

LEARNING CAUSAL RELATIONS FROM SUBSAMPLED TIME-SERIES WITH ONLY TWO TIME-SLICES

Anonymous authors

Paper under double-blind review

ABSTRACT

This paper studies the summary causal graph on subsampled time series with instantaneous effects using only two time-slices, in which time-slices (cross-sectional observations) are sampled at a coarser timescale than the causal timescale of the underlying system. Given the presence of unmeasured time-slices, conventional causal discovery methods designed for standard time series data would produce significant errors about the system’s causal structure. To address these issues, one promising approach is to construct a topological ordering and then prune unnecessary edges to approximate the true causal structure. Existing topology-based methods often yield non-unique orderings with many spurious edges, reducing accuracy and efficiency in downstream search tasks, while using interventional data for more precise ordering is frequently costly, unethical, or infeasible. Therefore, we explore how the more readily available two time-slices data can replace intervention data to improve topological ordering. Based on a conditional independence criterion using two time-slices as auxiliary instrumental variables, we propose a novel Descendant Hierarchical Topology algorithm with Conditional Independence Test (DHT-CIT) to more efficiently learn causal relations in subsampled time series data. Empirical results on both synthetic and real-world datasets demonstrate the superiority of our DHT-CIT algorithm.

1 INTRODUCTION

Learning causal relations from time series data is a fundamental problem in many fields of science (Granger, 1969; 1980; Lütkepohl, 2005; Hyvärinen et al., 2010; Runge et al., 2019; Bussmann et al., 2021; Löwe et al., 2022; Assaad et al., 2022). While most existing methods have well-identified the causal graph from standard time series data, in many applications, the time series sampling process may be slower than the timescale of causal processes, resulting in numerous previous time-slices (cross-sectional observations) being missing or unreliable (Peters et al., 2017; Hyttinen et al., 2016). In the subsampled time series scenarios, since incorrectly modeling causal structures at the coarser timescale, conventional causal discovery methods designed for standard time series data would produce significant errors about the system’s causal structure. While some recent methods can identify parts of the causal information (i.e., an equivalence class or causal ordering) from subsampled time series data, they still rely on prior knowledge of the subsampling rate and structural functions (Gong et al., 2015; Plis et al., 2015; Hyttinen et al., 2016). Therefore, learning causal structure from subsampled time series (Fig. 1(a,b)) is still challenging without efficient solutions yet.

Inspired by extension works (Runge et al., 2019; Malinsky & Spirtes, 2018; Hyvärinen et al., 2010), a naive thought is to directly extend conventional non-temporal methods to the subsampled time-series setting, incorporating additional assumptions. However, this remains challenging due to the super-exponential search space and the lack of guarantees for identifiable results. To address these issues, one more promising approach is to construct a topological ordering and then prune unnecessary edges to approximate the true causal structure (Teyssier & Koller, 2005; Peters et al., 2014; Loh & Bühlmann, 2014; Park & Klabjan, 2017). For example, SCORE and DiffAN (Rolland et al., 2022; Sanchez et al., 2022) use the Hessian of the data log-likelihood to iteratively identify leaf nodes and generate a complete topological ordering (Fig. 1(c)) to approximate the true causal structure. However, these methods typically generate non-unique topological ordering with numerous spurious edges, which poses potential difficulties in downstream search tasks and may lead to errors in learning summary causal graphs (Fig. 1(a) (Rolland et al., 2022; Sanchez et al., 2022)).

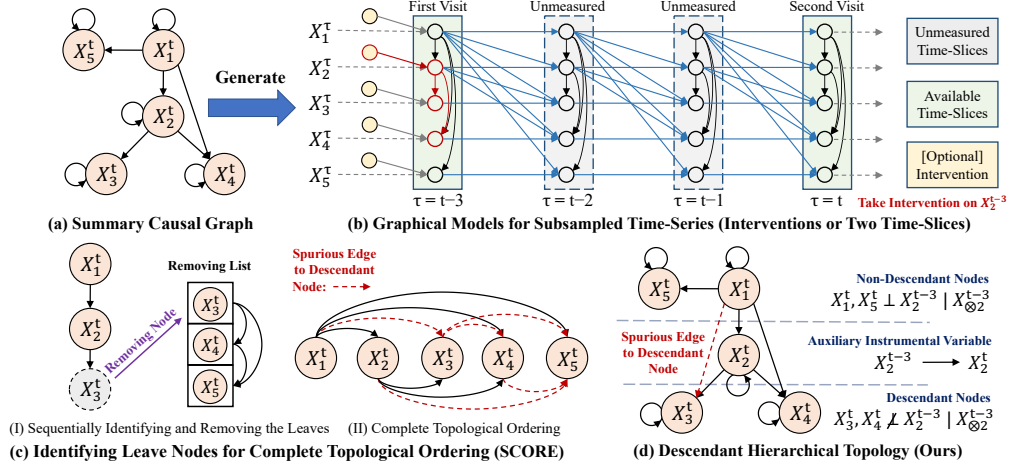


Figure 1: (a, b) Subsampled Time Series with Only Two Time-Slices. (c, d) The SCORE and DHT-CIT architectures. In healthcare, doctors typically compare earlier X^{t-3} and current X^t patient records to identify causes of outcome of interest. Patient visits may be recorded less frequently than the causal timescale of the underlying system, leaving X^{t-2} and X^{t-1} unrecorded.

To construct more precise topological orderings for learning DAGs of Summary Causal Graph, a typical solution is to conduct intervention experiments on each node and then make a comparison to identify which variables have been influenced by the intervention. As marked in red in Fig. 1(b), intervening on X_2^{t-3} enables quick identification of its descendants (X_3^{t-3} , X_4^{t-3}) and improves the topological ordering. However, full interventions are often expensive, unethical, or even infeasible (Wang et al., 2017; Yang et al., 2018). Therefore, in the subsampled time series setting, we explore using just two time-slices to replace interventions to improve topological ordering, in which a time-slice denotes sampled cross-sectional observations on multiple samples captured at a time point on the timeline of system. As illustrated in Fig. 1(b), where each node is influenced by its ancestors and itself, each variable in the earlier time-slice would transmit self-perturbations to both itself and its descendants in the subsequent time-slice, serving as simulated interventions.

Unlike traditional time series studies where all previous time-slices within the observation windows are accessible, in many subsampled time series scenarios, we have access only to limited reliable time-slices. This situation is quite common in healthcare, as shown in Fig. 1(b), where doctors typically use limited time-slices to analyze a patient’s condition and determine treatments. Therefore, in this paper, we study the summary causal graph on subsampled time series with instantaneous effects using only two time-slices. Based on a conditional independence criterion using the earlier time-slice as auxiliary instrumental variables, we propose a Descendant Hierarchical Topology algorithm with Conditional Independence Test (DHT-CIT) to quickly identify (non-)descendants for each node (Fig. 1(d)), ultimately constructing a more efficient unique Descendant Hierarchical Topology with merely a few spurious edges. Subsequently, we prune unnecessary edges to approximate the true causal graph. We rigorously prove that it is both sound and complete under some graph constraints.

The contribution of our paper is four-fold:

- We study the Directed Summary Causal Graph on Subsampled Time Series with Instantaneous Effects using only two time-slices, in which time-slices (cross-sectional observations) are sampled at a coarser timescale than the causal timescale of the underlying system.
- Theoretically, we prove that by incorporating a previous state or implementing interventions for each node, a single conditional independence test per node is sufficient to distinguish between its descendants and non-descendant nodes in the topological ordering.
- Methodologically, we propose DHT-CIT, a novel identifiable topology-based algorithm for a unique descendant hierarchical topology. The search space for underlying graphs over the learned descendant hierarchical topology is much smaller than conventional methods.
- Empirically, our algorithm demonstrates superior performance in synthetic data experiments and partially identifies causal relationships in a real-world study on cardiovascular mortality rate (CMR), providing valuable insights for personalized policy decisions.

2 RELATED WORK ON TEMPORAL DATA

Standard methods for inferring causal structure from conventional time series typically focus either on estimating a transition model at the measurement timescale (e.g., Granger causality (Granger, 1969; 1980)) or they integrate a model of measurement timescale with 'instantaneous' or 'contemporaneous' causal relations to capture interactions within and between d -variate time series from observational data (Lütkepohl, 2005; Hyvärinen et al., 2010; Luo et al., 2015; Nauta et al., 2019; Runge et al., 2019; Runge, 2020; Bussmann et al., 2021; Löwe et al., 2022; Assaad et al., 2022). However, these methods depend on modeling causal structures at the system timescale and assuming causal sufficiency. Both of these conditions might not hold in Subsampled Time Series with only two time-slices, as there could be numerous unmeasured time slices latent in the time series, either before or between these two observed time-slices (Peters et al., 2017).

Subsampled processes with a few time-slices in time series setting are ubiquitous and inherent in the real world, however, causal discovery over Subsampled Time Series is not as well explored. With the prior of the degree of undersampling, (Gong et al., 2015) uses Expectation-Maximization algorithm to recover the linear temporal causal relations from the subsampled data. (Tank et al., 2019) take structural vector autoregressive models for parameter identifiability and estimation. The identifiability of both works is achieved only for linear data. As for nonlinear data, (Danks & Plis, 2013) only can extract constraints on summary graphs from the strongly connected components (SCC). Inspired by works (Gong et al., 2015; Plis et al., 2015), (Hytinen et al., 2016) proposes a constraint optimization approach to identify a small part of the causal information (i.e., an equivalence class) from subsampled time series data, but requires no instantaneous effects. Causal discovery from subsampled time series is still challenging without an efficient solution yet.

Recently, promising topology-based methods tackle the causal discovery problem by finding a certain topological ordering of the nodes and then pruning the spurious edges (Teyssier & Koller, 2005; Peters et al., 2014; Loh & Bühlmann, 2014; Park & Klabjan, 2017; Ghoshal & Honorio, 2018; Ahammad et al., 2021; Sanchez et al., 2022; Reisach et al., 2023). In this paper, we study the Directed Summary Causal Graph on Subsampled Time Series with instantaneous effects using only two time-slices and explore using two time-slices as a substitute for intervention data to improve causal ordering of Summary Causal Graph in Subsampled Time Series. More related works about non-temporal data and their variants can be found in Appendix A and C.

3 PROBLEM SETUP

Standard Time Series: Let $\mathbf{X} = \{\mathbf{X}_i^\tau\}_{d \times t}$ denote the full d -variate time series with all time slices $\mathbf{X}^\tau$ at t time points, where $\mathbf{X}_i^\tau, i \in \{1, 2, \dots, d\}$ and $\tau \in \{1, 2, \dots, t\}$, is a random vector comprising observations of n samples. For simplicity of notation, we will not discuss each sample individually. Instead, we refer to X_i^τ as a random variable when discussing the causal structure. The true causal structure of the summary causal graph is represented by a DAG $\mathcal{G}$. For each X_i^τ , we use the notation $\mathbf{pa}_i^\tau$ to refer to the set of parents of X_i^τ at time τ in $\mathcal{G}$. Similarly, we define $\mathbf{ch}_i^\tau$ for the set of child nodes, $\mathbf{an}_i^\tau$ for the ancestors set, $\mathbf{sib}_i^\tau$ for the siblings set, and $\mathbf{de}_i^\tau$ for the descendants set. As shown in Fig. 1(a,b), following the summary causal graph, the causal structure can be expressed in the functional relationship, for $i \in \{1, 2, \dots, d\}$ and $\tau \in \{1, 2, \dots, t\}$:

$$X_i^\tau = f_i(\mathbf{pa}_i^\tau, X_i^{\tau-1}, \mathbf{pa}_i^{\tau-1}) + \epsilon_i^\tau, \quad (1)$$

where $f_i(\cdot)$ is a twice continuously differentiable function capturing the instantaneous effects from its parents $\mathbf{pa}_i^\tau$ at time τ and a non-zero time-lagged effect from variable $X_i^{\tau-1}$ at time $\tau - 1$; and ϵ_i^τ is an *Additive Noise* at time-slice τ .

Subsampled Time Series with Two Time-Slices: In this paper, we focus on learning the DAG $\mathcal{G}$ of the summary causal graph (Fig. 1(a)) on subsampled time series (Fig. 1(b)) with instantaneous effects using only two time-slices $\mathcal{D} = \{\mathbf{X}^{t_a}, \mathbf{X}^{t_b}\}_{1 < t_a < t_b < t}$. Based on a conditional independence criterion using earlier time-slice $\mathbf{X}^{t_a}$ as conditional instrumental variables¹, we can easily

¹Given $\mathbf{pa}_i^\tau$, the auxiliary variable X_i^τ (i.e., conditional instruments) is conditional independent with its non-descendants $\{\mathbf{an}_i^\tau, \mathbf{sib}_i^\tau\}$ and only indirectly affects its descendant nodes $\mathbf{de}_i^\tau$ (i.e., outcomes) at the current state through its association with X_i^t (i.e., treatments).

distinguish descendants and non-descendants for each node in the current time-slice $\mathbf{X}^{t_b}$. We rigorously prove that it is both sound and complete under some graph constraints, i.e., Markov property, acyclic summary causal graph, stationary full-time graph. More discussion about advantages and limitations of using two time-slices data could be found in Appendix G.

Assumption 1 (Markov Property). *The Markov property of a time series assumes the future time-slice $\mathbf{X}^{t+1}$ depends only on current state $\mathbf{X}^t$ and does not depend on past history $\mathbf{X}^{1 \cdots t-1}$.*

Assumption 2 (Acyclic Summary Causal Graph, Section 5.2.1 in (Assaad et al., 2022)). *The summary causal graph of a multivariate time series is considered acyclic if the lagged effect of each variable solely affects its own value and its descendants, without any influence on its non-descendants.*

Assumption 3 (Consistency Throughout Time, Definition 7 in (Assaad et al., 2022)). *A causal graph $\mathcal{G}$ for a multivariate time series $\mathbf{X}$ is said to be consistent throughout time if all the causal relationships remain constant throughout time, also referred to as stationary full-time graph.*

Following these assumptions, the topological ordering is known to be identifiable from observational data (Peters et al., 2014; Bühlmann et al., 2014), and it is possible to recover the DAG of summary graph underlying the additive noise models (Eq. equation 1). Furthermore, in Appendix F, we relax the Markov Assumption to a high-order Markov Assumption and extend our algorithm to subsampled time series with high-order lagged effect.

4 ALGORITHM

In this section, we will first introduce the complete topological ordering from classical topology-based approaches (Rolland et al., 2022; Sanchez et al., 2022) and show how two time-slice data help identify a unique descent hierarchical topology. Then, based on a conditional independence criterion using the previous time-slice as auxiliary instrumental variables, we propose a novel identifiable topology-based algorithm (DHT-CIT) for two time-slices, which is applicable to any type of noise. The search space over the learned descendant hierarchical topology is much smaller than that of advanced approaches. Then, the underlying summary graph can be found by pruning the unnecessary edges with a well-defined pruning method (Bühlmann et al., 2014).

4.1 FROM COMPLETE TO DESCENDANT HIERARCHICAL TOPOLOGY

As shown in Fig. 1(c), the conventional typology-based approach SCORE (Rolland et al., 2022; Sanchez et al., 2022) sequentially identifies and removes leaf nodes to generate a complete topological ordering based on the Hessian’s diagonal of the data log-likelihood.

Definition 1 (Complete Topological Ordering). *The complete topological ordering ($\pi(\mathbf{X}) = (X_{\pi_1}, X_{\pi_2}, \dots, X_{\pi_d})$, π_i is the reordered index of node) is a sorting of all nodes in a DAG such that for any pair of nodes X_{π_i} and X_{π_j} , if there exists a directed edge from X_{π_i} to X_{π_j} , then $i > j$.*

However, a complete topological ordering is a dense graph with $d(d-1)/2$ edges, containing numerous spurious edges, many of which point to non-descendants unnecessarily. Moreover, these methods (Rolland et al., 2022; Sanchez et al., 2022) may not always produce a unique solution, making it challenging to eliminate false edges and resulting in errors when learning summary causal graph. Fortunately, obtaining two time-slices data can help identify a unique hierarchical topological ordering (Fig. 1(d)), i.e., Descendant Hierarchical Topology, in which each edge only points from an ancestor node to its descendant nodes and not to any non-descendant nodes.

Definition 2 (Hierarchical Topological Ordering). *In the hierarchical topological ordering e.g., $\Pi(\mathbf{X}) = (\{X_{\pi_1}\}_{L_1}, \{X_{\pi_2}, X_{\pi_3}\}_{L_2}, \dots)$, each layer is denoted by L_i and the located layer of X_j are represented as l_j . If there is a directed edge from X_{π_i} to X_{π_j} , then $l_{\pi_i} > l_{\pi_j}$.*

Definition 3 (Descendant Hierarchical Topology). *In the descendant hierarchical topology, each node X_i^t identifies other nodes as either non-descendant nodes or descendant nodes, and each node X_i^t establishes direct edges pointing to its descendant nodes $\mathbf{de}_i^t$, i.e., $X_i^t \rightarrow \mathbf{de}_i^t, i \in \{1, 2, \dots, d\}$.*

For a given causal graph, there may be multiple complete topological orderings (CTO) and hierarchical topological orderings (HTO). However, the descendant hierarchical topology is unique and contains fewer non-essential edges compared to CTO and HTO. This improvement eliminates spurious edges pointing to non-descendant nodes in the learned Descendant Hierarchical Topology and reduces the search space during the pruning stage of topology-based methods.

4.2 CONDITIONAL INDEPENDENCE CRITERION ARISING FROM TWO TIME-SLICES

Two time-slices help topological ordering for learning summary causal graphs. Based on a conditional independence criterion using previous time-slice $\mathbf{X}^{t_a}$ as auxiliary instrumental variables, we can easily distinguish descendants and non-descendants for each node in the current time-slice $\mathbf{X}^{t_b}$.

Theorem 1 (Descendant-oriented Conditional Independence Criteria). *Given two time-slice observations $\mathcal{D} = \{\mathbf{X}^{t_a}, \mathbf{X}^{t_b}\}_{t_a < t_b}$ satisfying Assumptions 1, 2, and 3, for variables $X_i^{t_a}$ and $X_i^{t_b}$, where $i \in \{1, 2, \dots, d\}$, we can conclude that $X_j^{t_b}$ is a descendant node of $X_i^{t_b}$ iff $X_i^{t_a} \not\perp\!\!\!\perp X_j^{t_b} \mid \mathbf{an}_i^{t_a}$.*

Proof. According to the non-zero time-lagged effect of $X_i^{t_a}$ on $X_i^{t_b}$ (Eq. (1)), we have the causal path: (a) $X_i^{t_a} \rightarrow X_i^{t_b}$. Under Assumption 1, we have: (b) $\mathbf{X}^\tau \not\rightarrow X_i^{t_b}$ for all time-slices $\tau < t_a < t_b$. Under Assumption 2, we have: (c) $X_i^{t_a} \not\rightarrow \mathbf{an}_i^{t_b}$ for $t_a < t_b$. Under Assumption 3, we observe a path: (d) $X_i^{t_a} \rightarrow X_j^{t_a} \rightarrow X_j^{t_b}$. Given (a), (b), (c) and (d), if $X_j^{t_b} \in \mathbf{an}_i^{t_b}$, then there are only two summary paths between $X_i^{t_a}$ and $X_j^{t_b}$: $X_i^{t_a} \leftarrow \mathbf{an}_i^{t_a} \rightarrow X_j^{t_b}$ and $X_i^{t_a} \rightarrow \{X_i^{t_b}, \mathbf{de}_i^{t_b}\} \leftarrow X_j^{t_b}$. Hence, once we control the conditional set $\mathbf{an}_i^{t_a}$, i.e., cut off all backdoor path, then the confounding effect between $X_i^{t_a}$ and $X_j^{t_b}$ would be eliminated and $X_i^{t_a} \perp\!\!\!\perp X_j^{t_b} \mid \mathbf{an}_i^{t_a}$. Similarly, if $X_j^{t_b} \in \mathbf{sib}_i^{t_b}$, then the summary backdoor path is $X_i^{t_a} \leftarrow \mathbf{an}_i^{t_a} \rightarrow \mathbf{an}_j^{t_b} \rightarrow X_j^{t_b}$. In summary, if $X_j^{t_b}$ is a non-descendant node of $X_i^{t_b}$, then $X_i^{t_a} \perp\!\!\!\perp X_j^{t_b} \mid \mathbf{an}_i^{t_a}$. In turn, given the condition $X_i^{t_a} \not\perp\!\!\!\perp X_j^{t_b} \mid \mathbf{an}_i^{t_a}$, $X_j^{t_b}$ is a descendant node of $X_i^{t_b}$. $\square$

However, since the causal graph is unknown, we are unable to directly determine the ancestor nodes $\mathbf{an}_i^{t_a}$. Therefore, we choose $\mathbf{X}_{\otimes i}^{t_a}$ to include all variables at time τ , except for $X_i^{t_a}$ and any variables that are independent of $X_i^{t_a}$, as the conditional set. This means that $X_i^{t_a} \not\perp\!\!\!\perp X_j^{t_a}$ for each variable $X_j^{t_a} \in \mathbf{X}_{\otimes i}^{t_a}$. As $\mathbf{X}_{\otimes i}^{t_a}$ occurs prior to time t , it does not introduce any additional backdoor paths to non-descendant nodes at time t , nor can it block the path $X_i^{t_a} \rightarrow X_i^{t_b} \rightarrow X_j^{t_b}$. Thus, we can directly modify the conditional set in Theorem 1 to $\mathbf{X}_{\otimes i}^{t_a}$.

Corollary 1. *Given a two time-slice observations $\mathcal{D} = \{\mathbf{X}^{t_a}, \mathbf{X}^{t_b}\}_{\tau < t}$, for the variables $X_i^{t_a}$ and $X_i^{t_b}$, where $i = 1, 2, \dots, d$, we have $X_j^{t_b}$ is a descendant node of $X_i^{t_b}$ iff $X_i^{t_a} \not\perp\!\!\!\perp X_j^{t_b} \mid \mathbf{X}_{\otimes i}^{t_a}$.*

Based on the corollary 1, we can distinguish between descendant and non-descendant nodes of each variable X_i by conducting a single conditional independence test per variable ($X_i^{t_a} \not\perp\!\!\!\perp \mathbf{de}_i^{t_b} \mid \mathbf{X}_{\otimes i}^{t_a}$). Additionally, if we perform a random intervention on $X_i^{t_a}$, the conditional set in corollary 1 will be empty because the random intervention is independent of the other variables. As a result, the conditional independence test in corollary 1 can be replaced with a simple independence test, which will effectively speed up the search for descendant hierarchical topology. [The difference between the theorems of this paper with those of traditional methods are placed in Appendix B.](#)

4.3 DHT-CIT ALGORITHM

4.3.1 IDENTIFYING DESCENDANT HIERARCHICAL TOPOLOGY

Based on the Conditional Independence Criteria in Corollary 1 and using previous time-slice $\mathbf{X}_{\otimes i}^{t_a} = \{X_j^{t_a} \mid X_j^{t_a} \perp\!\!\!\perp X_i^{t_a}\}$ as AIVs, we can identify descendants $\mathbf{de}_i^{t_b}$ of each variable $X_i^{t_b}$ by a single conditional independence test per variable ($X_i^{t_a} \not\perp\!\!\!\perp \mathbf{de}_i^{t_b} \mid \mathbf{X}_{\otimes i}^{t_a}$). For every $i, j \in \{1, 2, \dots, d\}$, we calculate the conditional independence significance P using the conditional HSIC test from (Zhang et al., 2011) with Gaussian kernel. We determine that X_i is a descendant of X_j if the reported p -value is less than or equal to a threshold α , i.e., $X_i^{t_a} \not\perp\!\!\!\perp X_j^{t_b} \mid \mathbf{X}_{\otimes i}^{t_a}$. Then, we can obtain the adjacency matrix of the unique descendant hierarchical topology by:

$$\mathbf{P} = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,d} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ p_{d,1} & p_{d,2} & \cdots & p_{d,d} \end{pmatrix}, \mathbf{A}^{TP} = \begin{pmatrix} \mathbb{I}(p_{1,1} \leq \alpha) & \mathbb{I}(p_{1,2} \leq \alpha) & \cdots & \mathbb{I}(p_{1,d} \leq \alpha) \\ \mathbb{I}(p_{2,1} \leq \alpha) & \mathbb{I}(p_{2,2} \leq \alpha) & \cdots & \mathbb{I}(p_{2,d} \leq \alpha) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{I}(p_{d,1} \leq \alpha) & \mathbb{I}(p_{d,2} \leq \alpha) & \cdots & \mathbb{I}(p_{d,d} \leq \alpha) \end{pmatrix}. \quad (2)$$

where $p_{i,j} = \mathbf{HSIC}(X_i^{t_a}, X_j^{t_b} \mid \mathbf{X}_{\otimes i}^{t_a})$, α is a hyper-parameter denoting significance threshold, and $\mathbb{I}(\cdot)$ is the indicator function. If the p -value is less than α , the result is considered significant and an

edge is added in the descendant hierarchical topology. In statistical hypothesis testing, α is typically set to 0.05 or 0.01. In this paper, we set the hyper-parameter $\alpha = 0.01$ as the default.

Despite the significant advancements in the development of conditional independence (Zhang et al., 2011; Runge, 2018; Bellot & van der Schaar, 2019), testing for conditional independence remains a challenging task, particularly in high-dimensional cases. The conditional independence test (HSIC) may result in biased topological ordering with cycles. To address this issue, we propose topological layer adjustment as a double guarantee for acyclic constraints in causal discovery.

4.3.2 ADJUSTING THE TOPOLOGICAL ORDERING

To correct the conditional independence test and avoid cycles in the topological ordering, we propose topological layer adjustment to rectify the cycle graph in ordering.

Identifying Leaf Layer. We systematically identify the leaf nodes of the descendant hierarchical topology layer by layer, specifically by selecting nodes that do not have any descendant nodes. We denote all variables at time-slice τ , except for X_i^τ , as $\mathbf{X}_{-i}^\tau$, where $\tau \in \{1, 2, \dots, t\}$. At k -th leaf layer $\mathbf{L}_k$, if $X_i^{t_b} \perp \mathbf{X}_{-i}^{t_a} \mid \mathbf{X}_{\otimes i}^{t_a}$, then $X_i^{t_b}$ is a leaf node at time-slice $\tau = t_b$ and $X_i \in \mathbf{L}_k$. By repeating this operation, we can iteratively $k := k + 1$ and identify the current leaf layer $\mathbf{L}_k$:

$$X_i^{t_b} \in \mathbf{L}_k, \text{ if } a_{i,j}^{TP} = 0 \text{ for all } j \in M_{i,k}, \quad (3)$$

where $X_{M_{i,k}} = \{\mathbf{X}^{t_a}/X_i^{t_a}, \mathbf{L}_{1:k-1}\}$ denotes all variables at time-slice t_a , except for $X_i^{t_a}$ and the variables in lower layer $\mathbf{L}_{1:k-1}$. Then $M_{i,k}$ is the index of variables $\{\mathbf{X}^{t_a}/X_i^{t_a}, \mathbf{L}_{1:k-1}\}$.

Ensuring Acyclic Constraints. By repeating the above procedure, we can sequentially leaf nodes layer-by-layer until we encounter cycles in the topological ordering, which makes it impossible to identify any leaf node as all nodes have at least one descendant node at this time. To ensure acyclic constraints and rectify the edges in descendant hierarchical topology, if the causal relationship between the unprocessed nodes in topological ordering forms a DAG, we locate the maximum p -value that is less than α and change it to a value of 2α , deleting the corresponding edge in the topology

$$p_{i^*,j^*} := 2\alpha \quad \text{and} \quad a_{i^*,j^*}^{TP} = 0, \quad (i^*, j^*) = \arg \max_{i,j} (p_{i,j} \leq \alpha), \quad (4)$$

we repeat this operation until a new leaf node is identified. By adjusting the p -value, the layer sorting leads to a more precise hierarchical topological ordering $\mathbf{A}^{TP} = \{a_{i,j}^{TP}\}_{d \times d}$. This ensures that the topological ordering of the graph is acyclic and improves the accuracy of topological ordering.

4.3.3 PRUNING SPURIOUS EDGES

Based on a conditional independence criterion and two time-slices data, we propose a DHT-CIT algorithm to construct a more efficient descendant hierarchical topology with merely a few spurious edges (Fig. 1(d)). Theoretically, the conditional independence between the hierarchical topological layer ordering allows for a pruning process that only requires one higher layer’s nodes, current layers’ nodes and two lower layers’ nodes as the conditional set, or the node’s non-descendants and one lower layer’s nodes as the conditional set, to examine if a spurious edge exists between a node and its descendant nodes. However, classical methods such as CAM appear to perform better in practice (Bühlmann et al., 2014), which use significance testing based on generalized additive models and select cause if the p -values are less than or equal to 0.001. Like (Rolland et al., 2022), we use CAM to prune spurious edges. The pseudo-code is placed in Algorithm 1 in Appendix D.

5 NUMERICAL EXPERIMENTS

5.1 BASELINES AND EVALUATION

In the experiments, we provide a broad range of time series variants of conventional non-temporal methods that utilize a concatenation of the two cross-sectional data and the temporal edge as additional information to initialize the adjacency matrix and remove temporal edges that are not the same variable. Then, we apply the proposed algorithm (**DHT-CIT**) to both synthetic and real-world data and compare its performance to the following baselines: constraint-based methods, **PC**

and **FCI** (Spirtes et al., 2000); score-based methods, **GOLEM** (Ng et al., 2020), **NOTEARS** with MLP (Zheng et al., 2020), and **ReScore** (Zhang et al., 2023); traditional time-series method, Granger (Shojaie & Michailidis, 2010), VARLiNGAM (Hyvärinen et al., 2010), and **CD-NOD** (Huang et al., 2020); topology-based methods, **CAM** (Bühlmann et al., 2014) and **SCORE** (Rolland et al., 2022). The discussions about the rationale behind the chosen baselines are deferred to Appendix C.

To evaluate the performance of the proposed **DHT-CIT**, we compute the Structural Hamming Distance (**SHD**) between the output and the true graphs, which evaluates the differences in terms of node, edge, and connection counts in the two graphs. Besides, we use Structural Intervention Distance (**SID**) to count the minimum number of interventions required to transform the output DAG into the true DAG, or vice versa. The accuracy of the identified edges can also be evaluated through the use of commonly adopted metrics **F1-Score** and L2-distance (**Dis.**) between two graphs.

In topology-based methods, SCORE often generates a complete topological ordering with $d(d-1)/2$ edges. However, many of these edges are unnecessary and point to non-descendants, requiring pruning, which can decrease the accuracy and efficiency of the pruning process. The proposed DHT-CIT improves upon existing topology-based methods by eliminating numerous spurious edges in the learned Descendant Hierarchical Topology and reducing the search space in the pruning stage. As a comparison among topology-based methods, we count the number of spurious edges that needed to be pruned for each method, which is represented by **#Prune**.

5.2 EXPERIMENTS ON SYNTHETIC DATA

Datasets. We test our algorithm on synthetic data generated from a *additive non-linear noise model* (Eq. 1) under Assumptions 1, 2 and 3. Given d nodes and e edges, we generate the causal graph $\mathcal{G}$ using Erdos-Renyi model (Erdős & Rényi, 2011). In main experiments, we generate the data with Gaussian Noise for every variable $X_i^\tau, i = 1, 2, \dots, d$ at time $\tau = 1, 2, \dots, t$:

$$X_i^\tau = \text{Sin}(\mathbf{pa}_i^\tau, X_i^{\tau-1}) + \frac{1}{10} \text{Sin}(\mathbf{w} \cdot \mathbf{pa}_i^{\tau-1}) + \epsilon_i^\tau, \mathbf{X}^0 \sim \mathcal{N}(0, \mathbf{I}_d), \epsilon^\tau \sim \mathcal{N}(0, 0.4 \cdot \mathbf{I}_d), \quad (5)$$

where $\text{Sin}(\mathbf{pa}_i^\tau) = \sum_{j \in \text{pa}(X_i)} \sin(X_j^\tau)$, $\mathbf{I}_d$ is a d -th order identity matrix, and $\mathbf{w}$ is a random 0-1 vector that controls the number and existence of time-lagged edges from $\mathbf{pa}_i^{\tau-1}$. The experiments on varying time-lagged edges are deferred to Fig. 2, while in the main experiments, we set the number of time-lagged edges from $\mathbf{pa}_i^{\tau-1}$ as 0. To evaluate the performance of our DHT-CIT across various scenarios, we vary the number of nodes (d) and edges (e) to generate larger and denser graphs, which we refer to as **Sin- d - e** . Moreover, to test the algorithm’s robustness against different noise types, we also generate data with Laplace noise ($X_i^0, \epsilon^\tau \sim \text{Laplace}(0, 1/\sqrt{2})$) and Uniform noise ($X_i^0, \epsilon^\tau \sim U(-1, 1)$). In each experiment setting, we perform 10 replications, each with a sample size 1000, to report the mean and the standard deviation of the mentioned metrics.

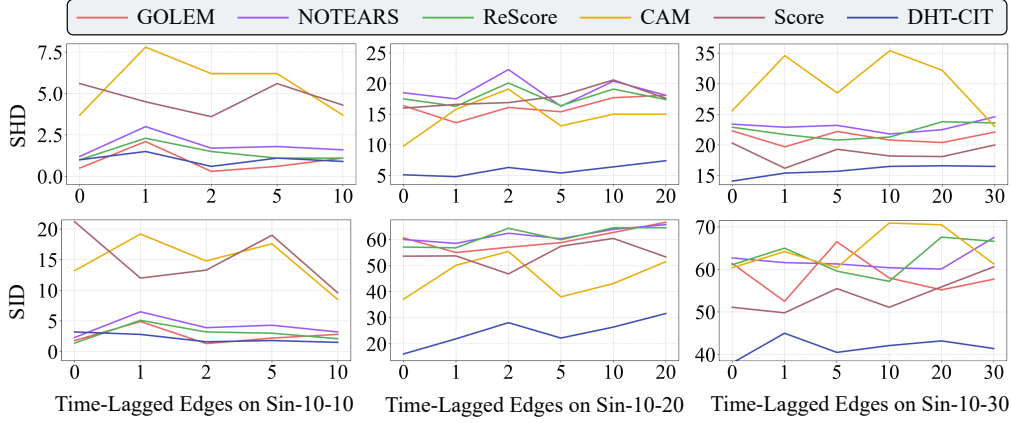
Additional, experiments on exploring **complex non-linear relationships** and on **large graph with 50/100-dimension variables** are deferred to Appendix E.1 and E.2.

Investigating the Influence of Underlying DAG’s Size and Sparsity in Sin- d - e Experiments. The results of the synthetic experiments are shown in Tab. 1 and Fig. 2. From the results on sparse graphs (**Sin-10-10** and **Sin-20-20**) in Tab. 1, we have the following observation: (1) In non-linear time series data, causal sufficiency and time dependency can be violated. Despite concatenating two time-slices, the time series variants of PC and FCI still fail to identify causal graphs. (2) The three methods (GOLEM, NOTEARS, and ReScore) specifically designed for sparse graphs have shown excellent performance, surpassing the proposed DHT-CIT on Sin-10-10 with observational data ($\mathcal{D} = \{\mathbf{X}^1, \mathbf{X}^2\}$). However, their performance significantly deteriorates on denser graphs (Fig. 2). (3) Traditional time series algorithms like Granger and VARLiNGAM, which rely on modeling at the standard measurement timescale, struggle with identifying the summary causal graph of subsampled time series. They even underperform compared to temporal variants of conventional methods designed for non-temporal data. While the CD-NOD method shows promising results in small-sized datasets, it only offers an equivalence class of the causal graph, which constrains the exploration of true causality. (4) As a topology-based method, SCORE was able to recover nearly true causal structure when applied to interventional data ($\mathcal{D} = \{\mathbf{X}^0, \mathbf{X}^1\}$). However, its performance decreases when dealing with observational data ($\mathcal{D} = \{\mathbf{X}^1, \mathbf{X}^2\}$) due to the presence of complex causal relationships from previous states. (5) The proposed DHT-CIT builds a descendant hierarchical topology with merely a few spurious edges. The search space over the learned descen-

Table 1: The results (mean $\pm$ std) on sparse graph Sin- d - e with simulated interventional data ($\mathcal{D} = \{X^0, X^1\}$) or observational data ($\mathcal{D} = \{X^1, X^2\}$).

	Sin-10-10 Graph with Interventional Data ($\mathcal{D} = \{X^0, X^1\}$)					Sin-10-10 Graph with Observational Data ($\mathcal{D} = \{X^1, X^2\}$)				
Method	SHD $\downarrow$	SID $\downarrow$	F1-Score $\uparrow$	Dis. $\downarrow$	#Prune $\downarrow$	SHD $\downarrow$	SID $\downarrow$	F1-Score $\uparrow$	Dis. $\downarrow$	#Prune $\downarrow$
PC	5.90 $\pm$ 3.28	34.7 $\pm$ 20.8	0.77 $\pm$ 0.11	2.32 $\pm$ 0.74	-	12.8 $\pm$ 5.03	43.6 $\pm$ 9.94	0.56 $\pm$ 0.12	3.51 $\pm$ 0.72	-
FCI	9.70 $\pm$ 2.87	58.9 $\pm$ 17.3	0.67 $\pm$ 0.07	3.08 $\pm$ 0.48	-	15.3 $\pm$ 3.77	71.0 $\pm$ 11.5	0.54 $\pm$ 0.09	3.89 $\pm$ 0.46	-
GOLEM	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	1.00 $\pm$ 0.00	0.00 $\pm$ 0.00	-	0.50 $\pm$ 0.80	1.80 $\pm$ 2.70	0.97 $\pm$ 0.03	0.38 $\pm$ 0.59	-
NOTEARS	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	1.00 $\pm$ 0.00	0.00 $\pm$ 0.00	-	1.20 $\pm$ 0.60	2.30 $\pm$ 1.20	0.94 $\pm$ 0.02	1.02 $\pm$ 0.30	-
ReScore	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	1.00 $\pm$ 0.00	0.00 $\pm$ 0.00	-	1.00 $\pm$ 0.63	1.40 $\pm$ 1.36	0.95 $\pm$ 0.03	0.88 $\pm$ 0.47	-
Granger	26.5 $\pm$ 5.22	75.6 $\pm$ 13.8	0.20 $\pm$ 0.08	5.12 $\pm$ 0.52	-	31.3 $\pm$ 11.6	66.8 $\pm$ 30.8	0.21 $\pm$ 0.04	5.48 $\pm$ 1.10	-
VarLiNGAM	35.0 $\pm$ 0.00	69.4 $\pm$ 3.20	0.36 $\pm$ 0.00	5.91 $\pm$ 0.00	-	35.0 $\pm$ 0.00	69.4 $\pm$ 3.20	0.36 $\pm$ 0.00	5.91 $\pm$ 0.00	-
CD-NOD	30.3 $\pm$ 3.16	11.3 $\pm$ 13.3	0.86 $\pm$ 0.14	1.28 $\pm$ 1.16	-	5.40 $\pm$ 0.92	15.5 $\pm$ 4.70	0.74 $\pm$ 0.04	2.32 $\pm$ 0.19	-
CAM	5.00 $\pm$ 6.27	14.9 $\pm$ 18.5	0.78 $\pm$ 0.27	1.53 $\pm$ 1.72	80.00 $\pm$ 0.00	3.70 $\pm$ 2.95	13.2 $\pm$ 10.6	0.84 $\pm$ 0.13	1.79 $\pm$ 0.74	80.00 $\pm$ 0.00
SCORE	1.20 $\pm$ 3.46	4.2 $\pm$ 10.7	0.95 $\pm$ 0.14	0.43 $\pm$ 1.06	35.30 $\pm$ 0.95	5.60 $\pm$ 3.92	21.2 $\pm$ 16.1	0.78 $\pm$ 0.14	2.25 $\pm$ 0.78	35.80 $\pm$ 0.98
DHT-CIT	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	1.00 $\pm$ 0.00	0.00 $\pm$ 0.00	9.00 $\pm$ 2.65	1.00 $\pm$ 1.22	3.20 $\pm$ 3.70	0.95 $\pm$ 0.05	0.68 $\pm$ 0.72	13.20 $\pm$ 4.30
	Sin-20-20 Graph with Interventional Data ($\mathcal{D} = \{X^0, X^1\}$)					Sin-20-20 Graph with Observational Data ($\mathcal{D} = \{X^1, X^2\}$)				
Method	SHD $\downarrow$	SID $\downarrow$	F1-Score $\uparrow$	Dis. $\downarrow$	#Prune $\downarrow$	SHD $\downarrow$	SID $\downarrow$	F1-Score $\uparrow$	Dis. $\downarrow$	#Prune $\downarrow$
PC	10.7 $\pm$ 5.70	61.2 $\pm$ 35.6	0.79 $\pm$ 0.10	3.18 $\pm$ 0.83	-	21.5 $\pm$ 6.75	98.2 $\pm$ 31.8	0.61 $\pm$ 0.11	4.59 $\pm$ 0.69	-
FCI	20.1 $\pm$ 3.03	181 $\pm$ 49.9	0.66 $\pm$ 0.05	4.47 $\pm$ 0.35	-	30.5 $\pm$ 4.09	237 $\pm$ 59.1	0.54 $\pm$ 0.05	5.51 $\pm$ 0.37	-
GOLEM	0.60 $\pm$ 1.50	2.50 $\pm$ 5.20	0.98 $\pm$ 0.04	0.32 $\pm$ 0.70	-	1.30 $\pm$ 1.10	5.60 $\pm$ 4.40	0.97 $\pm$ 0.03	0.93 $\pm$ 0.66	-
NOTEARS	0.20 $\pm$ 0.40	1.00 $\pm$ 2.0	0.99 $\pm$ 0.01	0.20 $\pm$ 0.40	-	2.60 $\pm$ 1.49	6.00 $\pm$ 3.40	0.94 $\pm$ 0.03	1.55 $\pm$ 0.46	-
ReScore	0.90 $\pm$ 2.70	3.90 $\pm$ 11.7	0.98 $\pm$ 0.06	0.30 $\pm$ 0.90	-	2.00 $\pm$ 0.77	5.10 $\pm$ 2.90	0.95 $\pm$ 0.01	1.38 $\pm$ 0.28	-
Granger	103 $\pm$ 17.2	371 $\pm$ 8.19	0.10 $\pm$ 0.01	10.1 $\pm$ 0.89	-	104 $\pm$ 20.7	368 $\pm$ 8.82	0.10 $\pm$ 0.03	10.1 $\pm$ 0.01	-
VarLiNGAM	170 $\pm$ 0.00	339 $\pm$ 3.20	0.19 $\pm$ 0.00	13.0 $\pm$ 0.00	-	170 $\pm$ 0.00	339 $\pm$ 3.20	0.19 $\pm$ 0.00	13.0 $\pm$ 0.00	-
CD-NOD	exceed 48h	-	-	-	-	-	-	-	-	-
CAM	4.50 $\pm$ 3.03	15.8 $\pm$ 14.2	0.89 $\pm$ 0.07	1.86 $\pm$ 1.07	360.0 $\pm$ 0.00	10.3 $\pm$ 6.50	41.6 $\pm$ 34.7	0.79 $\pm$ 0.12	3.07 $\pm$ 0.98	360.0 $\pm$ 0.00
SCORE	0.20 $\pm$ 0.63	0.90 $\pm$ 2.70	0.99 $\pm$ 0.02	0.14 $\pm$ 0.45	170.1 $\pm$ 0.32	7.40 $\pm$ 2.41	31.3 $\pm$ 21.7	0.85 $\pm$ 0.04	2.68 $\pm$ 0.47	172.1 $\pm$ 0.22
DHT-CIT	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	1.00 $\pm$ 0.00	0.00 $\pm$ 0.00	16.44 $\pm$ 3.81	1.00 $\pm$ 1.32	3.10 $\pm$ 4.40	0.98 $\pm$ 0.03	0.51 $\pm$ 0.61	30.60 $\pm$ 7.70

* CD-NOD on Sin-20-20 takes over 48 hours and #Prune on one-stage methods is not meaningful. We don't discuss these results and represent them with '-'.

Figure 2: Exploring varying time-lagged edges and denser graph in observations $\mathcal{D} = \{X^1, X^2\}$.

dent hierarchical topology is much smaller than that of SCORE. On average, compared to SCORE, the number of pruned edges in DHT-CIT decreases 24.4 for **Sin-10-10** and 147.6 for **Sin-20-20**. As the underlying DAG's size increases, DHT-CIT achieves unbiased causal discovery on interventional data, but there may be a slight decrease on observational data, i.e., merely one error edge on average, but its F1-Score still exceeds 95%.

Exploring Varying Time-Lagged Edges and Denser Graph in Observations $\mathcal{D} = \{X^1, X^2\}$. In the experiments on denser graphs with more edges ($e = 2d$ and $e = 3d$), we gradually increase the number of time-lagged edges from other variables X_{-i}^{t-1} from 0 to d . As shown in Figure 3, most well-performed baselines on sparse graphs exhibit a substantial decrease in performance when applied to denser graphs. However, our DHT-CIT algorithm outperforms the best baseline on denser graphs. On the Sin-10-20 dataset, we achieve a 48% increase in SHD, a 48% increase in SID, and a 15% boost in F1-Score. On the Sin-10-30 dataset, we achieve a 30% increase in SHD, a 43% increase in SID, and a 7% boost in F1-Score. Our DHT-CIT is robust to varying time-lagged edges.

Scaling to Different Noise Types. To evaluate algorithm's robustness against various noise types, **Sin-10-10** data was generated with Laplace and Uniform noise. The results (Tab. 2) demonstrate the superior and robust performance of DHT-CIT against different types of noise, with the accuracy consistently comparable to that under Gaussian noise.

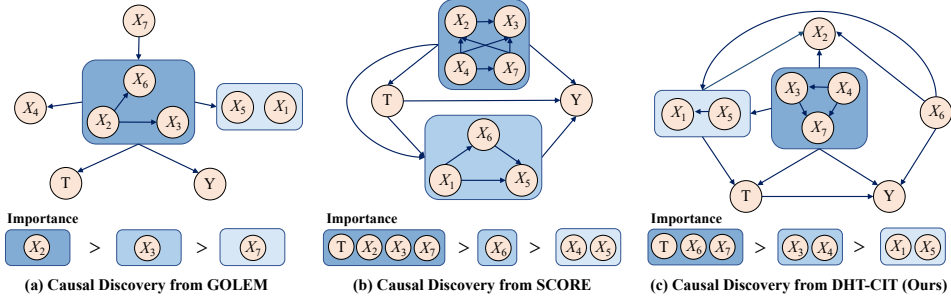


Figure 3: Causal Discovery on PM-CMR Dataset.

Table 2: The experiments on different noise type.

	Sin-10-10 data with Laplace noise ($\mathcal{D} = \{X^1, X^2\}$)				
Method	SHD↓	SID↓	F1-Score↑	Dis↓	#Prune↓
GOLEM	1.50±1.20	2.80 ±2.52	0.92±0.05	1.00±0.70	-
NOTEARS	1.60±0.06	3.70±3.10	0.92±0.03	1.23±0.26	-
ReScore	2.00±1.34	3.00±2.41	0.90±0.06	1.29±0.57	-
CAM	5.30±2.83	14.0±8.01	0.78±0.12	2.23±0.57	80.0±0.00
SCORE	3.90±1.70	9.90±6.01	0.84±0.06	1.93±0.43	35.5±0.92
DHT-CIT	1.20 ±1.99	3.60±6.55	0.94 ±0.04	0.59 ±0.92	0.80 ±1.40
	Sin-10-10 data with Uniform noise ($\mathcal{D} = \{X^1, X^2\}$)				
Method	SHD↓	SID↓	F1-Score↑	Dis↓	#Prune↓
GOLEM	2.60±1.80	6.80±3.94	0.89±0.06	1.46±0.68	-
NOTEARS	2.00±1.34	4.80±1.30	0.91±0.05	1.29±0.57	-
ReScore	1.70±0.90	3.70±2.90	0.92±0.04	1.21±0.48	-
CAM	8.90±7.15	21.4±12.0	0.68±0.22	2.14±0.73	80.0±0.00
SCORE	5.10±3.42	13.6±8.30	0.80±0.11	2.14±0.73	35.0±0.00
DHT-CIT	1.00 ±2.19	1.10 ±2.47	0.96 ±0.09	0.44 ±0.90	0.70 ±1.55

Table 3: Average running time(s).

	Sin-10-10	Sin-10-20	Sin-10-30	Sin-20-20
CD-NOD	5433s	> 2h	> 2h	> 5h
CAM	97.2s	92.8s	111.3s	543.6s
GOLEM	44.9s	45.8s	47.6s	63.0s
SCORE	40.7s	38.1s	45.2s	193.1s
NOTEARS	33.6s	35.6s	36.4s	747.2s
ReScore	24.1s	23.2s	24.2s	29.2s
PC	21.1s	20.7s	21.1s	32.8s
FCI	18.7s	18.7s	18.8s	30.1s
VarLiNGAM	17.9s	18.3s	18.4s	27.3s
Granger	8.4s	8.5s	8.4s	36.4s
DHT-CIT	54.7s	77.1s	58.2s	224.4s

Training Cost Analysis. We implement 10 replications to study the average running time(s) of our DHT-CIT in a single execution in Tab. 3. Despite the increasing complexity and training cost, the DHT-CIT model maintains superior performance and scalability to larger and denser graphs, with a single execution time below 300 seconds, which is still within an acceptable threshold.

5.3 EXPERIMENTS ON REAL-WORLD DATA

The **PM-CMR** (Wyatt et al., 2020) is a public time series data that is commonly used to study the impact of the particle ($\text{PM}_{2.5}$, T) on the cardiovascular mortality rate (CMR, Y) in 2132 counties in the US from 1990 to 2010. Additionally, the dataset includes 7 variables ($X_{1:7}$) related to the city status, which are potential common causes of both $\text{PM}_{2.5}$ and CMR. The corresponding description of variables is detailed in Tab. 6 in Appendix E.3. With the prior knowledge, i.e., $T \leftarrow X_{1:7} \rightarrow Y$ and $T \rightarrow Y$, we draw two time-slices in 2000 & 2010 to evaluate the performance of the proposed DHT-CIT and two well-performed baselines (GOLEM and SCORE). As illustrated in Fig. 3, both GOLEM and SCORE do not generate true summary causal graph, and only our DHT-CIT achieves more accurate causal relationships in real-world data. GOLEM shows there is no direct edge from T to Y and SCORE shows that T is the parent node of $\{X_1, X_5, X_6\}$, which contradicts the prior knowledge. Only our DHT-CIT algorithm recovers the dense causal graph, i.e., $T \leftarrow X_{1:7} \rightarrow Y$ and $T \rightarrow Y$. The results are consistent with the experiments on denser graphs: both GOLEM and SCORE are only applicable to sparse graphs, whereas our DHT-CIT maintains superior performance and scalability to larger and denser graphs. More detailed results are deferred to Appendix E.3.

6 CONCLUSION

In the subsampled time series with only two time-slices, conventional causal discovery methods designed for standard time series data would produce significant errors about the system’s causal structure. To address this issue, we use some auxiliary instrumental variables in two time-slices as interventions to improve topological ordering and propose a novel DHT-CIT algorithm to learn a unique descendant hierarchical topology with merely a few spurious edges for identifying DAG of summary causal graph. The proposed DHT-CIT algorithm considerably eases the assumptions typically made in traditional time series studies about modeling causal structures at the system timescale, requiring causal sufficiency, and observing all time slices within the observation windows.

REFERENCES

- Tanzin Ahammad, Mohammad Hasan, and Md Zahid Hassan. A new topological sorting algorithm with reduced time complexity. In *Proceedings of the 3rd International Conference on Intelligent Computing and Optimization 2020 (ICO 2020)*, pp. 418–429. Springer, 2021.
- Charles K Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.
- Alexis Bellot and Mihaela van der Schaar. Conditional independence testing using generative adversarial networks. *Advances in Neural Information Processing Systems*, 32, 2019.
- 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.
- Bart Bussmann, Jannes Nys, and Steven Latré. Neural additive vector autoregression models for causal discovery in time series. In *Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings 24*, pp. 446–460. Springer, 2021.
- Wei Chen, Kun Zhang, Ruichu Cai, Biwei Huang, Joseph Ramsey, Zhifeng Hao, and Clark Glymour. Fritl: A hybrid method for causal discovery in the presence of latent confounders. *arXiv preprint arXiv:2103.14238*, 2021.
- David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.
- David Danks and Sergey Plis. Learning causal structure from undersampled time series. 2013.
- P Erdős and A Rényi. On the evolution of random graphs. In *The Structure and Dynamics of Networks*, pp. 38–82. Princeton University Press, 2011.
- Asish Ghoshal and Jean Honorio. Learning linear structural equation models in polynomial time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pp. 1466–1475. PMLR, 2018.
- Mingming Gong, Kun Zhang, Bernhard Schölkopf, Dacheng Tao, and Philipp Geiger. Discovering temporal causal relations from subsampled data. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1898–1906, Lille, France, 07–09 Jul 2015. PMLR.
- Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.
- Clive WJ Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352, 1980.
- Uzma Hasan, Emam Hossain, and Md Osman Gani. A survey on causal discovery methods for temporal and non-temporal data. *arXiv preprint arXiv:2303.15027*, 2023.
- Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1):2409–2464, 2012.
- Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *The Journal of Machine Learning Research*, 21(1):3482–3534, 2020.
- Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, pp. 340–349, 2014.
- Antti Hyttinen, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. Causal discovery from subsampled time series data by constraint optimization. In *Conference on Probabilistic Graphical Models*, pp. 216–227. PMLR, 2016.

- Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.
- Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C Mozer, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.
- Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=rklbKA4YDS>.
- Yan Li, Rui Xia, Chunchen Liu, and Liang Sun. A hybrid causal structure learning algorithm for mixed-type data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7435–7443, 2022.
- Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1):3065–3105, 2014.
- Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*, pp. 509–525. PMLR, 2022.
- Ling Luo, Wei Liu, Irena Koprinska, and Fang Chen. Discovering causal structures from time series data via enhanced granger causality. In *AI 2015: Advances in Artificial Intelligence: 28th Australasian Joint Conference, Canberra, ACT, Australia, November 30–December 4, 2015, Proceedings 28*, pp. 365–378. Springer, 2015.
- Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.
- Daniel Malinsky and Peter Spirtes. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD workshop on causal discovery*, pp. 23–47. PMLR, 2018.
- Atalanti A Mastakouri, Bernhard Schölkopf, and Dominik Janzing. Necessary and sufficient conditions for causal feature selection in time series with latent common causes. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7502–7511. PMLR, 18–24 Jul 2021.
- Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *2nd Conference on Causal Learning and Reasoning*, 2023. URL <https://openreview.net/forum?id=rV00Bx90deu>.
- Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, 2019.
- 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.
- Young Woong Park and Diego Klabjan. Bayesian network learning via topological order. *The Journal of Machine Learning Research*, 18(1):3451–3482, 2017.
- Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.
- Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.
- Sergey Plis, David Danks, and Jianyu Yang. Mesochronal structure learning. In *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, volume 31. NIH Public Access, 2015.
- Joseph Ramsey, Jiji Zhang, and Peter L Spirtes. Adjacency-faithfulness and conservative causal inference. *arXiv preprint arXiv:1206.6843*, 2012.
- Alexander G Reisach, Myriam Tami, Christof Seiler, Antoine Chambaz, and Sebastian Weichwald. Simple sorting criteria help find the causal order in additive noise models. *arXiv preprint arXiv:2303.18211*, 2023.
- Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pp. 18741–18753. PMLR, 2022.
- Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pp. 938–947. PMLR, 2018.
- Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1388–1397. PMLR, 2020.
- Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019.
- Pedro Sanchez, Xiao Liu, Alison Q O’Neil, and Sotirios A Tsaftaris. Diffusion models for causal discovery via topological ordering. *arXiv preprint arXiv:2210.06201*, 2022.
- Ali Shojaie and George Michailidis. Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523, 2010.
- Liam Solus, Yuhao Wang, and Caroline Uhler. Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika*, 108(4):795–814, 2021.
- Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.
- Xiaohai Sun, Dominik Janzing, Bernhard Schölkopf, and Kenji Fukumizu. A kernel-based causal learning algorithm. In *Proceedings of the 24th international conference on Machine learning*, pp. 855–862, 2007.
- Alex Tank, Emily B Fox, and Ali Shojaie. Identifiability and estimation of structural vector autoregressive models for subsampled and mixed-frequency time series. *Biometrika*, 106(2):433–452, 2019.
- Marc Teyssier and Daphne Koller. Ordering-based search: a simple and effective algorithm for learning bayesian networks. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pp. 584–590, 2005.
- Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.
- Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. *Advances in Neural Information Processing Systems*, 30, 2017.
- Lauren H Wyatt, Geoffrey Colin L Peterson, Tim J Wade, Lucas M Neas, and Ana G Rappold. Annual pm2. 5 and cardiovascular mortality rate data: Trends modified by county socioeconomic status in 2,132 us counties. *Data in brief*, 30:105–318, 2020.

- Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, pp. 5541–5550. PMLR, 2018.
- An Zhang, Fangfu Liu, Wenchang Ma, Zhibo Cai, Xiang Wang, and Tat-Seng Chua. Boosting causal discovery via adaptive sample reweighting. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=LNpMtk15AS4>.
- Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.
- Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *UAI*, pp. 804–813. AUAI Press, 2011.
- Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018a. URL <https://proceedings.neurips.cc/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf>.
- Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018b.
- Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse non-parametric dags. In *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425. PMLR, 2020.
- Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=S1g2skStPB>.

A RELATED WORK ON NON-TEMPORAL DATA

Constraint-based methods typically rely on conditional independence tests to identify causal relationships by testing the independence between variables given a set of conditions (Sun et al., 2007; Hyttinen et al., 2014), such as PC, FCI, SGS, and ICP (Spirtes et al., 2000; Zhang, 2008; Ramsey et al., 2012; Peters et al., 2016). Score-based methods (Tsamardinos et al., 2006; Ke et al., 2019; Zhu et al., 2020) search through the space of all possible causal structures with the aim of optimizing a specified metric, and rely on local heuristics to enforce the acyclicity, such as GES, and GIES (Chickering, 2002; Hauser & Bühlmann, 2012). Continuous-optimization methods (Zheng et al., 2018a; Lachapelle et al., 2020) view the search as a constrained optimization problem and apply first-order optimization methods to solve it, such as GraNDAG, GOLEM, NOTEARS, ReScore (Lachapelle et al., 2020; Ng et al., 2020; Zheng et al., 2018b; 2020; Zhang et al., 2023). Hybrid methods combine the advantages of both types of methods (Tsamardinos et al., 2006; Chen et al., 2021; Li et al., 2022; Hasan et al., 2023). GSP and IGSP algorithms (Solus et al., 2021; Wang et al., 2017) evaluate the score of each DAG structure using some information criterion and search for the optimal solution by iteratively changing permutations. Nevertheless, most constraint-based methods (e.g. PC, FCI) typically find causal structures within an equivalence class, resulting in a limited understanding of the underlying causal relationships. Score-based methods (e.g. NOTEARS, GES) rely on local heuristics to enforce acyclicity constraints, which can be insufficient for effectively handling large datasets. Additionally, the causal graphs produced by minimizing a specific score function are not guaranteed to be entirely accurate.

Recently, topology-based methods tackle the causal discovery problem by finding a certain topological ordering of the nodes and then pruning the spurious edges in topological ordering (Teyssier & Koller, 2005; Peters et al., 2014; Loh & Bühlmann, 2014; Park & Klabjan, 2017; Ghoshal & Honorio, 2018; Ahammad et al., 2021; Sanchez et al., 2022; Reisach et al., 2023). Examples of topology-based methods include CAM, SCORE and NoGAM (Bühlmann et al., 2014; Rolland et al., 2022; Montagna et al., 2023). These methods encounter a less combinatorial problem as the set of permutations is much smaller than the set of directed acyclic graphs. While these methods restrict the number and direction of potential edges in the learned DAG, they often generate numerous spurious edges that need to be pruned. In this paper, we focus on only two time-slices for learning causal relations and concatenate two time-slices data with the temporal edge to study a causal graph. Although the topology-based approach is widely applicable to cross-sectional data, its application to two time-slices studies is not routine and the data opportunities that two time-slices provide for topology-based methods are also overlooked.

B THE DIFFERENCE OF OUR THEOREMS WITH TRADITIONAL METHODS

Our theorem differs from that of Peters et al. (2017): Theorems 10.1, 10.2, 10.3, and 10.4 in Section 10 of Peters et al. (2017) rely on all previous time-slices within the observation windows that are accessible, i.e., $\mathbf{X}_{\text{past}(t)}$ could be observed, which are not satisfied in subsampled time-series. In Section 10.2.1, Peters et al. (2017) also noted that current methods for subsampled time-series require well-modeled interventions. Without well-defined interventions, it remains challenging to learn causal structures from subsampled time series without efficient solutions. In our Theorem 1, given two time-slice observations $\mathcal{D} = \{\mathbf{X}^{t_a}, \mathbf{X}^{t_b}\}_{t_a < t_b}$, $X_j^{t_b}$ is a descendant node of $X_i^{t_a}$ iff $X_i^{t_a} \not\perp\!\!\!\perp X_j^{t_b} \mid \text{an}_i^{t_a}$. We rigorously prove that it is both sound and complete under some graph constraints, i.e., acyclic summary causal graph, stationary full-time graph, and Markov property.

Our theorem differs from that of Mastakouri et al. (2021): Mastakouri et al. (2021) focuses on the detection of direct and indirect causes of a given target time series, rather than full graph discovery. Their theory requires that each observed candidate time series must be a non-descendant node of the target time series, and they need to conduct two conditional independence tests for each observed candidate to identify the direct causes of the target. In contrast, our theory not only constructs the Descendant Hierarchical Topology of the Directed Summary Causal Graph but also

requires only one conditional independence test per component. Furthermore, the full graph studied by Mastakouri et al. (2021) does not include instantaneous effects.

C THE DISCUSSION ABOUT THE COMPARED BASELINES

In this paper, we study the Directed Summary Causal Graph on Subsampled Time Series with instantaneous effects using only two time-slices, which is different from Standard Time Series Setting. Traditional methods designed for time series setting typically depend on modeling causal structures at the system timescale and assuming causal sufficiency, they require multiple time-slices with equal time intervals to estimate causal graphs. While Gong et al. (2015); Plis et al. (2015); Hyttinen et al. (2016) can identify a part of the causal information (i.e., an equivalence class) from subsampled time series data, they either rely on linear models or require the absence of instantaneous effects, which does not align with our setting. Most importantly, the algorithms designed for subsampled time series from these studies do not have reproducible open-source code available.

Since there are no available standard algorithms designed for two time-slice data, we develop variants of conventional non-temporal methods to the two time-slices setting proposed in this paper and compare the proposed algorithm with them: constraint-based methods, **PC** and **FCI** (Spirtes et al., 2000); score-based methods, **GOLEM** (Ng et al., 2020), **NOTEARS** with MLP (Zheng et al., 2020), and **ReScore** (Zhang et al., 2023); time-series method, **CD-NOD** (Huang et al., 2020); topology-based methods, **CAM** (Bühlmann et al., 2014) and **SCORE** (Rolland et al., 2022).

How to apply the non-temporal algorithms to the time series setting: In the main experiments, we provide a broad range of time series variants of conventional non-temporal methods - namely PC, FCI, GOLEM, NOTEARS with MLP, ReScore + NOTEARS, CAM, and SCORE - that utilize a concatenation of the two cross-sectional data and the temporal edge (where the previous variable of the same components leads to the subsequent variable) as prior information (as initializing the adjacency matrix, and removing temporal edges that are not the same variable). Given $\mathcal{D} = \{\mathbf{X}^{t_a}, \mathbf{X}^{t_b}\}_{t_a < t_b}$, since the subsampled time series is a stationary stochastic process and the summary causal graph is acyclic, after removing the nodes $\mathbf{X}^{t_a}$ in the learned graph, the DAGs on $\mathbf{X}^{t_b}$ learned by the above causal discovery variants would approximate the summary causal graphs of interest.

Score-based methods GOLEM, NOTEARS and ReScore are designed to recover the whole DAG by applying first-order optimization methods to solve a constrained optimization problem. While they may produce significant errors regarding the lagged effect, the causal graph learned on $\mathbf{X}^{t_b}$ is expected to approximate the true graph. Similarly, CAM relies on an additive structure to estimate a topological order by greedily maximizing data likelihood, and Score derives a topological order by approximating the score’s Jacobian. Consequently, the causal graphs learned on $\mathbf{X}^{t_b}$ using these methods are also considered reliable in our setting. However, the time series variants of PC and FCI may struggle to identify causal graphs due to potential violations of causal sufficiency and time dependency. Although the identifiability results of these variants in two time-slices are not guaranteed, our experiments show that they significantly outperform traditional time series algorithms.

Traditional temporal algorithms designed for time series setting: Besides, we also provide a dynamic time series method CD-NOD (Huang et al., 2020), which applies the PC algorithm for causal discovery on an augmented dataset that includes a time label to capture unobserved changing factors. Furthermore, we also incorporate three traditional causal discovery methods designed for standard time series data: Granger causality (Granger, 1969; Shojaie & Michailidis, 2010), VARLiNGAM (Hyvärinen et al., 2010) as baselines in our main experiments².

D THE MOTIVATION AND PSEUDO-CODE OF OUR PROPOSED DHT-CIT

In many applications, the time series sampling process may be slower than the timescale of causal processes, resulting in numerous previous time-slices being missing or unreliable. In the presence of unmeasured time-slices, relying solely on a single time-slice is insufficient for identifying causal relations. Therefore, our motivation is to use just two reliable time-slices to explore the summary causal graph of subsampled time series, rather than depending on all previous time-slices that are

²For CD-NOD, Granger causality and VARLiNGAM, we use the latest implementation from the causal-learn package.

available and reliable (the limitation of traditional methods). In this paper, we demonstrate that if two valid time slices at two arbitrary moments are available, the variables in the earlier slice can be used as conditional instrumental variables to replace interventions and improve topological ordering. This method significantly relaxes the assumption inherent in traditional time series studies that depend on modeling causal structures at the system timescale, causal sufficiency, and all time slices in the observation windows could be observed (Granger, 1969; 1980; Luo et al., 2015; Nauta et al., 2019; Runge et al., 2019; Runge, 2020; Bussmann et al., 2021; Löwe et al., 2022; Assaad et al., 2022). However, it is important to note that if no previous time-slice data is available or reliable, our approach, like other causal discovery algorithms, will not produce identifiable results.

Under Assumptions 1, 2 and 3, we show how two time-slice help topological ordering for learning causal relations. In such cases, we propose DHT-CIT, a novel topological sorting algorithm that utilizes conditional independence tests per node to distinguish between its descendant and non-descendant nodes and build a unique descendant hierarchical topology with a few spurious edges for identifying summary causal graph. Algorithm 1 shows the pseudo-code of our DHT-CIT.

Hardware used: Ubuntu 16.04.3 LTS operating system with 2 * Intel Xeon E5-2660 v3 @ 2.60GHz CPU (40 CPU cores, 10 cores per physical CPU, 2 threads per core), 256 GB of RAM, and 4 * GeForce GTX TITAN X GPU with 12GB of VRAM.

Software used: Python 3.8 with cdt 0.6.0, ylearn 0.2.0, causal-learn 0.1.3, GPy 1.10.0, igraph 0.10.4, scikit-learn 1.2.2, networkx 2.8.5, pytorch 2.0.0.

Algorithm 1 DHT-CIT: Descendant Hierarchical Topology with Conditional Independence Test

Input: Two time-slices $\mathcal{D} = \{X^{t_a}, X^{t_b}\}_{t_a < t_b}$ with d nodes; two significance threshold $\alpha = 0.01$ and $\beta = 0.001$ for conditional independence test and pruning process; the layer index $k = 0$.

Output: One adjacency matrix of Descendant Hierarchical Topology A^{TP} , one DAG $\mathcal{G}$.

Components: Conditional independence test $\text{HSIC}(\dots)$; and pruning process $\text{CAM}(\dots)$.

Stage 1 - Identifying Descendant Hierarchical Topology:

for $i = 1$ **to** d **do**

Construct the conditional set $X_{\otimes i}^{t_a}$ via an independence test $X_{\otimes i}^{t_a} = \{X_j^{t_a} \mid X_j^{t_a} \perp X_i^{t_a}\}$

for $j = 1$ **to** d **do**

$p_{i,j} = \text{HSIC}(X_i^{t_a}, X_j^{t_b} \mid X_{\otimes i}^{t_a})$

$a_{i,j}^{TP} = \mathbb{I}(p_{i,j} \leq \alpha)$

end for

end for

We obtain $P = \{p_{i,j}\}_{d \times d}$ and $A^{TP} = \{a_{i,j}^{TP}\}_{d \times d}$

Stage 2 - Adjusting the Topological Ordering:

while The causal relationship between the unprocessed nodes is a directed cyclic graph **do**

$k := k + 1$

$X_{M_{i,k}} = \{X_i^{t_a} / X_i^{t_b}, L_{1:k-1}\}$

$X_i^{t_b} \in L_k$, if $a_{i,j}^{TP} = 0$ for all $j \in M_{i,k}$

while $L_k = \emptyset$ **do**

$p_{i^*,j^*} := 2\alpha$ and $a_{i^*,j^*}^{TP} = 0$, $(i^*, j^*) = \arg \max_{i,j} (p_{i,j} \leq \alpha)$

$X_i^{t_b} \in L_k$, if $a_{i,j}^{TP} = 0$ for all $j \in M_{i,k}$

end while

We obtain $P = \{p_{i,j}\}_{d \times d}$ and $A^{TP} = \{a_{i,j}^{TP}\}_{d \times d}$

end while

Stage 3 - Pruning Spurious Edges:

We obtain $\mathcal{G} = \text{CAM}(\mathcal{D}, A^{TP}, \beta)$

Return: A^{TP} and $\mathcal{G}$

E ADDITIONAL EXPERIMENTS ON SYNTHETIC AND REAL-WORLD DATASETS

E.1 THE EXPERIMENTS ON MORE COMPLEX NON-LINEAR RELATIONSHIPS

Datasets. We test our algorithm on synthetic data generated from a *additive non-linear noise model* (Eq. 1) under Assumptions 1, 2 and 3. Given d nodes and e edges, we generate the causal graph $\mathcal{G}$

Table 4: The experiments on Sigmoid-10-10 & Poly-10-10 with observations ($\mathcal{D} = \{X^1, X^2\}$)

Sigmoid-10-10 data with observational data ($\mathcal{D} = \{X^1, X^2\}$)					
Method	SHD↓	SID↓	F1-Score↑	Dis.↓	#Prune↓
GOLEM	4.30±2.19	18.4±7.92	0.78±0.11	2.00±0.51	-
NOTEARS	12.5±5.40	45.3±17.9	0.46±0.21	3.44±0.78	-
ReScore	12.2±4.30	45.6±14.4	0.45±0.17	3.43±0.63	-
CAM	3.70±3.43	10.4±7.86	0.82±0.17	1.55±1.20	80.00±0.00
SCORE	9.90±3.81	32.8±11.6	0.56±0.16	3.09±0.61	38.90±1.60
DHT-CIT	0.67 ±1.12	1.80 ±2.99	0.96 ±0.06	0.46 ±0.72	8.67 ±2.92
Poly-10-10 data with observational data ($\mathcal{D} = \{X^1, X^2\}$)					
Method	SHD↓	SID↓	F1-Score↑	Dis.↓	#Prune↓
GOLEM	19.00±4.00	59.4±13.6	0.20±0.12	4.33±0.45	-
NOTEARS	17.8±5.36	56.4±16.9	0.23±0.18	4.16±0.64	-
ReScore	17.7±4.73	57.3±14.1	0.22±0.15	4.16±0.56	-
CAM	8.00±4.69	19.8±7.88	0.63±0.21	2.68±0.95	80.00±0.00
SCORE	18.90±4.33	40.4±10.9	0.23±0.13	4.32±0.52	42.20±1.48
DHT-CIT	3.22 ±3.15	10.8 ±5.69	0.84 ±0.15	1.51 ±1.03	11.33 ±3.87

using Erdos-Renyi model (Erdős & Rényi, 2011). In main experiments, we generate the data with Gaussian Noise for every variable X_i^τ , $i = 1, 2, \dots, d$ at time $\tau = 1, 2, \dots, t$:

$$X_i^\tau = f_i(\mathbf{pa}_i^\tau, X_i^{\tau-1}, \mathbf{pa}_i^{\tau-1}) + \epsilon_i^\tau, \epsilon_i^\tau \sim \mathcal{N}(0, 0.4 \cdot \mathbf{I}_d), \quad (6)$$

where f_i is a twice continuously differentiable arbitrary function in each component, and $\mathbf{I}_d$ is a d order identity matrix. To simulate real-world data as much as possible, we design 3 different non-linear functions non-linear($\cdot$) to discuss the performance of the HTS-CIT algorithm, for example:

$$\text{Sin}(\mathbf{pa}_i^\tau) = \sum_{j \in \text{pa}(X_i)} \sin(X_j^\tau), \quad (7)$$

$$\text{Sigmoid}(\mathbf{pa}_i^\tau) = \sum_{j \in \text{pa}(X_i)} \frac{3}{1 + \exp(-X_j^\tau)}, \quad (8)$$

$$\text{Poly}(\mathbf{pa}_i^\tau) = \sum_{j \in \text{pa}(X_i)} \frac{1}{10} (X_j^\tau + 2)^2. \quad (9)$$

In this paper, we use **Sin- d -e** to denote the synthetic dataset generated by non-linear function Sin($\cdot$):

$$X_i^\tau = \text{Sin}(\mathbf{pa}_i^\tau, X_i^{\tau-1}) + \frac{1}{10} \text{Sin}(\mathbf{w} \cdot \mathbf{pa}_i^{\tau-1}) + \epsilon_i^\tau. \quad (10)$$

where $\text{Sin}(\mathbf{pa}_i^\tau) = \sum_{j \in \text{pa}(X_i)} \sin(X_j^\tau)$, $\mathbf{I}_d$ is a d -th order identity matrix, and $\mathbf{w}$ is a random 0-1 vector that controls the number and existence of time-lagged edges from $\mathbf{pa}_i^{\tau-1}$. Similarly, we define Sigmoid- d -e and Poly- d -e.

Results. To simulate real-world data as much as possible, we design 2 additional non-linear functions to test the performance of our DHT-CIT, i.e., Sigmoid- d -e & Poly- d -e. The results (Tab. 4) demonstrate that our DHT-CIT remains superior for other complex nonlinear functions with low error edges in identifying causal graphs. In addition, the number of spurious edges that must be pruned in the topological ordering is also minimal compared to CAM and SCORE.

E.2 THE EXPERIMENTS ON LARGE GRAPHS WITH HIGH-DIMENSION VARIABLES

Datasets. Followed the data generation process (Eq. equation 5) in Section 5.2 in the main text. Given d nodes and e edges, we generate the causal graph $\mathcal{G}$ using the Erdos-Renyi model.

$$X_i^\tau = \text{Sin}(\mathbf{pa}_i^\tau, X_i^{\tau-1}) + \frac{1}{10} \text{Sin}(\mathbf{w} \cdot \mathbf{pa}_i^{\tau-1}) + \epsilon_i^\tau, \mathbf{X}^0 \sim \mathcal{N}(0, \mathbf{I}_d), \epsilon_i^\tau \sim \mathcal{N}(0, 0.4 \cdot \mathbf{I}_d) \quad (11)$$

where $\text{Sin}(\mathbf{pa}_i^\tau) = \sum_{j \in \text{pa}(X_i)} \sin(X_j^\tau)$, $\mathbf{I}_d$ is a d -th order identity matrix, and $\mathbf{w}$ is a random 0-1 vector that controls the number and existence of time-lagged edges from $\mathbf{pa}_i^{\tau-1}$. In this experiment,

Table 5: The experiments on Sin-50-50 & Sin-100-100 datasets.

Sin-50-50 data with Gauss noise ($\mathcal{D} = \{X^1, X^2\}$)						
Method	SHD↓	SID↓	F1-Score↑	Dis.↓	#Prune↓	Running Time(s)↓
GOLEM	87.9±11.0	846.5±166	0.24±0.10	9.35±0.58	-	1049.1s
ReScore	83.5±7.17	1044±65.4	0.31±0.06	9.13±0.39	-	455.2s
SCORE	17.4±6.17	91.4±49.7	0.85±0.04	4.11±0.74	1175±0.32	143.3s
DHT-CIT(50% Intervention)	21.3±6.85	102.±64.3	0.81±0.05	4.56±0.64	175.±8.17	829.1s
DHT-CIT(80% Intervention)	18.5±5.62	97.2±39.0	0.84±0.05	4.11±0.57	86.6±7.52	625.7s
DHT-CIT(100% Intervention)	16.4 ±4.40	88.4 ±37.1	0.86 ±0.04	4.00 ±0.58	58.8 ±8.52	327.1s

Sin-100-100 data with Gauss noise ($\mathcal{D} = \{X^1, X^2\}$)						
Method	SHD↓	SID↓	F1-Score↑	Dis.↓	#Prune↓	Running Time(s)↓
GOLEM	160.6±17.2	1898±764.1	0.25±0.08	12.6±0.68	-	3904.2s
ReScore	163.3±13.7	4009±549.3	0.34±0.05	12.7±0.55	-	578.2s
SCORE	32.6±4.71	211.4±39.7	0.86±0.02	5.70±0.42	4450±0.47	149.2s
DHT-CIT(50% Intervention)	37.6±6.30	219.4±64.3	0.83±0.02	6.11±0.49	306.4±17.6	3204.2s
DHT-CIT(80% Intervention)	29.5±6.26	180.9±47.2	0.87±0.02	5.40±0.56	248.0±27.0	1187.2s
DHT-CIT(100% Intervention)	26.7 ±6.80	160.5 ±55.3	0.88 ±0.03	5.13 ±0.63	197.1 ±24.8	761.2s

we set the number of time-lagged edges from other variables as 0. To evaluate our DHT-CIT on a larger graph with 50/100 nodes, we generate large graph **Sin-50-50** and **Sin-100-100**.

Although theoretically, DHT-CIT can achieve unbiased estimation, it is limited by the performance of conditional independence tests. For the conditional instrumental variables described above, we calculate the conditional independencies using the conditional independence HSIC test from (Zhang et al., 2011) with Gaussian kernel. However, as the data dimension increases, the accuracy of the HSIC test decreases, leading to incorrect topological orderings generated by DHT-CIT. To mitigate this issue, given two-time slices ($\mathcal{D} = \{X^1, X^2\}$), we implement random intervention to some nodes in the previous states of two time-slices, and then apply DHT-CIT to identify potential directed acyclic graphs. Based on the percentage of intervened nodes in the previous state, we refer to it as **DHT-CIT(50% Intervention)**, **DHT-CIT(80% Intervention)**, and **DHT-CIT(100% Intervention)**. Note that the time consumption of NOTEARS-MLP and CAM increases substantially (exceeds 5000s) as the graph size increases, thus, we do not implement NOTEARS-MLP and CAM.

Results. From the results on larger graphs (**Sin-50-50** and **Sin-100-100**) in Tab. 5, we have the following observation: (1) GOLEM and ReScore fail to identify the true DAG on larger graphs; (2) In terms of pruning efficiency, DHT-CIT outperforms SCORE by providing a more effective hierarchical topological ordering. The number of edges to be pruned in the topological ordering learned by SCORE is at least 10 times greater than that of the proposed DHT-CIT, which greatly increases the workload for subsequent pruning processes; (3) With at least 50% intervened nodes in the previous states, DHT-CIT(50% Intervention) can produce results that are comparable to the most advanced methods SCORE. As the proportion of intervened nodes in the previous states increases (exceed 80%), our approach (DHT-CIT(80% Intervention) and DHT-CIT(100% Intervention)) will gradually outperform SCORE. Two time-slices with random intervention will help to improve the identification of the topological ordering of the underlying DAG.

E.3 THE EXPERIMENTS ON REAL-WORLD DATASET

The **PM-CMR**³ (Wyatt et al., 2020) is a public time series data that is commonly used to study the impact of the particle (PM_{2.5}, T) on the cardiovascular mortality rate (CMR, Y) in 2132 counties in the US from 1990 to 2010. Additionally, the dataset includes 7 variables ($X_{1:7}$) related to the city status, which are potential common causes of both PM_{2.5} and CMR. The corresponding description of variables is detailed in Tab. 6. With the prior knowledge, i.e., $T \leftarrow X_{1:7} \rightarrow Y$ and $T \rightarrow Y$, we draw two time-slices in 2000 & 2010 to evaluate the performance of the proposed DHT-CIT and two well-performed baselines (GOLEM and SCORE).

Results. With the prior knowledge, i.e., $T \leftarrow X_{1:7} \rightarrow Y$ and $T \rightarrow Y$, we draw two time-slices in 2000 & 2010 to evaluate the performance of the proposed DHT-CIT and two well-performed baselines (GOLEM and SCORE). As illustrated in Fig. 3 and 4, both GOLEM and SCORE do

³PM-CMR: https://pasteur.epa.gov/uploads/10.23719/1506014/SES_PM25_CM2_data.zip

Table 6: The Description for Real Variables on PM-CMR Dataset.

Variable	Description
$PM_{2.5}(T)$	Annual county PM2.5 concentration, $\mu g/m^3$
$CMR(Y)$	Annual county cardiovascular mortality rate, deaths/100,000 person-years
$Unemploy(X_1)$	Civilian labor force unemployment rate in 2010
$Income(X_2)$	Median household income in 2009
$Female(X_3)$	Family households - female householder, no spouse present in 2010 / Family households in 2010
$Vacant(X_4)$	Vacant housing units in 2010 / Total housing units in 2010
$Owner(X_5)$	Owner-occupied housing units - percent of total occupied housing units in 2010
$Edu(X_6)$	Educational attainment - persons 25 years and over - high school graduate (includes equivalency) in 2010
$Poverty(X_7)$	Families below poverty level in 2009

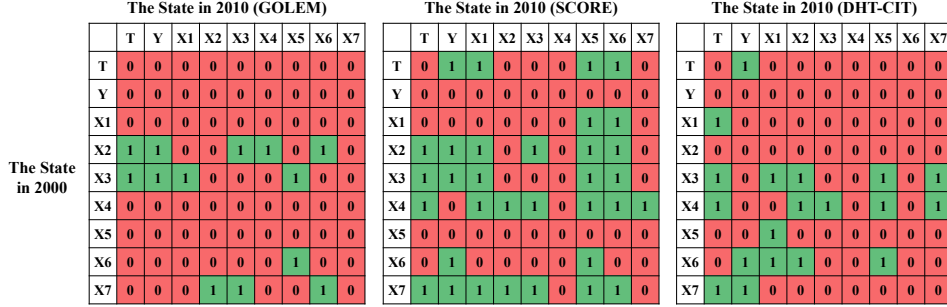


Figure 4: Learned Adjacency Matrix on PM-CMR Dataset.

not generate true summary causal graphs, and only our DHT-CIT achieves more accurate causal relationships in real-world data. GOLEM shows there is no direct edge from T to Y and SCORE shows that T is the parent node of $\{X_1, X_5, X_6\}$, which contradicts the prior knowledge. Only our DHT-CIT algorithm recovers the dense causal graph, i.e., $T \leftarrow X_{1:7} \rightarrow Y$ and $T \rightarrow Y$. The results are consistent with the experiments on denser graphs: both GOLEM and SCORE are only applicable to sparse graphs, whereas our DHT-CIT maintains superior performance and scalability to larger and denser graphs.

Notably, our algorithm demonstrates significant improvement on denser graphs (Fig. 2). In comparison to the best baseline, our algorithm boasts a 48% increase in SHD, a 48% increase in SID, and a 15% boost in F1-Score on **Sin-10-20**, and boasts a 30% increase in SHD, a 43% increase in SID, and a 7% boost in F1-Score on **Sin-10-30**. Most previous baselines were only applicable to sparse graphs, whereas our algorithm exhibits substantial improvements on dense graphs. Therefore, we believe that DHT-CIT provides a more precise DAG for the PM-CMR dataset. Therefore, to effectively combat cardiovascular disease, it is recommended that cities disseminate information about its dangers, promote prevention, and provide medical care for low-income families.

F RELAXING MARKOV ASSUMPTION TO HIGH-ORDER MARKOV MODELS

Notably, in this paper, we can relax the Markov Assumption to a high-order Markov Assumption. This means that the future time-slice X^{t+1} depends only on states $X^{t \cdots t-q+1}$ and does not directly depend on states $X^{1 \cdots t-q}$. Then, with $q+1$ time-slices ($X^{t_a \cdots t_a-q+1}$ and X^{t_b}), we can use $\mathbf{an}^{t_a \cdots t_a-q+1}$ to replace the condition set $\mathbf{an}^{t_a}$ to infer the Descendant-oriented Conditional Independence Criteria (Theorem 1). However, in this paper, we focus exclusively on the two-time-slices algorithm to demonstrate our theorem and algorithm. We have included a section to discuss the extension of the Markov Assumption to a high-order Markov Assumption in the Appendix.

G THE ADVANTAGES AND LIMITATIONS OF USING TWO TIME-SLICES DATA

Advantage: (1) Enhanced Topological Ordering: Traditional topology-based methods typically produce non-unique topological orderings with numerous spurious edges, resulting in decreased accuracy and efficiency in downstream search tasks. By using two time-slices as auxiliary instrumental variables, we can learn causal relations more efficiently, with a reduced search space and fewer spurious edges. (2) Feasibility in Intervention-Limited Contexts: Using interventional data can quickly identify (non-)descendants for each node and construct a more precise topological ordering. In scenarios where interventions are infeasible, unethical, or too costly, using two time-slices to replace intervention can be a practical alternative. (3) Reduced Data Requirements: In time series scenarios, traditional methods depend on the modeling causal structures at the system timescale, causal sufficiency, and all time slices in the observation windows could be observed. In this paper, we propose exploring limited time-slices, i.e., two reliable time-slices, to ease the data requirements. (4) Scaling to Non-linear and Non-Gaussian Models: We use two time-slices as conditional instrumental variables to simulate exogenous interventions. When applied to a variable, these simulated interventions only affect the variable’s value, and the permutation would propagate to its descendant nodes. Then, we can apply conditional independence tests to capture these intervention-related permutations for identifying each variable’s descendants, without requiring any structural or distributional assumptions about the data.

Limitations: Our DHT-CIT algorithm strictly relies on the Acyclic Summary Causal Graph and Consistency Throughout Time Assumptions, which are common in time series studies [Assaad et al., 2022]. Additionally, if no previous time-slice data is available or reliable, both the two time-slices approach and other causal discovery algorithms will not yield identifiable results.

H FUTURE APPLICATIONS

H.1 POTENTIAL APPLICATIONS OF OUR DHT-CIT ALGORITHM

The proposed DHT-CIT can be integrated as a module into any existing topology-based method to enhance the topological ordering. Additionally, our DHT-CIT algorithm is capable of identifying the true causal graph from the Markov equivalence classes that are typically learned using traditional methods. For instance, our DHT-CIT algorithm can effectively use the Descendant Hierarchical Topology to orient the undirected edges outputted by a constraint-based algorithm such as PCMC1 (Runge et al., 2019; Runge, 2020).

H.2 GENERALIZING THE DHT-CIT ALGORITHM TO OTHER DOMAINS

As long as the common time series causal assumptions in (Assaad et al., 2022) and the q -order Markov Assumption are satisfied, we can directly extend our algorithm to other domains and applications with $q+1$ time-slices ($\mathbf{X}^{t_a \cdots t_a - q + 1}$ and $\mathbf{X}^{t_b}$). For example, we can analyze the causal graph of city status variables in PM-CMR (Wyatt et al., 2020), and explore the relationships between various factors affecting soil moisture. In human genomics and gene expression, we also can establish two-time-slices causal relationships (surjections: where each expression variable can find a corresponding conditional instrumental variable in the genomic sequence variables). The challenges arise as different time series data may adhere to various high-order Markov Assumptions, which we need to identify. Additionally, sometimes the temporal transfer of events/processes might conceal causal relationships, requiring further extraction, such as the two-time-slices causal relationships between genomic and gene expression data. The DHT-CIT algorithm provides an excellent tool and opportunity for identifying the topological ordering in the aforementioned forms of data. We haven’t covered experiments in all areas in this paper due to the high cost of data acquisition. Future updates on these datasets will be shared on our project pages.