Double Machine Learning based Structure Identification from Temporal Data

Anonymous authors
Paper under double-blind review

Abstract

Learning the causes of time-series data is a fundamental task in many applications, spanning from finance to earth sciences or bio-medical applications. Common approaches for this task are based on vector auto-regression, and they do not take into account unknown confounding between potential causes. However, in settings with many potential causes and noisy data, these approaches may be substantially biased. Furthermore, potential causes may be correlated in practical applications or even contain cycles. To address these challenges, we propose a new double machine learning based method for structure identification from temporal data (DR-SIT). We provide theoretical guarantees, showing that our method asymptotically recovers the true underlying causal structure. Our analysis extends to cases where the potential causes have cycles, and they may even be confounded. We further perform extensive experiments to showcase the superior performance of our method. Code: `https://anonymous.4open.science/r/TMLR_submission_DR_SIT-6B46/`

1 Introduction

One of the primary objectives when working with time series data is to uncover the causal structures between different variables over time. Learning these causal relations and their interactions is of critical importance in many disciplines, e.g., healthcare (Anwar et al., 2014), climate studies (Stips et al., 2016; Runge et al., 2019a), epidemiology (Hernán et al., 2000; Robins et al., 2000), finance (Hiemstra & Jones, 1994), ecosystems (Sugihara et al., 2012), and many more. Interventional data is not often accessible in many of these applications. For instance, in healthcare scenarios, conducting trials on patients may raise ethical concerns, or in the realm of earth and climate studies, randomized controlled trials are not feasible.

In general, understanding the underlying causal graph using only observational data is a cumbersome task due to many reasons: i) observational data, as opposed to interventional data, capture correlation-type relations instead of cause-effect ones. ii) unobserved confounders introduce biases and deceive the algorithms to falsely infer causal relations instead of the true structure, e.g., the existence of a hidden common confounder iii) the number of possible underlying structures grows super exponentially with the number of variables creating major statistical and computation barriers iv) the identifiability problem, since multiple causal models can result in the same observational distribution, thus making it impossible to uniquely determine the true structure. To overcome these problems and determine the underlying structure, additional assumptions are imposed. Typical assumptions include faithfulness, linearity of relations, or even noise-free settings, which limit the types of causal relationships that can be discovered (Pearl, 2000; Peters et al., 2017; Spirtes et al., 2000; Glymour et al., 2019). Almost all of these challenges extend to the problem of identifying the underlying causal structure from observational time-series datasets.

Subsequently, in many instances, the emphasis is placed on particular target variables of interest and their causal features. Causal features of a target are defined as the set of variables that conditioned on them, the target variable is independent of the rest variables. Causal feature selection enables training models which are much simpler, more interpretable, and more robust (Aliferis et al., 2010; Janzing et al., 2020). However, learning the causal structures between variables and a specific target is still a demanding task, and many current approaches for causal feature selection face limitations by making unrealistic simplifying

assumptions about the data-generating process or by lacking computational and/or statistical scalability (Yu et al., 2021; 2020). These challenges become particularly amplified in the context of time series data, where the number of variables grows linearly with the length of the data trajectories. As a result, to mitigate these problems, additional assumptions, e.g., stationarity or no hidden confounders are included (Moraffah et al., 2021; Bussmann et al., 2021; Runge, 2018) and/or weaker notions of causality[1] such as Granger causality have been studied extensively (Granger, 1988; 1969; Marinazzo et al., 2011; Tank et al., 2018; Bussmann et al., 2021; Khanna & Tan, 2019; Runge, 2018; Hasan et al., 2023).

Existing causal feature selection from time-series data algorithms often assume some level of faithfulness or causal sufficiency (Runge et al., 2019b; Runge, 2018; 2020). Oftentimes, they overlook the presence of unknown confounding factors among potential causes (Moraffah et al., 2021). Moreover, most cannot adapt to cyclic settings (Entner & Hoyer, 2010), which is relatively ubiquitous in many domains (Bollen, 1989). Furthermore, many algorithms employ the popular vector auto-regression framework to model time-dependence structures (Bussmann et al., 2021; Lu et al., 2016; Chen et al., 2009; Weichwald et al., 2020; Hyvärinen et al., 2010), which again is restrictive. To overcome these problems, we propose an efficient algorithm for doubly robust structure identification from temporal data.

Our contributions Our key contributions can be summarized as follows:

1. We provide an efficient and easy-to-implement doubly robust structure identification from temporal data algorithm (DR-SIT) with theoretical guarantees enjoying $\sqrt{n}$-consistency.
2. We provide an extensive technical discussion relating Granger causality to Pearl's framework for time series and show under which assumptions our approach can be used for feature selection or full causal discovery. As a consequence of this, ours is the first paper to propose a non-parametric Granger causality test that achieves the semi-parametric $\sqrt{n}$-rate.
3. We provide theoretical insights demonstrating that our algorithm doesn't need faithfulness or causal sufficiency while allowing for general non-linear cyclic structures and also the presence of hidden confounders among the covariates for the task of finding the direct causes of the target variable. An important problem in many applications.
4. In extensive experiments we illustrate that our approach is significantly more robust, significantly faster, and more performative than state-of-the-art baselines.

2 Related Work

Causal Structure Learning for Timeseries A longstanding line of work intends to tailor the existing causal structure learning and Markov blanket discovery for i.i.d. data to the temporal setting. To name a few, Entner & Hoyer (2010) adapted the Fast Causal Inference algorithm (Spirtes et al., 2000) to time-series data. While the approach shares the benefit of being able to deal with hidden confounders, it is not possible to account for cyclic structures. Runge et al. (2019b) introduced PCMCI, as an adjusted version of PC (Spirtes et al., 2000) with an additional false positive control phase which is able to recover time-lagged causal dependencies. PCMCI+ modified the approach further to additionally be able to find contemporaneous causal edges (Runge, 2020). LPCMCI extends the scope by catering to the case of hidden confounders (Gerhardus & Runge, 2020). Even though methods in this category are able to provide theoretical guarantees for learning the causal structure, DR-SIT has several advantages over them: i) For all of these methods, the faithfulness assumption is a key ingredient while DR-SIT does not need it. ii) These methods are based on conditional independence testing which is widely recognized to be a cumbersome statistical problem (Bergsma, 2004; Kim et al., 2022). Shah & Peters (2020) have established that no conditional independence (CI) test can effectively control the Type-I error for all CI settings. Moreover in practice, conducting many conditional independence tests from lengthy time-series is burdensome. iii) Even having access to a perfect conditional independence test oracle, severe computational challenges exist (Chickering, 1996; Chickering et al., 2004)[2] (please refer to Appendix I for a detailed comparison).

[1] weaker than Pearl's structural equation model (Pearl, 2000).

[2] Learning Bayesian Networks with conditional independence tests is NP-Hard.

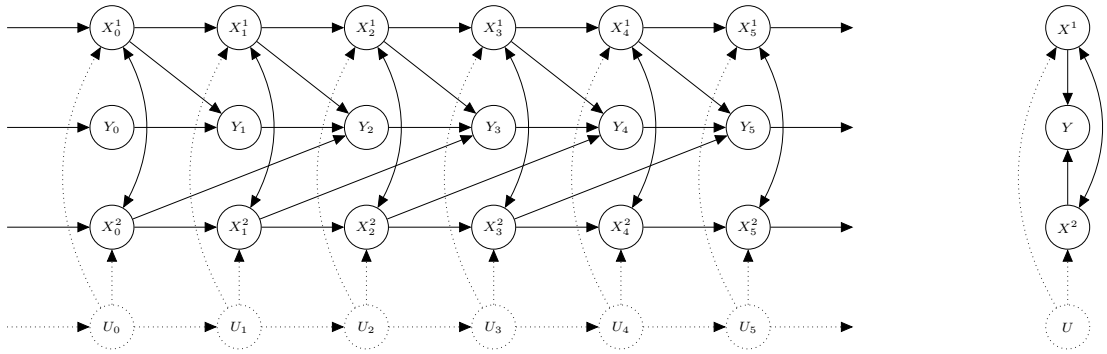


Figure 1: Example of a full-time graph (left), and the corresponding summary graph (right). Note that time series $\boldsymbol{X}^1$ and $\boldsymbol{X}^2$ causally influence the outcome $\boldsymbol{Y}$ with different lags. The time series $\boldsymbol{U}$ is unobserved, and it acts as a confounder for $\boldsymbol{X}^1$ and $\boldsymbol{X}^2$. Moreover, there is a cycle between the confoundings variables $\boldsymbol{X}^1$ and $\boldsymbol{X}^2$ of the outcome variable $\boldsymbol{Y}$.

Another line of work relies heavily on the vector auto-regression framework. VarLiNGAM (Hyvärinen et al., 2010) generalizes LiNGAM (Shimizu et al., 2006) to time-series and similar to that it assumes a linear non-Gaussian acyclic model for the data. In the work of Huang et al. (2019), a time-varying linear causal model is assumed, allowing for causal discovery even in the presence of hidden confounders. More recently, deep neural networks are used to train vector auto-regression. Tank et al. (2018) adapted neural networks (named cMLP and cLSTM) for Granger causality by imposing group-sparsity regularizers. In a similar fashion, Khanna & Tan (2019) used recurrent neural networks. In another work, Bussmann et al. (2021) designed neural additive vector auto-regression (NAVAR), a neural network approach to model non-linearities. In contrast to previous works, they extract Granger-type causal relations by injecting the necessary sparsity directly into the architecture of the neural networks. This line of work is quite limited to ours as they consider confining structural assumptions over the underlying causal structural equations; details on these assumptions are discussed next to Axiom (A) in Section 3.

Double machine learning The use of double robustness in causality problems has a long history mainly concentrated on estimating average treatment effect (Robins et al., 1994; Funk et al., 2011; Benkeser et al., 2017; Bang & Robins, 2005; Słoczyński & Wooldridge, 2018). Chernozhukov et al. (2018) introduced the DML framework to achieve double robustness for structural parameters. Upon that, Soleymani et al. (2022) defined an orthogonalized score to infer the direct causes of partially linear models. Their approach is fast and allows for the assumption of complicated underlying structures but unfortunately, it is limited to only linear direct causal effects. While this was later extended to the non-linear case (Quinzan et al., 2023), we propose a doubly robust approach for identifying causal structures from temporal data under the general assumptions discussed in Section 3.

3 Framework

3.1 Problem Description

We are given i.i.d. observations from a univariate target time series $\boldsymbol{Y} \coloneqq \{Y_t\}_{t\in\mathbb{Z}}$ and i.i.d. observations from a multivariate time series $\boldsymbol{X} \coloneqq \{X_t^1, \ldots, X_t^m\}_{t\in\mathbb{Z}}$ of potential causes. We assume that the outcome time series is specified by some of the potential causes by a deterministic function with posterior additive noise, and no instantaneous effects. We can formalize this model as

Axiom (A) $Y_T = f(\mathsf{pa}_T(\boldsymbol{Y}), T) + \varepsilon_T,$

with ε_T exogenous independent noise and $\mathsf{pa}_T(\boldsymbol{Y}) \subseteq \{X_t^1, \ldots, X_t^m\}_{t\in\mathbb{Z}}$ is a subset of random variables of the multivariate time series $\boldsymbol{X}$. Note that the independence of the additive noise is important for identifiability. In fact, if there are dependencies between the noise and the history, then one might run into identifiability problems. We refer the reader to Appendix B for a counterexample. We are interested in identifying the time

series $\boldsymbol{X}^i$ that directly affects the outcome Y. That is, we wish to identify time series $\boldsymbol{X}^i$ such that it holds $X_t^i \subseteq \mathsf{pa}_T(\mathbf{Y})$ for some time steps t, T.

We further use the following assumptions:

Axiom (B) there are no causal effects backward in time;
Axiom (C) there are no instantaneous causal effect between $\boldsymbol{Y}$ and any of the potential causes $\boldsymbol{X}^i$.

Both Axiom (B) and Axiom (C) appear in previous related work (see, e.g., Peters et al. (2013); Mastakouri et al. (2021); Löwe et al. (2022)). Note that these axioms allow for cycles and hidden common confounders between the potential causes. Axiom (B) is a natural assumption as a system is called causal when the output of the system only depends on the past, not the future (Peters et al., 2017; Pearl, 2000). Axiom (C) poses additional restrictions on the class of models that we consider, since instantaneous effects may be relevant in some cases and applications (Lippe et al., 2022a)[3]. However, without this assumption, it is impossible to learn causes *from observational data*. The necessity of Axiom (C) for causal discovery is a well-known fact (Peters et al., 2013)[4].

3.2 Generality over Previous Work

Our framework retains some degree of generality over previous related work. In fact, Axiom (A)-Axiom (C) allow for hidden confounding and cycles between the potential causes (see Figure 1), providing a more general framework than the full autoregressive model studied, e.g., by Hyvärinen et al. (2010); Peters et al. (2013); Löwe et al. (2022); Wu et al. (2020). Furthermore, in contrast to several previous works (Khemakhem et al., 2020; Gresele et al., 2021; Lachapelle et al., 2022; Lippe et al., 2022b; Yao et al., 2021), we do not assume that the variables $Y_T, X_T^1, \ldots, X_T^m$ are independent conditioned on the observed variables at previous time steps. Importantly, we also do not assume causal faithfulness in the full time graph, or weaker notions such as causal minimality.[5] This is a major improvement over some previous works, e.g., Mastakouri et al. (2021); Gong et al. (2022), since there is no reason to assume that faithfulness or causal minimality hold in practice. Theorem 1 by Gong et al. (2022) provides an identifiability result for a model with history-dependent noise and instantaneous effects. This result, however, requires causal minimality.

Furthermore, as discussed in Section 2 , vector auto-regression methods enforce heavy structural assumptions on the underlying causal structural equations. NAVAR (Bussmann et al., 2021) assumes that each variable is influenced by its causal parents exclusively in an additive way and higher-order interactions among them are precluded. In mathematical terms, they follow $Y_T = \beta + \sum_{X \in \mathsf{pa}_T(\boldsymbol{Y})}^{N} f_X(X_{t-\kappa:t-1}) + \varepsilon_t$, where β is a bias term, κ is the time lag, and ε_t is an independent noise variable. VarLiNGAM (Hyvärinen et al., 2010) takes a step further and uses much more restrictive assumptions on the functional form of the structural equations, confining to f_X being a linear transformation of $X_{t-\kappa:t-1}$. These structural assumptions are quite limited as they do not include even simple functionals, e.g., $Y_t = X_{t-1}^1 \times X_{t-1}^2$ or $Y_t = \log(X_{t-1}^1 + X_{t-1}^2)$, whereas Axiom (A) is able to capture any interaction between the covariates and timestep T which is a significant improvement over existing works.

3.3 Causal Structure

We are interested in direct causal effects, which are defined by distribution changes due to interventions on the DGP. An intervention amounts to actively manipulating the generative process of a potential cause $\boldsymbol{X}^i$ at some time step t, without altering the remaining components of the DGP. Then, a time series $\boldsymbol{X}^i$ has a

[3]An example of cases where time series exhibit instantaneous causal effects is given by dynamical systems (Mogensen et al., 2018; Rubenstein et al., 2016). In dynamical systems, a variable may instantaneously affect another variable of the model. Instantaneous effects have been studied in previous related work (Gong et al., 2022) but due to identifiability issues, they rely on stronger assumptions such as faithfulness.

[4]In general, Peters et al. (2013) show that causal discovery is impossible with instantaneous effects. Please refer to Appendix C for an example of this non-identifiability. However, Peters et al. (2013) also provide a special case in which the causal structure is identifiable with instantaneous effects. This special case occurs when the random variables of the model are jointly Gaussian, and the instantaneous effects are linear. Our framework extends to this special case.

[5]Recall that a distribution is faithful to a causal diagram if no conditional independence relations are present, other than the ones entailed by the Markov property.

direct effect on $\boldsymbol{Y}$ if performing an intervention on some temporal variable X_t^i will alter the distribution of Y_T, for some time steps t, T. For a precise definition of these concepts see Appendix A.

Following, e.g., Mastakouri et al. (2021), we define the *full time* graph $\mathcal{G}$ as a directed graph whose edges represent all direct causal effects among the variables at all time steps. Given the outcome Y_T at a given time step, we refer to the parent nodes in the full time graph as its *causal parents*. We further define the *summary* graph whose nodes are $\boldsymbol{X}^i$ and $\boldsymbol{Y}$, and with directed edges representing causal effects between the time series. We refer the reader to Figure 1 for a visualization of these graphs. Note that the causes of $\boldsymbol{Y}$ correspond to the parent nodes of $\boldsymbol{Y}$ in the summary graph.

3.4 Granger Causality

Granger causality (Granger, 1988) is one of the most commonly used approaches to infer causal relations from observational time-series data. Its central assumption is that "cause-effect relationships cannot work against time". Informally, if the prediction of the future of a target time-series $\boldsymbol{Y}$ can be improved by knowing past elements of another time-series $\boldsymbol{X}^i$, then $\boldsymbol{X}^i$ "Granger causes" $\boldsymbol{Y}$. Formally, we say that $\boldsymbol{X}^i$ Granger causes $\boldsymbol{Y}$ if it holds

$$\mathbb{P}\left(Y_T = y \mid \boldsymbol{X}_T^i = \boldsymbol{x}, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right) \neq \mathbb{P}\left(Y_T = y \mid \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right) \tag{1}$$

for a non-zero probability event $\{\boldsymbol{X}_T^i = \boldsymbol{x}, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\}$, where $\boldsymbol{I}_T$ stands for the set $\{\boldsymbol{X}_T^1, \boldsymbol{X}_T^2, \ldots, \boldsymbol{X}_T^m\}$ and $\boldsymbol{I}_T^{\setminus i}$ represents the set $\boldsymbol{I}_T \setminus \{\boldsymbol{X}_T^i\}$.

Granger causality is commonly used to identify causes. Assuming stationary, multivariate Granger causality analysis usually fits a vector autoregressive model to a given dataset. This model can be then used to determine the causes of a target $\boldsymbol{Y}$. However, it is important to note that Granger causality does not imply true causality in general. This limitation was acknowledged by Granger himself in Granger (1988).

4 Methodology

4.1 Double Machine Learning (DML)

DML is a general framework for parameter estimation, which uses debiasing techniques to achieve $\sqrt{n}$-consistency (see, e.g., Rotnitzky et al. (2020); Chernozhukov et al. (2022)). In DML, we consider the problem of estimating a parameter θ_0 as a solution of an equation of the form $\mathbb{E}\left[\mathcal{L}(\theta_0, \boldsymbol{\eta}_0)\right] = 0$. The score function $\mathcal{L}$ depends on two terms, the true parameter θ_0 that we wish to estimate, and a nuisance parameter $\boldsymbol{\eta}_0$. We do not directly care about the correctness of our estimate of $\boldsymbol{\eta}_0$, as long as we get a good estimator of θ_0. The nuisance parameter η_0 may induce an unwanted bias in the estimation process, resulting in slow convergence. To overcome this problem, we use score functions that fulfill the Mixed Bias Property (MBP) (Rotnitzky et al., 2020), and learn the desired parameter θ_0 using DML.

The MBP is a property that ensures that small changes of the nuisance parameter do not significantly affect the score function computed around the true parameters $(\theta_0, \boldsymbol{\eta}_0)$ (see Definition 1 by Rotnitzky et al. (2020)). I this work, we construct a score with the MBP as follows Chernozhukov et al. (2022); Rotnitzky et al. (2020). For a fixed time step T, let $\boldsymbol{X}$ be a random variable in the set of random variables $\boldsymbol{V}$, g any real-valued function of $\boldsymbol{X}$ such that $\mathbb{E}\left[g^2(\boldsymbol{X})\right] < \infty$. We consider parameters of the form $\theta_0 \coloneqq \mathbb{E}\left[m(\boldsymbol{V}; g)\right]$, where $m(\boldsymbol{V}; g)$ is a linear moment functional in g. The celebrated Riesz Representation Theorem ensures that, under certain conditions, there exists a function α_0 of $\boldsymbol{X}$ such that $\mathbb{E}\left[m(\boldsymbol{V}; g)\right] = \mathbb{E}\left[\alpha_0(\boldsymbol{X}) g(\boldsymbol{X})\right]$. The function α_0 is called the *Riesz Representer* (RR). Chernozhukov et al. (2021) shows that the Riesz representer can be estimated from samples. Using the RR, we can derive a score function for the parameter θ_0 with $g_0(\boldsymbol{X}) = \mathbb{E}\left[Y \mid \boldsymbol{X}\right]$ that fulfills the MBP. This function is defined as

$$\varphi(\theta, \boldsymbol{\eta}) \coloneqq m(\boldsymbol{V}; g) + \alpha(\boldsymbol{X}) \cdot (Y - g(\boldsymbol{X})) - \theta. \tag{2}$$

Here, $\boldsymbol{\eta} \coloneqq (\alpha, g)$ is a nuisance parameter consisting of a pair of square-integrable functions. As shown by Chernozhukov et al. (2022), the score function Eq. (2) yields $\mathbb{E}\left[\varphi(\theta_0, \boldsymbol{\eta})\right] = -\mathbb{E}\left[(\alpha(\boldsymbol{X}) - \alpha_0(\boldsymbol{X}))(g(\boldsymbol{X}) - g_0(\boldsymbol{X}))\right]$, which gives the MBP as in Definition 1 of Rotnitzky et al. (2020).

Algorithm 1 The DR-SIT

split the dataset D into k random disjoint subsets D_j of the same size;

for each dataset partition j **do**
  estimate $\hat{\boldsymbol{\eta}}_j^0 := (\hat{\alpha}_j^0, \hat{g}_j^0)$ on dataset $\mathsf{D} \setminus \mathsf{D}_j$, with $\hat{g}_j^0$ an estimate for g_0^0, and $\hat{\alpha}_j^0$ an estimate of the RR α_0^0 of $m_0(\boldsymbol{V}; g^0)$ as in Theorem 4.3;
end for
$\hat{\theta}^0 \leftarrow k^{-1} \sum_{j=1}^k \hat{\mathbb{E}}_{\mathsf{D}_j} \left[m_0(\boldsymbol{V}; \hat{g}_j^0) + \hat{\alpha}_j^0(\boldsymbol{X}) \cdot (Y - \hat{g}_j^0(\boldsymbol{X})) \right]$;

for each potential cause $\boldsymbol{X}^i$ **do**
  for each dataset partition j **do**
    estimate $\hat{\boldsymbol{\eta}}_j^i := (\hat{\alpha}_j^i, \hat{g}_j^i)$ on dataset $\mathsf{D} \setminus \mathsf{D}_j$, with $\hat{g}_j^i$ an estimate for g_0^i, and $\hat{\alpha}_j^i$ an estimate of the RR α_0^i of $m_i(\boldsymbol{V}; g^i)$ as in Theorem 4.3;
  end for
  $\hat{\theta}^i \leftarrow k^{-1} \sum_{j=1}^k \hat{\mathbb{E}}_{\mathsf{D}_j} \left[m_i(\boldsymbol{V}; \hat{g}_j^i) + \hat{\alpha}_j^i(\boldsymbol{X}) \cdot (Y - \hat{g}_j^i(\boldsymbol{X})) \right]$;
  perform a paired Student's t-test to determine if $\hat{\theta}^i \approx \hat{\theta}^0$, and select time series $\boldsymbol{X}^i$ as a cause if the null-hypotheses is rejected;
end for

return the selected time series;

For score functions that fulfill the MBP, we can use DML to partly remove the bias induced by the nuisance parameter. The DML is defined as follows.[6]

Definition 4.1 (DML, following Definition 3.2 by (Chernozhukov et al., 2018))**.** Given a dataset $\mathtt{D}$ of n observations, split the dataset D into k random disjoint subsets D_j of the same size. Consider a score function $\varphi(\theta, \boldsymbol{\eta})$ that fulfills the MBP as in (2). Construct estimators $\hat{\boldsymbol{\eta}}_j = (\hat{\alpha}_j, \hat{g}_j)$ for the nuisance parameter of the score using datasets $\mathtt{D} \setminus \mathtt{D}_j$. Then, construct an estimator $\hat{\theta}$ of the paraemeter θ as the solution to the following equation:

$$k^{-1} \sum_{j=1}^{k} \hat{\mathbb{E}}_{\mathtt{D}_j} \left[\varphi(\theta, \hat{\boldsymbol{\eta}}_j) \right] = 0,$$

where $\hat{\mathbb{E}}_{\mathtt{D}_j}[\,\cdot\,]$ is the empirical expected value over $\mathtt{D}_j$.

4.2 Granger Causality Implies True Causation under Axiom (A)-Axiom (C)

As discussed in Section 3.4, Granger causality does not imply true causality in general. In our case, however, Granger causality corresponds to true causation, as stated in the following result.

Theorem 4.2. *Consider a causal model as in Axiom (A)-Axiom (C). Then, a time series $\boldsymbol{X}^i$ is a potential cause of $\boldsymbol{Y}$ if and only if (iff.) $\boldsymbol{X}^i$ Granger causes $\boldsymbol{Y}$.*

Proof of this result is given in Appendix D. Importantly, Theorem 4.2 does not require causal faithfulness. We remark that Peters et al. (2013) shows that Granger causality implies true causation for fully autoregressive models (see also Löwe et al. (2022)). This result is based on the identifiability of additive noise models, in which all relevant variables are observed (Peters et al., 2011). Our framework, however, is more general than Peters et al. (2013); Löwe et al. (2022), since it allows for confounding among the covariates (see Figure 1). In the special case of a fully autoregressive model, Theorem 4.2 is equivalent to previous results (Peters et al., 2013; Löwe et al., 2022).

[6]We remark that DML requires a weaker assumption on the score function than the MBP, namely the Neyman Orthogonality Condition Neyman & Scott (1965); Chernozhukov et al. (2018). For simplicity, we give a description of DML only in terms of the MBP for linear moment functionals. However, Definition 4.1 can be generalized.

Table 1: AUROC score for the DR-SIT and common algorithms for Granger causality. Models with "✦" sign use deep neural networks. We remark that our estimator is much simpler than deep nets, and it has low sample complexity. We observe that our method achieves state-of-the-art performance on three tasks (E.Coli 1, E.Coli 2, Yeast 2) and that it obtains comparable performance on the remaining tasks (Yeast 1, Yeast 3). When reported, the error bounds represent 1 standard deviation of the AUROC score over 5 experiment repetitions

Method	**E.Coli 1**	**E.Coli 2**	**Yeast 1**	**Yeast 2**	**Yeast 3**
cMLP✦	0.644	0.568	0.585	0.506	0.528
cLSTM✦	0.629	0.609	0.579	0.519	0.555
TCDF✦	0.614	0.647	0.581	0.556	0.557
SRU✦	0.657	0.666	0.617	0.575	0.55
eSRU✦	0.66	0.629	0.627	0.557	0.55
DYNO.	0.590	0.547	0.527	0.526	0.510
PCMCI$^+$	0.530	0.519	0.530	0.510	0.512
Rhino✦ Reprod.	0.671 ±0.014	0.640 ±0.022	**0.656** ±0.011	0.565 ±0.012	0.549 ±0.004
Rhino+g✦ Reprod.	0.665 ±0.023	0.646 ±0.032	0.649 ±0.011	0.582 ±0.011	**0.571** ±0.010
DR-SIT (ours)	**0.704**±0.005	**0.680**±0.004	0.653±0.001	**0.585**±0.003	0.544±0.003

4.3 Testing Granger Causality with DML

Our approach to identifying potential causes consists of performing a statistical test to determine if Eq. 1 holds. Due to Theorem 4.2, a straightforward approach would then consist of using a conditional independence test, to select or discard a time series $\boldsymbol{X}^i$ as a cause of the outcome $\boldsymbol{Y}$. However, conditional independence testing is challenging in high-dimensional settings. Furthermore, kernel-based conditional independence tests (Fukumizu et al., 2007; Zhang et al., 2011; Park & Muandet, 2020; Sheng & Sriperumbudur, 2020) are computationally expensive. Instead, we provide a new statistical test based on DML. Our approach is based on the observation that under Axiom (A)-Axiom (C), the condition in Eq. 1 can be written in terms of simple linear moment functionals. The following theorem holds.

Theorem 4.3. *Consider the notation as in Eq. 1, and fix a time step T. For any square-integrable random variable $g^0(\boldsymbol{X}^i_T, \boldsymbol{I}^{\setminus i}_T)$, consider the moment functional $m_0(\boldsymbol{V}; g^0) := Y_T \cdot g^0$. Similarly, for any square-integrable random variable $g^i(\boldsymbol{I}^{\setminus i}_T)$, consider the moment functional $m_i(\boldsymbol{V}; g^i) := Y_T \cdot g^i$. Assuming Axiom (A)-Axiom (C), $\boldsymbol{X}^i$ Granger causes $\boldsymbol{Y}$ iff. it holds*

$$\mathbb{E}\left[m_0(\boldsymbol{V}; g^0_0)\right] - \mathbb{E}\left[m_i(\boldsymbol{V}; g^i_0)\right] \neq 0, \tag{3}$$

with $g^0_0(\boldsymbol{x}, \boldsymbol{i}) = \mathbb{E}[Y_T \mid \boldsymbol{X}^i_T = \boldsymbol{x}, \boldsymbol{I}^{\setminus i}_T = \boldsymbol{i}]$, and $g^i_0(\boldsymbol{i}) = \mathbb{E}[Y_T \mid \boldsymbol{I}^{\setminus i}_T = \boldsymbol{i}]$.

The proof is deferred to Appendix E. Intuitively, the parameter $\theta := \mathbb{E}[Y_T.\mathbb{E}[Y_T|I]]$ quantifies the cross-correlation between the target variable Y_T and its conditional mean $\mathbb{E}[Y_T|I]$, given the set of variables I. In general terms, variations in the parameter θ due to changes in the set I suggest the presence of a causal relationship between alterations in the set I and the target variable Y_T.

By this theorem, we can identify the causal parents of Y by testing Eq. 3. This boils down to learning parameters $\theta^0_0 := \mathbb{E}\left[m_0(\boldsymbol{V}; g^0_0)\right]$, $\theta^i_0 := \mathbb{E}\left[m_i(\boldsymbol{V}; g^i_0)\right]$, and then performing a sample test to verify if it holds $\theta^0 - \theta^i \neq 0$. Since both $m_0(\boldsymbol{V}; g^0_0)$ and $m_i(\boldsymbol{V}; g^i_0)$ we can use DML as in Definition 4.1 to estimate θ^0_0 and θ^i_0. Under mild convergence conditions on the nuisance parameters for $m_0(\boldsymbol{V}; g^0_0)$ and $m_i(\boldsymbol{V}; g^i_0)$, DML ensures fast convergence in distribution to the true parameters.

5 The Algorithm

5.1 Overview

Our algorithm learns causal relationships between time series, by testing Granger causality using DML, as outlined in Section 4.3. We refer to our approach as the DR-SIT (Structure Identification from Temporal Data). Our method is presented in Algorithm 1. This algorithm essentially performs the following steps:

(1) Select a potential cause $\boldsymbol{X}^i$ to test if $\boldsymbol{X}^i$ Granger causes $\boldsymbol{Y}$. Split the dataset D into k disjoint sets D_j with $k \geq 2$.

(2) Estimate $\hat{\boldsymbol{\eta}}_j^0 \coloneqq (\hat{\alpha}_j^0, \hat{g}_j^0)$ on dataset $\mathsf{D} \setminus \mathsf{D}_j$, with $\hat{g}_j^0$ an estimate for g_0^0, and $\hat{\alpha}_j^0$ an estimate of the RR α_0^0 of $m_0(\boldsymbol{V}; g^0)$ as in Theorem 4.3. Similarly, provide an estimate $\hat{\boldsymbol{\eta}}_j^i \coloneqq (\hat{\alpha}_j^i, \hat{g}_j^i)$ on dataset $\mathsf{D} \setminus \mathsf{D}_j$ for the pair $\boldsymbol{\eta}_0^i = (g_0^i, \alpha_0^i)$, with g_0^i as in Theorem 4.3 and α_0^i the RR α_0^i of $m_i(\boldsymbol{V}; g^i)$. This step corresponds to Line 3 and Line 8 of Algorithm 1.

(3) Provide an estimate $\hat{\theta}^0 \approx \mathbb{E}[m_0(\boldsymbol{V}; g^0)]$, by solving the equation $k^{-1} \sum_{j=1}^k \hat{\mathbb{E}}_{\mathsf{D}_j} \left[\varphi_0(\theta, \hat{\boldsymbol{\eta}}_j)\right] = 0$ with a score of the form $\varphi_0(\theta, \hat{\boldsymbol{\eta}}_j) \coloneqq m_0(\boldsymbol{V}; \hat{g}_j^0) + \hat{\alpha}_j^0(\boldsymbol{X}) \cdot (Y - \hat{g}_j^0(\boldsymbol{X})) - \theta$. This step corresponds to Line 5 of Algorithm 1. Provide an estimate $\hat{\theta}^i \approx \mathbb{E}[m_i(\boldsymbol{V}; g^i)]$ in a similar fashion, as in Line 10 of Algorithm 1.

(4) Use a two-sample Student's t-test to determine if $\mathbb{E}\left[\theta^0 - \theta^i\right]$ is approximately zero. Select $\boldsymbol{X}^i$ as a cause of $\boldsymbol{Y}$ if the null hypothesis is rejected. This step corresponds to Line 11 and 8 of Algorithm 1.

5.2 Strong Consistency Guarantees

In this paragraph, we provide an explanation for Step (4) of our algorithm. Under mild structural conditions on g_j^0, g_j^i and α_j^0, α_j^i Chernozhukov et al. (2022; 2018); Rotnitzky et al. (2020), the quantity $\theta^0 - \theta^i$ has $\sqrt{n}$-consistency. Hence, it holds

$$\sqrt{n}\left(\hat{\theta}^0 - \hat{\theta}^i\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right), \tag{4}$$

if and only if $\mathbb{E}\left[m_0(\boldsymbol{V}; g_0^0)\right] - \mathbb{E}\left[m_i(\boldsymbol{V}; g_0^i)\right] = 0$. Then, by Theorem 4.3, Eq. 4 holds iff. $\boldsymbol{X}^i$ does not Granger causes $\boldsymbol{Y}$. The notation in (4) means that the difference $\hat{\theta}^0 - \hat{\theta}^i$ converges in distribution to a zero-mean Gaussian for an increasing number of samples. That is, for all $\epsilon > 0$ and $\zeta > 0$ there exists $\delta > 0$, such that given $n > \delta$ samples it holds $\mathbb{P}\left(|\hat{\theta}^0 - \hat{\theta}^i| > \epsilon\right) \leq 1 - \Phi\left(\epsilon \frac{\sqrt{n}}{\sigma}\right) + \frac{\zeta}{2}$, if and only if $\mathbb{E}\left[m_0(\boldsymbol{V}; g_0^0)\right] - \mathbb{E}\left[m_i(\boldsymbol{V}; g_0^i)\right] = 0$. Here, Φ is the CDF of the standard Normal distribution. By this inequality, we can use a two-sample Student's t-test to determine if $\boldsymbol{X}^i$ Granger causes $\boldsymbol{Y}$, as in Line 11 of Algorithm 1. We refer the reader to Chernozhukov et al. (2022; 2018); Rotnitzky et al. (2020) for a detailed survey on strong consistency for DML and its relationship with the MBP.

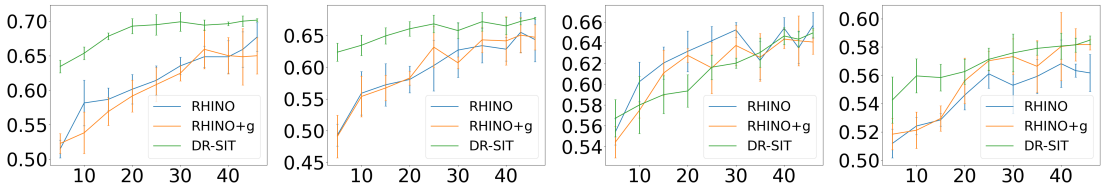
5.3 Complexity of Algorithm 1

Much of the run time of our algorithm consists of performing a regression to learn $\boldsymbol{\eta}_j^0$ and $\boldsymbol{\eta}_j^i$ on dataset D_j. Denote with d the time complexity of performing such a regression. For a given k-partition of the dataset and a fix potential cause $\boldsymbol{X}^i$, we can upper-bound the time complexity of our algorithm as $\mathcal{O}(dk)$. Here, d depends on the specific techniques used for the regression. Non-parametric regression can be performed efficiently in the problem size. We analyse the computational complexity in Appendix I further and show runtime plots in 6.3. We further discuss the run time and extension for full causal discovery in Appendix H.

We further improve efficiency in practice as follows: Instead of computing $\boldsymbol{\eta}_0^i = (g_0^i, \alpha_0^i)$ directly, we apply a zero-masking layer to the NNs used to estimate g_j^0 and α_j^0 for the features $\boldsymbol{X}^i$. This masking layer tells the sequence-processing layers that the input values for features $\boldsymbol{X}^i$ should be skipped. We then compute θ_0^i using the resulting surrogate models $\tilde{g}_j^i$ and $\tilde{\alpha}_j^i$. Please refer to Appendix F on an intuition behind zero-masking, and why zero-masking may not hurt the estimations. Using zero-masking dramatically improves run time, since it allows to perform only two regressions through the entire run time of Algorithm 1.

6 Experiments

In this section we provide an overview of the main experimental results. **We provide additional extensive experiments in Appendix J.**



(a) Task: E.Coli 1 (b) Task: Ecoli 2 (c) Task: Yeast 1 (d) Task: Yeast 2

Figure 2: This figure demonstrates the consistent performance of DR-SIT w.r.t number of training observations (i.e trajectories) compared to state-of-the-art methods Rhino and Rhino+g. For the same plots in all 5 tasks depicted in greater resolution, see Fig. 4 in Appendix J.2.

Table 2: Run time and Hardware used for our method (DR-SIT) and the state-of-the-art baseline Rhino.

Category\Method	**Rhino**	**DR-SIT (ours)**
Runtime	18 mins 40 sec $\pm$ 30 sec	57.12 sec $\pm$ 1.6 sec
Hardware	1 NVIDIA A100 GPU + AMD EPYC 7402 24-Core CPU	11th Gen Core i5-1140F CPU

6.1 Semi-Synthetic Experiments

The Dream3 dataset. Following previous related work (Tank et al., 2018; Khanna & Tan, 2019; Nauta et al., 2019; Bussmann et al., 2021; Gong et al., 2022), we evaluate performance with the Dream3 benchmark (Prill et al., 2010; Marbach et al., 2009). The Dream3 benchmark is a collection of gene expression level measurements across five different networks, where each network comprises 100 genes. The data in Dream3 are organized as time series. Specifically, for each of the five networks 46 trajectories are given; each trajectory details the progression of the 100 genes over 21 time steps after an initial perturbation in the gene values.

Description of the experiments. Our goal is to infer the causal structure of each network. This gives us a total of five tasks, i.e., E.Coli 1, E.Coli 2, Yeast 1, Yeast 2, and Yeast 3. We run our algorithm on this dataset and we use the area under the ROC curve (AUROC) as the performance metric. Following (Gong et al., 2022), we compare against the following baselines: cMLP (Tank et al., 2018), cLSTM (Tank et al., 2018), TCDF (Nauta et al., 2019), SRU (Khanna & Tan, 2019), eSRU (Khanna & Tan, 2019), Dynotears (Pamfil et al., 2020), Rhino+g (Gong et al., 2022), and Rhino (Gong et al., 2022).

Results. Shown in Table 1. The results for cMLP, cLSTM, TCDF, SRU and SRU are taken directly from Khanna & Tan (2019); Gong et al. (2022), where error bounds are not reported. Regarding Rhino+g and Rhino, we partially reproduce the experiments by Gong et al. (2022), using their source code. Our implementation of Rhino+g and Rhino differs from Gong et al. (2022) only in the choice for the hyper-parameters, which is the same on all five tasks. We specifically use the following hyper-parameters for Rhino: Node Embedd. = 16, Instantaneous eff. = False, Node Embedd. (flow) = 16, lag = 2, $\lambda_s = 19$, Auglag = 30. And we use the following for Rhino+g: Node Embedd. = 16, Instantaneous eff. = False, lag = 2, $\lambda_s = 15$, Auglag = 60. This is the setting that is reported for the Ecoli1 subtask and found in their corresponding code implementation. We opted for this approach because in the experiment by Gong et al. (2022), it seems that Rhino overfitted to the dataset. We learn η_j^i as in Line 3 of Algorithm 1, using a simple kernel ridge regression model with polynomial kernels of degree three combined with zero masking.

Overall, we observe that our method gives competitive results. Specifically, DR-SIT outperforms all the other benchmarks on three tasks (E.Coli 1, E.Coli 2, Yeast 2). Furthermore, DR-SIT obtains comparable performance on the remaining tasks (Yeast 1,Yeast 3). We also would like to emphasize that our estimator is much simpler than deep nets such as Rhino or Rhino+g. As such, it has lower sample complexity and lower run time than the other algorithms.

6.2 Low sample regime

One appealing property of DR-SIT is the strong consistency of the estimators (see Section 5). Due to this property, the estimators $\hat{\theta}^0$ and $\hat{\theta}^i$ exhibit fast convergence to the true parameters. Hence, our algorithm enjoys competitive results in low sample complexity settings, than richer models such as Rhino (Gong et al., 2022). We illustrate a compelling example of this, by comparing Rhino and Rhino+g against DR-SIT. In this example, DR-SIT learns nuisance parameters (Line 3 and Line 8 of Algorithm 1) using a simple kernel ridge regression estimator with polynomial kernels of degree three. In Fig. 2 we explicitly compare the performance of DR-SIT, Rhino, and Rhino+g as the number of trajectories used in training varies and demonstrate significantly improved performance especially in low sample settings.

6.3 Run Time and Hardware

To give a taste of the computational efficiency of our method, here we report the average runtime of DR-SIT and the competitive rival Rhino for experiment Table 1 across all five tasks (E.Coli 1, E.Coli 2, Yeast 1, Yeast 2 and Yeast 3) in Table 2. Despite having access only to a single CPU in contrast to the GPU-equipped execution of Rhino our method is almost 20x faster. This is because of the fact that DR-SIT algorithm will provide reasonable results even when employing simple fast efficient estimators (in this case a kernel regression). For a visual presentation of the AUROC performance progression vs runtime for various combinations of tasks and training sample sizes, see Figs. 5 to 9 in appendix.

7 Discussion

In this work, we propose an efficient algorithm for doubly robust structure identification from temporal data. We further provide asymptotical guarantees that our method is able to discover the direct causes even when there are cycles or hidden confounding and that our algorithm has $\sqrt{n}$-consistency. We extensively discuss the relations of the approach between the popular frameworks of Granger and Pearl's causality as well as relate and extend approaches from debiased machine learning to structure discovery from temporal data. We hope that our approach enables important real-world applications in bio-medicine where robustness to confounding, sample efficiency, and ease of use are important for causal discovery from observational time series.

Impact Statement

In many safety-critical applications, one needs to infer causal structure from observational data e.g. in healthcare, economics, or geosciences since interventions are unethical or simply impossible. For these cases, one requires algorithms that have theoretical guarantees, are robust to hidden confounding, and are computationally efficient. We hope that in the long term, our work can thus contribute to better machine learning-driven decision-making in safety-critical environments.

While theoretical guarantees and extensive experiments are critical components for the evaluation of algorithms, especially in safety-critical environments this can potentially lead to a false sense of security and trust. Constant monitoring and assessment by domain experts are needed before and during the deployment of any machine learning algorithm, especially in safety-critical environments.

One key limitation of our approach as well as of all causal discovery approaches from observational data is the reliance on additional assumptions. These assumptions are required since causal discovery from observational data is impossible without them. We related our assumptions to previous ones in the Framework Section 3. These assumptions of our approach need to be checked before applying the approach but are less restrictive than for comparable baselines. Moreover, since our approach is inspired by debiased machine learning, the focus is on inferring the true or "debiased" underlying structure rather than obtaining low variance estimates of the structure.

References

Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(1), 2010.

Abdul Rauf Anwar, Kidist Gebremariam Mideska, Helge Hellriegel, Nienke Hoogenboom, Holger Krause, Alfons Schnitzler, Günther Deuschl, Jan Raethjen, Ulrich Heute, and Muthuraman Muthuraman. Multimodal causality analysis of eyes-open and eyes-closed data from simultaneously recorded eeg and meg. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2825–2828. IEEE, 2014.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

David Benkeser, Marco Carone, MJ Van Der Laan, and Peter B Gilbert. Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880, 2017.

Wicher Pieter Bergsma. Testing conditional independence for continuous random variables. *Report Eurandom*, 2004.

Kenneth A Bollen. *Structural equations with latent variables*, volume 210. Wiley, 1989.

Bart Bussmann, Jannes Nys, and Steven Latré. Neural additive vector autoregression models for causal discovery in time series. In *Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings 24*, pp. 446–460. Springer, 2021.

Gang Chen, J Paul Hamilton, Moriah E Thomason, Ian H Gotlib, Ziad S Saad, and Robert W Cox. Granger causality via vector auto-regression tuned for fmri data analysis. In *Proc Intl Soc Mag Reson Med*, volume 17, pp. 1718, 2009.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

Victor Chernozhukov, Whitney K. Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. Automatic debiased machine learning via neural nets for generalized linear regression. *arXiv:2104.14737*, 2021.

Victor Chernozhukov, Whitney Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. RieszNet and ForestRiesz: Automatic debiased machine learning with neural nets and random forests. In *Proc. of ICML*, pp. 3901–3914, 2022.

David Maxwell Chickering. Learning bayesian networks is np-complete. *Learning from data: Artificial intelligence and statistics V*, pp. 121–130, 1996.

Max Chickering, David Heckerman, and Chris Meek. Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.

Doris Entner and Patrik O Hoyer. On causal discovery from time series data using fci. *Probabilistic graphical models*, pp. 121–128, 2010.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, pp. 489–496, 2007.

Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767, 2011.

Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. *Advances in Neural Information Processing Systems*, 33:12615–12625, 2020.

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

Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. *CoRR*, abs/2210.14706, 2022.

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

Clive WJ Granger. Some recent development in a concept of causality. *Journal of econometrics*, 39(1-2): 199–211, 1988.

Luigi Gresele, Julius Von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? In *Proc. of NeurIPS*, volume 34, pp. 28233–28248, 2021.

Uzma Hasan, Emam Hossain, and Md Osman Gani. A survey on causal discovery methods for temporal and non-temporal data. *arXiv preprint arXiv:2303.15027*, 2023.

Miguel Ángel Hernán, Babette Brumback, and James M Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, pp. 561–570, 2000.

Craig Hiemstra and Jonathan D Jones. Testing for linear and nonlinear granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5):1639–1664, 1994.

Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. Causal discovery and forecasting in nonstationary environments with state-space models. In *International conference on machine learning*, pp. 2901–2910. PMLR, 2019.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11:1709–1731, 2010.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pp. 2907–2916. PMLR, 2020.

Saurabh Khanna and Vincent YF Tan. Economy statistical recurrent units for inferring nonlinear granger causality. *arXiv preprint arXiv:1911.09879*, 2019.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *Proc. of AISTATS*, pp. 2207–2217, 2020.

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

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Proc. of CLearR*, pp. 428–484, 2022.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. icitris: Causal representation learning for instantaneous temporal effects. *CoRR*, abs/2206.06169, 2022a.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *Proc. of ICML*, pp. 13557–13603, 2022b.

Sindy Löwe, David Madras, Richard Z. Shilling, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Proc. of CLeaR*, pp. 509–525, 2022.

Zheng Lu, Chen Zhou, Jing Wu, Hao Jiang, and Songyue Cui. Integrating granger causality and vector auto-regression for traffic prediction of large-scale wlans. *KSII Transactions on Internet and Information Systems (TIIS)*, 10(1):136–151, 2016.

D. Marbach, T. Schaffter, C. Mattiussi, and D. Floreano. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology*, 16(2):229–239, 2009.

Daniele Marinazzo, Wei Liao, Huafu Chen, and Sebastiano Stramaglia. Nonlinear connectivity by granger causality. *Neuroimage*, 58(2):330–338, 2011.

Atalanti-Anastasia Mastakouri, Bernhard Schölkopf, and Dominik Janzing. Necessary and sufficient conditions for causal feature selection in time series with latent common causes. In *Proceedings of ICML*, volume 139, pp. 7502–7511, 2021.

Søren Wengel Mogensen, Daniel Malinsky, and Niels Richard Hansen. Causal learning for partially observed stochastic dynamical systems. In *UAI*, pp. 350–360, 2018.

Raha Moraffah, Paras Sheth, Mansooreh Karami, Anchit Bhattacharya, Qianru Wang, Anique Tahir, Adrienne Raglin, and Huan Liu. Causal inference for time series analysis: Problems, methods and evaluation. *Knowledge and Information Systems*, 63:3041–3085, 2021.

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

Jerzy Neyman and Elizabeth L. Scott. Asymptotically optimal tests of composite hypotheses for randomized experiments with noncontrolled predictor variables. *Journal of the American Statistical Association*, 60: 699–721, 1965. ISSN 1537274X.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pp. 1595–1605. PMLR, 2020.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In *Advances in Neural Information Processing Systems*, 2020.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. In *Proc. of UAI*, pp. 589–598, 2011.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *Proc. of NIPS*, pp. 154–162, 2013.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

R. J. Prill, D. Marbach, J. Saez-Rodriguez, P. K. Sorger, L. G. Alexopoulos, X. Xue, N. D. Clarke, G. Altan-Bonnet, and G. Stolovitzky. Towards a rigorous assessment of systems biology models: the dream3 challenges. *PloS one*, 5(2):e9202, 2010.

Francesco Quinzan, Ashkan Soleymani, Patrick Jaillet, Cristian R Rojas, and Stefan Bauer. Drcfs: Doubly robust causal feature selection. In *International Conference on Machine Learning*, pp. 28468–28491, 2023.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, pp. 550–560, 2000.

A Rotnitzky, E Smucler, and J M Robins. Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238, 08 2020.

Paul K Rubenstein, Stephan Bongers, Bernhard Schölkopf, and Joris M Mooij. From deterministic odes to dynamic structural causal models. *arXiv preprint arXiv:1608.08028*, 2016.

Jakob Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310, 2018.

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

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553, 2019a.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019b.

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *arXiv:1804.07203v6*, 2020.

Tianhong Sheng and Bharath K. Sriperumbudur. On distance and kernel measures of conditional independence, 2020.

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

Tymon Słoczyński and Jeffrey M Wooldridge. A general double robustness result for estimating average treatment effects. *Econometric Theory*, 34(1):112–133, 2018.

Ashkan Soleymani, Anant Raj, Stefan Bauer, Bernhard Schölkopf, and Michel Besserve. Causal feature selection via orthogonal search. *Transactions on Machine Learning Research*, 2022.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Adolf Stips, Diego Macias, Clare Coughlan, Elisa Garcia-Gorriz, and X San Liang. On the causal structure between co2 and global temperature. *Scientific reports*, 6(1):21691, 2016.

George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *science*, 338(6106):496–500, 2012.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily Fox. Neural granger causality for nonlinear time series. *stat*, 1050:16, 2018.

Sebastian Weichwald, Martin E Jakobsen, Phillip B Mogensen, Lasse Petersen, Nikolaj Thams, and Gherardo Varando. Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values. In *NeurIPS 2019 Competition and Demonstration Track*, pp. 27–36. PMLR, 2020.

D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

Tailin Wu, Thomas M. Breuel, Michael Skuhersky, and Jan Kautz. Discovering nonlinear relations with minimum predictive information regularization. *CoRR*, abs/2001.01885, 2020.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data, 2021.

Kui Yu, Xianjie Guo, Lin Liu, Jiuyong Li, Hao Wang, Zhaolong Ling, and Xindong Wu. Causality-based feature selection: Methods and evaluations. *ACM Computing Surveys (CSUR)*, 53(5):1–36, 2020.

Kui Yu, Lin Liu, and Jiuyong Li. A unified view of causal and non-causal feature selection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(4):1–46, 2021.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Conference on Uncertainty in Artificial Intelligence*, pp. 804–813, 2011.

A Direct Causal Effects and Interventions

In this section, we clarify the notion of an intervention and direct causal effects. We will also introduce some notation that will be used later on in the proofs.

We consider interventions by which a random variable X_t^j is set to a constant $X_t^j \leftarrow x$. We denote with $Y_T \mid do(X_t^i = x)$ the outcome time series $\boldsymbol{Y}$ at time step T, after performing an intervention as described above. We can likewise perform multiple joint interventions, by setting a group of random variables $\boldsymbol{I}$ at different time steps, to pre-determined constants specified by an array $\boldsymbol{i}$. We use the symbol $Y_T \mid do(\boldsymbol{I} = \boldsymbol{i})$ to denote the resulting post-interventional outcome, and we denote with $\mathbb{P}\left(Y_T = y \mid do(\boldsymbol{I} = \boldsymbol{i})\right)$ the probability of the event $\{Y_T \mid do(\boldsymbol{I} = \boldsymbol{i}) = y\}$.

Using this notation, a time series $\boldsymbol{X}^i$ has a direct effect on the outcome $\boldsymbol{Y}$, if performing different interventions on the variables $\boldsymbol{X}^i$, while keeping the remaining variables fixed, will alter the probability distribution of the outcome $\boldsymbol{Y}$. Formally, define the sets of random variables $\boldsymbol{I}_T \coloneqq \{X_t^1, \ldots, X_t^n, Y_t\}_{t<T}$, which consists of all the information before time step T. Similarly, define the random variable $\boldsymbol{X}_T^i \coloneqq \{X_t^i\}_{t<T}$, consisting of all the information of time series $\boldsymbol{X}^i$ before time step T. Define the variable $\boldsymbol{I}_T^{\setminus i} \coloneqq \boldsymbol{I}_T \setminus \boldsymbol{X}_T^i$, which consists of all the variables in $\boldsymbol{I}_T$ except for $\boldsymbol{X}_T^i$. Then, a time series $\boldsymbol{X}^i$ has a direct effect on the outcome $\boldsymbol{Y}$ if it holds

$$\mathbb{P}\left(Y_T = y \mid do\left(\boldsymbol{X}_T^i = \boldsymbol{x}', \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right)\right) \neq \mathbb{P}\left(Y_T = y \mid do\left(\boldsymbol{X}_T^i = \boldsymbol{x}'', \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right)\right) \tag{5}$$

We say that a time series $\boldsymbol{X}^i$ causes $\boldsymbol{Y}$, if there is a direct effect between $\boldsymbol{X}^i$ and $\boldsymbol{Y}$ as in Eq. 5, for any time step T.

B Necessity of the Statistical Independence of ε_t

We provide a counterexample, to show that if there are dependencies between the noise and the historical data, then the causal structure may not be identifiable from observational data. To this end, we consider a first dataset $\{X_t, Y_t\}_{t\in\mathbb{Z}}$, defined as

$$X_{t-1}, Y_t \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

Here, $\mathcal{N}(\mathbf{0}, \Sigma)$ is a zero-mean joint Gaussian distribution with covariance matrix

$$\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

We also consider a second dataset $\{W_t, Z_t\}_{t\in\mathbb{Z}}$, defined as

$$W_t \sim \mathcal{N}(0, 1), \quad Z_t = W_{t-1}$$

The parameter Σ is defined as above. Both datasets entail the same joint probability distribution. However, the causal diagrams change from one dataset to the other. Hence, the causal structure cannot be recovered from observational data, if the posterior additive noise ε_t is correlated with some of the covariates.

C Necessity of No Instantaneous Causal Effects between $\boldsymbol{Y}$ and the Potential Causes $\boldsymbol{X}^i$

Here, we provide a counterexample to show that without the no instantaneous causal effect, the causal structure may not be identifiable from observational data. Consider the following two models:

- Model 1: We consider time series $\{X_t\}$, $\{Y_t\}$ of the form $X_t = \mathbb{E}[X_{t-1}] + c$ and $Y_t = \mathbb{E}[Y_{t-1} - X_{t-1}] + X_t$. In this model, c is a random variable drawn from a Gaussian distribution with a mean of 0 and covariance of 1.
- Model 2: We consider time series $\{X_t\}$, $\{Y_t\}$ of the form $Y_t = \mathbb{E}[Y_{t-1}] + c$ and $X_t = \mathbb{E}[X_{t-1} - Y_{t-1}] + Y_t$. In this model, c is a random variable drawn from a Gaussian distribution with a mean of 0 and covariance of 1.

Both models entail the same joint distribution. However, in Model 1, X has a causal effect on Y, whereas in Model 2, Y has a causal effect on X. Hence, in this example, the causal structure is not identifiable.

D Proof of Theorem 4.2

We prove the following result.

Theorem 4.2. *Consider a causal model as in Axiom (A)-Axiom (C). Then, a time series $\boldsymbol{X}^i$ is a potential cause of $\boldsymbol{Y}$ if and only if (iff.) $\boldsymbol{X}^i$ Granger causes $\boldsymbol{Y}$.*

Proof. We first prove that it holds

$$\mathbb{P}\left(Y_T = y \mid do(X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i})\right) = \mathbb{P}\left(Y_T = y \mid X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right) \tag{6}$$

for any non-zero event $\{Y_T = y\}$. To this end, define the group $\boldsymbol{P}$ consisting of all the causal parents of the outcome. Note that $\boldsymbol{P} \subseteq \{X_t^i, \boldsymbol{I}_T^{\setminus i}\}$. By Axiom (A), the outcome can be described as $Y = f(\boldsymbol{P}) + \varepsilon$, where ε is independent of $\{X_t^i, \boldsymbol{I}_T^{\setminus i}\}$. Hence by Rule 2 of the do-calculus (see Pearl (2000), page 85) Eq. 6 holds, since Y becomes independent of $\{X_t^i, \boldsymbol{I}_T^{\setminus i}\}$ once all arrows from $\boldsymbol{P}$ to Y are removed from the graph of the DGP.

We now prove the claim using Eq. 6. To this end, assume that Eq. 6 holds and suppose that X^i does not Granger causes Y, i.e., it holds

$$\mathbb{P}\left(Y_T = y \mid X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right) = \mathbb{P}\left(Y_T = y \mid \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right), \tag{7}$$

for any non-zero event $\{Y_T = y\}$. Then,

$$\begin{aligned}
\mathbb{P}\left(Y_T = y \mid do(X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i})\right) &= \mathbb{P}\left(Y_T = y \mid X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right) && \text{[Eq. 6]}\\
&= \mathbb{P}\left(Y_T = y \mid \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right) && \text{[Eq. 7]}\\
&= \mathbb{P}\left(Y_T = y \mid X_t^i = x', \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right) && \text{[Eq. 7]}\\
&= \mathbb{P}\left(Y_T = y \mid do(X_t^i = x', \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i})\right). && \text{[Eq. 6]}
\end{aligned}$$

Hence, causality implies Granger causality.

We now prove that Granger causality implies causality. To this end, suppose that X^i is not a potential cause of Y. By the definition of direct effects, it holds

$$\begin{aligned}
&\mathbb{P}\left(Y_T = y \mid do(X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i})\right)\\
&\qquad = \mathbb{E}\left[\mathbb{P}\left(Y_T = y \mid do(X_t^i = x', \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i})\right) \mid \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right].
\end{aligned} \tag{8}$$

Hence,

$$\begin{aligned}
&\mathbb{P}\left(Y_T = y \mid do(X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i})\right)\\
&\qquad = \mathbb{P}\left(Y_T = y \mid X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right) && \text{[Eq. 6]}\\
&\qquad = \mathbb{E}\left[\mathbb{P}\left(Y_T = y \mid do(X_t^i, \boldsymbol{I}_T^{\setminus i})\right) \mid \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right] && \text{[Eq. 8]}\\
&\qquad = \mathbb{E}\left[\mathbb{P}\left(Y_T = y \mid X_t^i, \boldsymbol{I}_T^{\setminus i}\right) \mid \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right] && \text{[Eq. 6]}\\
&\qquad = \mathbb{P}\left(Y_T = y \mid \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right),
\end{aligned}$$

and the claim follows. □

E Proof of Theorem 4.3

We prove the following result.

Theorem 4.3. *Consider the notation as in Eq. 1, and fix a time step T. For any square-integrable random variable $g^0(\boldsymbol{X}_T^i, \boldsymbol{I}_T^{\setminus i})$, consider the moment functional $m_0(\boldsymbol{V}; g^0) \coloneqq Y_T \cdot g^0$. Similarly, for any square-integrable random variable $g^i(\boldsymbol{I}_T^{\setminus i})$, consider the moment functional $m_i(\boldsymbol{V}; g^i) \coloneqq Y_T \cdot g^i$. Assuming Axiom (A)-Axiom (C), $\boldsymbol{X}^i$ Granger causes $\boldsymbol{Y}$ iff. it holds*

$$\mathbb{E}\left[m_0(\boldsymbol{V}; g_0^0)\right] - \mathbb{E}\left[m_i(\boldsymbol{V}; g_0^i)\right] \neq 0, \tag{3}$$

with $g_0^0(\boldsymbol{x}, \boldsymbol{i}) = \mathbb{E}[Y_T \mid \boldsymbol{X}_T^i = \boldsymbol{x}, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}]$, and $g_0^i(\boldsymbol{i}) = \mathbb{E}[Y_T \mid \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}]$.

In order to prove Theorem 4.3, we use the following auxiliary lemma.

Lemma E.1. *Consider a causal model as in Axiom (A)-Axiom (C). Then, the following conditions are equivalent:*

1. $\mathbb{E}\left[Y_T \mid X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right] = \mathbb{E}\left[Y_T \mid X_t^i = x', \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right]$ *a.s. ;*
2. $\mathbb{P}\left(Y_T = y \mid X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right) = \mathbb{P}\left(Y_T = y \mid X_t^i = x', \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right)$ *a.s.*

Proof. Clearly, Item 2 implies Item 1.

We now prove the converse, i.e., we show that Item 1 implies Item 2. To this end, define the group $\boldsymbol{P}_T$ consisting of all the causal parents of Y_T. Note that it holds $\boldsymbol{P}_T \subseteq \{\boldsymbol{I}_T^{\setminus i}, X_t^i\} \subseteq \{\boldsymbol{I}_T^{\setminus i}, X_t^i\}$. Hence, the joint intervention $\{X_t^i, \boldsymbol{i}_{T-1}^{i,t}\} \leftarrow \{x, \boldsymbol{i}\}$ define an intervention on the parents $\boldsymbol{P}_T \leftarrow \boldsymbol{p}$. Further, we can write the potential outcome as

$$Y_T \mid do(X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}) = f(\boldsymbol{p}) + \varepsilon. \tag{9}$$

Similarly, the joint intervention $\{X_t^i, \boldsymbol{I}_{T-1}^{i,t}\} \leftarrow \{x, \boldsymbol{i}\}$, define an intervention on the parents $\boldsymbol{P}_T \leftarrow \boldsymbol{p}'$. We can write the potential outcome as

$$Y_T \mid do(X_t^i = x', \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}) = f(\boldsymbol{p}') + \varepsilon. \tag{10}$$

Hence, it holds

$$\begin{aligned}
f(\boldsymbol{p}) + \mathbb{E}[\varepsilon] &= \mathbb{E}\left[Y_T \mid do(X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i})\right] && \text{[Eq. 9]} \\
&= \mathbb{E}\left[Y_T \mid X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right] && \text{[Eq. 6, Theorem 4.2]} \\
&= \mathbb{E}\left[Y_T \mid X_t^i = x', \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right] && \text{[by assumption]} \\
&= \mathbb{E}\left[Y_T \mid do(X_t^i = x', \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i})\right] && \text{[Eq. 6, Theorem 4.2]} \\
&= f(\boldsymbol{p}') + \mathbb{E}[\varepsilon]. && \text{[Eq. 10]}
\end{aligned}$$

By Axiom (A), the variable ε is exogenous independent noise. From the chain of equations above it follows that $f(\boldsymbol{p}) = f(\boldsymbol{p}')$. Hence,

$$\begin{aligned}
\mathbb{P}\left(Y_T = y \mid do(X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i})\right) &= \mathbb{P}\left(f(\boldsymbol{p}) + \varepsilon = y\right) \\
&= \mathbb{P}\left(f(\boldsymbol{p}') + \varepsilon = y\right) = \mathbb{P}\left(Y_T = y \mid do(X_t^i = x', \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i})\right)
\end{aligned} \tag{11}$$

We conclude that it holds

$$\begin{aligned}
&\mathbb{P}\left(Y_T = y \mid X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right) \\
&\quad = \mathbb{P}\left(Y_T \mid do(X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i})\right) && \text{[Eq. 6, Theorem 4.2]} \\
&\quad = \mathbb{P}\left(Y_T = y \mid do(X_t^i = x', \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i})\right) && \text{[Eq. 11]} \\
&\quad = \mathbb{P}\left(Y_T = y \mid X_t^i = x', \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right), && \text{[Eq. 6, Theorem 4.2]}
\end{aligned}$$

as claimed. □

We can now prove the main result.

Proof of Theorem 4.3. We first prove that X^i Granger causes Y iff. it holds

$$\mathbb{E}\left[\left(\mathbb{E}\left[Y_T \mid X_t^i, \boldsymbol{I}_T^{\setminus i}\right] - \mathbb{E}\left[Y_T \mid \boldsymbol{I}_T^{\setminus i}\right]\right)^2\right] \neq 0. \tag{12}$$

First, suppose that Eq. 12 does not hold. Then, it holds $[Y_T \mid X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}] = \mathbb{E}[Y_T \mid X_t^i = x', \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}]$, a.s. Combining this equation with Lemma E.1 yields

$$\begin{aligned}
\mathbb{P}\left(Y_T = y \mid X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right) &= \mathbb{E}\left[\mathbb{P}\left(Y_T = y \mid X_t^i, \boldsymbol{I}_T^{\setminus i}\right) \mid \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right] \\
&= \mathbb{P}\left(Y_T = y \mid \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right),
\end{aligned}$$

a.s. Hence, if X^i Granger causes Y, then Eq. 12 holds.

$$\mathbb{E}\left[Y_T \mid X_t^i = x, \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right] \neq \mathbb{E}\left[Y_T \mid X_t^i = x', \boldsymbol{I}_T^{\setminus i} = \boldsymbol{i}\right], \tag{13}$$

for a triple $\{x, x', \mathbf{w}\}$. By combining Eq. 13 with Lemma E.1 we see that Eq. 12 implies causality. However, by Theorem 4.2 Granger causality is equivalent to causality in this case.

We now prove the claim. By the tower property of the expectation Williams (1991) that

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathbb{E}\left[Y_T \mid X_t^i, \boldsymbol{I}_T^{\setminus i}\right] - \mathbb{E}\left[Y_T \mid \boldsymbol{I}_T^{\setminus i}\right]\right)^2\right] \\
&= \mathbb{E}\left[\left(\mathbb{E}\left[Y_T \mid X_t^i, \boldsymbol{I}_T^{\setminus i}\right] - \mathbb{E}\left[Y_T \mid \boldsymbol{I}_T^{\setminus i}\right]\right)^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\mathbb{E}\left[Y_T \mid X_t^i, \boldsymbol{I}_T^{\setminus i}\right] - \mathbb{E}\left[Y_T \mid \boldsymbol{I}_T^{\setminus i}\right]\right)^2 \mid \boldsymbol{I}_T^{\setminus i}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\mathbb{E}\left[Y_T \mid X_t^i, \boldsymbol{I}_T^{\setminus i}\right]^2 - \mathbb{E}\left[Y_T \mid X_t^i, \boldsymbol{I}_T^{\setminus i}\right]\mathbb{E}\left[Y_T \mid \boldsymbol{I}_T^{\setminus i}\right]\right) \mid \boldsymbol{I}_T^{\setminus i}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_T \mid X_t^i, \boldsymbol{I}_T^{\setminus i}\right]^2\right] - \mathbb{E}\left[\mathbb{E}\left[\left(\mathbb{E}\left[Y_T \mid X_t^i, \boldsymbol{I}_T^{\setminus i}\right]\mathbb{E}\left[Y_T \mid \boldsymbol{I}_T^{\setminus i}\right]\right) \mid \boldsymbol{I}_T^{\setminus i}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_T \mid X_t^i, \boldsymbol{I}_T^{\setminus i}\right]^2\right] - \mathbb{E}\left[\mathbb{E}\left[Y_T \mid \boldsymbol{I}_T^{\setminus i}\right]^2\right] \\
&= \mathbb{E}\left[Y_T\mathbb{E}\left[Y_T \mid X_t^i, \boldsymbol{I}_T^{\setminus i}\right]\right] - \mathbb{E}\left[Y_T\mathbb{E}\left[Y_T \mid \boldsymbol{I}_T^{\setminus i}\right]\right],
\end{aligned}$$

as claimed. □

F Intuition on Zero-Masking

We provide intuition why masking is a reasonably good idea. Assume that the function $\widehat{f}$ is a ϵ-close estimator of the true function f^* in the $\mathcal{L}^2(P_{X_1,X_2,\ldots,X_m})$ norm, where the functions $\widehat{f}, f^*$ and the joint probability distribution $P_{X_2,\ldots,X_m}$ are defined on the set $\{X_1, X_2, \ldots, X_m\}$,

$$\|\widehat{f} - f^*\|_{\mathcal{L}^2(P_{X_1,X_2,\ldots,X_m})} \leq \epsilon$$

Now, let's mask the random variable X_1. We are interested to see how close is the estimator $\mathbb{E}_{X_1}\widehat{f}$ to the true function $\mathbb{E}_{X_1}f^*$ in the $\mathcal{L}^2(P_{X_2,\ldots,X_m})$ norm, where the functions $\mathbb{E}_{X_1}\widehat{f}, \mathbb{E}_{X_1}f^*$ and the marginal probability distribution $P_{X_2,\ldots,X_m}$ are defined on the rest of variables $\{X_2, \ldots, X_m\}$, By Jensen's inequality, we infer that for any realization of $X_2 = x_2, X_3 = x_3, \ldots, X_m = x_m$,

$$\begin{aligned}(\mathbb{E}_{X_1}\widehat{f}(X_1, x_2, \ldots, x_m) - \mathbb{E}_{X_1}f^*(X_1, x_2, \ldots, x_m))^2 \\ \leq \mathbb{E}_{X_1}[(\widehat{f}(X_1, x_2, \ldots, x_m) - f^*(X_1, x_2, \ldots, x_m))^2]\end{aligned}$$

Plugging it in the ϵ-closeness assumption leads to,

$$\|\mathbb{E}_{X_1}\widehat{f} - \mathbb{E}_{X_1}f^*\|_{\mathcal{L}^2(P_{X_2,\ldots,X_m})} \leq \|\widehat{f} - f^*\|_{\mathcal{L}^2(P_{X_1,X_2,\ldots,X_m})} \leq \epsilon,$$

which guarantees that $\mathbb{E}_{X_1}\widehat{f}$ is also ϵ-closs to $\mathbb{E}_{X_1}f^*$ and hence it's a good estimator. In the sequel, a natural solution would be to estimate $\mathbb{E}_{X_1}\widehat{f}$ by taking averages of $\widehat{f}$ over different samples of X_1. However, for the linear regression problem that the estimator has a linear structure of the input, it is straightforward to show that it is enough to evaluate $\widehat{f}$ at $\mathbb{E}[X_1]$. And finally due to the zero-centering step of data preprocessing, $\mathbb{E}[X_1] = 0$. Thus, the aforementioned procedure is equivalent to zero-masking.

G A Note on the Number of Partitions

The number of partitions k affects the performance of our algorithm in practice since a larger number of partitions will help in removing a bias in the estimates. However, in our experiments, we observe that a small number of partitions is sufficient to achieve good results. Furthermore, an excessive number of random partitions may have a detrimental effect, due to the possible small number of samples in each partition. Hence, we believe that the number of partitions will not drastically affect performance in practice. Reasonable choices of k for our experiments range between 3-7, hence $k = \mathcal{O}(1)$ w.r.t. parameters of the problem. Thus, the resulting runtime can be reported as $\mathcal{O}(md)$.

H Extension to Full Causal Discovery

It is possible to use Algorithm 1 for full causal discovery, for fully-observed acyclic auto-regressive models with no instantaneous effects (see Peters et al. (2013); Löwe et al. (2022) for a precise definition of this restricted framework). In fact, under these more restrictive assumptions, we can identify the causes of each random variable of the model by testing Granger causality. For these models, we can learn the full summary graph, by identifying the causes of each variable with Algorithm 1. The resulting run time can be quantified as $\mathcal{O}(mdk)$, where d is the time complexity of performing a regression, as outlined above, m is the number of time series considered, and k is the number of dataset partitions used for double cross-fitting. Since $k = \mathcal{O}(1)$ w.r.t. parameters of the problem (see Appendix G), the runtime of DR-SIT is $\mathcal{O}(md)$.

I Computational Complexity and Comparison

As discussed in Section 2, compared to DR-SIT, conditional independence-based approaches such as PCMCI (Runge et al., 2019b), PCMCI+ (Runge, 2020), and LPCMCI (Gerhardus & Runge, 2020) face exponential computational barriers. It is widely known that even endowed with a perfect infinite sample

Table 3: Table with runtime means and standard deviations for DR-SIT and PCMCI+ (in seconds).

	10	20	30	40	50	100	200	400
DR-SIT	42 ± 13	35 ± 6	29 ± 1	30 ± 4	35 ± 15	41 ± 6	62 ± 6	77 ± 10
PCMCI+	4.2 ± 0.4	18.6 ± 0.8	48.6 ± 1.4	99.2 ± 10	216.4 ± 42.2	1091 ± 65	5678 ± 264	≈ 8 hours

independence testing oracle, learning Bayesian Networks becomes NP-Hard (Chickering et al., 2004; Chickering, 1996). Consequently, computational challenges arise not only due to the nature of the conditional independence tests themselves but also from the computational intractability of searching through the exponentially large space of possible network structures. Hence, the runtime of the $\mathcal{O}(m)$ number of regressions that DR-SIT demands is negligible compared to the exponential number of conditional independence tests from lengthy time-series. To support this argument in practice, we provide a runtime comparison between DR-SIT and PCMCI+ w.r.t. the number of nodes m in Table 3.

J Additional Experiments

J.1 Synthetic Experiments

Dataset generation. We first set the number of potential causes m, and we fix the lag Δ and the number of time steps for the dataset T. We create a covariate adjacency 3-dimensional tensor Σ of dimensions $\Delta \times m \times m$. This tensor has 0-1 coefficients, where $\Sigma_{k,i,j} = 1$ if X^i_{t-k} has a casual effect on X^j_t for all time steps t. Similarly, we create an adjacency matrix $\Sigma^{\boldsymbol{Y}}$ for the outcome $\boldsymbol{Y}$. $\Sigma^{\boldsymbol{Y}}$ is a binary $\Delta \times m$ array, such that $\Sigma^{\boldsymbol{Y}}_{k,j} = 1$ if X^j_{t-k} has a causal effect on Y_t for all time steps t. The entries of Σ and $\Sigma^{\boldsymbol{Y}}$ follow the Bernouli distribution with parameter $p = 0.5$. Note that the resulting casual structure fulfills Axiom (B)-Axiom (C).

We then create m transformations that are used to produce the potential causes $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^m$. Each one of these transformations is modeled by an MLP with 1 hidden layer and 200 hidden units. We use Tanh nonlinearities (included also in the output layer) in order to control the scale of the values. The final output value is further scaled up so that all transforms generate values in the range $[-10, 10]$. The input layer of each MLP is coming from the corresponding causal parents of the corresponding time series, as calculated from Σ.

In order to generate the potential causes $\boldsymbol{X}^i$, we use Σ and the MLP transforms. The first value of each time series is randomly generated from a uniform distribution in $[-10, 10]$. Then, each $\boldsymbol{X}^i$ is produced by applying the appropriate transform to its causal parents, as determined by Σ, and a zero-mean unit variance Gaussian noise is added. We generate the target time series in a similar fashion. Each variable Y_t is produced by applying the MLP transform to its causal parents, as determined by the target adjacency matrix $\Sigma^{\boldsymbol{Y}}$. We then add zero-mean Gaussian noise. The scale of the posterior additive noise for the outcome is referred to as the *noise-to-signal ratio* (NTS).

Description of the experiments. We are given a dataset as described above with m potential causes and a fixed NTS for the generation of the outcome $\boldsymbol{Y}$. We determine which series $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^n$ are the causal parents of the outcome, using Algorithm 1. For a given choice of m and NTS, we repeat the experiments five times, and we report on the sample mean for the Accuracy (see Table 4). We additionally report on the F1 Score and CSI Score in Table 6-5 in the Appendix. This experiment is repeated for an increasing number of potential causes, and increasing noise-to-signal ratio, to evaluate the performance of Algorithm 1 on challenging datasets.

In this set of experiments, we learn η^i_j as in Line 3 of Algorithm 1, using for the regression task an MLP model. We found that this simple approach, combined with zero-masking, dramatically reduces the false positives of the Student's t-test in Line 11 of Algorithm 1.

Results. Table 4 shows the performance of our method with respect to the accuracy, for an increasing number of potential causes and increasing NSR. We observe that our method maintains a good score. These

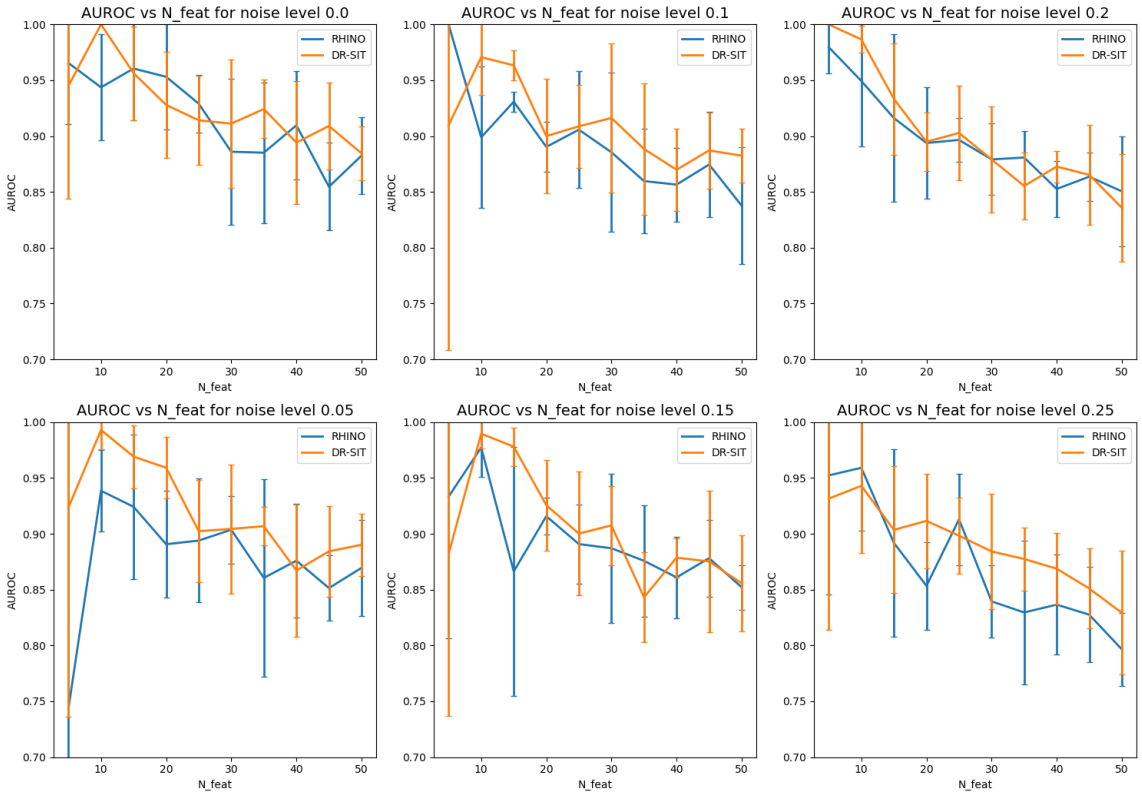


Figure 3: AUROC metric for RHINO and DR-SIT for various noise levels on the synthetic dataset.

Table 4: Accuracy of our method for increasing number of potential causes m, and different noise-to-signal ration (NSR). We observe that our method maintains good accuracy, even in challenging settings with many potential causes and high noise.

	Accuracy						
m	**NSR = 0**	**NSR = 0.05**	**NSR = 0.1**	**NSR = 0.15**	**NSR = 0.2**	**NSR = 0.25**	**NSR = 0.3**
5	0.60 ± 0.09	0.74 ± 0.25	0.66 ± 0.19	0.90 ± 0.11	0.80 ± 0.06	0.82 ± 0.10	0.94 ± 0.12
10	0.99 ± 0.02	0.92 ± 0.08	0.97 ± 0.04	0.79 ± 0.18	0.89 ± 0.06	0.94 ± 0.04	0.90 ± 0.08
15	0.89 ± 0.05	0.89 ± 0.05	0.88 ± 0.10	0.85 ± 0.02	0.79 ± 0.11	0.79 ± 0.09	0.81 ± 0.03
20	0.83 ± 0.07	0.73 ± 0.05	0.72 ± 0.07	0.77 ± 0.10	0.73 ± 0.05	0.75 ± 0.05	0.69 ± 0.06
25	0.76 ± 0.04	0.72 ± 0.10	0.71 ± 0.06	0.63 ± 0.08	0.68 ± 0.04	0.71 ± 0.07	0.66 ± 0.04
30	0.76 ± 0.04	0.74 ± 0.04	0.72 ± 0.07	0.70 ± 0.10	0.66 ± 0.05	0.68 ± 0.04	0.64 ± 0.07
35	0.68 ± 0.02	0.65 ± 0.07	0.72 ± 0.06	0.65 ± 0.03	0.66 ± 0.05	0.61 ± 0.07	0.64 ± 0.09
40	0.64 ± 0.03	0.69 ± 0.05	0.67 ± 0.05	0.62 ± 0.03	0.63 ± 0.06	0.58 ± 0.07	0.60 ± 0.07
45	0.65 ± 0.04	0.68 ± 0.08	0.58 ± 0.04	0.61 ± 0.03	0.64 ± 0.04	0.59 ± 0.04	0.61 ± 0.06
50	0.68 ± 0.05	0.63 ± 0.05	0.64 ± 0.06	0.63 ± 0.05	0.66 ± 0.05	0.59 ± 0.08	0.64 ± 0.07

results are partially confirmed when we look at the F1 and CSI scores (see Table 6-5 in the Appendix). In fact, we see that the DR-SIT is stable for an increasingly higher posterior noise.

In Figure 3 we plot the AUROC performance of DR-SIT against RHINO on the synthetic dataset, confirming the competitive performance of STID. In order to calculate the AUROC for DR-SIT, we sort our predictions (for existence of an edge) on the standard deviation of $Z := Y_T \cdot \tilde{g}_j^i + \tilde{\alpha}_j^i \cdot (Y_T - \tilde{g}_j^i) - Y_T \cdot g_j^0 - \alpha_j^0 \cdot (Y_T - g_j^0)$, which is simply the difference of the doubly robust statistics for θ^i and θ^0 for a datapoint in partition D_j.

We provide the CSI Score in Table 5 and the F1 Score in Table 6 for the experiments on synthetic datasets. Recall that in this set of experiments we considered synthetic datasets with varying numbers of potential causes (m) and noise-to-signal ratio (NSR). Increasing m and NSR give more challenging settings. We observe that our method maintains good CSI and F1 scores, for increasing NSR.

Table 5: CSI Score of Algorithm 1 for increasing number of potential causes m, and different noise-to-signal ration (NSR). Again, we observe that our method is robust to increasing NSR.

	CSI Score						
m	**NSR = 0**	**NSR = 0.05**	**NSR = 0.1**	**NSR = 0.15**	**NSR = 0.2**	**NSR = 0.25**	**NSR = 0.3**
5	0.57 ± 0.07	0.71 ± 0.28	0.63 ± 0.19	0.86 ± 0.12	0.69 ± 0.10	0.73 ± 0.12	0.91 ± 0.17
10	0.98 ± 0.04	0.86 ± 0.14	0.95 ± 0.06	0.69 ± 0.21	0.80 ± 0.09	0.87 ± 0.07	0.82 ± 0.14
15	0.80 ± 0.07	0.77 ± 0.11	0.75 ± 0.16	0.68 ± 0.06	0.59 ± 0.19	0.57 ± 0.15	0.59 ± 0.03
20	0.66 ± 0.12	0.52 ± 0.12	0.46 ± 0.07	0.56 ± 0.10	0.46 ± 0.08	0.51 ± 0.07	0.39 ± 0.09
25	0.51 ± 0.06	0.45 ± 0.12	0.43 ± 0.09	0.37 ± 0.09	0.38 ± 0.05	0.41 ± 0.08	0.33 ± 0.06
30	0.47 ± 0.05	0.50 ± 0.07	0.46 ± 0.12	0.38 ± 0.08	0.35 ± 0.08	0.34 ± 0.05	0.31 ± 0.07
35	0.42 ± 0.04	0.31 ± 0.06	0.39 ± 0.07	0.29 ± 0.07	0.30 ± 0.09	0.22 ± 0.09	0.26 ± 0.09
40	0.32 ± 0.06	0.38 ± 0.05	0.33 ± 0.08	0.29 ± 0.06	0.24 ± 0.06	0.20 ± 0.07	0.19 ± 0.11
45	0.33 ± 0.10	0.34 ± 0.07	0.20 ± 0.02	0.22 ± 0.05	0.25 ± 0.03	0.20 ± 0.05	0.19 ± 0.06
50	0.33 ± 0.03	0.29 ± 0.06	0.29 ± 0.06	0.23 ± 0.04	0.26 ± 0.06	0.21 ± 0.08	0.26 ± 0.07

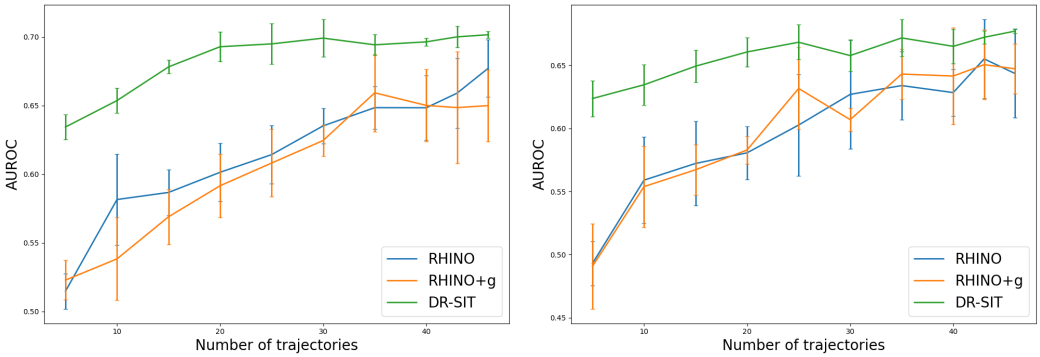
Table 6: F1 Score of the DR-SIT for increasing number of potential causes m, and different noise-to-signal ration (NSR). Interestingly, our method maintains a good F1 score for increasing NSR.

	F1 Score						
m	**NSR = 0**	**NSR = 0.05**	**NSR = 0.1**	**NSR = 0.15**	**NSR = 0.2**	**NSR = 0.25**	**NSR = 0.3**
5	0.73 ± 0.06	0.79 ± 0.20	0.75 ± 0.12	0.92 ± 0.07	0.81 ± 0.07	0.84 ± 0.08	0.95 ± 0.11
10	0.99 ± 0.02	0.92 ± 0.09	0.98 ± 0.03	0.80 ± 0.15	0.89 ± 0.06	0.93 ± 0.04	0.90 ± 0.09
15	0.89 ± 0.05	0.86 ± 0.08	0.85 ± 0.11	0.81 ± 0.04	0.73 ± 0.15	0.72 ± 0.12	0.74 ± 0.03
20	0.79 ± 0.08	0.67 ± 0.10	0.62 ± 0.07	0.71 ± 0.09	0.63 ± 0.07	0.67 ± 0.06	0.56 ± 0.09
25	0.68 ± 0.05	0.61 ± 0.11	0.60 ± 0.09	0.53 ± 0.11	0.55 ± 0.05	0.57 ± 0.08	0.49 ± 0.07
30	0.64 ± 0.05	0.66 ± 0.06	0.62 ± 0.10	0.54 ± 0.09	0.51 ± 0.09	0.50 ± 0.06	0.46 ± 0.09
35	0.59 ± 0.05	0.47 ± 0.07	0.56 ± 0.07	0.45 ± 0.08	0.46 ± 0.11	0.35 ± 0.11	0.41 ± 0.11
40	0.49 ± 0.07	0.55 ± 0.05	0.49 ± 0.08	0.45 ± 0.07	0.38 ± 0.08	0.33 ± 0.09	0.30 ± 0.15
45	0.49 ± 0.11	0.50 ± 0.08	0.34 ± 0.03	0.36 ± 0.06	0.40 ± 0.04	0.33 ± 0.06	0.32 ± 0.08
50	0.50 ± 0.03	0.44 ± 0.07	0.44 ± 0.08	0.38 ± 0.05	0.41 ± 0.07	0.34 ± 0.11	0.40 ± 0.09

J.2 Performance in Low-Sample Regimes

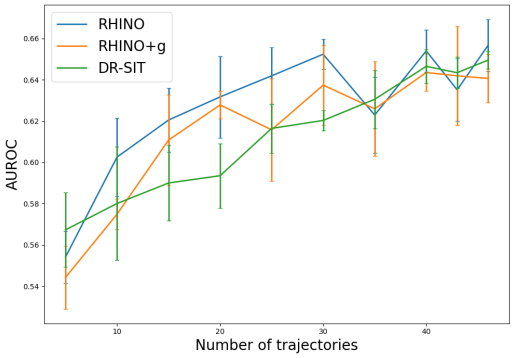
The double robustness property enables our algorithm to rely on simple estimators with low statistical complexity. As a result, our method shows more consistent performance in low-sample regimes as opposed to existing approaches that are based on overparameterized models demanding so many data points (Figure 4).

Moreover, in Figs. 5 to 9 we compare the progression of AUROC score vs the total runtime for DR-SIT vs RHINO in various combinations of tasks and training set sizes. The hardware specifications are described in Section 6.3 and the training hyperparameter settings for RHINO in Section 6.1. The runtime of DR-SIT is always less than 1 minute (so afterwards AUROC curve is plotted as a constant) where each epoch for RHINO takes about 30 seconds independently of the training dataset size.

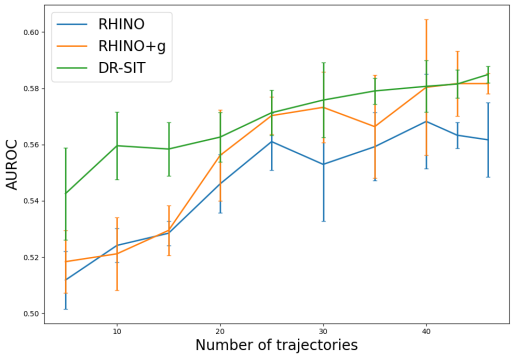


(a) Task: E.Coli 1

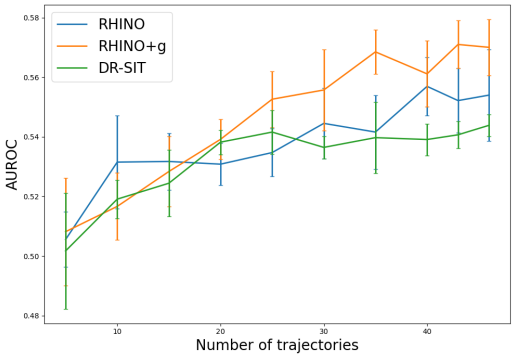
(b) Task: E.Coli 2



(c) Task: Yeast 1



(d) Task: Yeast 2



(e) Task: Yeast 3

Figure 4: This figure demonstrates the consistent performance of DR-SIT w.r.t number of observations compared to state-of-the-art methods Rhino and Rhino+g. Note that Rhino and Rhino+g are built on neural networks. DR-SIT significantly outperforms Rhino and Rhino+g in E.Coli 1 and E.Coli 2 and shows competitive results in Yeast 1. Thanks to the double robustness property of DR-SIT, the dependence of our algorithm on the estimator is much lower than the well-established approaches. In this regard, DR-SIT with a simple kernel regression with polynomial kernels has superior performance compared to state-of-the-art methods Rhino and Rhino+g. This superiority gets magnified in the low number of observation regimes due to the high sample complexity required by Rhino and Rhino+g.

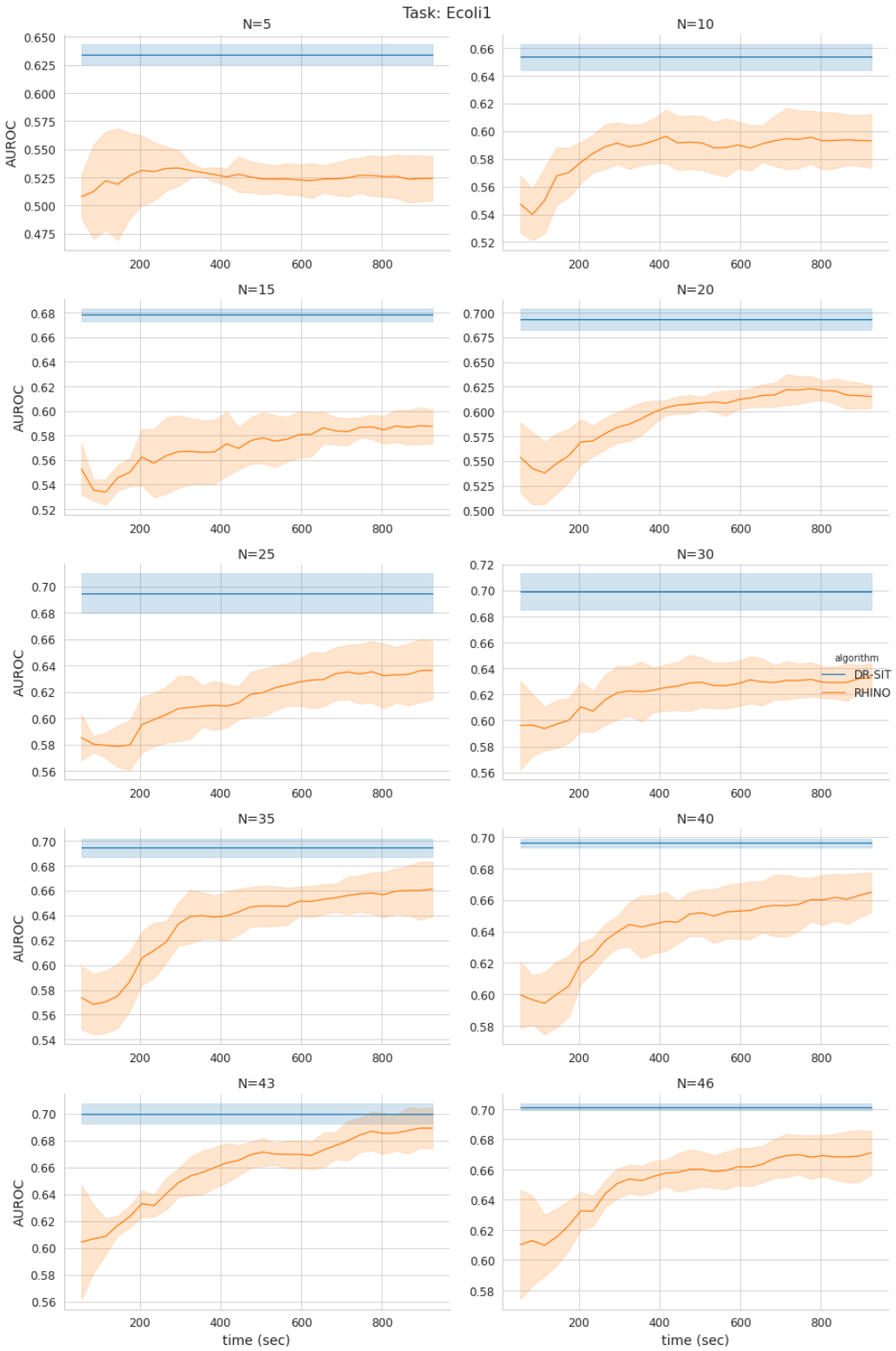


Figure 5: AUROC vs time (in secs) for DR-SIT (blue) vs RHINO (orange) for Ecoli 1 task and various numbers of training observations (number of trajectories).

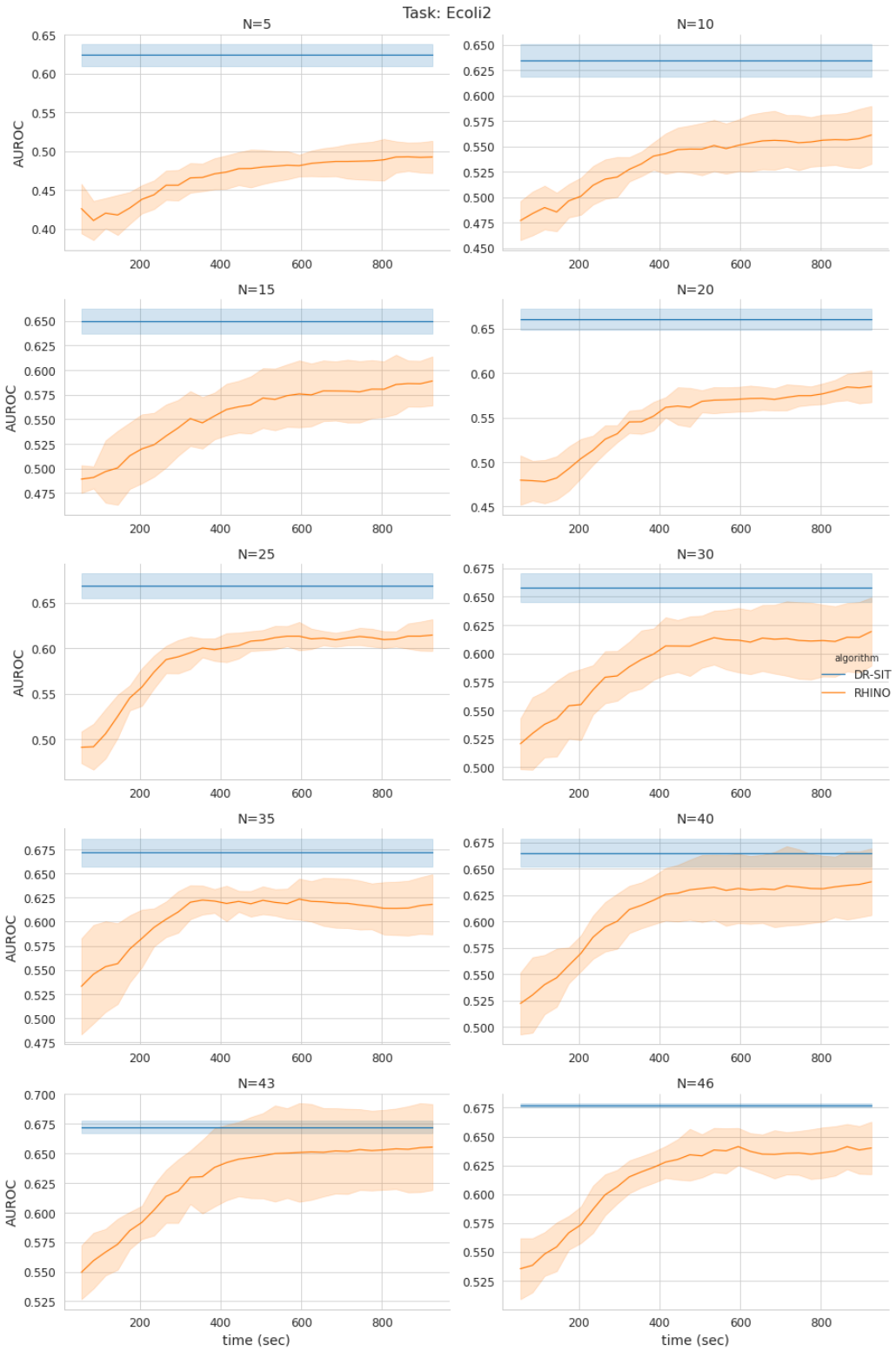


Figure 6: AUROC vs time (in secs) for DR-SIT (blue) vs RHINO (orange) for Ecoli 2 task and various numbers of training observations (number of trajectories).

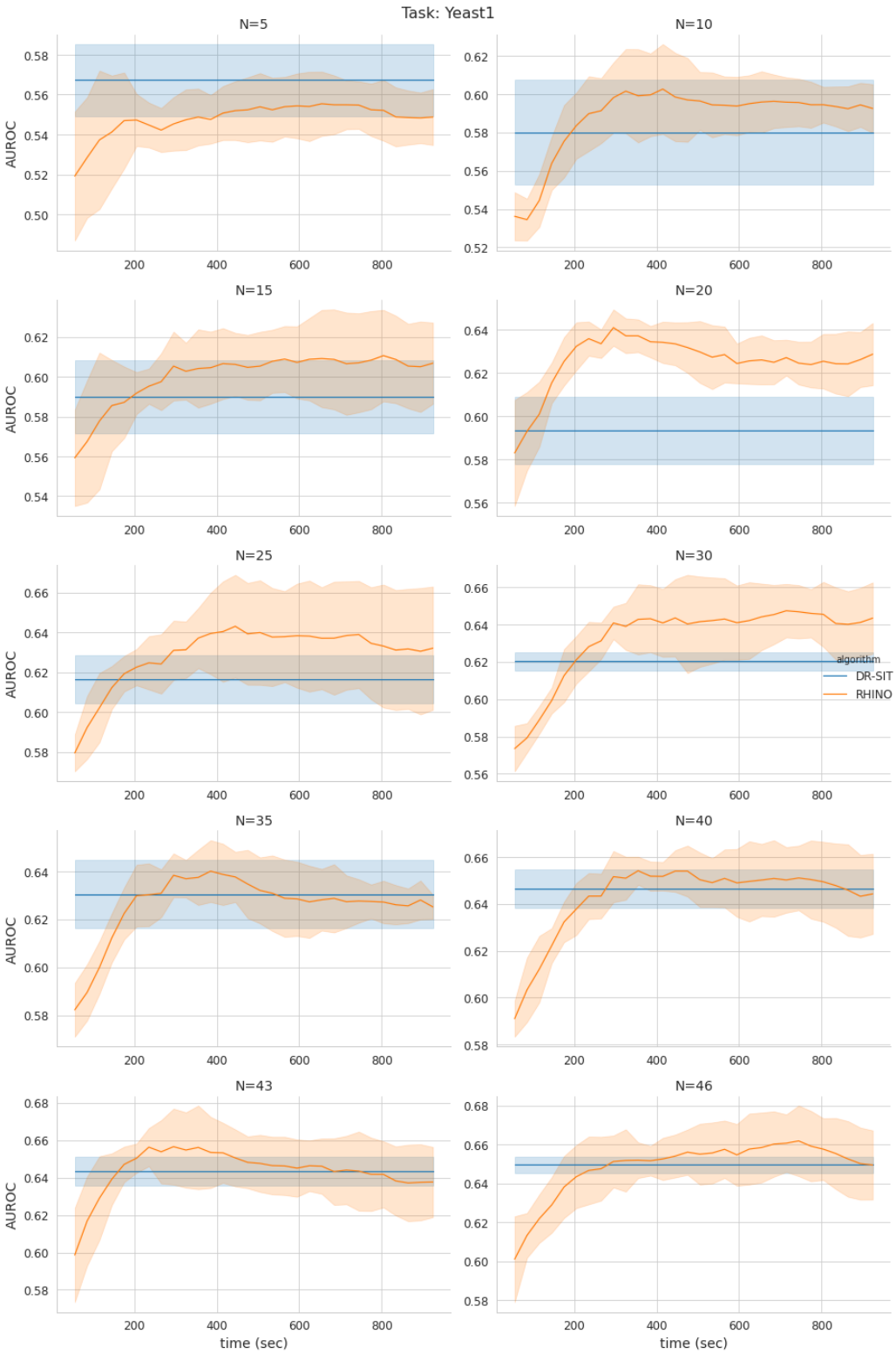


Figure 7: AUROC vs time (in secs) for DR-SIT (blue) vs RHINO (orange) for Yeast 1 task and various numbers of training observations (number of trajectories).

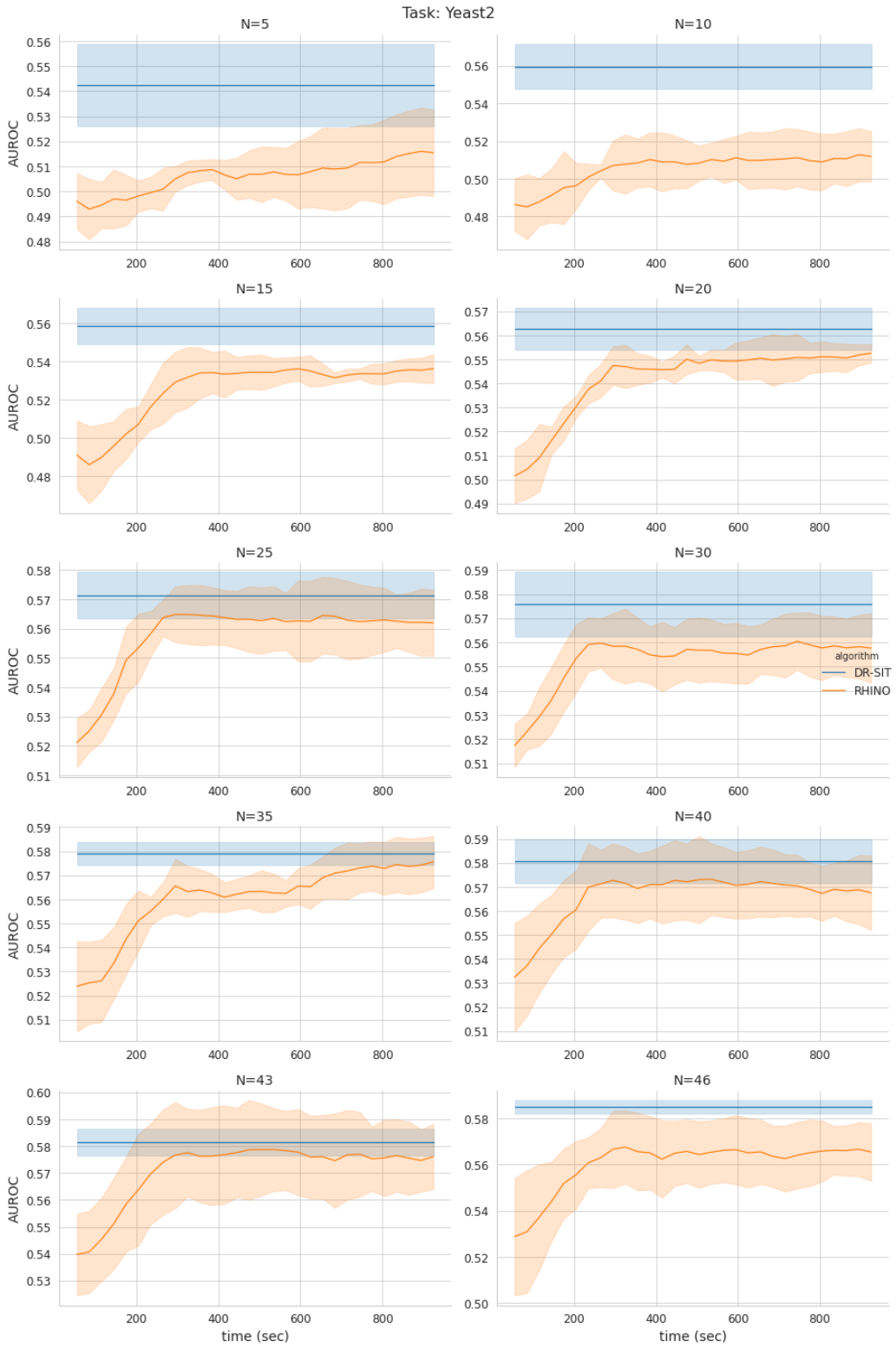


Figure 8: AUROC vs time (in secs) for DR-SIT (blue) vs RHINO (orange) for Yeast 2 task and various numbers of training observations (number of trajectories).

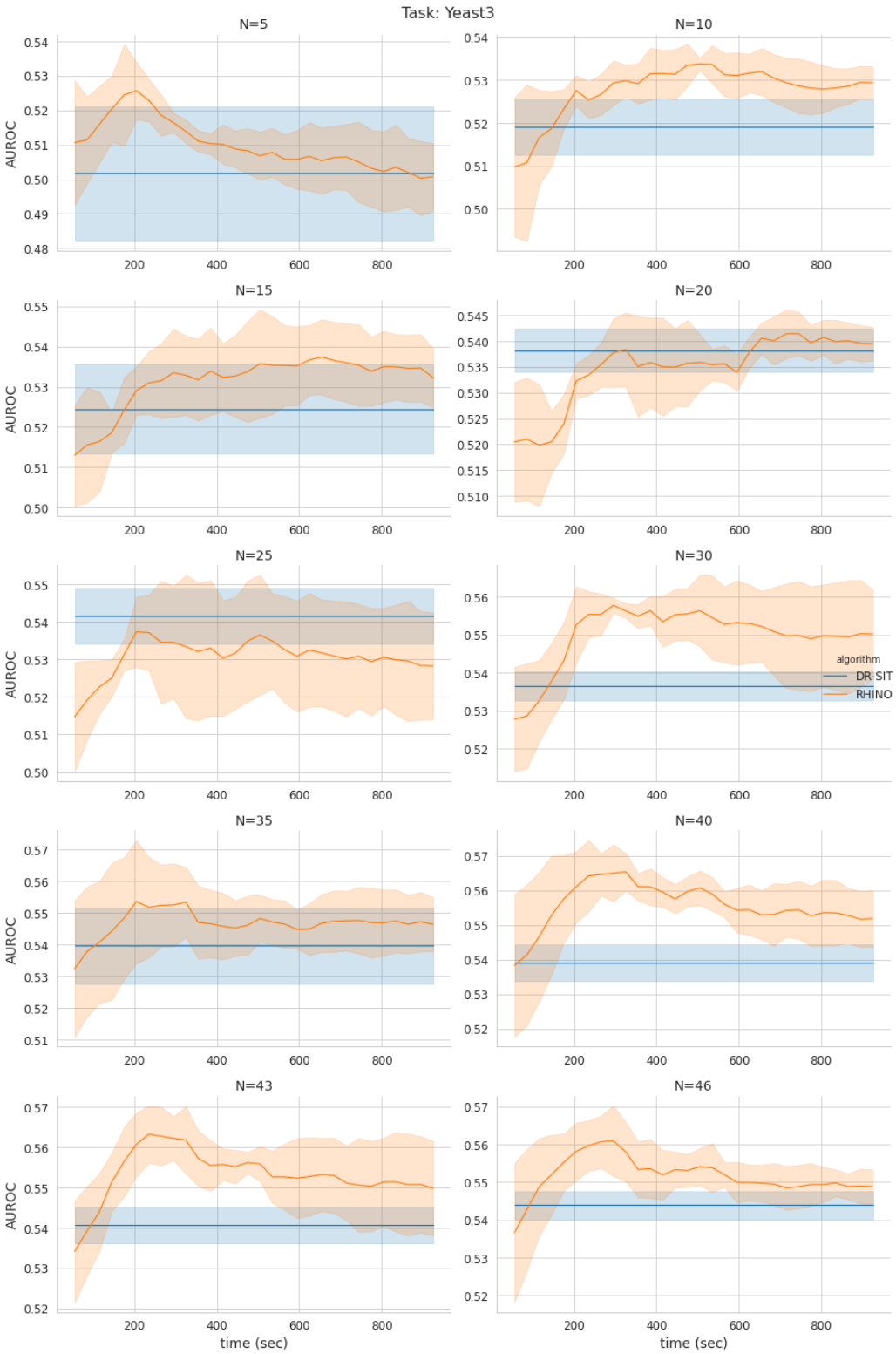


Figure 9: AUROC vs time (in secs) for DR-SIT (blue) vs RHINO (orange) for Yeast 3 task and various numbers of training observations (number of trajectories).