# Learning Causal Relations from Subsampled Time Series with Two Time-Slices

**Anpeng Wu** [1] **Haoxuan Li** [2] **Kun Kuang** [1] **Keli Zhang** [3] **Fei Wu** [1 4 5]

## Abstract

This paper studies the causal relations from subsampled time series, in which measurements are sparse and sampled at a coarser timescale than the causal timescale of the underlying system. In such data, because there are numerous missing time-slices (i.e., cross-sections at each time point) between two consecutive measurements, conventional causal discovery methods designed for standard time series data would produce significant errors. To learn causal relations from subsampled time series, a typical solution is to conduct different interventions and then make a comparison. However, full interventions are often expensive, unethical, or even infeasible, particularly in fields such as health and social science. In this paper, we first explore how readily available two-time-slices data can replace intervention data to improve causal ordering, and propose a novel **D**escendant **H**ierarchical **T**opology algorithm with **C**onditional **I**ndependence **T**est (**DHT-CIT**) to learn causal relations from subsampled time series using only two time-slices. Specifically, we develop a conditional independence criterion that can be applied iteratively to test each node from time series and identify all of its descendant nodes. 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). Most existing methods can well-identify causal structures from time series data by modeling these structures at the system's timescale, under the assumption of causal sufficiency. However, in practice, these methods may face challenges when measurements are sparse and sampled at a coarser timescale than the causal timescale of the system, as shown in Figure 1(a,b), there would be numerous missing time-slices[1] between two consecutive measurements (Gong et al., 2015; Plis et al., 2015b; Hyttinen et al., 2016; Peters et al., 2017). In such data, it has been demonstrated that full graph discovery is unidentifiable without prior knowledge of the subsampling rate and structural functions (Gong et al., 2015; Plis et al., 2015b; Hyttinen et al., 2016).

Recently, many works have been developed to study the causal ordering of the summary causal graph (Definition 3.1), also known as topological ordering, as depicted in Figure 1(c.II), in which a node in the ordering can only be a parent to nodes that appear after it in the same ordering (Dahlhaus & Eichler, 2003; Teyssier & Koller, 2005; Peters et al., 2014; Loh & Bühlmann, 2014; Park & Klabjan, 2017). To construct precise causal orderings for learning DAGs of summary causal graph[2], 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 Figure 1(b), intervening on $X_2^{t-3}$ would change the distribution of its descendants ($X_3^{t-3}, X_4^{t-3}$), enabling quick identification of them. Subsequently, we can construct a causal ordering in which each node is connected to its descendant nodes. The acyclicity constraint is automatically maintained because the causal ordering only represents ancestral and descendant relationships. However, while intervention experiments can significantly aid in accurately identifying the causal ordering, full interventions are often costly, ethically problematic, or even unfeasible (Wang et al., 2017; Yang et al., 2018).

Besides, while existing ordering-based methods could be well-generalized to the subsampled time series setting, these methods typically generate non-unique topological ordering with numerous spurious edges. For example, SCORE

[1]Department of Computer Science and Technology, Zhejiang University, Hangzhou, China [2]Center for Data Science, Peking University, Beijing, China [3]Huawei Noah's Ark Lab, Huawei, Shenzhen, China [4]Shanghai Institute for Advanced Study, Zhejiang University, Shanghai, China [5]Shanghai AI Laboratory, Shanghai, China. Correspondence to: Kun Kuang <kunkuang@zju.edu.cn>, Fei Wu <wufei@zju.edu.cn>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. 

[1]We refer to cross-sections at each time point as one time-slice.

[2]In discussing the summary graph's DAG, we omit self-loops.

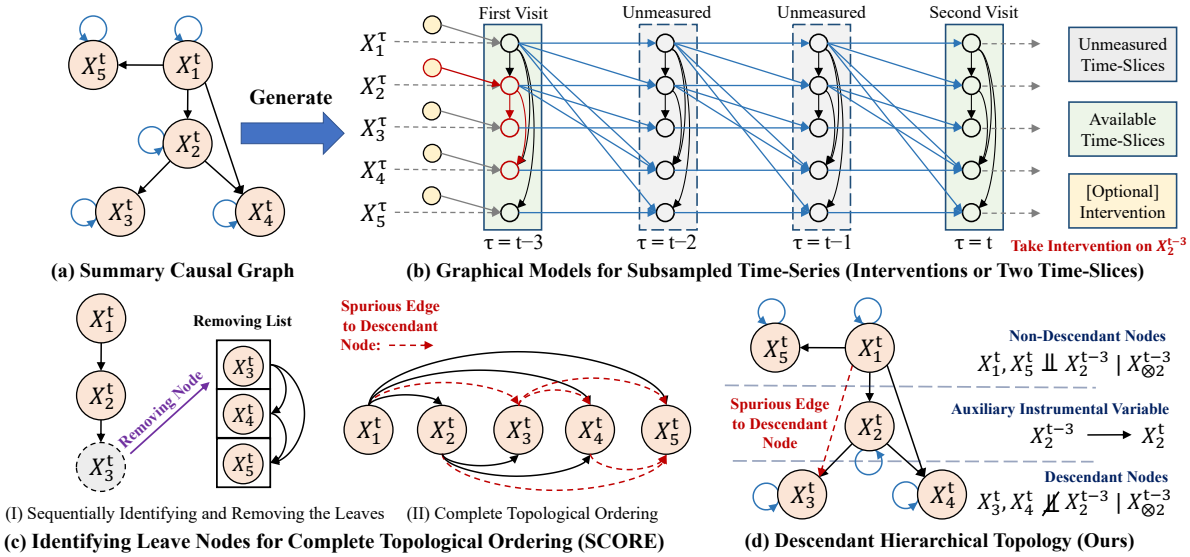



*Figure 1.* (a, b) Subsampled Time Series with Only Two Time-Slices. (c, d) The SCORE and DHT-CIT Algorithms. In clinical study, doctors typically compare earlier $\boldsymbol{X}^{t-3}$ and current $\boldsymbol{X}^{t}$ patient records to identify causes of outcome of interest. Patient visits may be recorded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

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 (Figure 1(c)) to approximate the true causal structure. However, the generated complete topological ordering is not only a non-unique topological ordering but also includes numerous spurious edges, which poses potential difficulties in downstream pruning tasks to identify directed acyclic graphs (DAGs) of summary causal graphs (Figure 1(a) (Rolland et al., 2022; Sanchez et al., 2022)). Therefore, in the subsampled time series setting, when full interventions are impractical, this paper explores a novel conditional independence criterion to learn descendant hierarchical topology (DHT) using only two time-slices.

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, where doctors typically use limited time-slices to analyze a patient's condition and determine treatments. Therefore, in this paper, we study the causal ordering for the summary causal graph on subsampled time series using only two time-slices. Inspired by the intervention in Figure 1(b), we find that each node is influenced by its ancestors and itself, then the earlier time-slice would transmit perturbations to both itself and its descendants in the subsequent time-slice, serving as simulated interventions. Treating the earlier perturbations as auxiliary instrumental variables, we propose a **D**escendant **H**ierarchical **T**opology algorithm with **C**onditional **I**ndependence **T**est (**DHT-CIT**) to quickly identify (non-)descendants for each node (Figure 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 summary causal graph. Empirical results on both synthetic and real-world datasets demonstrate the superiority of our DHT-CIT algorithm.

## 2. Related Work

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 (Gong et al., 2015; 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. For nonlinear data, by analyzing loop lengths and strongly connected components (SCCs) in compressed graphs, Danks & Plis (2013) provide theorems and algorithms to infer partial information about the structure of the compressed graph. Leveraging these findings and some additional structural insights about constraint satisfaction problems, Abavisani et al. (2023) generalizes the search-based RASL algorithm (Plis et al., 2015a) and proposes a new sRASL algorithm for learning the true directed causal structure from subsampled time-series. Inspired by works (Gong et al., 2015; Plis et al., 2015b), Hyttinen 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.

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 are placed in Appendix A.

## 3. Problem Setup

**Standard Time Series.** Let $\boldsymbol{X} = \{X_i^\tau\}_{d\times t}$ denote the full $d$-variate time series with all time slices $\boldsymbol{X}^\tau$ at $t$ time points, where $X_i^\tau, i \in \{1, 2, \cdots, d\}$ and $\tau \in \{1, 2, \cdots, t\}$, is a random vector comprising observations of $n$ samples. For simplicity of notation, we will not discuss each sample individually. Instead, we refer to $X_i^\tau$ 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 $X_i^\tau$, we use the notation $\mathbf{pa}_i^\tau$ to specify the partial set of parents of $X_i^\tau$ (as well as for $X_i^{\tau+1}$) at time $\tau$. Similarity, 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 Figure 1(a,b), following the summary causal graph (Definition 3.1), the causal structure can be expressed in the functional relationship, for $i \in \{1, 2, \cdots, d\}$ and $\tau \in \{1, 2, \cdots, t\}$:

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

where $f_i(\cdot)$ is a twice continuously differentiable function, which embeds the instantaneous effects from its parents $\mathbf{pa}_i^\tau$ at time $\tau$ and time-lagged effects from previous variable $\boldsymbol{X}^{\tau-1}$; and $\epsilon_i^\tau$ denotes the *Additive Noise* term at time $\tau$.

**Definition 3.1** (Summary Causal Graph)**.** The summary causal graph is the directed graph with nodes $X_1, ..., X_d$ containing an arrow from $X_j$ to $X_k$ for $j \neq k$ whenever there is a direct arrow from $X_j^{t_a}$ to $X_k^{t_b}$ for some $t_a \leq t_b$, and an optional self-loop arrow in $X_i$ for all $i \in \{1, \cdots, d\}$.

**Subsampled Time Series with Two Time-Slices.** In this paper, we focus on learning the directed acyclic graphs $\mathcal{G}$ of the summary causal graph (Figure 1(a)) on subsampled time series (Figure 1(b)) with instantaneous effects using only two time-slices $\mathcal{D} = \{\boldsymbol{X}^{t_a}, \boldsymbol{X}^{t_b}\}_{1<t_a<t_b<t}$. Based on a conditional independence criterion using earlier time-slice $\boldsymbol{X}^{t_a}$ as conditional instrumental variables, we can easily distinguish descendants and non-descendants for each node in the current time-slice $\boldsymbol{X}^{t_b}$. We rigorously prove that it is complete under graph constraints, i.e., Markov property, acyclic summary causal graph, stationary full-time graph. More discussion about advantages and limitations of two time-slices is deferred to Appendix I.

**Assumption 3.2** (Markov Property)**.** The Markov property of time series assumes the future slice $\boldsymbol{X}^{t+1}$ depends on current state $\boldsymbol{X}^t$ but does not depend on history $\boldsymbol{X}^{1\cdots t-1}$.

This is just for illustration, and later we can relax this assumption as high-order Markov assumption subsampled time series with high-order lagged effect in Appendix F.3.

**Assumption 3.3** (Acyclic Summary Causal Graph, Section 5.2.1 in Assaad et al. (2022)))**.** The summary causal graph of a 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 3.4** (Consistency Throughout Time, Definition 7 in Assaad et al. (2022)))**.** A causal graph $\mathcal{G}$ for a multivariate time series $\boldsymbol{X}$ 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 the summary graph underlying the additive noise models (Eq. (1)). Further discussion on the assumptions made in this paper is provided in Appendix F.

## 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. Descendant Hierarchical Topology

As shown in Figure 1(c), the conventional topology-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 4.1** (Complete Topological Ordering)**.** The complete topological ordering ($\pi(\boldsymbol{X}) = (X_{\pi_1}, X_{\pi_2}, \cdots, X_{\pi_d})$, $\pi_i$ is the reordered index of node) is a sorting of all nodes in a DAG such that for any pair of nodes $X_{\pi_i}$ and $X_{\pi_j}$, if there exists a directed edge from $X_{\pi_i}$ to $X_{\pi_j}$, 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, as shown in Figure 1(d), obtaining two time-slices data can help identify a unique hierarchical topological ordering, 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 4.2** (Hierarchical Topological Ordering)**.** In the hierarchical topological ordering e.g., $\Pi(\boldsymbol{X}) = (\{X_{\pi_1}\}_{\boldsymbol{L}_1}, \{X_{\pi_2}, X_{\pi_3}\}_{\boldsymbol{L}_2}, \cdots)$, each layer is denoted by $\boldsymbol{L}_i$ and the located layer of $X_j$ are represented as $l_j$. If there is a directed edge from $X_{\pi_i}$ to $X_{\pi_j}$, then $l_{\pi_i} > l_{\pi_j}$.

**Definition 4.3** (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 descendants $\mathbf{de}_i^t$, i.e., $X_i^t \to \mathbf{de}_i^t, i \in \{1, 2, \cdots, d\}$.

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. Descendant Conditional Independence Criteria

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

**Theorem 4.4** (Descendant-Oriented Conditional Independence Criteria)**.** *Given observations $\mathcal{D} = \{\boldsymbol{X}^{t_a}, \boldsymbol{X}^{t_b}\}_{t_a<t_b}$ satisfying Assumptions 3.2, 3.3, and 3.4, for variables $X_i^{t_a}$ and $X_i^{t_b}$, where $i \in \{1, 2, \cdots, d\}$, 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.* From the the non-zero *time-lagged effect* and Assumptions 3.2, 3.3, and 3.4, we can infer that:
(a) The effect of $X_i^{t_a}$ on $X_i^{t_b}$ is non-zero, i.e., $X_i^{t_a} \dashrightarrow X_i^{t_b}$;
(b) Under Markov property, $\boldsymbol{X}^{\tau} \not\to X_i^{t_b}$ for $\tau < t_a < t_b$;
(c) Under acyclic assumption, $X_i^{t_a} \not\to \mathbf{an}_i^{t_b}$ for $t_a < t_b$;
(d) Under stationary time series, $X_i^{t_a} \dashrightarrow X_j^{t_a} \dashrightarrow X_j^{t_b}$.
Under conditions (a), (b), (c) and (d), if $X_j^{t_b} \in \mathbf{an}_i^{t_b}$, then there are only two causal paths between $X_i^{t_a}$ and $X_j^{t_b}$: $X_i^{t_a} \dashleftarrow \mathbf{an}_i^{t_a} \dashrightarrow X_j^{t_b}$ and $X_i^{t_a} \dashrightarrow \{X_i^{t_b}, \mathbf{de}_i^{t_b}\} \dashleftarrow X_j^{t_b}$. Hence, once we cut off all backdoor paths by controlling the conditional set $\mathbf{an}_i^{t_a}$, then the confounding effect between $X_i^{t_a}$ and $X_j^{t_b}$ would be eliminated, leading to $X_i^{t_a} \perp\!\!\!\perp X_j^{t_b} \mid \mathbf{an}_i^{t_a}$. Similarity, if $X_j^{t_b} \in \mathbf{sib}_i^{t_b}$, then the summary backdoor path is $X_i^{t_a} \dashleftarrow \mathbf{an}_i^{t_a} \dashrightarrow \mathbf{an}_j^{t_b} \dashrightarrow X_j^{t_b}$. 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 $X_i^{t_b}$. □

However, since the causal graph is unknown, we are unable to directly determine the ancestor nodes $\mathbf{an}_i^{t_a}$. Therefore, we select all variables at time $\tau$ except for $X_i^{t_a}$ and any variables that are independent of $X_i^{t_a}$, as the conditional set $\boldsymbol{X}_{\otimes i}^{t_a}$. This means that $X_i^{t_a} \not\perp\!\!\!\perp X_j^{t_a}$ for each variable $X_j^{t_a} \in \boldsymbol{X}_{\otimes i}^{t_a}$. As the events in conditional set $\boldsymbol{X}_{\otimes i}^{t_a}$ occurs before current time $t$, $\boldsymbol{X}_{\otimes i}^{t_a}$ does not introduce additional backdoor paths to non-descendant nodes at time $t$, nor can it block the path $X_i^{t_a} \dashrightarrow X_i^{t_b} \dashrightarrow X_j^{t_b}$. Thus, we can reformulate the Theorem 4.4 using the conditional set $\boldsymbol{X}_{\otimes i}^{t_a}$.

**Corollary 4.5.** *Given observations $\mathcal{D} = \{\boldsymbol{X}^{t_a}, \boldsymbol{X}^{t_b}\}_{t_a<t_b}$, for variables $X_i$ and $X_j$ where $i, j \in \{1, 2, \cdots, d\}$, $X_j$ is a descendant node of $X_i$ iff $X_i^{t_a} \not\perp\!\!\!\perp X_j^{t_b} \mid \boldsymbol{X}_{\otimes i}^{t_a}$.*

Based on the corollary 4.5, we can distinguish between descendant and non-descendant nodes of each variable $X_i$ by conducting a single conditional independence test per variable ($X_i^{t_a} \not\perp\!\!\!\perp \mathbf{de}_i^{t_b} \mid \boldsymbol{X}_{\otimes i}^{t_a}$). Additionally, if we perform a random intervention on $X_i^{t_a}$, the conditional set in corollary 4.5 will be empty because the random intervention is independent of the other variables. As a result, the conditional independence test in corollary 4.5 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. Descendant Hierarchical Topology

Based on the conditional independence criteria in corollary 4.5, using previous time-slice $\boldsymbol{X}^{t_a}_{\otimes i} = \{X^{t_a}_j \mid X^{t_a}_j \perp\!\!\!\perp X^{t_a}_i\}$ as AIVs, we can identify descendants $\mathbf{de}^{t_b}_i$ of each variable $X^{t_a}_i$ by a single conditional independence test per variable ($X^{t_a}_i \not\perp\!\!\!\perp \mathbf{de}^{t_b}_i \mid \boldsymbol{X}^{t_a}_{\otimes i}$). For every $i, j \in \{1, 2, \cdots, d\}$, we calculate the conditional independence significance $\boldsymbol{P}$ using the conditional HSIC test with Gaussian kernel (Zhang et al., 2011). We determine that $X_i$ is a descendant of $X_j$ if the reported $p$-value is less than or equal to a threshold $\alpha$, i.e., $X^{t_a}_i \not\perp\!\!\!\perp X^{t_b}_j \mid \boldsymbol{X}^{t_a}_{\otimes i}$, and the adjacency matrix of the unique descendant hierarchical topology can be obtained via

$$\boldsymbol{P} = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,d} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ p_{d,1} & p_{d,2} & \cdots & p_{d,d} \end{pmatrix}, \tag{2}$$

$$\boldsymbol{A}^{TP} = \begin{pmatrix} \mathbb{I}(p_{1,1} \le \alpha) & \mathbb{I}(p_{1,2} \le \alpha) & \cdots & \mathbb{I}(p_{1,d} \le \alpha) \\ \mathbb{I}(p_{2,1} \le \alpha) & \mathbb{I}(p_{2,2} \le \alpha) & \cdots & \mathbb{I}(p_{2,d} \le \alpha) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{I}(p_{d,1} \le \alpha) & \mathbb{I}(p_{d,2} \le \alpha) & \cdots & \mathbb{I}(p_{d,d} \le \alpha) \end{pmatrix}, \tag{3}$$

where $p_{i,j} = \mathbf{HSIC}(X^{t_a}_i, X^{t_b}_j \mid \boldsymbol{X}^{t_a}_{\otimes i})$, $\alpha$ is a hyper-parameter denoting significance threshold, and $\mathbb{I}(\cdot)$ is the indicator function. If the $p$-value is less than $\alpha$, the result is considered significant and an edge is added in the descendant hierarchical topology. In statistical hypothesis testing, $\alpha$ 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 testing (Zhang et al., 2011; Runge, 2018; Bellot & van der Schaar, 2019), it remains a complex task, especially in high-dimensional scenarios. This complexity can lead to biased topological ordering that includes cycles. To address this challenge, our approach introduces topological ordering adjustment as a dual safeguard to ensure acyclicity in the causal discovery process.

#### 4.3.2. Topological Ordering Adjustment

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 leaf nodes of the descendant hierarchical topology layer by layer. Specifically, those nodes that do not have any descendants are classified as leaf nodes. We iteratively identify all leaf nodes of the descendant hierarchical topology as a leaf layer, and then delete them from the topology. Firstly, we denote all variables at time-slice $\tau$, except for $X^\tau_i$, as $\boldsymbol{X}^\tau_{-i}$ for $\tau \in \{1, 2, \cdots, t\}$. At $k$-th leaf layer $\boldsymbol{L}_k$, if $X^{t_b}_i \perp\!\!\!\perp \boldsymbol{X}^{t_a}_{-i} \mid \boldsymbol{X}^{t_a}_{\otimes i}$, then $X^{t_b}_i$ is a leaf node at time-slice $\tau = t_b$ and $X_i \in \boldsymbol{L}_k$. By repeating this operation, we can iteratively $k := k + 1$ and identify the current leaf layer $\boldsymbol{L}_k$:

$$X^{t_b}_i \in \boldsymbol{L}_k, \text{ if } a^{TP}_{i,j} = 0 \text{ for all } j \in M_{i,k}, \tag{4}$$

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

**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 $\alpha$ and change it to $2\alpha$, deleting the corresponding edge in the topology

$$\begin{aligned} p_{(i^*,j^*)} := 2\alpha \quad \text{and} \quad a^{TP}_{i^*,j^*} = 0, \\ (i^*, j^*) = \arg\max_{i,j}(p_{i,j} \le \alpha). \end{aligned} \tag{5}$$

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 $\boldsymbol{A}^{TP} = \{a^{TP}_{i,j}\}_{d \times d}$. This ensures that the graph's topological ordering is acyclic and improves the accuracy of learned graphs.

#### 4.3.3. Pruning Spurious Edges

Based on a conditional independence criterion and two time-slices data, as depicted in Figure 1(d), we propose a DHT-CIT algorithm to construct a more efficient descendant hierarchical topology with merely a few spurious edges. Theoretically, conditional independence in hierarchical topological layer ordering enables a pruning process that requires only a limited set of nodes - either nodes from one higher layer, the current layer, and two lower layers, or the node's non-descendants and one lower layer's nodes - to determine the existence of spurious edges between nodes. In contrast, classical methods like CAM, which rely on significance testing with generalized additive models and a $p$-value threshold of 0.001, often show superior practical performance (Bühlmann et al., 2014). Consequently, following Rolland et al. (2022), we utilize CAM for pruning spurious edges in our approach. The detailed pseudo-code for this method is provided in Algorithm 1 in Appendix D.

*Table 1.* The results (mean$_{\pm std}$ ) on Sin-$d$-$e$ using observational data ($\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^2\}$).

| | **Sin-10-10 Graph with Observational Data** ( $\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^2\}$ ) | | | | | **Sin-20-20 Graph with Observational Data** ( $\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^2\}$ ) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | **SHD**↓ | **SID**↓ | **F1-Score**↑ | **Dis.**↓ | **#Prune**↓ | **SHD**↓ | **SID**↓ | **F1-Score**↑ | **Dis.**↓ | **#Prune**↓ |
| PC | $12.8_{\pm 5.03}$ | $43.6_{\pm 9.94}$ | $0.56_{\pm 0.12}$ | $3.51_{\pm 0.72}$ | - | $21.5_{\pm 6.75}$ | $98.2_{\pm 31.8}$ | $0.61_{\pm 0.11}$ | $4.59_{\pm 0.69}$ | - |
| FCI | $15.3_{\pm 3.77}$ | $71.0_{\pm 11.5}$ | $0.54_{\pm 0.09}$ | $3.89_{\pm 0.46}$ | - | $30.5_{\pm 4.09}$ | $237._{\pm 59.1}$ | $0.54_{\pm 0.05}$ | $5.51_{\pm 0.37}$ | - |
| GOLEM | $\mathbf{0.50}_{\pm 0.80}$ | $1.80_{\pm 2.70}$ | $\mathbf{0.97}_{\pm 0.03}$ | $\mathbf{0.38}_{\pm 0.59}$ | - | $1.30_{\pm 1.10}$ | $5.60_{\pm 4.40}$ | $0.97_{\pm 0.03}$ | $0.93_{\pm 0.66}$ | - |
| NOTEARS | $1.20_{\pm 0.60}$ | $2.30_{\pm 1.20}$ | $0.94_{\pm 0.02}$ | $1.02_{\pm 0.30}$ | - | $2.60_{\pm 1.49}$ | $6.00_{\pm 3.40}$ | $0.94_{\pm 0.03}$ | $1.55_{\pm 0.46}$ | - |
| ReScore | $1.00_{\pm 0.63}$ | $\mathbf{1.40}_{\pm 1.36}$ | $0.95_{\pm 0.03}$ | $0.88_{\pm 0.47}$ | - | $2.00_{\pm 0.77}$ | $5.10_{\pm 2.90}$ | $0.95_{\pm 0.01}$ | $1.38_{\pm 0.28}$ | - |
| Granger | $31.3_{\pm 11.6}$ | $66.8_{\pm 30.8}$ | $0.21_{\pm 0.04}$ | $5.48_{\pm 1.10}$ | - | $104_{\pm 20.7}$ | $368_{\pm 8.82}$ | $0.10_{\pm 0.03}$ | $10.1_{\pm 1.01}$ | - |
| VarLiNGAM | $35.0_{\pm 0.00}$ | $69.4_{\pm 3.20}$ | $0.36_{\pm 0.00}$ | $5.91_{\pm 0.00}$ | - | $170_{\pm 0.00}$ | $339_{\pm 3.20}$ | $0.19_{\pm 0.00}$ | $13.0_{\pm 0.00}$ | - |
| CD-NOD | $5.40_{\pm 0.92}$ | $15.5_{\pm 4.70}$ | $0.74_{\pm 0.04}$ | $2.32_{\pm 0.19}$ | - | - | - | - | - | - |
| CAM | $3.70_{\pm 2.95}$ | $13.2_{\pm 10.6}$ | $0.84_{\pm 0.13}$ | $1.79_{\pm 0.74}$ | $80.00_{\pm 0.00}$ | $10.3_{\pm 6.50}$ | $41.6_{\pm 34.7}$ | $0.79_{\pm 0.12}$ | $3.07_{\pm 0.98}$ | $360.0_{\pm 0.00}$ |
| SCORE | $5.60_{\pm 3.92}$ | $21.2_{\pm 16.1}$ | $0.78_{\pm 0.14}$ | $2.25_{\pm 0.78}$ | $35.80_{\pm 0.98}$ | $7.40_{\pm 2.41}$ | $31.3_{\pm 21.7}$ | $0.85_{\pm 0.04}$ | $2.68_{\pm 0.47}$ | $172.1_{\pm 0.22}$ |
| **DHT-CIT** | $1.00_{\pm 1.22}$ | $3.20_{\pm 3.70}$ | $0.95_{\pm 0.05}$ | $0.68_{\pm 0.72}$ | $\mathbf{13.20}_{\pm 4.30}$ | $\mathbf{1.00}_{\pm 1.32}$ | $\mathbf{3.10}_{\pm 4.40}$ | $\mathbf{0.98}_{\pm 0.03}$ | $\mathbf{0.51}_{\pm 0.61}$ | $\mathbf{30.60}_{\pm 7.70}$ |

* CD-NOD on **Sin-20-20** takes over 5 hours and **#Prune** on one-stage methods is not meaningful. We don't discuss these results and represent them with '-'.

*Table 2.* The results (mean$_{\pm std}$ ) on Sigmoid-10-10 & Poly-10-10 data.

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

## 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-seires 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. Additionally, this paper primarily aims to enhance causal ordering for learning causal relations from subsampled time series. To this end, we compare topology-based methods by quantifying the number of spurious edges requiring pruning, denoted as **#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 3.2, 3.3 and 3.4. Given $d$ nodes and $e$ edges, we generate the causal graph $\mathcal{G}$ using Erdos-Renyi model (Erdös & Rényi, 2011). In main experiments, we generate the data with Gaussian Noise for every variable $X_i^\tau$, $i = 1, 2, \cdots, d$ at time $\tau = 1, 2, \cdots, t$:

$$X_i^\tau = \text{Sin}\left(\mathbf{pa}_i^\tau, X_i^{\tau-1}\right) + \tfrac{1}{10}\text{Sin}\left(\mathbf{w} \cdot \mathbf{pa}_i^{\tau-1}\right) + \epsilon_i^\tau, \quad \boldsymbol{X}^0 \sim \mathcal{N}\left(0, \mathrm{I}_d\right), \boldsymbol{\epsilon}^\tau \sim \mathcal{N}\left(0, 0.4 \cdot \mathrm{I}_d\right), \tag{6}$$

where $\text{Sin}(\mathbf{pa}_i^\tau) = \sum_{j \in \text{pa}(X_i)} \sin(X_j^\tau)$, $\mathrm{I}_d$ 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 $\mathbf{pa}_i^{\tau-1}$. To evaluate the performance of our DHT-CIT across various scenarios using observations $\mathcal{D} = \{\boldsymbol{X}^{t_a}, \boldsymbol{X}^{t_b}\}$ with subsampleing rate $u = t_b - t_a$, we vary the number of nodes ($d$) and edges ($e$) to generate larger and denser graphs, which we refer to as **Sin-$d$-$e$**. To simulate real-world data

*Table 3.* The experiments on different noise types.

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

as much as possible, we design 2 additional non-linear functions to test the performance of our DHT-CIT, i.e., Sigmoid-$d$-$e$ with $\text{Sigmoid}(X) = \frac{3}{1+\exp(-X)}$ and Poly-$d$-$e$ with $\text{Poly}(X) = \frac{1}{10}(X+2)^2$. Moreover, to test the algorithm's robustness against different noise types, we also generate data with Laplace noise ($X_i^0, \boldsymbol{\epsilon}^\tau \sim \text{Laplace}(0, 1/\sqrt{2})$) 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 error.

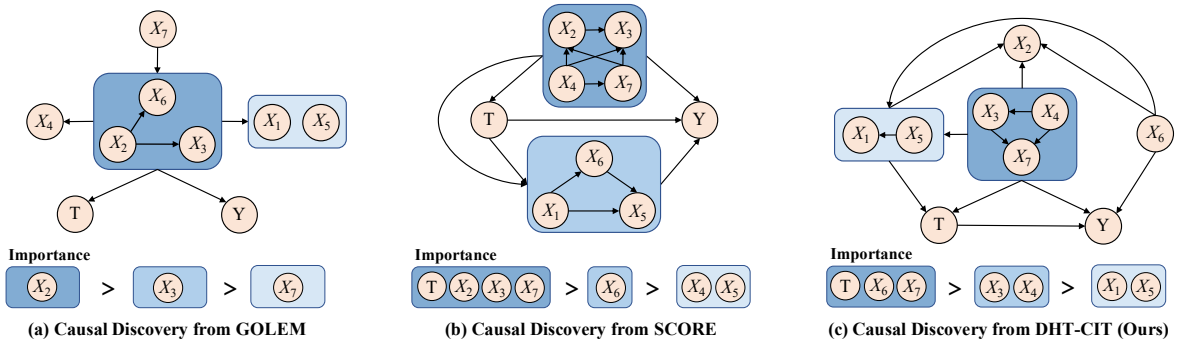
The experiments on varying time-lagged edges are deferred to Figure 3 in Appendix, while in the main experiments, we set the number of time-lagged edges from $\mathbf{pa}_i^{\tau-1}$ as 0. Additional, experiments on exploring **time costly** and **large graph** are deferred to Appendix H.2 and H.3.

**Studying Two Time-Slices without Sub-Sampling on Sparse Graph and Different Non-Linear Functions.** From the results on sparse graphs (**Sin-10-10** and **Sin-20-20**) in Table 1, we have the following observation: (1) In two time-slices settings, the time series variants of PC and FCI algorithms are limited to identifying Markov equivalence classes and struggle with non-linear relations. (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 ($\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^2\}$). However, their performance deteriorates on Sin-20-20 and relies on causal sufficiency and time dependency. (3) Traditional time series algorithms like Granger and VARLiNGAM, which require multiple time slices, struggle with identifying the summary causal graph using only two time-slices. 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. However, its performance decreases when dealing with observational data ($\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^2\}$) 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 descendant 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%.

Moreover, the results in Table 2 verify that our DHT-CIT excels in identifying causal graphs for complex nonlinear functions with fewer erroneous edges. Compared to CAM and SCORE, our DHT-CIT significantly lowers the number of spurious edges needing pruning in topological ordering, which demonstrates the scalability and superiority of our DHT-CIT method across diverse functions.

**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 in Table 3 demonstrate the superior and robust performance of DHT-CIT against different types of noise, with the accuracy consistently comparable to that under Gaussian noise.

**Studying Two Time-Slices $\mathcal{D} = \{\boldsymbol{X}^{t_a}, \boldsymbol{X}^{t_b}\}$ with Varying Sub-Sampling Rate $u = t_b - t_a$ on Denser Graphs.** Due to the limited space and constraints of most models, in this section, we only compare our method against the best baseline GLOEM and the advanced SCORE method in denser graphs from two time-slices $\mathcal{D} = \{\boldsymbol{X}^{t_a}, \boldsymbol{X}^{t_b}\}$ with varying sub-sampling rate $u = t_b - t_a$. We evaluate their efficacy using SID, SHD, and #Prune metrics. From the results on Sigmoid-d-e with varying numbers of nodes and edges (Sub-sampling Rate $u = 3$, Table 4), we have the following observations: as the graph becomes denser, that is, as the number of edges increases, the performance of DHT-CIT gradually declines, leading to an increase in erroneous edges. Conversely, SCORE tends to outperform DHT-CIT in extremely dense graphs (Sigmoid-10-40). Because, in extremely dense graphs, the causal relationships in the summary graph become similar to the complete graph identified by SCORE, while the performance of DHT-CIT declines due to the increasing complexity of conditional independence tests. Therefore, our DHT-CIT method is more suited to relatively sparse graphs, for example, graphs with 10 nodes and up to 30 edges (the complete topology graph has 45 edges), or graphs with 20 nodes and up to 100 edges (the complete topology graph has 190 edges). In extremely




(a) Causal Discovery from GOLEM (b) Causal Discovery from SCORE (c) Causal Discovery from DHT-CIT (Ours)

*Figure 2.* Causal Discovery on the PM-CMR Dataset.

*Table 4.* The experiments (mean$_{\pm std}$) on Sigmoid-d-e using observations $D = \{X^{t_a}, X^{t_b}\}$ with Subsampling Rate $u = t_b - t_a$.

| | **Sigmoid-10-20 on $\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^4\}$** | | | **Sigmoid-10-30 on $\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^4\}$** | | | **Sigmoid-10-40 on $\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^4\}$** | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | **SHD↓** | **SID↓** | **#Prune↓** | **SHD↓** | **SID↓** | **#Prune↓** | **SHD↓** | **SID↓** | **#Prune↓** |
| GOLEM | $10.70_{\pm 2.93}$ | $63.70_{\pm 11.85}$ | - | $26.90_{\pm 4.83}$ | $71.40_{\pm 8.59}$ | - | $35.20_{\pm 3.25}$ | $67.00_{\pm 11.20}$ | - |
| SCORE | $15.10_{\pm 3.65}$ | $53.10_{\pm 12.95}$ | $31.20_{\pm 1.60}$ | $14.80_{\pm 5.71}$ | $46.40_{\pm 8.04}$ | $20.80_{\pm 1.60}$ | $23.60_{\pm 2.24}$ | $38.40_{\pm 13.46}$ | $10.30_{\pm 2.00}$ |
| **DHT-CIT** | $\mathbf{6.30}_{\pm 2.90}$ | $\mathbf{25.30}_{\pm 13.66}$ | $\mathbf{13.50}_{\pm 2.91}$ | $\mathbf{14.10}_{\pm 4.46}$ | $\mathbf{38.80}_{\pm 11.02}$ | $\mathbf{11.10}_{\pm 1.58}$ | $\mathbf{23.60}_{\pm 2.24}$ | $\mathbf{41.20}_{\pm 6.90}$ | $\mathbf{6.80}_{\pm 2.48}$ |
| | **Sigmoid-20-20 on $\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^4\}$** | | | **Sigmoid-20-60 on $\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^4\}$** | | | **Sigmoid-20-100 on $\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^4\}$** | | |
| Method | **SHD↓** | **SID↓** | **#Prune↓** | **SHD↓** | **SID↓** | **#Prune↓** | **SHD↓** | **SID↓** | **#Prune↓** |
| GOLEM | $26.0_{\pm 5.60}$ | $138.0_{\pm 47.15}$ | - | $60.10_{\pm 5.49}$ | $322.3_{\pm 23.84}$ | - | $100.0_{\pm 5.32}$ | $336.4_{\pm 16.19}$ | - |
| SCORE | $8.40_{\pm 6.20}$ | $39.10_{\pm 38.82}$ | $173.2_{\pm 1.99}$ | $37.10_{\pm 8.14}$ | $257.9_{\pm 34.38}$ | $144.7_{\pm 4.27}$ | $57.5_{\pm 11.00}$ | $266.4_{\pm 48.18}$ | $112.4_{\pm 3.10}$ |
| **DHT-CIT** | $\mathbf{0.70}_{\pm 0.90}$ | $\mathbf{3.20}_{\pm 3.49}$ | $\mathbf{30.10}_{\pm 9.84}$ | $\mathbf{22.10}_{\pm 3.75}$ | $\mathbf{173.5}_{\pm 38.71}$ | $\mathbf{58.8}_{\pm 6.52}$ | $\mathbf{53.5}_{\pm 8.43}$ | $\mathbf{233.3}_{\pm 31.78}$ | $\mathbf{75.4}_{\pm 6.05}$ |
| | **Sigmoid-10-20 on $\mathcal{D} = \{\boldsymbol{X}^2, \boldsymbol{X}^4\}$** | | | **Sigmoid-10-20 on $\mathcal{D} = \{\boldsymbol{X}^2, \boldsymbol{X}^6\}$** | | | **Sigmoid-10-20 on $\mathcal{D} = \{\boldsymbol{X}^2, \boldsymbol{X}^{10}\}$** | | |
| Method | **SHD↓** | **SID↓** | **#Prune↓** | **SHD↓** | **SID↓** | **#Prune↓** | **SHD↓** | **SID↓** | **#Prune↓** |
| GOLEM | $17.40_{\pm 4.96}$ | $58.40_{\pm 13.81}$ | - | $21.60_{\pm 4.50}$ | $67.20_{\pm 11.41}$ | - | $21.80_{\pm 4.87}$ | $69.70_{\pm 10.66}$ | - |
| SCORE | $12.20_{\pm 1.78}$ | $47.20_{\pm 5.21}$ | $29.20_{\pm 0.40}$ | $15.80_{\pm 3.76}$ | $54.70_{\pm 10.17}$ | $30.00_{\pm 1.20}$ | $22.30_{\pm 4.43}$ | $67.60_{\pm 10.34}$ | $32.30_{\pm 2.40}$ |
| **DHT-CIT** | $\mathbf{8.30}_{\pm 3.82}$ | $\mathbf{26.00}_{\pm 12.77}$ | $\mathbf{0.46}_{\pm 0.72}$ | $\mathbf{8.30}_{\pm 1.55}$ | $\mathbf{38.10}_{\pm 5.19}$ | $\mathbf{11.20}_{\pm 2.28}$ | $\mathbf{14.60}_{\pm 3.75}$ | $\mathbf{36.20}_{\pm 12.86}$ | $\mathbf{13.60}_{\pm 0.92}$ |

dense graphs, we recommend using the topology obtained by the SCORE. When the density of the graph in real-world applications is uncertain, a better choice is to integrate the topological ordering from SCORE as prior knowledge, then apply our DHT-CIT algorithm to refine and enhance this ordering, i.e., DHT-CIT+SCORE (see Appendix G).

On Sigmoid-10-20 simulations, as the subsampling rate $u$ increases, indicating more unobserved time slices within the observation window, the number of error edges identified by DHT-CIT increases. However, it still outperforms all current state-of-the-art (SOTA) methods in performance. This demonstrates the scalability, superiority, and robustness of DHT-CIT in handling varying subsampling rates.

### 5.3. Experiments on Real-World Data

The **PM-CMR** (Wyatt et al., 2020) is a public time series dataset 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 ($X_{1:7}$={unemploy, income, female, vacant, owner, education, poverty}) 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 Table 8 in Appendix H.4. 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 Figure 2, 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 H.4.

## 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 without prior knowledge of the subsampling rate and structural functions. To address this issue, we treat the perturbations from the earlier time-slice as simulated intervention (i.e., auxiliary instrumental variables) 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.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (62441605, 62376243, 62037001, U20A20387, 623B2002), and the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010).

## Impact Statement

This paper introduces a novel Descendant Hierarchical Topology algorithm with Conditional Independence Test, named DHT-CIT, which can learn causal relationships between variables from subsampled time series data. This advancement in machine learning enhances decision-making processes in various fields such as medicine, finance, and social science. For instance, in healthcare, particularly in chronic disease monitoring or drug studies, DHT-CIT identifies causes and risk factors for diseases from electronic health record data, enabling doctors to provide better treatments. Additionally, based on the causal relationships discovered, we can design more efficient counterfactual reasoning models for machine learning. The limitation of our DHT-CIT algorithm is that it relies heavily on acyclic summary causal graphs and consistency throughout time assumptions. If the time series contains cyclic causal relationships or demonstrates inconsistencies over time, the algorithm may not accurately identify the causal relationships.

## References

Abavisani, M., Danks, D., and Plis, S. GRACE-c: Generalized rate agnostic causal estimation via constraints. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=B_pCIsX8KL_.

Ahammad, T., Hasan, M., and Zahid Hassan, M. 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.

Assaad, C. K., Devijver, E., and Gaussier, E. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.

Bellot, A. and van der Schaar, M. Conditional independence testing using generative adversarial networks. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Bussmann, B., Nys, J., and Latré, S. 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.

Chen, W., Zhang, K., Cai, R., Huang, B., Ramsey, J., Hao, Z., and Glymour, C. FRITL: A hybrid method for causal discovery in the presence of latent confounders. *arXiv preprint arXiv:2103.14238*, 2021.

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

Dahlhaus, R. and Eichler, M. Causality and graphical models in time series analysis. *Oxford Statistical Science Series*, pp. 115–137, 2003.

Danks, D. and Plis, S. Learning causal structure from undersampled time series. 2013.

Erdös, P. and Rényi, A. On the evolution of random graphs. In *The Structure and Dynamics of Networks*, pp. 38–82. Princeton University Press, 2011.

Ghoshal, A. and Honorio, J. Learning linear structural equation models in polynomial time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pp. 1466–1475. PMLR, 2018.

Gong, M., Zhang, K., Schoelkopf, B., Tao, D., and Geiger, P. Discovering temporal causal relations from subsampled data. In *International Conference on Machine Learning*, pp. 1898–1906. PMLR, 2015.

Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

Granger, C. W. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352, 1980.

Hasan, U., Hossain, E., and Gani, M. O. A survey on causal discovery methods for temporal and non-temporal data. *arXiv preprint arXiv:2303.15027*, 2023.

Hauser, A. and Bühlmann, P. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(1):2409–2464, 2012.

Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., and Schölkopf, B. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(1):3482–3534, 2020.

Hyttinen, A., Eberhardt, F., and Järvisalo, M. Constraint-based causal discovery: Conflict resolution with answer set programming. In *Conference on Uncertainty in Artificial Intelligence*, pp. 340–349. AUAI Press, 2014.

Hyttinen, A., Plis, S., Järvisalo, M., Eberhardt, F., and Danks, D. Causal discovery from subsampled time series data by constraint optimization. In *Conference on Probabilistic Graphical Models*, pp. 216–227. PMLR, 2016.

Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

Ke, N. R., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Schölkopf, B., Mozer, M. C., Pal, C., and Bengio, Y. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

Lachapelle, S., Brouillard, P., Deleu, T., and Lacoste-Julien, S. Gradient-based neural dag learning. In *International Conference on Learning Representations*, 2020.

Li, Y., Xia, R., Liu, C., and Sun, L. 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.

Loh, P.-L. and Bühlmann, P. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

Löwe, S., Madras, D., Zemel, R., and Welling, M. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*, pp. 509–525. PMLR, 2022.

Luo, L., Liu, W., Koprinska, I., and Chen, F. Discovering causal structures from time series data via enhanced granger causality. In *AI 2015: Advances in Artificial Intelligence: 28th Australasian Joint Conference*, pp. 365–378. Springer, 2015.

Lütkepohl, H. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

Mastakouri, A. A., Schölkopf, B., and Janzing, D. Necessary and sufficient conditions for causal feature selection in time series with latent common causes. In *International Conference on Machine Learning*, pp. 1898–1906. PMLR, 2015.

Montagna, F., Noceti, N., Rosasco, L., Zhang, K., and Locatello, F. Causal discovery with score matching on additive models with arbitrary noise. In *Conference on Causal Learning and Reasoning*, 2023.

Nauta, M., Bucur, D., and Seifert, C. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, 2019.

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

Park, Y. W. and Klabjan, D. Bayesian network learning via topological order. *Journal of Machine Learning Research*, 18(1):3451–3482, 2017.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

Peters, J., Bühlmann, P., and Meinshausen, N. 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.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Plis, S., Danks, D., Freeman, C., and Calhoun, V. Rate-agnostic (causal) structure learning. *Advances in neural information processing systems*, 28, 2015a.

Plis, S., Danks, D., and Yang, J. Mesochronal structure learning. In *Conference on Uncertainty in Artificial Intelligence*, 2015b.

Ramsey, J., Spirtes, P., and Zhang, J. Adjacency-faithfulness and conservative causal inference. pp. 401–408, 2006.

Reisach, A. G., Tami, M., Seiler, C., Chambaz, A., and Weichwald, S. Simple sorting criteria help find the causal order in additive noise models. *arXiv preprint arXiv:2303.18211*, 2023.

Rolland, P., Cevher, V., Kleindessner, M., Russell, C., Janzing, D., Schölkopf, B., and Locatello, F. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pp. 18741–18753. PMLR, 2022.

Runge, J. 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.

Runge, J. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1388–1397. PMLR, 2020.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019.

Sanchez, P., Liu, X., O'Neil, A. Q., and Tsaftaris, S. A. Diffusion models for causal discovery via topological ordering. 2022.

Shojaie, A. and Michailidis, G. Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523, 2010.

Solus, L., Wang, Y., and Uhler, C. Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika*, 108(4):795–814, 2021.

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

Sun, X., Janzing, D., Schölkopf, B., and Fukumizu, K. A kernel-based causal learning algorithm. In *International Conference on Machine Learning*, pp. 855–862. PMLR, 2007.

Tank, A., Fox, E. B., and Shojaie, A. Identifiability and estimation of structural vector autoregressive models for subsampled and mixed-frequency time series. *Biometrika*, 106(2):433–452, 2019.

Teyssier, M. and Koller, D. Ordering-based search: a simple and effective algorithm for learning bayesian networks. In *Conference on Uncertainty in Artificial Intelligence*, pp. 584–590. AUAI Press, 2005.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

Wang, Y., Solus, L., Yang, K., and Uhler, C. Permutation-based causal inference algorithms with interventions. *Advances in Neural Information Processing Systems*, 30, 2017.

Wyatt, L. H., Peterson, G. C. L., Wade, T. J., Neas, L. M., and Rappold, A. G. 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.

Yang, K., Katcoff, A., and Uhler, C. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, pp. 5541–5550. PMLR, 2018.

Zhang, A., Liu, F., Ma, W., Cai, Z., Wang, X., and Chua, T.-S. Boosting causal discovery via adaptive sample reweighting. In *International Conference on Learning Representations*, 2023.

Zhang, J. 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.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. In *Conference on Uncertainty in Artificial Intelligence*, pp. 804–813. AUAI Press, 2011.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425. PMLR, 2020.

Zhu, S., Ng, I., and Chen, Z. Causal discovery with reinforcement learning. In *International Conference on Learning Representations*, 2020.

## 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., 2006; 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., 2018; 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., 2018; 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 Algorithms With Traditional Methods

**Our algorithm differs from that of Zhang et al. (2011):** In this paper, we study the causal structure from subsampled time series using the proposed descendant-oriented conditional independence criteria, which can help us to determine descendants for each node in summary causal graphs using only two time slices. In this paper, the conditional HSIC proposed by Zhang et al. (2011) is a tool for testing conditional independence and cannot directly discern causal relations from subsampled time series. It could be replaced by any existing conditional hypothesis testing method within our criteria. Moreover, testing for conditional independence is inherently complex and may yield inaccurate results and incorrect ordering. Thus, our DHT-CIT introduces a topological ordering adjustment as a dual safeguard to ensure the acyclicity of the causal discovery process.

**Our algorithm 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., $\mathbf{X}_{\mathrm{past}(t)}$ 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 $\mathcal{D} = \{\boldsymbol{X}^{t_a}, \boldsymbol{X}^{t_b}\}_{t_a<t_b}$, $X_j^{t_b}$ is a descendant node of $X_i^{t_b}$ iff $X_i^{t_a} \not\perp\!\!\!\perp X_j^{t_b} \mid \mathbf{an}_i^{t_a}$.

**Our algorithm differs from that of Mastakouri et al. (2015):** Mastakouri et al. (2015) 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 be a non-descendant node of the target time series. They must conduct two conditional independence tests for each observed candidate to identify the direct causes of the target. In contrast, our theory constructs the descendant hierarchical topology of the directed summary causal graph and requires only one conditional independence test per component. Furthermore, the full graph studied by Mastakouri et al. (2015) 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. (2015b); 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. Besides, algorithms designed for subsampled time series in these studies lack reproducible open-source code.

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-seires 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 $\mathcal{D} = \{\boldsymbol{X}^{t_a}, \boldsymbol{X}^{t_b}\}_{t_a<t_b}$, since the subsampled time series is a stationary stochastic process and the summary causal graph is acyclic, after removing the nodes $\boldsymbol{X}^{t_a}$ in the learned graph, the learned DAGs on $\boldsymbol{X}^{t_b}$ would approximate the summary causal graphs.

Socre-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 $\boldsymbol{X}^{t_b}$ 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 $\boldsymbol{X}^{t_b}$ 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 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 experiments[3].

## 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 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 assumptions of 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). Notably, if no previous time-slice data is available or reliable, our approach, like other causal discovery algorithms, will not produce identifiable results.

Under Assumptions 3.2, 3.3 and 3.4, 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

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

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[4].

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

**Input:** Two time-slices $\mathcal{D} = \{\boldsymbol{X}^{t_a}, \boldsymbol{X}^{t_b}\}_{t_a<t_b}$ 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 $\boldsymbol{A}^{TP}$, one DAG $\mathcal{G}$.
**Components:** Conditional independence test $\mathbf{HSIC}(\dots)$; and pruning process $\mathbf{CAM}(\cdots)$.
**Stage 1 - Identifying Descendant Hierarchical Topology:**
**for** $i = 1$ **to** $d$ **do**
  Construct the conditional set $\boldsymbol{X}^{t_a}_{\otimes i}$ via an independence test $\boldsymbol{X}^{t_a}_{\otimes i} = \{X^{t_a}_j \mid X^{t_a}_j \perp\!\!\!\perp X^{t_a}_i\}$
  **for** $j = 1$ **to** $d$ **do**
    $p_{i,j} = \mathbf{HSIC}(X^{t_a}_i, X^{t_b}_j \mid \boldsymbol{X}^{t_a}_{\otimes i})$
    $a^{TP}_{i,j} = \mathbb{I}(p_{i,j} \leq \alpha)$
  **end for**
**end for**
We obtain $\boldsymbol{P} = \{p_{i,j}\}_{d\times d}$ and $\boldsymbol{A}^{TP} = \{a^{TP}_{i,j}\}_{d\times d}$
**Stage 2 - Adjusting the Topological Ordering:**
**while** The causal relationship between the unprocessed nodes is a directed cyclic graph **do**
  $k := k + 1$
  $X_{M_{i,k}} = \{X^{t_a}/X^{t_a}_i, \boldsymbol{L}_{1:k-1}\}$
  $X^{t_b}_i \in \boldsymbol{L}_k$, if $a^{TP}_{i,j} = 0$ for all $j \in M_{i,k}$
  **while** $\boldsymbol{L}_k = \emptyset$ **do**
    $p_{i^*,j^*} := 2\alpha \quad$ and $\quad a^{TP}_{i^*,j^*} = 0, \quad (i^*, j^*) = \arg\max_{i,j}(p_{i,j} \leq \alpha)$
    $X^{t_b}_i \in \boldsymbol{L}_k$, if $a^{TP}_{i,j} = 0$ for all $j \in M_{i,k}$
  **end while**
  We obtain $\boldsymbol{P} = \{p_{i,j}\}_{d\times d}$ and $\boldsymbol{A}^{TP} = \{a^{TP}_{i,j}\}_{d\times d}$
**end while**
**Stage 3 - Pruning Spurious Edges:**
We obtain $\mathcal{G} = \mathbf{CAM}(\mathcal{D}, \boldsymbol{A}^{TP}, \beta)$
**Return:** $\boldsymbol{A}^{TP}$ **and** $\mathcal{G}$

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.

## E. Explanations and Examples

### E.1. The Arrows but Self-Loops in the Summary Graph

In this paper, we use the notation $\mathbf{pa}^\tau_i$ to specify the partial set of parents of $X^\tau_i$ (as well as for $X^{\tau+1}_i$) at time $\tau$, and the nodes $\{X^{\tau-1}_i, \mathbf{pa}^{\tau-1}_i\}$ at time $\tau - 1$ also are parents to $X^\tau_i$, as shown in data generation function Eq. (1): $X^\tau_i = f_i\left(\mathbf{pa}^\tau_i, X^{\tau-1}_i, \mathbf{pa}^{\tau-1}_i\right) + \epsilon^\tau_i$. Here, $\mathbf{pa}^\tau_i \to X^\tau_i$ denotes the instantaneous effects at current time $\tau$, and $\{X^{\tau-1}_i, \mathbf{pa}^{\tau-1}_i\} \to X^\tau_i$ represents the lagged effects from both itself and its parents at previous time $\tau - 1$. Based on Eq. (1) $X^t_i = f_i\left(\mathbf{pa}^t_i, X^{t-1}_i, \mathbf{pa}^{t-1}_i\right) + \epsilon^\tau_i$, all arrows but self-loops in the summary graph not only represent the instantaneous interaction from $\mathbf{pa}^t_i \to X^t_i$, also captures the lagged effects from the nodes at previous time $\mathbf{pa}^{\tau-1}_i \to X^\tau_i$. The consistency of causal relation $f_i$ throughout time ensures that only the direct children, not the descendants, of a variable remain its children in the next time slice of the full-time series.

[4]The code of DHT-CIT is available at: https://github.com/anpwu/DHT-CIT.

### E.2. Clinical Study with Regular Sampling

In clinical studies, doctors and researchers often employ specific methods to collect samples at regular intervals by recruiting volunteers in schools, community centers, and other venues, or by offering some form of incentive policy. Researchers will then design studies requiring participants to undergo medical examinations or submit health information at fixed time intervals. Then, doctors typically compare earlier $\boldsymbol{X}^{t-3}$ and current $\boldsymbol{X}^{t}$ 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 $\boldsymbol{X}^{t-2}$ and $\boldsymbol{X}^{t-1}$ unrecorded. This situation is quite common in healthcare, where doctors typically use limited time-slices to analyze a patient's condition and determine treatments.

### E.3. Physical Example

In practical applications, there are extensive serialized events and periodic time series with acyclic summary causal graphs. For example, the local electricity transportation mechanism of urban power systems, and the rise and fall of the Earth's ocean surfaces. The gravitational forces exerted by the moon ($X_1^t$) and the sun ($X_2^t$) periodically influence the rise and fall of the Earth's ocean surfaces ($X_3^t$). In this scenario, the causal direction is unidirectional since the tidal movements on Earth cannot, in return, affect the motion states of the moon and the sun.

## F. Further Discussion on the Assumptions Used in This Paper

### F.1. Implications of Violating Assumptions in Our Algorithm

**Violation to Assumption 3.2**: If the first-order Markov assumption is violated, the current state $X^t$ would not be exclusively influenced by the immediately preceding state $X^{t-1}$ but would depend on a series of previous states $X^{t-q}, \cdots, X^{t-1}$. Consequently, in our Descendant-Oriented Conditional Independence Criteria (Theorem 4.4), the sufficient and necessary condition for $X_j^{t_b}$ being a descendant node of $X_i^{t_b}$ at time $t_b$ would be redefined as $X_i^{t_a} \not\perp X_j^{t_b} \mid \{\mathbf{an}_i^{t_a-q+1}, \cdots, \mathbf{an}_i^{t_a}\}$. We discuss the high-order Markov models in Appendix F.3.

**Violation to Assumption 3.3**: If the acyclic summary causal graph is violated, the data generation process would change from $X_i^\tau = f_i\left(\mathbf{pa}_i^\tau, X_i^{\tau-1}, \mathbf{pa}_i^{\tau-1}\right) + \epsilon_i^\tau$ to $X_i^\tau = f_i\left(\mathbf{pa}_i^\tau, X_i^{\tau-1}, \mathbf{pa}_i^{\tau-1}, \mathbf{de}_i^{\tau-1}\right) + \epsilon_i^\tau$ . Then, if the subsampling rate of the subsampled time series exceeds 2 (applicable for instantaneous effects) or the length of the loop containing $X_i^t$ (applicable for no instantaneous effects), any two nodes within this loop will be connected by a bidirectional arrow in the learned summary causal graph. This implies that the structure of the learned summary causal graph by two time-slice models will vary as the subsampling rate increases (Danks & Plis, 2013; Peters et al., 2017). While achieving perfect acyclic summary causal graphs in dynamical systems may be challenging due to external interventions and internal feedback, these assumptions offer a valuable framework for modeling system behavior in short-term stable and controlled environments.

**Violation of Assumption 3.4:** If the consistency throughout time is violated, then the learned causal graph might be a subgraph of a summary causal graph, but it will certainly be larger than the graph constructed by instantaneous effects at time $t_b$. In other words, the learned DAG will be a causal graph that lies between the DAG of instantaneous effects at time $t_b$ and the DAG of the true summary causal graph. Because the simulated interventions by two-time-slices only propagate to its descendant nodes at time $t_b$ through the link $X_i^{t_a} \to X_i^{t_a+1} \to \cdots \to X_i^{t_b} \to \mathbf{de}_i^{t_b}$, but the DAG of instantaneous effects at time $t_b$ might just be a subgraph of the summary causal graph. One solution is that, provided the observational windows are sufficiently extended beyond the longest cycle of changes in causal relations, it becomes feasible to accurately infer the Descendant Hierarchical Topology of the summary causal graph.

### F.2. Further Discussion on Acyclic Summary Causal Graph and Consistency Throughout Time Assumption

While perfect acyclic summary causal graphs and consistency throughout time may be challenging to achieve in dynamical and biological systems due to external interventions and internal feedback, these assumptions provide a useful framework for modeling and predicting system behavior, especially in relatively stable and controlled environments. Actually, we can divide the homeostatic systems into multiple short-term stages with an acyclic summary graph. For instance, the entire process of a cold virus entering the human body through the respiratory tract, lying dormant, proliferating, and attacking human cells, until the body recognizes the virus and initiates an immune response, can be simplified into an acyclic summary causal graph. These short-term acyclic graph models allow for the application of the proposed DTG-CIV algorithm to identify and analyze causal relationships within short-term windows.

Moreover, dynamical and biological systems do not encompass the entire real world; in practical applications, there are extensive serialized events and periodic time series with acyclic summary causal graphs. For example, the local electricity transportation mechanism of urban power systems, and the rise and fall of the Earth's ocean surfaces. The gravitational forces exerted by the moon ($X_1^t$) and the sun ($X_2^t$) periodically influence the rise and fall of the Earth's ocean surfaces ($X_3^t$). In this scenario, the causal direction is unidirectional since the tidal movements on Earth cannot, in return, affect the motion states of the moon and the sun. Therefore, studying time series with acyclic summary graphs is popular and remains a key area of interest in both academic studies and practical applications (please see the survey paper by (Assaad et al., 2022)).

### F.3. 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 $\boldsymbol{X}^{t+1}$ depends only on states $\boldsymbol{X}^{t\cdots t-q+1}$ and does not directly depend on states $\boldsymbol{X}^{1\cdots t-q}$. Then, with $q+1$ time-slices ($\boldsymbol{X}^{t_a\cdots t_a-q+1}$ and $\boldsymbol{X}^{t_b}$), 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 4.4). However, in this paper, we focus exclusively on the two-time-slices algorithm to demonstrate our theorem and algorithm.

### F.4. Discussion about Identical Causal Structure across Samples/Subgroups

Varying causal structures among different samples or subgroups is indeed an interesting topic. However, it is undeniable that in real life, there indeed exists a multitude of sequential events sharing the same causal mechanisms, such as the basic biological processes governing human health, the fundamental laws of physics that apply to various engineering problems, or the universal principles of economics that drive market behaviors. In fact, in studying the varying causal structures across samples, if we could observe all explanatory variables within the system, we might find that some overlooked variables, such as genes or different environments, could be sources of individual heterogeneity. In future work, we will explore the EM algorithm or heterogeneity detection algorithms to identify different causal structures across samples or subgroups.

## G. Relaxing Conditional Independence Test Limitations and Scaling to Large Graphs

While hypothesis testing using HSIC is indeed a sensible approach, we can relax this issue from three perspectives:

1. Our DHT-CIT algorithm introduces a topological ordering adjustment technique (refer to Section 4.3.2 for more details) as a dual safeguard to correct partial erroneous edges identified by the conditional HSIC test, ensuring the resulting summary graph remains acyclic. This provides tolerance for errors in the HSIC test.

2. The proposed DHT-CIT algorithm can be appended to any existing topology-based method to enhance topological ordering, as mentioned in Appendix J.1. By treating the topological ordering learned from existing methods as prior knowledge, the size of the conditional set required for the Descendant-Oriented Conditional Independence Criteria significantly reduces. Additionally, the search space size of the DHT-CIT algorithm decreases by at least half, substantially easing the difficulty of hypothesis testing. We refer to the combination of the SCORE and DHT-CIT algorithms as DHT-CIT+SCORE, which utilizes the topological ordering learned by SCORE as prior knowledge.

3. To further mitigate this issue, we implement random interventions to some nodes in the previous states of two time slices ($D = \{X^1, X^2\}$), then apply DHT-CIT to learn causal relations. Depending on the number of nodes intervened in the previous state, we have variants like DHT-CIT (10 Intervention), DHT-CIT(20 Intervention), and so on.

To verify the scalability of our DHT-CIT with interventions and DHT-CIT+SCORE, we conduct experiments in high-dimensional settings, such as Sin-50-50 Simulations with 50 nodes and 50 edges, Sin-50-100 Simulations with 100 edges and on Sin-100-100 Simulations with 100 nodes and 100 edges.

## H. Supplementary Experiments

### H.1. Exploring Varying Time-Lagged Edges and Denser Graph

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 $\boldsymbol{X}_{-i}^{\tau-1}$ 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

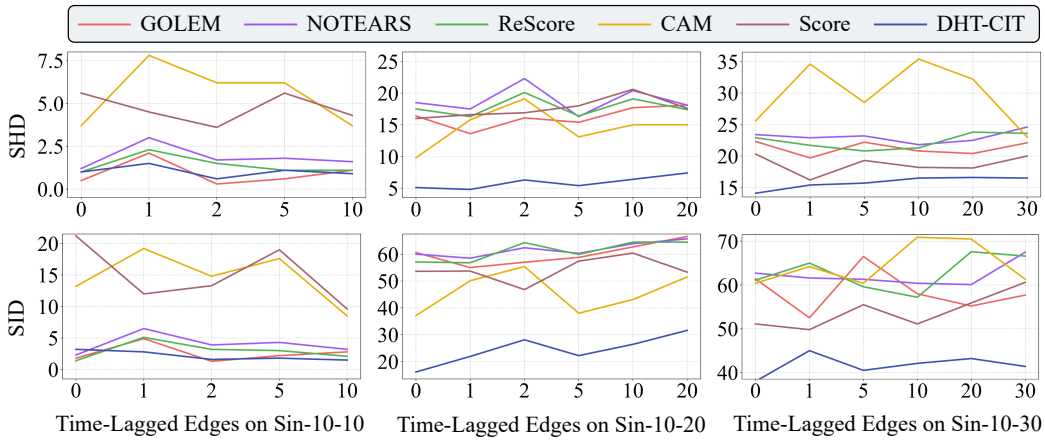



*Figure 3.* Exploring varying time-lagged edges and denser graph.

*Table 5.* Training time(s) of various methods in a single execution on different datasets.

| | Data-10-10 | Data-10-20 | Data-10-30 | Data-20-20 |
|---|---|---|---|---|
| CD-NOD | 5433s | $>$ 2h | $>$ 2h | $>$ 5h |
| CAM | 97.2s | 92.8s | 111.3s | 543.6s |
| GOLEM | 36.4s | 36.4s | 36.4s | 228.0s |
| NOTEARS | 33.6s | 35.6s | 36.4s | 747.2s |
| SCORE | 29.1s | 28.0s | 29.6s | 63.5s |
| 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 | 64.1s | 68.5s | 66.1s | 306.0s |
| DHT-CIT+SCORE | 33.7s | 33.6s | 34.0s | 140.0s |

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.

### H.2. Training Cost Analysis

In Data-$d$-$e$ Simulations, we implement 10 replications to study the average running time(s) for the proposed model in a single execution and sorted it by time spent on Data-10-10 in Table 5. From the experiments on denser and larger graphs in Tables 5, 6 & 7, we observed that although the time consumption of DHT-CIT+SCORE is always greater than that of SCORE, compared to DHT-CIT alone, DHT-CIT+SCORE reduces computational costs by 40%-66%. As shown in experiments on denser and larger graphs (Tables 5, 6 & 7), DHT-CIT+SCORE not only reduces computational costs but also improves performance on large graphs. Therefore, for graphs with over 20 nodes, DHT-CIT+SCORE is recommended to lower computational costs and achieve better performance, while for smaller graphs (under 20 nodes), the time cost of DHT-CIT (300 seconds) is manageable.

### H.3. The Experiments on Large Graphs With High-Dimension Variables

**Datasets**. Followed the data generation process (Eq. (6)) 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.

$$X_i^\tau = \mathrm{Sin}\left(\mathbf{pa}_i^\tau, X_i^{\tau-1}\right) + \frac{1}{10}\mathrm{Sin}\left(\mathbf{w}\cdot\mathbf{pa}_i^{\tau-1}\right) + \epsilon_i^\tau, \boldsymbol{X}^0 \sim \mathcal{N}\left(0, \mathrm{I}_d\right), \boldsymbol{\epsilon}^\tau \sim \mathcal{N}\left(0, 0.4\cdot\mathrm{I}_d\right) \tag{7}$$

where $\mathrm{Sin}(\mathbf{pa}_i^\tau) = \sum_{j\in \mathrm{pa}(X_i)} \sin(X_j^\tau)$, $\mathrm{I}_d$ is a $d$-th order identity matrix, and $\mathbf{w}$ is a random 0-1 vector that controls the

*Table 6.* The experiments on Sin-50-50 dataset.

| | **Sin-50-50 data with Gauss noise ( $\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^3\}$ )** | | | |
|---|---|---|---|---|
| Method | **SHD↓** | **SID↓** | **#Prune↓** | **Running Time(s)↓** |
| SCORE | $73.5_{\pm 28.71}$ | $8.80_{\pm 3.47}$ | $1175._{\pm 0.40}$ | 977s |
| DHT-CIT | $930._{\pm 147.3}$ | $99.4_{\pm 4.8}$ | $323.8_{\pm 26.54}$ | 2655s |
| DHT-CIT (10 Intervention) | $670._{\pm 106.8}$ | $51.4_{\pm 3.32}$ | $282.0_{\pm 23.71}$ | 2145s |
| DHT-CIT (20 Intervention) | $142._{\pm 18.45}$ | $14.6_{\pm 1.96}$ | $251.4_{\pm 19.11}$ | 1706s |
| DHT-CIT (25 Intervention) | $73.0_{\pm 27.77}$ | $6.80_{\pm 2.64}$ | $246.6_{\pm 16.66}$ | 1109s |
| DHT-CIT (50 Intervention) | $\mathbf{0.80}_{\pm 0.98}$ | $\mathbf{0.40}_{\pm 0.40}$ | $\mathbf{99.60}_{\pm 20.33}$ | **377s** |
| DHT-CIT+SCORE | $25.0_{\pm 5.18}$ | $3.80_{\pm 1.47}$ | $225.4_{\pm 10.52}$ | 1176s |
| DHT-CIT+SCORE (10 Intervention) | $3.00_{\pm 1.55}$ | $0.80_{\pm 0.40}$ | $223.2_{\pm 15.10}$ | 884s |
| DHT-CIT+SCORE (20 Intervention) | $4.40_{\pm 1.41}$ | $0.90_{\pm 0.50}$ | $187.4_{\pm 13.65}$ | 649s |

*Table 7.* The experiments on Sin-50-100 & Sin-100-100 datasets.

| | **Sin-50-100 data with Gauss noise ( $\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^3\}$ )** | | | |
|---|---|---|---|---|
| Method | **SID↓** | **SHD↓** | **#Prune↓** | **Running Time(s)↓** |
| SCORE | $247.0_{\pm 102.5}$ | $23.0_{\pm 8.56}$ | $1127._{\pm 2.06}$ | 1027s |
| DHT-CIT | $2039._{\pm 84.14}$ | $234._{\pm 2.71}$ | $397.6_{\pm 25.54}$ | 3217s |
| DHT-CIT (50 Intervention) | $203.0_{\pm 61.1}$ | $14.8_{\pm 3.90}$ | $\mathbf{149.0}_{\pm 27.00}$ | **357s** |
| DHT-CIT+SCORE | $97.4_{\pm 101.6}$ | $7.60_{\pm 5.28}$ | $352.6_{\pm 23.69}$ | 1249s |
| DHT-CIT+SCORE (10 Intervention) | $\mathbf{53.2}_{\pm 20.29}$ | $\mathbf{4.80}_{\pm 0.98}$ | $284.0_{\pm 26.58}$ | 1109s |
| | **Sin-100-100 data with Gauss noise ( $\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^3\}$ )** | | | |
| Method | **SID↓** | **SHD↓** | **#Prune↓** | **Running Time(s)↓** |
| SCORE | $381._{\pm 156.5}$ | $28.67_{\pm 4.5}$ | $4850_{\pm 0.2}$ | 4689s |
| DHT-CIT | $2377_{\pm 427}$ | $218._{\pm 12.9}$ | $787.0_{\pm 49.1}$ | 19655s |
| DHT-CIT (100 Intervention) | $\mathbf{5.33}_{\pm 7.50}$ | $\mathbf{1.00}_{\pm 1.41}$ | $\mathbf{347.0}_{\pm 9.10}$ | **1074s** |
| DHT-CIT+SCORE | $28.67_{\pm 9.53}$ | $4.67_{\pm 0.47}$ | $925.0_{\pm 83.3}$ | 6342s |
| DHT-CIT+SCORE (10 Intervention) | $14.67_{\pm 11.9}$ | $3.33_{\pm 1.25}$ | $797.3_{\pm 29.4}$ | 6108s |

number and existence of time-lagged edges from $\mathbf{pa}_i^{\tau-1}$. In this experiment, we set the number of time-lagged edges from other variables as 0. To evaluate our DHT-CIT on larger graphs, we generate large graphs **Sin-50-50**, **Sin-50-100**, and **Sin-100-100**. Given the limited effectiveness of many methods when applied to subsampled time series data $\mathcal{D} = \boldsymbol{X}^1, \boldsymbol{X}^3$, this section is dedicated to a comparative analysis between variants of our approach and the SCORE algorithm. Moreover, we focus exclusively on the four most crucial metrics: SHD, SID, #Prune, and Running Time(s).

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 with Gaussian kernel (Zhang et al., 2011). 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 ( $\mathcal{D} = \{\boldsymbol{X}^1, \boldsymbol{X}^2\}$ ), we relax this issue from three perspectives: topological ordering adjustment, interventions, and pre-training using SCORE methods. The detailed description is placed in Appendix G.

**Two types of DHT-CIT variants**: (1) We implement random interventions to some nodes in the previous states of two time slices ($D = \{X^1, X^2\}$). Depending on the number of nodes intervened in the previous state, we have variants like DHT-CIT (10 Intervention), DHT-CIT(20 Intervention), and so on. (2) We refer to the combination of the SCORE and DHT-CIT algorithms as DHT-CIT+SCORE, which utilizes the topological ordering learned by SCORE as prior knowledge.

**Results**. From the results on larger graphs (**Sin-50-50**, **Sin-50-100** and **Sin-100-100**) in Tables 6 & 7, we have the following observation: As the number of interventions on previous states of two-time slices increases, DHT-CIT with Intervention

*Table 8.* 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 |

can more accurately identify true causal relationships. When the number of interventions exceeds half, the performance of DHT-CIT with Intervention surpasses that of SCORE. Additionally, utilizing SCORE's topological ordering as prior knowledge, DHT-CIT+SCORE greatly outperforms SCORE and DHT-CIT in **Sin-50-50**, **Sin-50-100** and **Sin-100-100** simulations. This is partly because DHT-CIT improves the topological ordering learned by SCORE, and also graph knowledge from SCORE helps reduce the size of the conditional set and search space of DHT-CIT. As shown in Table 7, similar results can be found in **Sin-50-100** and **Sin-100-100** Simulations. This demonstrates the scalability of our DHT-CIT algorithm with interventions and the combined DHT-CIT+SCORE approach.

### H.4. The Experiments on Real-World Dataset

**Datasets**. The **PM-CMR**[5] (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 Table 8. 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 Figure 2 and 4, both GOLEM and SCORE do 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 (Figure 3). 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.

## I. The Advantages and Limitations of Using Two Time-Slices Data

**Advantages:** (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

[5]PM-CMR:https://pasteur.epa.gov/uploads/10.23719/1506014/SES_PM25_CMR_data.zip

The State in 2010 (GOLEM)

| The State in 2000 | T | Y | X1 | X2 | X3 | X4 | X5 | X6 | X7 |
|---|---|---|---|---|---|---|---|---|---|
| T | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Y | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| X1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| X2 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| X3 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| X4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| X5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| X6 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| X7 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |

The State in 2010 (SCORE)

| The State in 2000 | T | Y | X1 | X2 | X3 | X4 | X5 | X6 | X7 |
|---|---|---|---|---|---|---|---|---|---|
| T | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| Y | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| X1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| X2 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| X3 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| X4 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| X5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| X6 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| X7 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 |

The State in 2010 (DHT-CIT)

| The State in 2000 | T | Y | X1 | X2 | X3 | X4 | X5 | X6 | X7 |
|---|---|---|---|---|---|---|---|---|---|
| T | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Y | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| X1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| X2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| X3 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| X4 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| X5 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| X6 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| X7 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

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

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:** (1) Our DHT-CIT algorithm relies on the acyclic summary causal graph and consistency throughout time assumptions. While perfect acyclic summary causal graphs and consistency throughout time may be challenging to achieve in real applications due to external interventions and internal feedbacks, these assumptions provide a useful framework for modeling and predicting system behavior, especially in relatively stable and controlled environments. Sometimes, we can divide dynamic systems into multiple short-term stages with acyclic summary graph. (2) The two-slice model requires that sampling occurs at the same two timesteps across all observations. If multiple two-slice samples are collected from varying starting points, the DHT-CIT method would fail to accurately identify descendant nodes via the conditional independence tests. Fortunately, numerous volunteer recruitment activities or data selection strategies exist to assist us in obtaining regular sampled two-slice data. (3) The efficacy of the proposed DHT-CIT algorithm is contingent upon the conditional independence test. Despite the significant advancements in the development of conditional independence testing, it remains a complex task, especially in high-dimensional scenarios. This complexity may induce biased topological orders, including cycles. To address this challenge, our approach introduces topological ordering adjustment as a dual safeguard to ensure acyclicity in the causal discovery process.

## J. Future Applications

### J.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 PCMCI (Runge et al., 2019; Runge, 2020).

### J.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 ($\boldsymbol{X}^{t_a \cdots t_a - q+1}$ and

$\boldsymbol{X}^{t_b}$). 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.