding joint distribution $P(\mathcal{X})$, and assume that these variables satisfy *causal sufficiency* meaning that any causal parents for each $X_i \in \mathcal{X}$ lie within $\mathcal{X}$. We work with the assumption that the causal relationships between variables in $\mathcal{X}$ can be expressed through a Directed Acyclic Graph (DAG) $\mathcal{G}$ where an edge $X_i \to X_j$ implies that X_i is a causal parent of X_j. The set of causal parents of X_j is denoted by pa_j with size $|pa_j|$, and causal children with ch_j and size $|ch_j|$. Variable X_i is called a source if $|pa_i| = 0$ and we denote by $\mathcal{S}_\mathcal{X}$ the set of all source variables in $\mathcal{X}$. A path π_{ij} in $\mathcal{G}$ of length $|\pi_{ij}|$ is a sequence of directed edge traversals starting from X_i and ending at X_j with $\pi^\star_{ij}$ being the longest such path. We can decompose $\mathcal{G}$ into a hierarchical ordering of variables by first assigning all source variables to layer 0 and then assigning each X_j to layer λ_j as $\lambda_j = \max\{|\pi^\star_{ij}| \text{ s.t. } X_i \in \mathcal{S}_\mathcal{X}\}$. The set of all variables in layer l are denoted by $\Lambda_l = \{X_i | \lambda_i = l\}$ and the set of all variables from layer 0 up to layer l are denoted by bold face $\mathbf{\Lambda}_l$. Each $X_i \in \mathcal{G}$ is either convergent or isolated. Let $i \neq j \neq k$, the set of convergent variables is given by,

$$Con(\mathcal{X}) = \{X_i \mid \exists \{X_j, X_k\} \subseteq \mathcal{S}_\mathcal{X} \wedge |\pi^\star_{ji}|, |\pi^\star_{ki}| > 0\} \ .$$

Simply put, a convergent variable has an incoming path from multiple source variables. An isolated variable, on the other hand, has an incoming path from at most one source. We define the set of isolated variables as $Iso(\mathcal{X}) = \mathcal{X} \setminus Con(\mathcal{X})$.

Causal Discovery Given an i.i.d. sample $\mathbf{X} \in \mathbb{R}^{N \times M}$ containing N samples from the joint distribution over $\mathcal{X}$, the goal of causal discovery is to recover the underlying graph $\mathcal{G}$. Under the assumptions of causal faithfulness, the causal Markov condition (Spirtes et al., 2000), and causal sufficiency (Pearl, 2009), $\mathcal{G}$ can be identified up to its Markov equivalence class (MEC) (Glymour et al., 2019), i.e., the set of DAGs that entail the same conditional independence relations and are therefore indistinguishable from observational data. Constraint-based methods address this task using conditional independence tests ($X_i \perp\!\!\!\perp X_j \mid \mathcal{Z}$) with $\mathcal{Z} \subseteq \mathcal{X} \setminus \{X_i, X_j\}$. If $X_i \perp\!\!\!\perp X_j \mid \mathcal{Z}$ holds, we remove the edge between X_i and X_j; and retain it otherwise. Iterating over all pairs and conditioning sets yields an undirected skeleton, which is then partially oriented to recover the MEC (Spirtes et al., 2000).

Under additional assumptions such as additive noise models (Shimizu et al., 2006; Hoyer et al., 2009), one

can go beyond MEC and identify $\mathcal{G}$ exactly (Marx and Vreeken, 2019b; Mian et al., 2021; Mameche et al., 2022). A different line of work builds on the Algorithmic Markov Condition (AMC) (Janzing and Schölkopf, 2010), which grounds causal discovery in Kolmogorov complexity: for a distribution P, its complexity is the length of the shortest program outputting $P(x)$ up to precision q on input $\langle x, q\rangle$, $K(P) = \min_{p\in\{0,1\}^*}\left\{\|p\| : |U(p,x,q) - P(x)| \leq \frac{1}{q}\right\}$. AMC requires that the shortest description of P factorizes as

$$K(P(X_1,\ldots,X_m)) = \sum_{j=1}^{m} K(P(X_j \mid pa_j)) + \mathcal{O}(1),$$

so the true graph is the one minimizing Kolmogorov complexity. Since $K(\cdot)$ is uncomputable, it is often approximated via the Minimum Description Length (MDL) principle (Grunwald, 2004; Marx and Vreeken, 2021), with recent methods showing that minimizing a suitable MDL score $L(P(X_j \mid pa_j))$ can recover the true graph (Kaltenpoth and Vreeken, 2019; Mameche et al., 2023; Mian et al., 2024; Xu et al., 2025). Thus, given $\mathbf{X}$ and a Markov equivalence class, AMC-based methods search for that DAG attaining the lowest MDL score within the Markov equivalence class.

Missing Data Given $\mathbf{X}$, and a binary indicator matrix $\mathbf{R} \in \{0,1\}^{N\times M}$ with entries $\mathbf{R}_{n,m}$ and $\mathbf{X}_{n,m}$ denoting the n^{th} row and m^{th} column of $\mathbf{R}$ and $\mathbf{X}$, define

$$\tilde{\mathbf{X}} = \mathbf{X} \oplus \mathbf{R}\,,$$

where $\oplus$ is applied elementwise so that $\mathbf{X}_{n,m} \oplus \mathbf{R}_{n,m} =$ **?** if $\mathbf{R}_{n,m} = 0$, and $\mathbf{X}_{n,m}$ otherwise. Thus, applying $\mathbf{R}$ to $\mathbf{X}$ produces $\tilde{\mathbf{X}}$, in which entries with $\mathbf{R}_{n,m} = 0$ are unobserved (denoted by **?**). Learning $\mathcal{G}$ from $\tilde{\mathbf{X}}$ is challenging: a naive strategy that discards incomplete rows introduces bias in structure estimation (Gao et al., 2022) and could drastically reduce sample size, weakening the power of causal discovery algorithms (Tu et al., 2019).

Imputation methods aim to recover $\mathbf{X}$ from $\tilde{\mathbf{X}}$ (Van Buuren and Groothuis-Oudshoorn, 2011; Stekhoven and Bühlmann, 2012; Yoon et al., 2018). Each missing entry $\tilde{\mathbf{X}}_{n,m}$ is generally reconstructed by learning a function f such that $\tilde{\mathbf{X}}_{n,i} = f(\tilde{\mathbf{X}}_{n,-i})$, where $\tilde{\mathbf{X}}_{n,-i}$ denotes the n-th row of $\tilde{\mathbf{X}}$ with variable X_i removed. Thus, missing values are expressed as functions of remaining variables. These methods typically assume data either *Missing Completely At Random (MCAR)*, where missingness is independent of observed and unobserved values, or *Missing At Random (MAR)*, where missingness may depend on observed variables but not on the unobserved entry. Unbiased estimation of f can be obtained under both MCAR and MAR. In contrast, when data are *Missing Not At Random (MNAR)*, i.e., when missingness depends on the unobserved value itself, these methods are typically biased (Mohan et al., 2013; Mohan and Pearl, 2014).

Causal Discovery with Missing Data One straightforward approach to causal discovery with missing data is to first obtain a completed dataset $\hat{\mathbf{X}}$ via imputation and then apply a causal discovery algorithm. This approach is problematic for two reasons: (i) imputations may rely on spurious dependencies that do not reflect true causal relationships, leading to biased estimates, and (ii) the imputed values themselves are generated through a non-causal mechanism, thereby contaminating the causal signal in the data. Both issues undermine the consistency guarantees of standard causal discovery algorithms. A fully causal imputation, moreover, would require the generally unverifiable *source-safe missingness* assumption, which stipulates that all $X_i \in \mathcal{S}_{\mathcal{X}}$ are never missing (Kyono et al., 2021, Assm. 4).

This leads us to a cyclic problem — To impute without bias, we need to know the causal structure, but to estimate the causal structure we need a fully imputed dataset. The key question we hence need to answer is: how can we simultaneously discover the underlying causal structure from $\tilde{\mathbf{X}}$ while making causally consistent imputations in the process. This we discuss next.

4 CLUSTER DISCOVERY AND SOURCE IDENTIFICATION

To achieve simultaneous causal discovery and data imputation, we propose a framework that learns causal relationships among variables in a layerwise manner. At each stage, the discovered relationships guide the imputation mapping for the next layer. Iterating this process recovers both the global causal network and the mappings from causes to their effects, which are then used for imputation. The first step is to identify Λ_0, consisting of source variables, and then design an algorithm to move from Λ_l to Λ_{l+1}. In this section we focus on the first task.

We assume *causal sufficiency* and the *Markov condition*. Following prior work (Magliacane et al., 2018; Tu et al., 2019), we also assume *faithful observability*, which requires that any conditional independence present in observational data also holds in missing data. For the structural form, we adopt an *additive noise model (ANM)* (Hoyer et al., 2009) of the form $X_i = f_i(pa_i)+\epsilon_i$, where f_i is a nonlinear function of the parents pa_i, and ϵ_i is independent, zero-mean Gaussian noise, with $\epsilon_i \perp\!\!\!\perp X_i$ for all i and $\epsilon_i \perp\!\!\!\perp \epsilon_j$ for all

$i \neq j$. We further assume that MCAR or MAR missingness type. Unlike existing methods (Kyono et al., 2021), we do not impose the source-safe missingness assumption; instead, we show how to verify it under the algorithmic causal model.

We now formalize the assumptions and methodology for identifying source variables and validating the source-safe missingness condition. This proceeds via two steps: (i) cluster discovery and (ii) source identification, detailed below.

4.1 Cluster Discovery

In cluster discovery, we identify groups of pairwise dependent variables within $\mathcal{X}$. For all X_i, we define its independence set as $\mathbf{I}_i = \{X_j \mid X_i \perp\!\!\!\perp X_j \ \forall X_j \in \mathcal{X}\}$, and define the corresponding cluster as $\mathbf{C}_i = \mathcal{X} \setminus \mathbf{I}_i$.

We begin by performing pairwise independence tests $X_i \perp\!\!\!\perp X_j$ for all $X_i, X_j \in \mathcal{X}$. For each pair, we apply pairwise deletion to remove rows containing missing values in either variable, and then conduct the independence test. Any edge is removed if the independence constraint holds. Doing so is asymptotically consistent under MCAR (Mohan et al., 2013) whereas under MAR only spurious dependencies may arise but no new independences are introduced (Tu et al., 2019). Consequently, this procedure produces a super-skeleton of the underlying graph $\mathcal{G}$, i.e., a graph whose edge set is a superset of the edges in $\mathcal{G}$. We show that we can use this pairwise independence testing to estimate the partial variable ordering within clusters.

Lemma 1. *Let $X_i \in \mathcal{S}_\mathcal{X}$ with cluster $\mathbf{C}_i$. For any $X_j \in Iso(\mathbf{C}_i)$ and $X_k \in Con(\mathbf{C}_i)$, the following holds: under MCAR, $|\mathbf{I}_i| = |\mathbf{I}_j| > |\mathbf{I}_k|$; under MAR, $|\mathbf{I}_i| > |\mathbf{I}_j| \geq |\mathbf{I}_k|$.*

Proofs are given in the supplementary material. ntuitively, Lemma 1 shows that within a source-variable cluster, isolated variables satisfy more independence constraints than convergent variables, while source variables attain the maximum under MAR. This induces a partial ordering by decreasing independence counts, ensuring that isolated variables precede convergent ones. Given this ordering, we identify all source clusters $\mathbf{C}_s$ for $X_s \in \mathcal{S}_\mathcal{X}$ as follows:

Procedure 1 (Cluster Discovery)**.** *Let $\mathbf{I}^\star = \{|\mathbf{I}_i| : X_i \in \mathcal{X}\}$, and initialize $\mathcal{C}^\star = \emptyset$. Repeat until $\mathbf{I}^\star = \emptyset$:*

1. *Select $X_s = \arg\max_{X_i \in \mathcal{X}} \mathbf{I}^\star$.*
2. *Update $\mathcal{C}^\star \leftarrow \mathcal{C}^\star \cup \{\mathbf{C}_s\}$.*
3. *Remove all variables in $\mathbf{C}_s$ from $\mathbf{I}^\star$.*

Proc. 1 partitions the causal graph into exactly $|\mathcal{S}_\mathcal{X}|$ partially overlapping clusters, with overlaps arising from convergent variables. Let $\mathbf{C}_\gamma$ represent any cluster in $\mathcal{C}^\star$, each cluster contains one unique source variable, as stated below:

Lemma 2. *After Proc. 1, it holds that $|\mathcal{C}^\star| = |\mathcal{S}_\mathcal{X}|$, $|\mathbf{C}_\gamma \cap \mathcal{S}_\mathcal{X}| = 1$, and $\bigcup_{\mathbf{C}_\gamma \in \mathcal{C}^\star} (\mathbf{C}_\gamma \cap \mathcal{S}_\mathcal{X}) = \mathcal{S}_\mathcal{X}$.*

The discovery phase therefore produces as many clusters as source variables, each anchored by exactly one source. Intuitively, clusters obtained from Proc. 1 correspond to the set of nodes reachable by a directed path from a single source. We now describe how to identify the source variable within each cluster.

4.2 Source Identification

Given the clusters $\mathcal{C}^\star$ obtained in the previous step, our goal is to identify the source variable within each cluster. This can be done via pairwise comparisons between variables using any additive noise model (ANM) or algorithmic Markov complexity-based (AMC) causal inference method, under the following assumption.

Assumption 3. *For every directed path $\pi_{ij} = X_i \to \cdots \to X_j$ and for all $Y_\iota, Y_\kappa \in \pi_{ij}$, let $l = \min\{\lambda_\iota, \lambda_\kappa\}$ and let $\mathcal{Z}_{\iota\kappa} = \mathbf{\Lambda}_{l-1} \cap (\mathbf{C}_\iota \cup \mathbf{C}_\kappa)$, it holds that $K(\tilde{Y}_\kappa \mid \tilde{\mathcal{Z}}_{\iota\kappa}) + K(\tilde{Y}_\iota \mid \tilde{Y}_\kappa, \tilde{\mathcal{Z}}_{\iota\kappa}) < K(\tilde{Y}_\iota \mid \tilde{\mathcal{Z}}_{\iota\kappa}) + K(\tilde{Y}_\kappa \mid \tilde{Y}_\iota, \tilde{\mathcal{Z}}_{\iota\kappa})$ if and only if $\lambda_\kappa < \lambda_\iota$.*

Assumption 3 is minimal but essential. Intuitively, it ensures that missingness in $\tilde{\mathbf{X}}$ does not invert causal direction in a path: causal parents or ancestors remain more concise descriptors of their descendants than vice versa. Equivalently, it ensures that causal mechanisms within a cluster remain identifiable under missingness. A violation occurs, for instance, when a nonlinear mechanism appears linear after missingness, reducing the problem to the non-identifiable case of linear-Gaussian models (Peters et al., 2017).

We can now identify the source variable within each cluster using pairwise causal inference. Specifically, the source variable is the one minimizing joint Kolmogorov complexity with other cluster members:

Proposition 4. *Let $X_s \in \mathbf{C}_\gamma \cap \mathcal{S}_\mathcal{X}$ be the source variable in $\mathbf{C}_\gamma \in \mathcal{C}^\star$. Under Assumption 3,*

$$X_s = \arg\max_{X_j \in \mathbf{C}_\gamma} \sum_{X_k \in \mathbf{C}_\gamma \setminus X_j} \mathbb{I}[\, K(X_k \mid X_j) < K(X_j \mid X_k)\,],$$

where $\mathbb{I}[\cdot] = 1$ if the condition holds and 0 otherwise.

Proof Sketch Since causal direction is never inverted, the source variable "wins" the largest number of pairwise comparisons. Any descendant loses at least once to its causal parent, whereas the source has no parents and is therefore the unique maximizer.

Algorithm 1: LoGIC Algorithm for Unified Causal Discovery and Data Imputation

Input: Dataset $\tilde{\mathbf{X}}$, Missingness Mask $\mathbf{R}$, Independence Test $\mathcal{I}$, AMC-based Scoring Criterion L
Output: Estimated Causal Structure $\hat{\mathcal{G}}$, Partially Imputed Dataset $\tilde{\mathbf{X}}$

1 $visited \leftarrow \emptyset$
2 $\hat{\mathcal{G}} \leftarrow \emptyset$
3 $l \leftarrow 0$
4 $\mathcal{C}^\star \leftarrow$ DetermineClusters($\tilde{\mathbf{X}}$,$\mathcal{I}$)
5 $\Lambda_l \leftarrow$ IdentifySources($\tilde{\mathbf{X}}, \mathcal{C}^\star$,$L$)
6 $current\ layer \leftarrow \Lambda_l$
7 **while** $current\ layer \neq \emptyset$ **do**
8 $visited \leftarrow visited \cup current\ layer$
9 $\mathcal{C}^\star \leftarrow$ UpdateClusters($\tilde{\mathbf{X}}$,$\mathcal{I}$, $visited$)
10 $\Lambda_{l+1} = \{X_i \mid \hat{pa}_i \subseteq visited\}$
11 **foreach** X_i *in* Λ_{l+1} **do**
12 $pa_i \leftarrow L$.IdentifyParents(X_i, $visited$)
13 **foreach** X_j *in* pa_i **do**
14 $\hat{\mathcal{G}}$.Add($X_j \rightarrow X_i$)
15 $\hat{ch}_i = \arg\min_{X_j \in \mathcal{X} \setminus \mathbf{\Lambda}_{l+1}} K(X_i \mid X_j)$
16 $f \leftarrow$ LearnMapping($X_i, pa_i, \hat{ch}_i$)
17 $\tilde{\mathbf{X}}_{:,i} \leftarrow \tilde{\mathbf{X}}_{:,i} \cdot (R) + f(\tilde{\mathbf{X}}_{:,pa_i}, \tilde{\mathbf{X}}_{:,\hat{ch}_i}) \cdot (1-R)$
18 $current\ layer \leftarrow \Lambda_{l+1}$
19 $l \leftarrow l + 1$
20 **return** $\hat{\mathcal{G}}$,$\tilde{\mathbf{X}}$;

Using pairwise independence tests and any consistent AMC-based scoring criterion that upper bounds K, we recover the causal sources X_s from $\tilde{\mathbf{X}}$ and define $\mathcal{S}_\mathcal{X}$ as the set of sources identified across clusters. We then verify the source-safe missingness assumption by checking these variables for missingness, and omit imputation for their immediate descendants. Formally,

Definition 5 (**Source-Safe Row**). *A row $\tilde{\mathbf{X}}_{n,:}$ is source-safe if $\forall\ X_s \in \mathcal{S}_\mathcal{X},\ \mathbf{R}_{n,s} \neq 0$.*

This extends naturally: iterating layer-wise and treating each discovered layer with its ancestors as provisional sources enables recovery of the causal structure while imputing missing values when possible.

5 LAYER-ORDERED CAUSAL DISCOVERY AND IMPUTATION

In this section, we introduce LoGIC for **L**ayer-**O**rdered **G**raph structure discovery and data **I**mputation via **C**ausality. The method iteratively identifies causal layers and their parent variables, while simultaneously learning mappings to impute missing values using these causal parents. We provide an overview in Alg. 1 with steps detailed below.

LoGIC first identifies source variables as described in Sec. 4 (Lines 4-5). We do not impute missing values for sources, since causally consistent imputation would require stronger assumptions like invertible causal relationships and low-noise assumption (Bloebaum et al., 2018; Marx and Vreeken, 2019a). The algorithm initializes Λ_0 with the identified sources (Line 6) and then performs pairwise independence tests conditioned on sources to prune additional edges (Line 9). Because the source-safe assumption holds, the number of valid samples for conditional independence tests between X_i and X_j is the same as for unconditional tests. Let $\mathbf{C}_i^+$ denote the updated cluster of X_i after these tests, and let $\hat{pa}_i = \{X_j \mid K(X_j)+K(X_i \mid X_j) < K(X_i)+K(X_j \mid X_i),\ X_j \in \mathbf{C}_i^+\}$ be the set of potential parents of X_i. Given Λ_l, we identify Λ_{l+1} as follows.

Lemma 6. *Let Λ_l denote variables in the l-th layer of $\mathcal{G}$, and $\mathbf{\Lambda}_l = \bigcup_{\lambda=0}^{l} \Lambda_\lambda$. Under Assm. 3, $\Lambda_{l+1} = \{X_i \mid \hat{pa}_i \subseteq \mathbf{\Lambda}_l\}$.*

Given Λ_0, we construct Λ_1 by selecting variables that lose pairwise comparisons only to those in Λ_0. More generally, once Λ_{l+1} is identified, the causal parents of each $X_i \in \Lambda_{l+1}$ are obtained via variable selection from $\mathbf{\Lambda}_l$ using a consistent scoring criterion such as BIC (Schwarz, 1978) or AMC-based scores (Mian et al., 2021; Xu et al., 2025) (Line 12). This yields the true parent set $pa_i^\star \subseteq \hat{pa}_i$, which is added to the learned causal graph (Lines 13–14).

For imputation in the *i.i.d.* setting, the most relevant information for X_i comes from its causal parents (mechanism) and children (residual variation). For efficiency, we use the parents and the most informative child (Line 15), defined as

$$\hat{ch}_i = \underset{X_j \in \mathcal{X} \setminus \mathbf{\Lambda}_{l+1}}{\arg\min}\ K(X_i \mid X_j)\ ,$$

to impute missing values. We learn the mapping for X_i from pa_i and $\hat{ch}_i$ using observed rows, and apply it to rows where $X_i \in \Lambda_{l+1}$ is missing (Lines 16–17). If any parent of X_i is missing within a sample, we abstain from imputing X_i due to insufficient causal information. Formally, we define imputability using Definitions 7 and 8 as follows.

Definition 7 (**Resolvability**). *X_i is resolvable in row n if $\mathbf{R}_{n,i} \neq 0$, or X_i is imputable.*

Definition 8 (**Imputability**). *X_i is imputable in row n if all $X_j \in pa_i^\star$ in row n are resolvable.*

Simply put, resolvablility means that $\tilde{\mathbf{X}}_{i,m}$ is either observed or can be reconstructed from available information, while imputability requires that all causal parents of X_i are resolvable. Together, these definitions

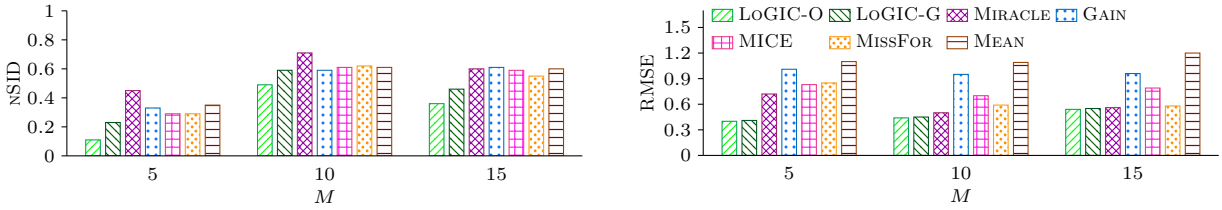

Figure 2: **[Lower is better]** nSID (left) and RMSE (right) for random graphs of sizes $M \in \{5, 10, 15\}$. LoGIC outperforms Miracle in terms of causal structure consistency while still providing accurate imputations.

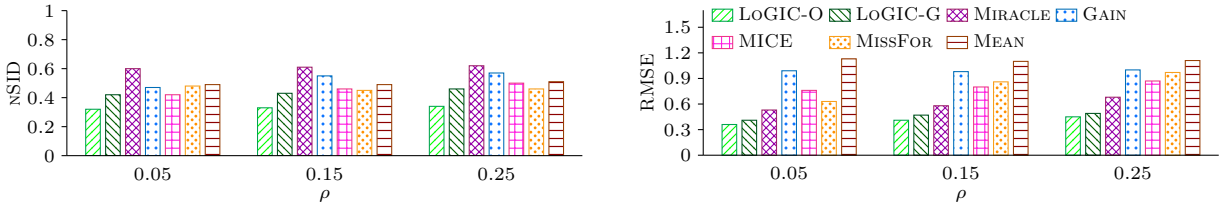

Figure 3: **[Lower is better]** nSID (left) and RMSE (right) for missingness rates $\rho \in \{0.05, 0.15, 0.25\}$. LoGIC shows consistent performance across different missingness rates.

induce a recursive criterion: X_i is imputed only if its upstream causal chain is accessible. Applied inductively, Definitions 7 and 8 prevent imputations of X_i or its descendants unless the required ancestral dependencies are observed or consistently imputed, thereby enforcing causal consistency. Iterating this procedure across all $X_i \in \mathcal{X}$ produces the estimated causal graph $\hat{\mathcal{G}}$ and a causally consistent imputed dataset, with explicit "cannot impute" declarations whenever no valid imputation is possible. LoGIC maximizes sample use by ensuring at most two variables i.e X_i and $\hat{ch}_i$, are missing at any time.

Our proposed procedure has worst-case complexity of $\mathcal{O}(M^2 + 2^{\Delta})$, where Δ is the maximum indegree in the causal structure. This compares favorably to the PC-algorithm and Greedy Equivalence Search, both of which have worst-case complexity in $\mathcal{O}(2^M)$.

We establish the following consistency guarantee for the estimated structure:

Theorem 9. *Let $\mathcal{I}$ be an independence test and L a consistent scoring criterion for data $\tilde{\mathbf{X}} = \mathbf{X} \oplus \mathbf{R}$ generated under M(C)AR missingness with $\mathbb{E}[\mathbf{R}] = \rho$. Under Assumption 3, as $n \to \infty$ and $\rho/n \to 0$,* LoGIC *with $\mathcal{I}$ and L recovers the true causal graph $\mathcal{G}$ from $\tilde{\mathbf{X}}$.*

Proof Sketch With $\mathcal{I}$, we construct a super-skeleton of the underlying graph. Since MAR reduces to MCAR for any source variable, clustering on sources removes spurious dependencies that would arise if we used other isolated variables. Restricting variable selection to members of X_i's cluster ensures we only discard false edges rather than add new ones. Consistency then follows from the guarantees of L.

6 EVALUATION

We present two variants of LoGIC and make our code available for research purposes [1]. LoGIC-O relies on an independence oracle for cluster discovery, serving as a benchmark with perfect cluster identification. LoGIC-G replaces this oracle with an MDL-based independence test adapted from the Globe score (Mian et al., 2021), an information-theoretic criterion. We then use Globe score for both discovery and imputation, since it comes with a practical advantage that the functional mappings learned during causal discovery transfer directly to imputation, providing a principled link between the two tasks.

As the only other causal-discovery-aware imputation method, we compare LoGIC to Miracle (Kyono et al., 2021). Baselines include column-wise mean imputation, Multivariate Imputation by Chained Equations (MICE) (Van Buuren and Groothuis-Oudshoorn, 2011), the tree-based MissFor (Stekhoven and Bühlmann, 2012), and the deep learning method Gain (Yoon et al., 2018). We postpone comparisons to MVPC (Tu et al., 2019) and MissDAG (Gao et al., 2022), which only perform causal discovery, to the supplementary material.

For experiments with synthetic data, we sample random Erdős–Rényi graphs with $M \in \{5, 10, 15\}$ and generate data via $X_i = f(pa_i) + \epsilon_i$, with f defined by Gaussian processes. All data are standardized to zero mean and unit variance. We experiment with missingness rates $\rho \in \{0.05, 0.15, 0.25\}$. While LoGIC is

[1] `https://github.com/osman-mian/LoGIC`

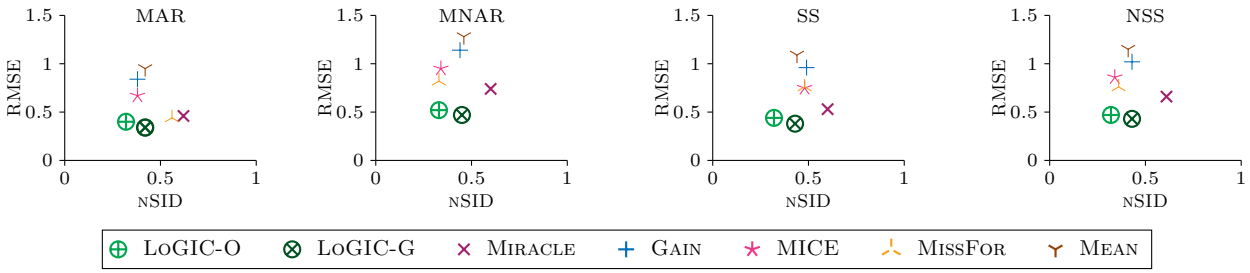

Figure 4: **[Closer to origin is better]** nSID vs. RMSE tradeoff for data missing at random (MAR), data missing not at random (MNAR), source-safe missingness (SS), and full missingness (NSS) averaged over missingness rates $\rho \in \{0.05, 0.15, 0.25\}$ and graph sizes $M \in \{5, 10, 15\}$. LoGIC gives the best trade-off between prediction accuracy and causal consistency.

		LoGIC-O	LoGIC-G
M	5	0.98	0.66
	10	0.87	0.65
	15	0.83	0.64
ρ	0.05	0.86	0.60
	0.15	0.87	0.57
	0.25	0.86	0.53

Table 1: **[Higher is Better]** F1-score for source identification for graphs of sizes $M \in \{5, 10, 15\}$ and missingness probabilities $\rho \in \{0.05, 0.15, 0.25\}$.

valid under MCAR and MAR, we also test on MNAR.

We use the $F1$ score to assess LoGIC's source recovery from missing data. As only LoGIC and Miracle learn a causal DAG, we estimate a post-hoc graph (up to Markov equivalence) on other baselines outputs using the PC algorithm (Spirtes et al., 2000) with the Hilbert–Schmidt Independence Criterion (Gretton et al., 2005). This allows us to evaluate whether imputation-only methods preserve causal structure. We measure imputation accuracy using Root Mean Squared Error (RMSE) and causal correctness using Structural Intervention Distance (SID) (Peters and Bühlmann, 2014), normalized to $[0, 1]$ as nSID. Since LoGIC may abstain from imputing when parent information is insufficient, RMSE is reported only on imputed entries. We further provide the Structural Hamming Distance (SHD) metric (Kalisch and Bühlman, 2007) in the supplementary material.

On synthetic data, we evaluate LoGIC's source recovery performance across varying M and ρ, and robustness under violations of source-safe as well as MAR assumptions. Due to space constraints, an ablation study on LoGIC's performance with and without the most informative child ($\hat{ch}_i$) is provided in the supplementary material. We conclude the evaluation with experiments on Lung-Cancer gene-expression data.

Source Identification. To validate our claims from Sec. 4, we assess how well LoGIC recovers true causal sources (Table 1). LoGIC-O achieves high precision and recall, reflected in a strong F1 score. LoGIC-G shows a slight drop due to finite-sample artifacts from the independence test, which can introduce errors in cluster identification. Overall, we see that both variants show stable performance for varying M and ρ.

Varying Graph Sizes. Fig. 2 shows that LoGIC consistently outperforms competing methods in both causal structure consistency and RMSE. Miracle frequently produces cyclic graphs, indicating difficulty capturing true causal structure, with its imputation accuracy likely driven by overfitting. Oracle-based LoGIC-O slightly outperforms LoGIC-G, yet the latter remains competitive, highlighting the robustness of the MDL-based independence criterion.

Different Missing Rates Fig. 3 shows that LoGIC and Miracle have a fairly consistent performance across different missing rates with LoGIC being visibly better in terms of underlying causal structure discovery. We observed that LoGIC is able to complete up to 60% of the missing data even at the highest possible missingness setting while keeping RMSE low by avoiding imputation in cases where a causally consistent imputation is not possible. We, hence, observe that the performance gap over RMSE improves in LoGIC's favor with increasing missingness rate.

Non-Source-Safe Missingness. We compare the "Source Safe" (SS) setting, where source variables are fully observed, to 'Not Source Safe" (NSS). As shown in Fig. 4, LoGIC achieves the best balance between causal consistency and imputation accuracy in SS, a key assumption underlying Miracle. While Miracle achieves similar RMSE, it exhibits causal inconsistencies due to incorrect graph recovery. In NSS, competing methods degrade in RMSE, as imputations along anti-causal directions produce inferior results.

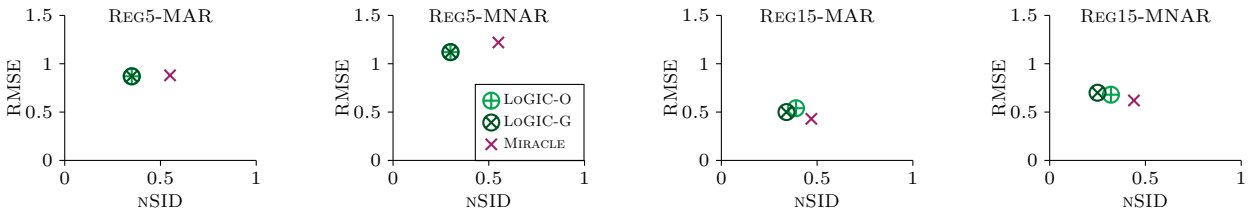

Figure 5: **[Closer to origin is better]** NSID vs. RMSE tradeoff for MAR and MNAR missingness settings for the REGED5 and REGED15 lung cancer gene-expression data averaged over missingness rates $\rho \in \{0.05, 0.15, 0.25\}$. LOGIC-O and LOGIC-G have identical performance for REGED5.

MAR vs MNAR We consider the case where data missingness depends on the missingness value itself. To do so, we assign missingness probability to each value based on its absolute magnitude and then sample the missingness mask. This implies that high magnitude values are more likely to be missing in data. We show the results in Fig.4 where we see that going from MAR to MNAR, all methods including LOGIC see a degradation in performance. Nonetheless, we see that overall LOGIC performs better than competitor approaches by a visible margin.

Real world Data We evaluate LOGIC on the lung cancer gene-expression dataset (REGED) (Statnikov et al., 2015), with 1000 variables and 20,000 samples. Using the first 5,000 samples, we extract two disjoint sub-networks of sizes 5 and 15 and evaluate under MAR and MNAR. As shown in Fig. 5, for REGED5, both LOGIC-O and LOGIC-G achieve identical performance and outperform MIRACLE. For REGED15 under MAR, MIRACLE attains lower RMSE but worse NSID, indicating that improved imputation likely arises from non-causal overfitting

7 DISUCSSION AND CONCLUSION

We propose a unified framework for causal discovery and data imputation. Building on the algorithmic model of causation (Janzing and Schölkopf, 2010), LOGIC enforces the source-safe assumption and constructs the causal graph layer-wise while imputing missing values. Experiments show that LOGIC outperforms existing approaches and yields a causal discovery method applicable to both fully observed and incomplete data, marking a promising step toward joint causal inference and data completion.

The current implementation uses MDL-based independence tests, though the framework is agnostic to this choice. Alternatives such as kernel-based measures (e.g., HSIC (Gretton et al., 2005)) or likelihood-ratio criteria may offer greater sensitivity in high-dimensional settings, provided consistency between the scoring criterion and independence test is maintained. Identifying combinations that balance statistical power and computational efficiency remains an open challenge.

Currently, LOGIC does not impute source variables due to the lack of theoretical guarantees. Recent work (Kaltenpoth and Vreeken, 2023) proposes detecting selection bias under missing data in exponential family distributions. Given the assumption of independent Gaussian noise, this idea could be integrated to design an EM procedure that detects selection bias in source variables and imputes to maximize the likelihood of downstream effects, potentially enabling principled source imputation under MNAR settings.

LOGIC may abstain from imputing when causally consistent imputation is impossible. While this can yield incomplete reconstructions, it preserves causal validity, a key safeguard in domains like medicine. For applications requiring complete datasets, LOGIC outputs can be combined with methods such as MIRACLE, at the cost of causal guarantees.

We instantiate LOGIC under the additive noise model (ANM) assumption (Hoyer et al., 2009) because it aligns with the GLOBE score of Mian et al. (2021) and supports a compatible independence test for cluster discovery. This, however, is only a design choice and the framework itself remains general: it extends to other model classes, such as post-nonlinear models (Zhang and Hyvarinen, 2012), as long as the independence test and causal scoring function share the same structural assumptions. This design allows for straightforward adaptation of the procedure by pairing each model class with its corresponding test and score, and using it for discovery and imputation.

While our focus has been on causal recovery and imputation, many applications demand robust predictions or representations for downstream tasks (Keyl et al., 2025). Extending LOGIC to jointly optimize causal consistency and predictive performance could offer a principled alternative to purely predictive imputers, reducing shortcut learning under distribution shifts.

Acknowledgments

This work is supported by the German Federal Ministry of Research, Technology and Space (DECIPHER-M, 01KD2420C) and the Cancer Research Center Cologne Essen (CCCE).

References

Bloebaum, P., Janzing, D., Washio, T., Shimizu, S., and Schoelkopf, B. (2018). Cause-effect inference by comparing regression errors. In *Proceedings of the Twenty-First AISTATS*, PMLR.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554.

Gao, E., Ng, I., Gong, M., Shen, L., Huang, W., Liu, T., Zhang, K., and Bondell, H. (2022). Missdag: Causal discovery in the presence of missing data with continuous additive noise models. *Advances in Neural Information Processing Systems*, 35:5024–5038.

Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524.

Gondara, L. and Wang, K. (2018). Mida: Multiple imputation using denoising autoencoders. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 260–272. Springer.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

Grunwald, P. (2004). A tutorial introduction to the minimum description length principle.

Grünwald, P. (2007). *The Minimum Description Length Principle*. MIT Press.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *NeurIPS*, volume 21. Curran.

Huang, B., Zhang, K., Lin, Y., Schölkopf, B., and Glymour, C. (2018). Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1551–1560.

Janzing, D. and Schölkopf, B. (2010). Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56:5168–5194.

Kalisch, M. and Bühlman, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *JMLR*, 8(3).

Kaltenpoth, D. and Vreeken, J. (2019). We are not your real parents: Telling causal from confounded using MDL. In *Proceedings of the 2019 SIAM International Conference on Data Mining, SDM 2019, Calgary, Alberta, Canada, May 2-4, 2019*, pages 199–207. SIAM.

Kaltenpoth, D. and Vreeken, J. (2023). Identifying selection bias from observational data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8177–8185.

Keyl, J., Keyl, P., Montavon, G., Hosch, R., Brehmer, A., Mochmann, L., Jurmeister, P., Dernbach, G., Kim, M., Koitka, S., et al. (2025). Decoding pan-cancer treatment outcomes using multimodal real-world data and explainable artificial intelligence. *Nature Cancer*, 6(2):307–322.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Kyono, T., Zhang, Y., Bellot, A., and van der Schaar, M. (2021). Miracle: Causally-aware imputation via learning missing data mechanisms. *Advances in Neural Information Processing Systems*, 34:23806–23817.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. (2018). Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions. In *Advances in Neural Processing Systems*.

Mameche, S., Kaltenpoth, D., and Vreeken, J. (2022). Discovering invariant and changing mechanisms from data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 1242–1252, New York, NY, USA. Association for Computing Machinery.

Mameche, S., Kaltenpoth, D., and Vreeken, J. (2023). Learning causal models under independent changes. In *Advances in Neural Information Processing Systems*, volume 36, pages 75595–75622. Curran Associates, Inc.

Marx, A. and Vreeken, J. (2019a). Identifiability of cause and effect using regularized regression. ACM.

Marx, A. and Vreeken, J. (2019b). Telling cause from effect by local and global regression. *Knowledge and Information Systems*, 60(3):1277–1305.

Marx, A. and Vreeken, J. (2021). Formally justifying MDL-based inference of cause and effect.

Mattei, P.-A. and Frellsen, J. (2019). Miwae: Deep generative modelling and imputation of incomplete

data sets. In *International conference on machine learning*, pages 4413–4423. PMLR.

Mian, O., Kamp, M., and Vreeken, J. (2023). Information-theoretic causal discovery and intervention detection over multiple environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9171–9179.

Mian, O., Mameche, S., and Vreeken, J. (2024). Learning causal networks from episodic data. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2224–2235.

Mian, O., Marx, A., and Vreeken, J. (2021). Discovering fully oriented causal networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Mohan, K. and Pearl, J. (2014). On the testability of models with missing data. In *Artificial Intelligence and Statistics*, pages 643–650. PMLR.

Mohan, K., Pearl, J., and Tian, J. (2013). Graphical models for inference with missing data. *Advances in neural information processing systems*, 26.

Morales-Alvarez, P., Lamb, A., Woodhead, S., Jones, S. P., Allamanis, M., and Zhang, C. (2021). Vicause: Simul-taneous missing value imputation and causal discovery. *Microsoft Res., Redmond, WA, USA, Tech. Rep. MSR-TR-2021-14*.

Pearl, J. (2009). *Causality*. Cambridge university press.

Peters, J. and Bühlmann, P. (2014). Structural intervention distance (sid) for evaluating causal graphs.

Peters, J., Janzing, D., and Schlkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.*, 15(1):2009–2053.

Pezzullo, A. (2022). Sources of bias in covid-19 infection fatality rate estimation. *European Journal of Public Health*, 32(Supplement_3):ckac129–128.

Ramsey, J., Glymour, M., Sanchez-Romero, R., and Glymour, C. (2017). A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics*, 3(2):121–129.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Shimizu, S., Hoyer, P. O., Hyvärinen, A. A., and Kerminen, A. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030.

Shpitser, I., Mohan, K., and Pearl, J. (2015). Missing data as a causal and probabilistic problem. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 802–811.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, prediction, and search*. MIT Press.

Squires, C., Wang, Y., and Uhler, C. (2020). Permutation-based causal structure learning with unknown intervention targets. pages 1039–1048. PMLR.

Statnikov, A., Ma, S., Henaff, M., Lytkin, N., Efstathiadis, E., Peskin, E. R., and Aliferis, C. F. (2015). Ultra-scalable and efficient methods for hybrid observational and experimental local causal pathway discovery. *Journal of Machine Learning Research*, 16:3219–3267.

Stekhoven, D. J. and Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Strobl, E. V., Visweswaran, S., and Spirtes, P. L. (2018). Fast causal inference with non-random missingness by test-wise deletion. *International journal of data science and analytics*, 6(1):47–62.

Tu, R., Zhang, C., Ackermann, P., Mohan, K., Kjellström, H., and Zhang, K. (2019). Causal discovery in the presence of missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1762–1770. Pmlr.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67.

Xu, S., Mameche, S., and Vreeken, J. (2025). Information-theoretic causal discovery in topological order. In *The 28th International Conference on Artificial Intelligence and Statistics*.

Yoon, J., Jordon, J., and Schaar, M. (2018). Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pages 5689–5698. PMLR.

Zhang, K. and Hyvarinen, A. (2012). On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Proofs

Lemma 1. *Let $X_i \in \mathcal{S}_{\mathcal{X}}$ with cluster $\mathbf{C}_i$. For any $X_j \in Iso(\mathbf{C}_i)$ and $X_k \in Con(\mathbf{C}_i)$, the following holds: under MCAR, $|\mathbf{I}_i| = |\mathbf{I}_j| > |\mathbf{I}_k|$; under MAR, $|\mathbf{I}_i| \geq |\mathbf{I}_j|$ and $|\mathbf{I}_i| > |\mathbf{I}_k|$.*

Proof. Let $X_j \in Iso(\mathbf{C}_i)$, and $X_k \in Con(\mathbf{C}_i)$. Let $\{\mathbf{C}_r, \mathbf{C}_s\}$ be such that $X_k \in \mathbf{C}_r$ and $X_k \in \mathbf{C}_s$.

MCAR Case Note that for MCAR, in the large sample limit, there can be no spurious dependence or independence relationships in $\tilde{\mathbf{X}}$ that are not present in $\mathbf{X}$. For first part of result i.e. $|\mathbf{I}_i| = |\mathbf{I}_j|$, assume that there exists X_j such that $|\mathbf{I}_i| \neq |\mathbf{I}_j|$. This implies that there exists $X \in \mathcal{X}$ such that $X_i \perp\!\!\!\perp X$ and $X_j \not\perp\!\!\!\perp X$, with $X_i \not\perp\!\!\!\perp X_j$ by defintion. The only structure that satisfies all three constraints is $X \to ... \to Y_j \leftarrow ... \leftarrow S_i$. This implies that X and X_i are in different cluster, which implies $X_j \in Con(\mathbf{C}_i)$, which is a contradiction.

The second part $|\mathbf{I}_j| > |\mathbf{I}_k|$ follows a similar argument. Assume that $|\mathbf{I}_j| \leq |\mathbf{I}_k|$ Then there must exists X such that $X \not\perp\!\!\!\perp X_j$, and $X \perp\!\!\!\perp X_k$. Since $X_j \not\perp\!\!\!\perp X_k$ is implied from X_j and X_k being in the same cluster, the only structure that satisfies all three constraints is $X \to ... \to X_j \leftarrow ... \leftarrow X_k$. This means that X and X_k are not part of the same cluster, which implies $X_j \in Con(\mathbf{C}_i)$, which is a contradiction.

MAR Case For the case of MAR, the independence test $\mathcal{I}$ could result in spurious dependences among variables Tu et al. (2019). As MAR assumption reduces down to MCAR for $X_i \in \mathcal{S}_{\mathcal{X}}$, MAR only affects independences of non-source variables. The proof then follows directly from the proof of MCAR setting. □

Lemma 2. *After Proc. 1, it holds that $|\mathcal{C}^\star| = |\mathcal{S}_{\mathcal{X}}|$, $|\mathbf{C}_\gamma \cap \mathcal{S}_{\mathcal{X}}| = 1$, and $\bigcup_{\mathbf{C}_\gamma \in \mathcal{C}^\star}(\mathbf{C}_\gamma \cap \mathcal{S}_{\mathcal{X}}) = \mathcal{S}_{\mathcal{X}}$.*

Proof. If follows directly via Proc. 1 that $\mathcal{X} = \bigcup_{\mathbf{C}_\gamma \in \mathcal{C}^\star} \mathbf{C}_\gamma$. Assume that $|\mathcal{C}^\star| < |\mathcal{S}_{\mathcal{X}}|$, then based on pigeonhole principle, there exists a cluster $\mathbf{C}_\gamma \in \mathcal{C}^\star$ such that $|\mathbf{C}_\gamma \cap \mathcal{S}_{\mathcal{X}}| > 1$. Let $\{X_r, X_s\}$ be two such sources in $\mathbf{C}_\gamma$ without loss of generality. This implies $X_r \not\perp\!\!\!\perp X_s$, which is a contradiction (ref. Lemma. 1).

Alternatively, assume $|\mathcal{C}^\star| > |\mathcal{S}_{\mathcal{X}}|$. Then there must exist a cluster $\mathbf{C}_\gamma \in \mathcal{C}^\star$ such that $\mathbf{C}_\gamma \cap \mathcal{S}_{\mathcal{X}} = \emptyset$. This implies that $\forall Y_j \in \mathbf{C}_\gamma$, and $\forall X_s \in \mathcal{S}_{\mathcal{X}}$: $Y_j \perp\!\!\!\perp X_s$. This means no $Y_j \in \mathbf{C}_\gamma$ is reachable by any source $X_s \in \mathcal{S}_{\mathcal{X}}$. Then by definition of a DAG, there must be atleast one variable $X \in \mathbf{C}_\gamma$ with no parents, implying that $X \in \mathcal{S}_{\mathcal{X}}$, which is a contradiction. □

Assumption 3. *For every directed path $\pi_{ij} = X_i \to \cdots \to X_j$ and for all $Y_\iota, Y_\kappa \in \pi_{ij}$, let $l = \min\{\lambda_\iota, \lambda_\kappa\}$ and let $\mathcal{Z}_{\iota\kappa} = \mathbf{\Lambda}_{l-1} \cap (\mathbf{C}_\iota \cup \mathbf{C}_\kappa)$, it holds that $K(\tilde{Y}_\kappa \mid \tilde{\mathcal{Z}}_{\iota\kappa}) + K(\tilde{Y}_\iota \mid \tilde{Y}_\kappa, \tilde{\mathcal{Z}}_{\iota\kappa}) < K(\tilde{Y}_\iota \mid \tilde{\mathcal{Z}}_{\iota\kappa}) + K(\tilde{Y}_\kappa \mid \tilde{Y}_\iota, \tilde{\mathcal{Z}}_{\iota\kappa})$ if and only if $\lambda_\kappa < \lambda_\iota$.*

Proposition 4. *Let $X_s \in \mathbf{C}_\gamma \cap \mathcal{S}_{\mathcal{X}}$ be the source variable in $\mathbf{C}_\gamma \in \mathcal{C}^\star$. Under Assumption 3,*

$$X_s = \underset{X_j \in \mathbf{C}_\gamma}{\arg\max} \sum_{X_k \in \mathbf{C}_\gamma \setminus X_j} \mathbb{I}[\, K(X_k \mid X_j) < K(X_j \mid X_k)\,],$$

where $\mathbb{I}[\cdot] = 1$ if the condition holds and 0 otherwise.

Proof. Let $|\mathbb{I}(X_i, \mathbf{C}_\gamma)| = \sum_{X_k \in \mathbf{C}_\gamma \setminus X_i} \mathbb{I}[K(X_k \mid X_j) < K(X_j \mid X_k)]$. Note that there can be no X_i such that $|\mathbb{I}(X_i, \mathbf{C}_\gamma)| > |\mathbb{I}(X_s, \mathbf{C}_\gamma)|$ as that would imply a direct violation of Assm. 3 by requiring that for atleast one non-source X_j, $K(X_s|X_j) < K(X_j|X_s)$. At best, assume there exists X_i with $\lambda_s < \lambda_i$ and $|\mathbb{I}(X_i, \mathbf{C}_\gamma)| = |\mathbb{I}(X_s, \mathbf{C}_\gamma)|$. From Assm. 3 it directly follows that $K(X_i|X_s) < K(X_s|X_i)$, meaning that there must exist atleast one X_j with $K(X_j|X_i) < K(X_i|X_j)$ and $K(X_j|X_s) > K(X_s|X_j)$, implying that $\lambda_j \leq \lambda_s$. $\lambda_j = \lambda_s$ implies that there two sources in a given cluster resulting in a contradiction (ref Lemm. 2). $\lambda_j < \lambda_s$ implies that $X_s \notin \mathcal{S}_{\mathcal{X}}$, which is again a contradiction. Hence X_s is the unique maximizer of $|\mathbb{I}(\cdot, \mathbf{C}_\gamma)|$ □

Definition 5 (**Source-Safe Row**). *A given row $\tilde{\mathbf{X}}_{n,:}$ is source-safe if $\forall\ X_s \in \mathcal{S}_{\mathcal{X}},\ \mathbf{R}_{n,s} \neq 0$.*

Lemma 6. *Let Λ_l denote variables in the l-th layer of $\mathcal{G}$, and $\mathbf{\Lambda}_l = \bigcup_{\lambda=0}^{l} \Lambda_\lambda$. Under Assumption 3, $\Lambda_{l+1} = \{X_i \mid \hat{pa}_i \subseteq \mathbf{\Lambda}_l\}$.*

Proof. Assume that $\exists X_i \in \Lambda_{l+1}$ s.t. $\hat{pa}_i \not\subseteq \mathbf{\Lambda}_l$. This implies that there must be atleast one X_j such that $l < \lambda_j < \lambda_i$, meaning $l < \lambda_j < l + 1$. For $l, \lambda_j \in \mathbb{Z}^+$, there is no λ_j that satisfies the latter inequality. □

Definition 7 (**Resolvability**). *X_i is resolvable in row n if $\mathbf{R}_{n,i} \neq 0$, or X_i is imputable.*

Definition 8 (**Imputability**). *X_i is imputable in row n if all $X_j \in pa_i^\star$ in row n are resolvable.*

Theorem 9. *Let $\mathcal{I}$ be an independence test and L a consistent scoring criterion for data $\tilde{\mathbf{X}} = \mathbf{X} \oplus \mathbf{R}$ generated under M(C)AR missingness with $\mathbb{E}[\mathbf{R}] = \rho$. Under Assumption 3, as $n \to \infty$ and $\rho/n \to 0$,* LoGIC *with $\mathcal{I}$ and L recovers the true causal graph $\mathcal{G}$ from $\tilde{\mathbf{X}}$.*

Proof. Following Lemma 2 and Prop. 4, layer Λ_0 consisting only of causal sources is correctly identified, while Lemm. 6 shows that layer Λ_{l+1} is recoverable given Λ_l. As $n \to \infty$ and $\rho/n \to 0$, we are guaranteed an unbiased effect estimate in MCAR and MAR settings Mohan et al. (2013). It then remains to show that given variables from Λ_0 to Λ_l i.e. $\mathbf{\Lambda}_l$, we will correctly identify the causal parents for each $X_i \in \Lambda_{l+1}$. This causal variable selection consistency is directly

enforced by using any consistent scoring criterion L such as BIC Schwarz (1978) or MDL-based scores such as GLOBE Mian et al. (2021). By applying the logic recursively, we arrive at the true graph $\mathcal{G}$. □

B Details and Additional Experiments

B.1 Experimental Setup

We implement LOGIC as cpu-parallellized approach in Python and make the code available. We present two variants of LOGIC. LOGIC-O relies on an independence oracle for cluster discovery, serving as a benchmark with perfect cluster identification. LOGIC-G replaces this oracle with an MDL-based independence test adapted from the GLOBE score Mian et al. (2021), an information-theoretic criterion.

Globe-based Conditional Independence Test. The GLOBE score, denoted by $L(\mathcal{D}, G)$, is a consistent AMC-based scoring criterion instantiated via the Minimum Description Length (MDL) principle. It was originally designed to identify the highest-scoring causal structure G for a given dataset $\mathcal{D}$, but it can also be adapted to emulate a conditional independence test. To assess conditional independence between two variables X_i and X_j given a conditioning set $\mathcal{Z}_{ij}$, we first compute the score gains δ_{ij} and δ_{ji} for the edge directions $X_i \to X_j$ and $X_j \to X_i$, respectively, following Mian et al. (2021) Mian et al. (2021):

$$\delta_{ij} = L(\mathcal{D}, G) - L(\mathcal{D}, G \cup \{X_i \to X_j\}) \, .$$

Intuitively, δ_{ij} measures the additional information that can be compressed in X_j when the edge $X_i \to X_j$ is included, compared to when it is excluded from the graph G. For a pairwise independence test, G is taken as an empty graph. For conditional independence tests, G includes edges $Z_k \to X_i$ and $Z_k \to X_j$ for all $Z_k \in \mathcal{Z}_{ij}$.

Next, we determine which of the two directions yields a larger gain:

$$\psi_{ij} = \max(\delta_{ij}, \delta_{ji}) \, .$$

Finally, we assess whether this gain is statistically significant using the no-hypercompression inequality. Let $s = \psi_{ij}$. According to the inequality Grünwald (2007), the probability of achieving a gain of s bits or more under the null model is at most 2^{-s}. If the observed gain ψ_{ij} is not significant—i.e., $2^{-s} > \alpha$, where α is a user-defined significance level—we conclude that the variables X_i and X_j are conditionally independent given $\mathcal{Z}_{ij}$.

Edge Thresholding for Miracle. MIRACLE is a continuous optimization-based framework for data imputation and causal discovery. To enforce acyclicity, it employs a differentiable constraint over the adjacency matrix of the form

$$h(W) = \operatorname{tr}\big(\mathrm{e}^{W \circ W}\big) - d \, ,$$

where W is a binary adjacency matrix representing the causal structure, with $W_{ij} = 1$ indicating the presence of an edge $X_i \to X_j$ in the underlying graph $\mathcal{G}$. The function $h(W)$ equals zero if and only if $\mathcal{G}$ is acyclic, and is strictly positive otherwise. The authors optimize their proposed likelihood objective jointly with this acyclicity regularization to recover the causal graph.

Due to the continuous nature of the optimization, the learned adjacency matrix W is rarely binary and must be thresholded using a user-defined cutoff. In MIRACLE, we observe a similar behavior: the learned W almost never contains exact zeros, implying that imputations performed by MIRACLE may rely on spurious dependencies not belonging to a true variable cluster. In our experiments, we apply a threshold of 0.1 to remove weak edges, yet still frequently observe pairwise cycles, i.e., both $X_i \to X_j$ and $X_j \to X_i$ with comparable weights. Consequently, selecting a single direction with high confidence remains non-trivial.

B.2 Additional Results

In this section we report our results for three additional experimental setting. For the first case we compare to Missing-value PC Tu et al. (2019) algorithm, which is a constraint-based causal discovery algorithm designed to work with MCAR, MAR and MNAR data. Next, we perform an ablation study to see how our practical choice of including the most informative child in data imputation affects the overall performance. Third, we report the metric structural hamming distance Kalisch and Bühlman (2007) over the learned causal structures to assess the accuracy of discovered causal edges. In the following we elaborate on the results of each of these experiments.

Comparison to Missing-Value PC (MvPC)

MvPC extends the PC causal discovery algorithm Spirtes et al. (2000) to handle missing data not only under MCAR (missing completely at random), but also under MAR and MNAR mechanisms. It first applies a test-wise deletion variant of PC (TD-PC) to estimate an initial skeleton, then identifies and corrects edges that are likely spurious due to missingness via either a permutation-based correction (PermC) or a density-ratio weighting correction (DRW). After

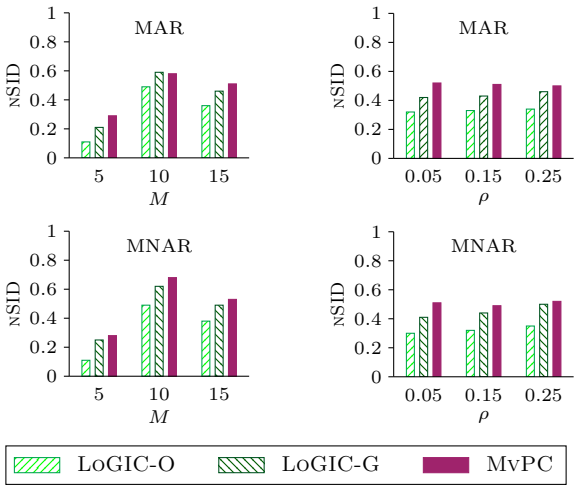

Figure B.1: **[Lower is better]** nSID over random graphs of sizes $M \in \{5, 10, 15\}$ (left) and varying missing rates $\rho \in \{0.05, 0.15, 0.25\}$ (right) for data generated under MAR (top) and data generated under MNAR (bottom).

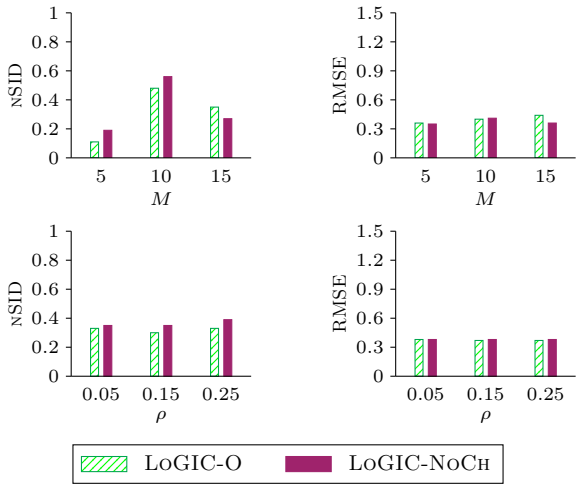

Figure B.2: **[Lower is better]** nSID (left) and RMSE (right) over random graphs of sizes $M \in \{5, 10, 15\}$ (top) and varying missing rates $\rho \in \{0.05, 0.15, 0.25\}$ (bottom) compared for LoGIC-O and its variant LoGIC-NoCh which does not use the most informative child for imputation.

these corrections, edge orientations are recovered as in standard PC, yielding a graph identifiable up to the Markov equivalence class. Although capable of causal discovery under missingness, MvPC does not perform data imputation, and thus we compare it to LoGIC solely on the quality of the inferred causal structure.

The results, shown in Fig. B.1, indicate that LoGIC recovers causal networks with lower nSID values than MvPC across both MAR and MNAR settings. We attribute this performance gap to MvPC's reliance on the assumption of linear data-generating processes, which biases its estimations and degrades accuracy. Nevertheless, we see that relative to Miracle, MvPC demonstrates superior structure recovery performance.

LoGIC without $\hat{ch}_i$

In LoGIC, we additionally incorporate the most informative child $\hat{ch}_i$ of each variable to enhance imputation quality. While conditioning on the parent set pa_i of X_i is theoretically sufficient to identify the underlying causal mechanism Janzing and Schölkopf (2010), including $\hat{ch}_i$ aids in estimating the sample-specific noise term, thereby yielding more accurate imputations. This component is not required for the consistency guarantees of our method. To assess its practical impact, we introduce a variant without the informative child, denoted as LoGIC-NoCh, and compare it with LoGIC, as shown in Fig. B.2.

Across nearly all cases, the overall RMSE remains comparable between the two variants. However, LoGIC-O consistently discovers causal structures with lower nSID values. This suggests that, without $\hat{ch}_i$, LoGIC may rely on spurious dependencies when predicting effects, thereby degrading structural accuracy in its quest to achieve the same prediction error. Incorporating $\hat{ch}_i$ thus provides a tangible advantage in finite-sample regimes by reducing noise-driven estimation errors and mitigating the influence of incidental correlations.

Structural Hamming Distance

As an additional metric, we report results using the Structural Hamming Distance (SHD) Kalisch and Bühlman (2007). The SHD between two graphs, $\mathcal{G}$ and $\mathcal{H}$, counts the number of edge insertions, deletions, or reversals required to transform one graph into the other. A lower SHD indicates greater structural similarity, though it provides no information about causal interpretability. For instance, reversing a single causal edge contributes only one unit to SHD, yet it may fundamentally alter the interventional semantics of the model. For methods that employ the PC algorithm to estimate causal structure up to a Markov equivalence class (Mean, MICE, Gain), we compute SHD relative to the ground-truth equivalence class. For methods that produce a DAG directly (LoGIC, Miracle), we compare against the true DAG. Given a graph with m nodes, the maximum possible SHD is $\frac{m(m-1)}{2}$; we therefore normalize it to the interval $[0, 1]$ and denote

		GAIN	MEAN	MICE	MIRACLE	LOGIC-G	LOGIC-O
M	**5**	0.28	0.21	0.18	0.37	0.31	0.26
	10	0.23	0.25	0.28	0.43	0.47	0.35
	15	0.13	0.13	0.12	0.36	0.32	0.22
ρ	**0.05**	0.14	0.18	0.16	0.39	0.36	0.28
	0.15	0.32	0.25	0.24	0.39	0.39	0.29
	0.25	0.36	0.21	0.22	0.39	0.35	0.27

Table B.1: [Lower is Better] Normalized Structural Hamming Distance (NSHD) for methods for graphs of sizes $M \in \{5, 10, 15\}$ and missingness probabilities $\rho \in \{0.05, 0.15, 0.25\}$ for data generated under MAR assumption.

		GAIN	MEAN	MICE	MIRACLE	LOGIC-G	LOGIC-O
M	**5**	0.29	0.19	0.19	0.34	0.32	0.22
	10	0.23	0.21	0.22	0.42	0.45	0.36
	15	0.15	0.15	0.14	0.30	0.35	0.25
ρ	**0.05**	0.18	0.17	0.15	0.37	0.34	0.27
	0.15	0.30	0.20	0.21	0.35	0.37	0.28
	0.25	0.36	0.19	0.22	0.34	0.41	0.28

Table B.2: [Lower is Better] Normalized Structural Hamming Distance (NSHD) for methods for graphs of sizes $M \in \{5, 10, 15\}$ and missingness probabilities $\rho \in \{0.05, 0.15, 0.25\}$ for data generated under MNAR assumption.

the result as NSHD for comparability across experiments.

Results for MAR are shown in Tab. B.1, those for MNAR are shown in Tab. B.2. In both settings, LOGIC consistently achieves lower NSHD values than MIRACLE. Although some baseline methods report lower NSHD, these values are computed over Markov equivalence classes rather than DAGs and are thus not directly comparable.