Rhino: Deep Causal Temporal Relationship Learning with history-dependent noise

Wenbo Gong, Joel Jennings, Cheng Zhang & Nick Pawlowski
Microsoft Research
Cambridge, UK
{t-gongwenbo, joeljennings, cheng.zhang, nick.pawlowski}@microsoft.com

Abstract

Discovering causal relationships between different variables from time series data has been a long-standing challenge for many domains. For example, in stock markets, the announcement of acquisitions from leading companies may have immediate effects on stock prices and increased uncertainty of the future market due to this past action. This requires the model to take non-linear relationships, instantaneous effects and the past-action dependent uncertainty into account. We name the latter as history-dependent noise. However, previous works do not offer a solution addressing all these problems together. In this paper, we propose a structural equation model, called Rhino, which combines vector auto-regression, deep learning and variational inference to model non-linear relationships with instantaneous effects and flexible history-dependent noise. Theoretically, we prove the structural identifiability for a generalization of Rhino. Our empirical results from extensive synthetic experiments and a real-world benchmark demonstrate better discovery performance compared to relevant baselines, with ablation studies revealing its robustness when the Rhino is misspecified.

1 Introduction

Time series data is a collection of data points recorded at different timestamps describing a pattern of chronological change. With the rapid growth of observational time series data generated by different domains (e.g. climate science, health care, etc.), there has been an increasing interest in identifying the causal relations between different variables and their interactions through time [19, 4, 22, 39, 30, 37, 14, 24]. We name this task temporal causal discovery.

Temporal causal discovery can be challenging for several reasons: (1) relations between variables can be non-linear in the real world; (2) with a slow sampling interval, everything happens in between will be aggregated into the same timestamp, i.e. instantaneous effect; (3) the noise may be non-stationary in the sense that its distribution depends on the past observations, i.e. history-dependent noise. For example, in stock markets, the announcements of some decisions from a leading company after the market closes may have complex effects (i.e. non-linearity) on its stock price immediately after the market opening (i.e. slow sampling interval and instantaneous effect) and its price volatility may also be changed (i.e. history-dependent noise).

To the best of our knowledge, no previous work addresses all these aspects in a satisfactory way. Especially, history-dependent noise has been rarely considered in past. A large category of the preceding works, called *Granger causality* [9], is based on the fact that cause-effect relationships can never go against time. Despite many recent advances [39, 34, 35, 1, 19, 37, 4, 5, 40], they all rely on the absence of instantaneous effects with a fixed noise distribution. Constraint-based methods can potentially address the aforementioned requirements [30, 31], but can only identify the causal graph up to a Markov equivalence class (MECs) without detailed functional relationships

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).

between variables. Structure equation models (SEMs) approach can incorporate both instantaneous and lagged effects as long as they are *structural identifiable*. Unfortunately, it is challenging to model history-dependent noise, since this additional flexibility may break the model structural identifiability. Thus, the key research question is whether the identifiability can be preserved even with complex historical dependencies in the temporal setting.

Motivated by these requirements, we propose a novel SEM-based temporal discovery called Rhino (*deep causal temporal relationship learning with history dependent noise*), which can model non-linear lagged and instantaneous effects with flexible history-dependent noise. Our contributions are:

- A novel formulation of SEMs, called Rhino, which combines vector auto-regression and deep learning to model non-linear lagged and instantaneous effects with history-dependent noise. We also propose a principled variational training framework.
- We prove that Rhino is structurally identifiable under similar assumptions as additive noise models (ANM). To achieve this, we provide general conditions for structural identifiability with history-dependent noise, of which Rhino is a special case.
- We conduct extensive synthetic experiments with ablation studies to demonstrate the advantages of Rhino and its robustness under model misspecification. Additionally, we compare its performance to a wide range of baselines in a real-world discovery benchmark.

2 Rhino: Relationship learning with history dependent noise

2.1 Model formulation

Before jumping into the formulations, we briefly introduce SEMs. Appendix B includes preliminaries about Granger causality and vector auto-regression, which is optional for understanding Rhino.

Structural Equation Models (SEMs) Consider a multivariate time-series $\boldsymbol{X}_t = \left(X_t^i\right)_{i \in \boldsymbol{V}}$ where $\boldsymbol{V}$ is a set of nodes with size D. An SEM describes the functional relationships between variables X_t^i across and within the time frame given a temporal causal graph $\boldsymbol{G}$. A general form is

$$X_t^i = f_{i,t}(\mathbf{Pa}_G^i(<t), \mathbf{Pa}_G^i(t), \epsilon_t^i), \tag{1}$$

where $\mathbf{Pa}_G^i(<t)$ contains the parent values specified by G in previous time (*lagged parents*); $\mathbf{Pa}_G^i(t)$ are the parents at the current time t (*instantaneous parents*); ϵ_t^i is the mutually independent exogenous noise and $f_{i,t}$ describes the functional relationships. The above SEM induces a joint distribution over the stationary time series $\{\boldsymbol{X}_t\}_{t=0}^T$ (see Assumption 1 in Appendix D for definition). However, Eq. (1) with this flexible form cannot be directly used for causal discovery due to the structural unidentifiability (Lemma 1, Zhang et al. [42]). One way to solve this is sacrificing the flexibility by restricting the functional class. For example, ANM, [12]

$$X^i = f_i(\mathbf{Pa}_G(X^i)) + \epsilon_i, \tag{2}$$

which have recently been used for causal reasoning with non-temporal data [8].

Next, we propose the formulation of Rhino. We assume that the multivariate temporal process is defined by the adjacency matrix $\boldsymbol{G}_{0:K}$ with maximum lag K, where $\boldsymbol{G}_{\tau \in [1,K]}$ specifies the lagged effects and $\boldsymbol{G}_0$ specifies the instantaneous parents. In the following, we interchange the usage of the notation $\boldsymbol{G}$ and $\boldsymbol{G}_{0:K}$ for brevity. We propose a novel SEM that incorporates non-linear functions, instantaneous effects, and flexible history-dependent noise, called Rhino:

$$X_t^i = f_i(\mathbf{Pa}_G^i(<t), \mathbf{Pa}_G^i(t)) + g_i(\mathbf{Pa}_G^i(<t), \epsilon_t^i) \tag{3}$$

where f_i is a general differentiable non-linear function, and g_i is a differentiable transform s.t. the transformed noise has a proper density. Rhino has an additive structure similar to a standard ANM. However, our formulation is much more flexible. By placing few restrictions on f_i, g_i, Rhino can capture functional non-linearity through f_i and noise history dependency through g_i.

We also propose flexible functional designs for f_i, g_i, which must respect the relations encapsulated in $\boldsymbol{G}$. Namely, if $X^j_{t-\tau} \notin \mathbf{Pa}^i_G(<t) \cup \mathbf{Pa}^i_G(t)$, then $\partial f_i / \partial X^j_{t-\tau} = 0$ and similarly for g_i. We design

$$f_i(\mathbf{Pa}^i_G(<t), \mathbf{Pa}^i_G(t)) = \zeta_i \left(\sum_{\tau=0}^{K} \sum_{j=1}^{D} G_{\tau,ji} \ell_{\tau j} \left(X^j_{t-\tau} \right) \right) \tag{4}$$

where ζ_i and $\ell_{\tau i}$ ($i \in [1, D]$ and $\tau \in [0, K]$) are neural networks. For efficient computation, we use weight sharing across nodes and lags: $\zeta_i(\cdot) = \zeta(\cdot, \boldsymbol{u}_{0,i})$ and $\ell_{\tau j}(\cdot) = \ell(\cdot, \boldsymbol{u}_{\tau,j})$, where $\boldsymbol{u}_{\tau,i}$ is the trainable embedding for node i at time $t - \tau$.

For the design of g_i, we use the conditional spline flow [38, 7, 25] with Gaussian noise to balance the flexibility and likelihood tractability. Their bin parameters are predicted using a hyper-network with a similar form to Eq. (4) to incorporate history dependency, where we sum over $\tau \in [1, K]$ to remove the instantaneous parents. Due to the invertibility of g_i, the noise likelihood conditioned on lagged parents is

$$p_{g_i}(g_i(\epsilon^i_t) | \mathbf{Pa}^i_G(<t)) = p_\epsilon(\epsilon^i_t) \left| \frac{\partial g_i^{-1}}{\partial \epsilon^i_t} \right|. \tag{5}$$

2.2 Variational Inference for Rhino

Rhino adopts a Bayesian view of causal discovery [10], which aims to learn a graph posterior distribution instead of inferring a single graph. For N observed multivariate time series $\boldsymbol{X}^{(1)}_{0:T}, \ldots, \boldsymbol{X}^{(N)}_{0:T}$, the joint likelihood of Rhino is

$$p(\boldsymbol{X}^{(1)}_{0:T}, \ldots, \boldsymbol{X}^{(N)}_{0:T}, \boldsymbol{G}) = p(\boldsymbol{G}) \prod_{n=1}^{N} p_\theta(\boldsymbol{X}^{(n)}_{0:T} | \boldsymbol{G}) \tag{6}$$

where θ are the model parameters.

Graph Prior Inspired by Geffner et al. [8], Zheng et al. [43], we propose the following unnormalised prior

$$p(\boldsymbol{G}) \propto \exp\left(-\lambda_s \|\boldsymbol{G}_{0:K}\|^2_F - \rho h^2(\boldsymbol{G}_0) - \alpha h(\boldsymbol{G}_0) - \lambda_p \|\boldsymbol{G}_{0:K} - \boldsymbol{G}^p_{0:K}\|^2_F \right) \tag{7}$$

where $h(\boldsymbol{G}) = \mathrm{tr}(e^{\boldsymbol{G} \odot \boldsymbol{G}}) - D$ is the DAG penalty proposed in [43]; $\odot$ is the Hadamard product; $\boldsymbol{G}^p$ is an optional domain-specific prior graph; λ_s, λ_p specify the strength of the graph sparseness and domain-specific prior terms respectively; α, ρ characterize the strength of the DAG penalty. Since the lagged connections can only follow the direction of time, only the instantaneous part, $\boldsymbol{G}_0$, can contain cycles, to which we apply DAG constraint.

Variational Objective To overcome the posterior intractability, we use independent Bernoulli distribution $q_\phi(\boldsymbol{G})$ (refer to Appendix G for details) to approximate it with varitional inference. The corresponding *evidence lower bound* (*ELBO*) is

$$\log p_\theta \left(\boldsymbol{X}^{(1)}_{0:T}, \ldots, \boldsymbol{X}^{(N)}_{0:T} \right) \geq \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{G})} \left[\sum_{n=1}^{N} \log p_\theta(\boldsymbol{X}^{(n)}_{0:T} | \boldsymbol{G}) + \log p(\boldsymbol{G}) \right] + H(q_\phi(\boldsymbol{G}))}_{\text{ELBO}(\theta,\phi)} \tag{8}$$

where $H(q_\phi(\boldsymbol{G}))$ is the entropy of $q_\phi(\boldsymbol{G})$. The parameters θ, ϕ are learned by maximizing the ELBO, where the Gumbel-softmax gradient estimator is used for ϕ [15, 20]. We also leverage the same augmented Lagrangian training procedure [11, 2], as Geffner et al. [8], to anneal α, ρ to make sure Rhino only produces DAGs.

Treatment effect estimation Apart from inferring temporal causal graphs, our model can be extended for causal inference tasks such as treatment effect estimation. See Appendix F for details.

3 Theoretical Considerations: Structural Identifiability

One of the key challenges for causal discovery with a flexible SEM is to show the structural identifiability. We present a theorem for Rhino that summarizes our main theoretical contribution.

Theorem 1 (Identifiability of Rhino)**.** Assuming Rhino satisfies the *causal stationarity*, *causal Markov property, causal minimality, causal sufficiency and the induced likelihood has a proper density* (see Appendix D), and we further assume (1) all functions and induced distributions are third-order differentiable; (2) function f_i is *non-linear* and *not invertible* w.r.t. any nodes in $\mathbf{Pa}^i_G(t)$; (3) the double derivative $(\log p_{g_i}(g_i(\epsilon^i_t)|\mathbf{Pa}^i_G(<t)))''$ w.r.t ϵ^i_t is zero at most at some discrete points, then Rhino defined in Eq. (3) is structural identifiable for *both bivariate and multivariate time series*.

Sketch of proof. This theorem is a summary of a collection of theorems proved in Appendix D. The strategy is instead of directly proving the identifiability of Rhino, we provide identifiability conditions for a general temporal SEM, followed by showing a generalization of Rhino satisfies these conditions. The identifiability of Rhino directly follows from it.

Prove bivariate identifiability conditions for general temporal SEMs The first step is to prove the bivariate identifiability conditions that a general temporal SEM (Eq. (1)) should satisfy (refer to Theorem 2 in Appendix D.1). In a nutshell, we proved the SEM is bivariate identifiable if (1) the model for initial conditions is identifiable; (2) the SEM is **identifiable w.r.t. instantaneous parents**. Remarkably, (2) implies we only need to pay attention to instantaneous parents for identifiability, and opens the door for flexible lagged parent dependency. This theorem assumes *causal stationarity, causal Markov, minimality, sufficiency and proper density* assumptions.

Identifiability of history-dependent post non-linear model Next, we propose a novel generalization of Rhino, called *history-dependent PNL*. Theorem 3 and Corollary 3.1 in Appendix D.2 prove it is bivariate identifiable w.r.t. instantaneous parents (i.e. satisfy the conditions in Theorem 2) with additional assumptions (1), (2) and (3) in Theorem 1. The history-dependent PNL is defined as

$$X^i_t = \nu_{it}\left(f_{it}\left(\mathbf{Pa}^i_G(<t), \mathbf{Pa}^i_G(t)\right) + g_{it}\left(\mathbf{Pa}^i_G(<t), \epsilon_{it}\right), \mathbf{Pa}^i_G(<t)\right),$$

where ν is invertible w.r.t. the first argument. The bivariate identifiability of Rhino directly follows from this, since Rhino is a special case with ν being the identity mapping.

Generalization to multivariate case In the end, inspired by Peters et al. [28], we prove the above bivariate identifiability can be generalized to the multivariate case. Refer to Theorem 4 in Appendix D.3 for details.

□

Other theoretical aspects To fully validate Rhino as a causal discovery method, we need to consider the soundness of the proposed variational training objective (Eq. (8)). In summary, Theorem 5 in Appendix E.1 shows that under the assumptions of Theorem 1, optimizing Eq. (8) can lead to the ground truth graph and data generating mechanism. Additionally, in Appendix E.2, we clarify the connections of Rhino to many existing works, and it is the most flexible member of this family.

4 Experiments

We evaluate Rhino on a synthetic experiment, followed by a real-world benchmark using fMRI data. Additionally, Appendix I.4 reports the performance of our model in another benchmark gene dataset, called DREAM3, where Rhino is also consistently better than baselines.

4.1 Synthetic data

We evaluate our method on a large set of synthetically generated datasets with known causal graphs. We use the main body of this paper to present the overall performance compared to relevant baselines and one ablation study on the robustness to lag mismatch. In Appendix H.3, we conduct extensive analysis, including (1) on different graph type; (2) ablation on history-dependency; (3) ablation

study on instantaneous effect. This set of datasets are generated by various settings (e.g. type of graphs, instantaneous/no instantaneous effect, etc.). 5 datasets are generated for each combination of settings with different seeds, yielding 160 datasets in total. In order to comprehensively test Rhino's robustness, we deliberately generated 75% of the datasets that mismatch the Rhino configurations. Details of the data generation can be found in Appendix H.1.

We compare Rhino to a wide range of baselines, including VARLiNGaM [14], PCMCI$^+$[31] and DYNOTEARS [24]. For the MECs from PCMCI$^+$, we enumerate all DAGs. For details on the methods, see Appendix H.2. Additionally, we include two variants of Rhino: (1) Rhino+g, where an independent Gaussian noise is used; (2) Rhino+s, where Gaussian ϵ_i is transformed by an independent spline.

Figure 1 presents the F_1 score of all methods aggregated over all datasets. Rhino achieves overall competitive or the best performance in terms of the full temporal adjacency matrix across all possible datasets, especially for lower dimensions. Comparing Rhino's lagged discovery to its two variants, the better score indicates the history-dependent noise is useful to the lagged graph discovery, contributing to the better overall F_1 performance (Appendix H.3 for ablation: with/without history dependency).

Despite of the strong performance from PCMCI$^+$, PCMCI$^+$ exceeds the maximum training time of **1 week** on 40 nodes, suggesting its computation bottleneck in high dimensions. DYNOTEARS achieves competitive results in low dimensions but suffers in high dimensions due to the limited modelling power from its linear nature.

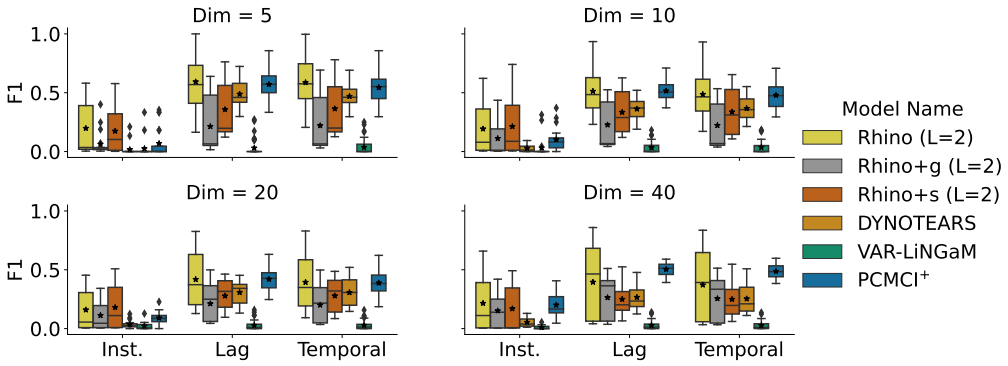

Figure 1: F_1-scores of Rhino (light yellow) compared to all baselines. The subplots corresponds to the different number of nodes. 'Inst.', 'Lag', and 'Temporal' refer to the metrics calculated for the instantaneous, the lagged, and the full temporal adjacency matrix, respectively. 'L=2' refers to lag 2.

We explore the behaviour of Rhino with different lag parameters other than the ground truth lag 2. From Table 1, worse training log-likelihoods suggest that Rhino with insufficient history ($\text{lag} = 1$) is unable to correctly model the data and this leads to a decrease in F_1 scores. Interestingly, Rhino is also robust with longer lags. Despite of the slightly better likelihood ($\text{lag} = 3$), it achieves comparable performance to the model with the correct lag. Also, from their similar F_1 Lag score, it suggests the extra adjacency matrix is nearly empty.

4.2 Netsim Brain Connectivity

In this section, we evaluate Rhino using *blood oxygenation level dependent* (BOLD) imaging data, which has been used as a benchmark for temporal causal discovery [19, 16, 3]. Each time series represents the BOLD signal simulated for $T = 200$ for a human subject, which describes $d = 15$ different brain regions. The goal of this task is to infer the connectivity between different brain regions. We assume that different human subjects share the same connectivity. We only use the data from human subject $2 - 6$ in *Sim-3.mat* from `https://www.fmrib.ox.ac.uk/datasets/netsim/index.html` and also include self-connections during evaluation. We consider the same baselines as in the synthetic experiments. Additionally, we also consider relevant Granger causality methods, including *cMLP*, *cLSTM* [37]; *TCDF*[23]; *SRU* and *eSRU* [16]. Appendix I.2 describes hyperparameter settings. Since the ground truth graph is a summary graph (see Definition I.1),

Dim		Rhino (L=1)	Rhino (L=2)	Rhino (L=3)
5	F_1 Lag	0.28 ± 0.13	0.59 ± 0.22	0.57 ± 0.24
	F_1 Temporal	0.34 ± 0.12	0.59 ± 0.20	0.56 ± 0.22
	LL	-4.14 ± 1.63	-3.83 ± 1.62	-3.75 ± 1.64
10	F_1 Lag	0.26 ± 0.08	0.51 ± 0.17	0.48 ± 0.19
	F_1 Temporal	0.28 ± 0.11	0.49 ± 0.18	0.45 ± 0.20
	LL	-7.97 ± 2.09	-7.21 ± 2.22	-7.01 ± 1.91
20	F_1 Lag	0.24 ± 0.12	0.42 ± 0.22	0.40 ± 0.21
	F_1 Temporal	0.25 ± 0.13	0.39 ± 0.22	0.37 ± 0.21
	LL	-15.62 ± 3.16	-14.70 ± 2.87	-14.72 ± 2.82
40	F_1 Lag	0.20 ± 0.18	0.40 ± 0.31	0.34 ± 0.30
	F_1 Temporal	0.20 ± 0.18	0.37 ± 0.30	0.32 ± 0.29
	LL	-31.44 ± 5.16	-30.10 ± 4.71	-30.20 ± 4.74

Table 1: Comparison of the causal discovery performance of Rhino with different lag-parameters ($L \in [1, 3]$). LL shows the log-likelihood of the training data.

Appendix I.1 details about the post-processing step on aggregating temporal graph to summary graph for Rhino, DYNOTEARS and PCMCI$^+$. We use the *area under the ROC curve* (AUROC) as the performance metric.

Table 2 shows the AUROCs for different methods. Remarkably, the proposed Rhino and its variants achieve significantly better AUROC compared to the baselines. Especially, Rhino obtains nearly optimal AUROC, demonstrating its robustness to the small dataset and good balances between true and false positive rates (see discussion in Appendix J). By comparing Rhino and Rhino+NoInst., we conclude that modelling instantaneous effects is important in real application, indicating the sampling interval is not frequent enough to explain everything as lagged effects. This can be double confirmed by comparing Rhino+NoInst with Granger causality, where it performs on par with the state-of-the-art baseline when disabling the instantaneous effect. Last but not least, by comparing Rhino+g with Rhino, we find that history-dependent noise is also helpful in this dataset.

Method	AUROC
cMLP	0.93
cLSTM	0.83
TCDF	0.91
SRU	0.80
eSRU	0.88
DYNO.	0.90
PCMCI$^+$	0.83 ± 0
VARLiNGaM	0.84 ± 0
Rhino+g	0.974 ± 0.002
Rhino+NoInst.	0.93 ± 0.006
Rhino	$\mathbf{0.99 \pm 0.001}$

Table 2: The AUROCs of the summary graph for Netsim dataset, where we take self-connections into consideration. Rhino+NoInst is the Rhino without the instantaneous effects. For Rhino, VARLiNGaM, PCMCI$^+$, the results are obtained by averaging over 5 different runs.

5 Conclusion

Inferring temporal causal graphs from observational time series is an important task in many scientific fields. Especially, some applications (e.g. education, climate science, etc.) require the modelling of non-linear relationships; instantaneous effects and history-dependent noise distributions at the same time. Previous works fail to offer an appropriate solution for all three requirements. Motivated by this, we propose Rhino, which combines vector auto-regression with deep learning and variational inference to perform causal temporal relationship learning with all three requirements. Theoretically, we prove the structural identifiability of Rhino with flexible history-dependent noise, and clarify its relations to existing works. Empirical evaluations demonstrate its superior performance and robustness when Rhino is misspecified, and the advantages of history-dependent noise mechanisms. This opens an exciting route of extending Rhino to handle non-stationary time-series and unobserved confounders in future work.

References

[1] Amornbunchornvej, C., Zheleva, E., and Berger-Wolf, T. Y. (2019). Variable-lag granger causality for time series analysis. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 21–30. IEEE.

[2] Andreani, R., Birgin, E. G., Martínez, J. M., and Schuverdt, M. L. (2008). On augmented lagrangian methods with general lower-level constraints. *SIAM Journal on Optimization*, 18(4):1286–1309.

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

[4] Bussmann, B., Nys, J., and Latré, S. (2021). Neural additive vector autoregression models for causal discovery in time series. In *International Conference on Discovery Science*, pages 446–460. Springer.

[5] Dang, X.-H., Shah, S. Y., and Zerfos, P. (2018). seq2graph: Discovering dynamic dependencies from multivariate time series with multi-level attention. *arXiv preprint arXiv:1812.04448*.

[6] Dolatabadi, H. M., Erfani, S., and Leckie, C. (2020). Invertible generative modeling using linear rational splines. In *International Conference on Artificial Intelligence and Statistics*, pages 4236–4246. PMLR.

[7] Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. (2019). Neural spline flows. *Advances in neural information processing systems*, 32.

[8] Geffner, T., Antoran, J., Foster, A., Gong, W., Ma, C., Kiciman, E., Sharma, A., Lamb, A., Kukla, M., Pawlowski, N., et al. (2022). Deep end-to-end causal inference. *arXiv preprint arXiv:2202.02195*.

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

[10] Heckerman, D., Meek, C., and Cooper, G. (2006). A bayesian approach to causal discovery. In *Innovations in Machine Learning*, pages 1–28. Springer.

[11] Hestenes, M. R. (1969). Multiplier and gradient methods. *Journal of optimization theory and applications*, 4(5):303–320.

[12] Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21.

[13] Huang, B., Zhang, K., Zhang, J., Ramsey, J. D., Sanchez-Romero, R., Glymour, C., and Schölkopf, B. (2020). Causal discovery from heterogeneous/nonstationary data. *J. Mach. Learn. Res.*, 21(89):1–53.

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

[15] Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

[16] Khanna, S. and Tan, V. Y. (2019). Economy statistical recurrent units for inferring nonlinear granger causality. *arXiv preprint arXiv:1911.09879*.

[17] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[18] Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

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

[20] Maddison, C. J., Mnih, A., and Teh, Y. W. (2016). The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

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

[22] Moraffah, R., Sheth, P., Karami, M., Bhattacharya, A., Wang, Q., Tahir, A., Raglin, A., and Liu, H. (2021). Causal inference for time series analysis: Problems, methods and evaluation. *Knowledge and Information Systems*, pages 1–45.

[23] Nauta, M., Bucur, D., and Seifert, C. (2019). Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):19.

[24] Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P., and Aragam, B. (2020). Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605. PMLR.

[25] Pawlowski, N., Coelho de Castro, D., and Glocker, B. (2020). Deep structural causal models for tractable counterfactual inference. *Advances in Neural Information Processing Systems*, 33:857–869.

[26] Peters, J., Janzing, D., and Schölkopf, B. (2013). Causal inference on time series using restricted structural equation models. *Advances in Neural Information Processing Systems*, 26.

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

[28] Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. (2012). Identifiability of causal graphs using functional models. *arXiv preprint arXiv:1202.3757*.

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

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

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

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

[33] Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10).

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

[35] Siggiridou, E. and Kugiumtzis, D. (2015). Granger causality in multivariate time series using a time-ordered restricted vector autoregressive model. *IEEE Transactions on Signal Processing*, 64(7):1759–1773.

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

[37] Tank, A., Covert, I., Foti, N., Shojaie, A., and Fox, E. (2018). Neural granger causality for nonlinear time series. *stat*, 1050:16.

[38] Trippe, B. L. and Turner, R. E. (2018). Conditional density estimation with bayesian normalising flows. *arXiv preprint arXiv:1802.04908*.

[39] Wu, T., Breuel, T., Skuhersky, M., and Kautz, J. (2020). Discovering nonlinear relations with minimum predictive information regularization. *arXiv preprint arXiv:2001.01885*.

[40] Xu, C., Huang, H., and Yoo, S. (2019). Scalable causal graph learning through a deep neural network. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1853–1862.

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

[42] Zhang, K., Wang, Z., Zhang, J., and Schölkopf, B. (2015). On estimation of functional causal models: general results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–22.

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See future work in Section 5 and discussion in Appendix F
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] For our contribution, we include the entire proof. For cited theorems, we point out readers to the original paper for proof.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] , see Appendix I, Appendix H.2 and Appendix H.1 for real world experiment details, synthetic experiment details, and data generation, respectively.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] , results for our model and baselines are averaged over 5 runs. The cited results from other papers only use 1 run due to their setup.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Related Work

Discovering causal relationships from time series has been a popular research question for several decades now. Assaad et al. [3] provides a comprehensive overview of causal discovery method for time series. In a nutshell, there are three main categories. The first category is Granger causality, where this field can be further split into (1) vector auto-regressive methods [39, 34, 35, 1] and (2) deep learning based approaches [19, 37, 4, 5, 40]. Despite recent advances, all Granger causality methods cannot handle instantaneous effects, which can be observed due to the aggregation effect in a slow-sampling system. Additionally, they also assume a fixed noise distribution without history dependency.

Using SEMs for time series discovery can mitigate the aforementioned two problems. VARLiNGaM [14] extends the identifiability theory of linear non-Gaussian ANM [33] to vector auto-regression for modelling time series data. DYNOTEARS [24] leverages the recently proposed NOTEARS framework [43] to continuously relax the DAG constraints for fully differentiable DAG structure learning. However, the above approach is still limited to linear functional forms. TiMINo [26] provides a general theoretical framework for temporal causal discovery with SEMs. Our theory leverages some of their proof techniques. Unfortunately, all the aforementioned methods assume no history dependency for the noise. On the other hand, Rhino can model (1) non-linear function relations; (2) instantaneous effect; (3) and history-dependent noise at the same time.

The third category is constraint-based approaches based on conditional independence tests. Due to its non-parametric nature, it can handle history-dependent noise. PCMCI [32] combines PC [36] and the momentary conditional independence test to discover the lagged parents from time series. PCMCI$^+$ [30, 31] further extends PCMCI to infer both lagged and instantaneous effects. CD-NOD [13] has recently been proposed to handle non-stationary heterogeneous data, where the data distribution can shift across time. Despite their generality, they can only infer MECs; cannot learn the explicit functional forms between variables; and require a stronger assumption than minimality (i.e. faithfulness).

B Preliminaries

In this section, we briefly introduce some preliminaries. In particular, we focus on Granger causality [9] and vector auto-regression.

Granger Causality Granger causality [9] has been extensively used for temporal causal discovery. It is based on the idea that the series $\boldsymbol{X}^j$ does not Granger cause $\boldsymbol{X}^i$ if the history, $\boldsymbol{X}^j_{<t}$, does not help the prediction of X^i_t for some t given the past of all other time series $\boldsymbol{X}^k$ for $k \neq j, i$.

Definition B.1 (Granger Causality [37, 19])**.** Given a multivariate stationary time series $\{\boldsymbol{X}_t\}_{t=0}^T$ and a SEM $f_{i,t}$ defined as

$$X^i_t = f_{i,t}(\mathbf{Pa}^i_G(<t)) + \epsilon^i_t, \tag{9}$$

$\boldsymbol{X}^j$ Granger causes $\boldsymbol{X}^i$ if $\exists l \in [1, t]$ such that $X^j_{t-l} \in \mathbf{Pa}^i_G(<t)$ and $f_{i,t}$ depends on X^j_{t-l}.

Granger causality is equivalent to causal relations for *directed acyclic graph* (*DAG*) if there are no latent confounders and instantaneous effects [26, 27]. Apart from the lack of instantaneous effects, they also ignore the history-dependent noise with independent ϵ^i_t.

Vector Auto-regressive Model Another line of research focuses on directly fitting the identifiable SEM to the observational data with instantaneous effects. One commonly-used approach is called vector auto-regression [14, 24]:

$$X^i_t = \beta^i + \sum_{\tau=0}^{K} \sum_{j=1}^{D} B_{\tau,ji} X^j_{t-\tau} + \epsilon^i_t \tag{10}$$

where β^i is the offset, K is the model lag, $\boldsymbol{B}_\tau \in \mathbb{R}^{D \times D}$ is the weighted adjacency matrix specifying the connections at time $t - \tau$ (i.e. if $B_{\tau,ji} = 0$ means no connection from $X^j_{t-\tau}$ to X^i_t) and ϵ^i_t is the independent noise. Under these assumptions, the above linear SEM is structurally identifiable, which is a necessary condition for recovering the ground truth graph [14, 26, 24]. However, the above

linear SEM with independent noise variables is too restrictive to fulfil the requirements described in Section 1.

C ELBO and likelihood derivation

The goal is to derive a lower bound for the joint likelihood $p_\theta(\boldsymbol{X}_{0:T}^{(1)}, \ldots, \boldsymbol{X}_{0:T}^{(N)})$.

$$
\begin{aligned}
& p_\theta(\boldsymbol{X}_{0:T}^{(1)}, \ldots, \boldsymbol{X}_{0:T}^{(N)}) \\
=& \log \int p_\theta \left(\boldsymbol{X}_{0:T}^{(1)}, \ldots, \boldsymbol{X}_{0:T}^{(N)} | \boldsymbol{G} \right) p(\boldsymbol{G}) d\boldsymbol{G} \\
=& \log \int \frac{q_\phi(\boldsymbol{G})}{q_\phi(\boldsymbol{G})} p_\theta \left(\boldsymbol{X}_{0:T}^{(1)}, \ldots, \boldsymbol{X}_{0:T}^{(N)} | \boldsymbol{G} \right) p(\boldsymbol{G}) d\boldsymbol{G} \\
\geq& \int q_\phi(\boldsymbol{G}) \log p_\theta \left(\boldsymbol{X}_{0:T}^{(1)}, \ldots, \boldsymbol{X}_{0:T}^{(N)} | \boldsymbol{G} \right) p(\boldsymbol{G}) d\boldsymbol{G} + H(q_\phi(\boldsymbol{G})) \qquad (11) \\
=& \mathbb{E}_{q_\phi(\boldsymbol{G})} \left[\sum_{n=1}^{N} \log p_\theta(\boldsymbol{X}_{0:T}^{(n)} | \boldsymbol{G}) + \log p(\boldsymbol{G}) \right] + H(q_\phi(\boldsymbol{G}))
\end{aligned}
$$

where Eq. (11) is obtained by using Jensen's inequality.

We can further simplify the likelihood $p_\theta(\boldsymbol{X}_{0:T}^{(n)} | \boldsymbol{G})$:

$$
\begin{aligned}
\log p_\theta(\boldsymbol{X}_{0:T}^{(n)} | \boldsymbol{G}) =& \log \prod_{t=0}^{T} p_\theta(\boldsymbol{X}_t^{(n)} | \boldsymbol{X}_{<t}^{(n)}, \boldsymbol{G}) \\
=& \sum_{t=0}^{T} \log p_\theta \left(\boldsymbol{X}_t^{(n)} | \boldsymbol{X}_{<t}^{(n)}, \boldsymbol{G} \right) \\
=& \sum_{t=0}^{T} \sum_{i=1}^{D} \log p_\theta \left(X_t^{i,(n)} | \mathbf{Pa}_G^i(<t), \mathbf{Pa}_G^i(t) \right) \qquad (12)
\end{aligned}
$$

where Eq. (12) is obtained through Markov factorization [18].

D Structural Identifiability

In this section, we will focus on proving the structural identifiability of Rhino. Before diving into the details, let us clarify the required assumptions.

Assumption 1 (Causal Stationarity [30])**.** The time series process $\boldsymbol{X}_t$ with a graph $\boldsymbol{G}$ is called *causally stationary* over a time index set $\mathcal{T}$ if and only if for all links $X_{t-\tau}^i \to X_t^j$ in the graph

$$X_{t-\tau}^i \not\perp\!\!\!\perp X_t^j | \boldsymbol{X}_{<t} \backslash \{X_{t-\tau}^i\} \text{ holds for all } t \in \mathcal{T}$$

This characterizes the nature of the time-series data generating mechanism, which validates the choice of the auto-regressive model.

Assumption 2 (Causal Markov Property [27])**.** Given a DAG $\boldsymbol{G}$ and a joint distribution p, this distribution is said to satisfy causal Markov property w.r.t. the DAG $\boldsymbol{G}$ if each variable is independent of its non-descendants given its parents.

This is a common assumptions for the distribution induced by an SEM. With this assumption, one can deduce conditional independence between variables from the graph.

Assumption 3 (Causal Minimality)**.** Consider a distribution p and a DAG $\boldsymbol{G}$, we say this distribution satisfies causal minimality w.r.t. $\boldsymbol{G}$ if it is Markovian w.r.t. $\boldsymbol{G}$ but not to any proper subgraph of $\boldsymbol{G}$.

Minimality is also a common assumption for SEMs [12, 41, 28], which can be regarded as a weaker version of *faithfulness* [27].

Assumption 4 (Causal Sufficiency)**.** A set of observed variables $\boldsymbol{V}$ is causally sufficient for a process $\boldsymbol{X}_t$ if and only if in the process every common cause of any two or more variables in $\boldsymbol{V}$ is in $\boldsymbol{V}$ or has the same value for all units in the population.

This assumption implies there are no latent confounders present in the time-series data.

Assumption 5 (Well-defined Density)**.** We assume the joint likelihood induced by the Rhino SEM (Eq. (3)) is absolutely continuous w.r.t. a Lebesgue or counting measure and $|\log p(\boldsymbol{X}_{0:T}; \boldsymbol{G})| < \infty$ for all possible $\boldsymbol{G}$.

This assumption is to make sure the induced distribution has a well-defined probability density function. It is also required for the equivalence of the global, local Markov property and Markov factorization property (Theorem 6.22 from [27]).

In the following, we will structure the entire proof into three steps:

1. Prove a general conditions that the *bivariate* time series model needs to satisfy for structural identifiability. This adapts from the theorem 1 in Peters et al. [26].
2. Prove that a generalized form of SEM, modified from the *post non-linear* (*PNL*) model [41], satisfies the conditions mentioned in step 1. The proposed Rhino (Eq. (3)) is a special case of the above SEM.
3. In the end, we generalize the above indentifiability to the *multivariate* case.

D.1 General Identifiability Conditions

First, we derive the conditions required for identifiability for a general bivariate time series SEM, defined as

$$X_t^i = f_{i,t}\left(\mathbf{Pa}_G^i(<t), \mathbf{Pa}_G^i(t), \epsilon_t^i\right). \tag{13}$$

We call the above SEM *transition model*, since it only defines the transition behavior rather than the initial conditions. We also need to incorporate a *source model*, which characterizes the initial conditions:

$$X_s^i = f_{i,s}(\mathbf{Pa}_G^i, \epsilon_s^i) \tag{14}$$

for $s \in [0, \mathcal{S}]$, where $\mathcal{S}$ is the length for the initial conditions and $\mathbf{Pa}_G^i$ contains the parents for node i. We define $p_s(\boldsymbol{X}_{0:\mathcal{S}})$ as the induced joint distribution for the initial conditions.

Now, we prove the following theorem.

Theorem 2 (Identifiability conditions for bivariate time series)**.** Assuming Assumption 1-5 are satisfied, given a bivariate temporal process $\boldsymbol{X}_{0:T}$ and $\boldsymbol{Y}_{0:T}$ that are governed by the above SEM (Eq. (13)) with source model (Eq. (14)), then the above SEM for the bivariate temporal process is structural identifiable if the following conditions are true:

1. Source model $f_{i,s}$ is structural identifiable for all $i = 1, \ldots, D$ and $s \in [0, \mathcal{S}]$.
2. The transition model (Eq. (13)) is *bivariate identifiable* w.r.t the *instantaneous parents*. Namely, if graph $\boldsymbol{G}$ induced conditional distributions $p(X_t, Y_t|\mathbf{Pa}_G^{X,Y}(<t))$, then $\nexists \boldsymbol{G}'$ such that $\boldsymbol{G} \neq \boldsymbol{G}'$ and the induced conditional $\bar{p}(X_t, Y_t|\overline{\mathbf{Pa}}_{G'}^{X,Y}(<t)) = p$ for all $t \in [\mathcal{S}+1, T]$.

where $\mathbf{Pa}_G^{X,Y}(<t)$ is the union of the lagged parents of X_t and Y_t under $\boldsymbol{G}$, and $\overline{\mathbf{Pa}}_{G'}^{X,Y}(<t)$ is the union of parents under $\boldsymbol{G}'$.

Proof. We prove this by contradiction. Assume we have an induced joint distribution $p(\boldsymbol{X}_{0:T}, \boldsymbol{Y}_{0:T})$ under $\boldsymbol{G}$, and corresponding $\bar{p}$ under $\boldsymbol{G}'$. We further assume the above two conditions in the theorem are met and $p = \bar{p}$ but $\boldsymbol{G} \neq \boldsymbol{G}'$.

Thus, we have $D_{\mathrm{KL}}[p\|\bar{p}]$ = 0. Due to the temporal nature of the model, we can further decompose it as the following:

$$
\begin{aligned}
&D_{\mathrm{KL}}[p\|\bar{p}] \\
=&\int p(\boldsymbol{X}_{0:T},\boldsymbol{Y}_{0:T})\log\frac{p(\boldsymbol{X}_{0:T},\boldsymbol{Y}_{0:T})}{\bar{p}(\boldsymbol{X}_{0:T},\boldsymbol{Y}_{0:T})}d\boldsymbol{X}_{0:T}d\boldsymbol{Y}_{0:T} \\
=&D_{\mathrm{KL}}[\underbrace{p(\boldsymbol{X}_{0:\mathcal{S}},\boldsymbol{Y}_{0:\mathcal{S}})}_{p_s}\,\|\,\underbrace{\bar{p}(\boldsymbol{X}_{0:\mathcal{S}},\boldsymbol{Y}_{0:\mathcal{S}})}_{\bar{p}_s}]+\int p(\boldsymbol{X}_{0:\mathcal{S}},\boldsymbol{Y}_{0:\mathcal{S}})D_{\mathrm{KL}}[p(\boldsymbol{X}_{\mathcal{S}+1:T},\boldsymbol{Y}_{\mathcal{S}+1:T}|\boldsymbol{X}_{0:\mathcal{S}},\boldsymbol{Y}_{0:\mathcal{S}})\| \\
&\bar{p}(\boldsymbol{X}_{\mathcal{S}+1:T},\boldsymbol{Y}_{\mathcal{S}+1:T}|\boldsymbol{X}_{0:\mathcal{S}},\boldsymbol{Y}_{0:\mathcal{S}})]d\boldsymbol{X}_{0:\mathcal{S}}d\boldsymbol{Y}_{0:\mathcal{S}} \\
=&D_{\mathrm{KL}}[p_s\|\bar{p}_s]+\sum_{t=\mathcal{S}+1}^{T}\mathbb{E}_{p(\boldsymbol{X}_{0:t-1},\boldsymbol{Y}_{0:t-1})}\left[D_{\mathrm{KL}}\left[p(X_t,Y_t|\boldsymbol{X}_{0:t-1},\boldsymbol{Y}_{0:t-1})\|\bar{p}(X_t,Y_t|\boldsymbol{X}_{0:t-1},\boldsymbol{Y}_{0:t-1})\right]\right] \\
=&0.
\end{aligned}
$$

This means we have $D_{\mathrm{KL}}[p_s\|\bar{p}_s]=0$ and $D_{\mathrm{KL}}\left[p(X_t,Y_t|\boldsymbol{X}_{0:t-1},\boldsymbol{Y}_{0:t-1})\|\bar{p}(X_t,Y_t|\boldsymbol{X}_{0:t-1},\boldsymbol{Y}_{0:t-1})\right]=0$ almost everywhere. Inspired by the strategy used in [26], We consider the following three scenarios:

Disagree on initial conditions We assume $\boldsymbol{G}$ and $\boldsymbol{G}'$ disagree on the initial conditions. From the condition 1, we know the source model $f_{i,s}$ is identifiable. Namely, we cannot find $\boldsymbol{G}\neq\boldsymbol{G}'$ with disagreement on initial conditions such that $D_{\mathrm{KL}}[p_s\|\bar{p}_s]=0$. This is a contradiction, meaning that $\boldsymbol{G}$ and $\boldsymbol{G}'$ must agree on the connections between initial set of nodes.

Disagree on lagged parents only This means for all $t\in[\mathcal{S}+1,T]$, the instantaneous connections at t for $\boldsymbol{G}$ and $\boldsymbol{G}'$ are the same, and $\exists t\in[\mathcal{S}+1,T]$ such that $\mathbf{Pa}_G^{X,Y}(<t)\neq\overline{\mathbf{Pa}}_{G'}^{X,Y}(<t)$. We can use a similar argument as the theorem 1 in Peters et al. [26]. W.l.o.g., we assume under $\boldsymbol{G}$, we have $X_{t-\tau}\to Y_t$ and there is no connections between them under $\boldsymbol{G}'$. Thus, from Markov conditions, we have

$$Y_t\perp\!\!\!\perp X_{t-\tau}|\boldsymbol{X}_{0:t-1}\cup\boldsymbol{Y}_{0:t-1}\cup\mathrm{ND}_t^Y\backslash\{Y_t,X_{t-\tau}\}$$

under $\boldsymbol{G}'$, where ND_t^Y is the non-descendants of node Y_t w.r.t instantaneous effect. However, from the causal minimality and proposition 6.16 in Peters et al. [27], we have

$$Y_t\not\perp\!\!\!\perp X_{t-\tau}|\boldsymbol{X}_{0:t-1}\cup\boldsymbol{Y}_{0:t-1}\cup\mathrm{ND}_t^Y\backslash\{Y_t,X_{t-\tau}\}$$

under $\boldsymbol{G}$. This means under this case, $D_{\mathrm{KL}}\left[p(X_t,Y_t|\boldsymbol{X}_{0:t-1},\boldsymbol{Y}_{0:t-1})\|\bar{p}(X_t,Y_t|\boldsymbol{X}_{0:t-1},\boldsymbol{Y}_{0:t-1})\right]\neq 0$, which is a contradiction.

Disagree also on instantaneous parents This scenarior means $\exists t\in[\mathcal{S}+1,T]$ such that they disagree on instantaneous parents. W.l.o.g. we assume $X_t\to Y_t$ under $\boldsymbol{G}$ and $Y_t\to X_t$ under $\boldsymbol{G}'$.

Let's define $\boldsymbol{X}_{0:t-1}\cup\boldsymbol{Y}_{0:t-1}=\boldsymbol{h}$, $\boldsymbol{h}_G^Y\subseteq\boldsymbol{h}$ contains the values of $\mathbf{Pa}_G^Y(<t)$ under $\boldsymbol{G}$, $\bar{\boldsymbol{h}}_{G'}^Y\subseteq\boldsymbol{h}$ contains the parent values under $\boldsymbol{G}'$, and $\boldsymbol{h}_G^X$, $\bar{\boldsymbol{h}}_{G'}^X$ accordingly. Thus, the induced conditional distributions from SEM (Eq. (13)) with $\boldsymbol{G}$, $\boldsymbol{G}'$ are

$$p(X_t,Y_t|\boldsymbol{h}_G^X\cup\boldsymbol{h}_G^Y)\quad\text{and}\quad\bar{p}(X_t,Y_t|\bar{\boldsymbol{h}}_{G'}^X\cup\bar{\boldsymbol{h}}_{G'}^Y)$$

From the Markov conditions, we have

$$p(X_t,Y_t|\boldsymbol{X}_{0:t-1},\boldsymbol{Y}_{0:t-1})=p(X_t,Y_t|\mathbf{Pa}_G^{X,Y}(<t))$$

Therefore, we have

$$
\begin{aligned}
&D_{\mathrm{KL}}\left[p(X_t,Y_t|\boldsymbol{h})\|\bar{p}(X_t,Y_t|\boldsymbol{h})\right] \\
=&0 \\
=&D_{\mathrm{KL}}[p(X_t,Y_t|\boldsymbol{h}_G^X\cup\boldsymbol{h}_G^Y)\|\bar{p}(X_t,Y_t|\bar{\boldsymbol{h}}_{G'}^X\cup\bar{\boldsymbol{h}}_{G'}^Y)]
\end{aligned}
$$

for arbitrary $\boldsymbol{h}$, which contradicts the strucutral identifiability w.r.t. the instantaneous parents.

In summary, with the two conditions, we cannot find $\boldsymbol{G}\neq\boldsymbol{G}'$ such that the induced joint $p(\boldsymbol{X}_{0:T},\boldsymbol{Y}_{0:T})=\bar{p}(\boldsymbol{X}_{0:T},\boldsymbol{Y}_{0:T})$, meaning that the SEMs defined as Eq. (13) and Eq. (14) are identifiable w.r.t. *bivariate time series*. □

Since one can use any identifiable static models to characterize the initial behavior of the time series, we will focus on condition 2 for the transition model. In the following, we will show that a generalization of PNL, called *history-dependent PNL*, satisfies condition 2 under assumptions.

D.2 Identifiability of history-dependent PNL

First, we propose a generalization of PNL [41] so that it can be history-dependent. For a multivariate temporal process $\boldsymbol{X}_{0:T}$, we propose *history-dependent PNL* as

$$X_t^i = \nu_{it}\left(f_{it}\left(\mathbf{Pa}_G^i(<t), \mathbf{Pa}_G^i(t)\right) + g_{it}\left(\mathbf{Pa}_G^i(<t), \epsilon_{it}\right), \mathbf{Pa}_G^i(<t)\right) \tag{15}$$

where ν_{it} is an invertible transformation w.r.t. the first argument. The main differences of the above SEM compared to typical PNL are (1) the invertible transformation ν_{it} can be history dependent; (2) the inner noise distribution can also be history-dependent.

Next, we show the main theorem about its bivariate identifiability w.r.t. it instantaneous parents.

Theorem 3 (History-dependent PNL Bivariate Identifiability)**.** Assume Assumption 1-5 are satisfied, all transformations in Eq. (15) and corresponding induced distributions are 3rd-order differentiable. Given a bivariate temporal process $\boldsymbol{X}_{0:T}$, $\boldsymbol{Y}_{0:T}$, then the history-dependent PNL defined as Eq. (15) is bivariate identifiable w.r.t its instantaneous parents (i.e. satisfy condition 2 in Theorem 2), except for some special cases.

Proof. W.l.o.g. at time $t \in [\mathcal{S}+1, T]$, we assume $X_t \to Y_t$ for instantaneous connection under $\boldsymbol{G}$ and $Y_t \to X_t$ under $\boldsymbol{G}'$. We fix a value $\boldsymbol{h}$ for their entire history $\boldsymbol{X}_{0:t-1} \cup \boldsymbol{Y}_{0:t-1} = \boldsymbol{h}$. With $\boldsymbol{h}$, we further define their lagged parents as $\mathbf{Pa}_G^X(<t) = \boldsymbol{h}_G^X \subseteq \boldsymbol{h}$, $\mathbf{Pa}_G^Y(<t) = \boldsymbol{h}_G^Y \subseteq \boldsymbol{h}$ under $\boldsymbol{G}$ and $\overline{\mathbf{Pa}}_{G'}^X(<t) = \bar{\boldsymbol{h}}_{G'}^X \subseteq \boldsymbol{h}$, $\overline{\mathbf{Pa}}_{G'}^Y(<t) = \bar{\boldsymbol{h}}_{G'}^Y$ under $\boldsymbol{G}'$.

Therefore, the SEM at time t can be written as

$$Y_t = \nu\left(f(\boldsymbol{h}_G^Y, X_t) + g(\boldsymbol{h}_G^Y, \epsilon_Y), \boldsymbol{h}_G^Y\right) \tag{16}$$

and

$$X_t = \bar{\nu}\left(\bar{f}(\bar{\boldsymbol{h}}_{G'}^X, Y_t) + \bar{g}(\bar{\boldsymbol{h}}_{G'}^X, \epsilon_X), \bar{\boldsymbol{h}}_{G'}^X\right) \tag{17}$$

under $\boldsymbol{G}$ and $\boldsymbol{G}'$, respectively. Let's assume that their induced distributions at time t are equal (i.e. violating the identifiable conditions):

$$\underbrace{\log p(X_t, Y_t | \boldsymbol{h}_G^X \cup \boldsymbol{h}_G^Y)}_{\text{under } \boldsymbol{G}} = \underbrace{\log \bar{p}(X_t, Y_t | \bar{\boldsymbol{h}}_{G'}^X \cup \bar{\boldsymbol{h}}_{G'}^Y)}_{\text{under } \boldsymbol{G}'}$$

From the Markov properties, the above equation is equivalent to

$$\log p(X_t, Y_t | \boldsymbol{h}) = \log \bar{p}(X_t, Y_t | \boldsymbol{h})$$

with a fixed value $\boldsymbol{h}$ of the entire history.

Now, let's define

$$\alpha_t = \bar{\nu}^{-1}(X_t) \quad \text{and} \quad \beta_t = \nu^{-1}(Y_t)$$

where we omits the dependence of $\bar{\nu}^{-1}$ to $\bar{\boldsymbol{h}}_{G'}^X$ and ν^{-1} to $\boldsymbol{h}_G^Y$. It is easy to observe that we have an invertible mapping between (X_t, Y_t) and (α_t, β_t). Thus, from the change of variable formula, we have

$$\log p(X_t, Y_t | \boldsymbol{h}) = \log p_{\alpha,\beta}(\alpha_t, \beta_t | \boldsymbol{h}) + \log |\boldsymbol{J}|$$

and

$$\log \bar{p}(X_t, Y_t | \boldsymbol{h}) = \log \bar{p}_{\alpha,\beta}(\alpha_t, \beta_t | \boldsymbol{h}) + \log |\boldsymbol{J}|$$

where $\boldsymbol{J}$ is the Jacobian matrix of the transformation. Thus, the equivalence of $\log p$ and $\log \bar{p}$ in the (X_t, Y_t) space can be translated to (α_t, β_t) space.

Thus, from Eq. (16), we have

$$\beta_t = \Phi(\alpha_t) + g(\boldsymbol{h}_G^Y, \epsilon_Y) \tag{18}$$

under $\boldsymbol{G}$. And from Eq. (17), we have

$$\alpha_t = \Psi(\beta_t) + \bar{g}(\bar{\boldsymbol{h}}_{G'}^X, \epsilon_X) \tag{19}$$

under $\boldsymbol{G}'$. This forms an *additive noise model* between α_t, β_t with history-dependent noise. Next, we can use a similar proof techniques as in Hoyer et al. [12]. Here, $\Phi(\cdot) = f(\boldsymbol{h}_G^Y, \cdot) \circ \bar{\nu}(\bar{\boldsymbol{h}}_{G'}^X, \cdot)$ and $\Psi(\cdot) = \bar{f}(\bar{\boldsymbol{h}}_{G'}^X, \cdot) \circ \nu(\boldsymbol{h}_G^Y, \cdot)$. We further define

$$\eta_1(\alpha_t) = \log p(\alpha_t|\boldsymbol{h}) \qquad \bar{\eta}_1(\beta_t) = \log \bar{p}(\beta_t|\boldsymbol{h})$$
$$\eta_2(g(\boldsymbol{h}_G^Y, \epsilon_Y)) = \log p_g(g(\boldsymbol{h}_G^Y, \epsilon_Y)|\boldsymbol{h}) \qquad \bar{\eta}_2(\bar{g}(\bar{\boldsymbol{h}}_{G'}^X, \epsilon_X)) = \log \bar{p}_g(\bar{g}(\bar{\boldsymbol{h}}_{G'}^X, \epsilon_X)|\boldsymbol{h})$$

Thus, under $\boldsymbol{G}$ (i.e. Eq. (18)), we have

$$\begin{aligned}\log p(\alpha_t, \beta_t|\boldsymbol{h}) =& \log p(\beta_t|\alpha_t, \boldsymbol{h}) + \log p(\alpha_t|\boldsymbol{h}) \\ =& \eta_2(\beta_t - \Phi(\alpha_t)) + \eta_1(\alpha_t)\end{aligned} \tag{20}$$

Similarly, under $\boldsymbol{G}'$ (i.e. Eq. (19)), we have

$$\log \bar{p}(\alpha_t, \beta_t) = \bar{\eta}_2(\alpha_t - \Psi(\beta_t)) + \bar{\eta}_1(\beta_t) \tag{21}$$

Based on Eq. (21), we have

$$\frac{\partial^2 \log \bar{p}}{\partial \alpha_t \partial \beta_t} = -\bar{\eta}_2'' \Psi' \quad \text{and} \quad \frac{\partial^2 \log \bar{p}}{\partial {\alpha_t}^2} = \bar{\eta}_2''$$

Thus, we have

$$\frac{\partial}{\partial \alpha_t} \left(\frac{\partial^2 \log \bar{p} / \partial \alpha_t \partial \beta_t}{\partial^2 \log \bar{p} / \partial {\alpha_t}^2} \right) = 0$$

Due to the equivalence of $\log \bar{p}$ and $\log p$, we apply the above operations to Eq. (20). After some algebraic manipulation, we obtained the following differential equations for $\eta_2'' \Phi' \neq 0$:

$$\eta_1''' - \frac{\eta_1'' \Phi''}{\Phi'} = \left(\frac{\eta_2' \eta_2'''}{\eta_2''} - 2\eta_2'' \right) \Phi'' \Phi' - \frac{\eta_2'''}{\eta_2''} \Phi' \eta_1'' + \eta_2' \left(\Phi''' - \frac{(\Phi'')^2}{\Phi'} \right). \tag{22}$$

Interestingly, this is exactly equivalent to Eq.(4) in Zhang and Hyvarinen [41]. The main difference is the definition of variables and transformations in here are all history-dependent.

Further, we can also observe that

$$\beta_t \perp\!\!\!\perp \bar{g}(\boldsymbol{h}_G^Y, \epsilon_Y) | \boldsymbol{X}_{0:t-1} \cup \boldsymbol{Y}_{0:t-1} = \boldsymbol{h}.$$

Since $\beta_t = \Phi(\alpha_t) + g(\boldsymbol{h}_G^Y, \epsilon_Y)$ and $\bar{g}(\bar{\boldsymbol{h}}_{G'}^X, \epsilon_X) = \alpha_t - \Psi(\beta_t)$, it is trivial to show the determinant of the Jacobian of the transformation (α_t, g) to $(\beta_t, \bar{g})$ is 1. Thus, by a similar argument in theorem 1 from Zhang and Hyvarinen [41], we can derive

$$\frac{1}{\Psi'} = \frac{\eta_1'' + \eta_2''(\Phi')^2 - \eta_2' \Phi''}{\eta_2'' \Phi'}$$

for $\eta_2'' \Phi' \neq 0$.

Thus, the above two differential equations has the same form as theorem 1 in Zhang and Hyvarinen [41] where the main difference is that all distributions and transformations involved in our case depends on history $\boldsymbol{h}$.

Therefore, we can directly cite the theorem 8 from Zhang and Hyvarinen [41], which proves that the above differential equations hold true only for 5 types of special cases. One can refer to Table 1 in Zhang and Hyvarinen [41] for details. □

Corollary 10 from Zhang and Hyvarinen [41] validates the choice of using nueral network for the transformation f. For completeness, we include it here with slight modification:

Corollary 3.1 (Identifiability with neural netowrk f)**.** Assuming the assumptions in Theorem 3 are true, and the double derivative $(\log p_g(g(\mathbf{Pa}_G^Y(<t), \epsilon_Y)|\boldsymbol{X}_{0:t-1} \cup \boldsymbol{Y}_{0:t-1}))''$ w.r.t ϵ_Y is zero at most at some discrete points. If function f is not invertible *w.r.t. the instantaneous parents*, then, the history-dependent PNL defined as Eq. (15) is *bivariate identifiable w.r.t. the instantaneous parents* (i.e. satisfy condition 2 in Theorem 2).

It is clear to see that Rhino (Eq. (3)) is a special case of the history-dependent PNL (Eq. (15)), where the outer history-dependent invertible transformation ν is the identity mapping. Thus, we can directly leverage Theorem 2 together with Theorem 3 to show Rhino is identifiable w.r.t bivariate time series, and Corollary 3.1 to validate our design choice (Eq. (4)).

D.3 Generalizing to multivariate time series

Previously, we prove the identifiability conditions for bivariate time series. In this section, we will generalize it to the multivariate case.

Theorem 4 (Generalization to multivariate time series)**.** Assuming the assumptions in Theorem 3 are satisfied, we further assume that the multivariate SEM defined in Eq. (15) satisfies: for each pair of node $i, j \in \boldsymbol{V}$, the SEM

$$X_t^i = \nu_{it}\left(f_{it}\left(\mathbf{Pa}_G^i(<t), \mathbf{Pa}_G^i(t)\backslash\{X_t^j\}, \underbrace{\cdot}_{X_t^j}\right) + g_{it}\left(\mathbf{Pa}_G^i(<t), \epsilon_{it}\right), \mathbf{Pa}_G^i(<t)\right)$$

is *bivariate identifiable* w.r.t. the input, and an identifiable source model is adopted. Then, the history-dependent PNL is *identifiable except for some special cases*.

Proof. For this proof, we can follow the strategy used in Theorem 2 and Peters et al. [26]. We categorize the difference of the graph $\boldsymbol{G}$ and $\boldsymbol{G}'$ into three types. Following the same analysis of the KL divergence of the two induced joint distributions, we can see that (1) $D_{\mathrm{KL}}[p_s \| \bar{p}_s] = 0$ and $D_{\mathrm{KL}}[p(\boldsymbol{X}_t|\boldsymbol{X}_{0:t-1}) \| \bar{p}(\boldsymbol{X}_t|\boldsymbol{X}_{0:t-1})] = 0$.

Disagree on initial conditions Since we assume that the source model is identifiable, this contradicts $D_{\mathrm{KL}}[p_s \| \bar{p}_s] = 0$.

Disagree on lagged parents only We notice that the analysis used in Theorem 2 for this disagreement can be directly translated to multivariate case. The only difference is that the notation Y_t, X_t is changed accordingly.

Disagree also on instantaneous parents For this case, with a fixed history value $\boldsymbol{h} = \boldsymbol{X}_{0:t-1}$, the aim is to compare the conditionals $D_{\mathrm{KL}}[p(\boldsymbol{X}_t|\boldsymbol{X}_{0:t-1} = \boldsymbol{h}) \| \bar{p}(\boldsymbol{X}_t|\boldsymbol{X}_{0:t-1} = \boldsymbol{h})]$. Thus, the problem becomes to how to generalize the bivariate identifiability for instantaneous parents to the multivariate case. We leverage the theorem 2 from Peters et al. [28], which proves the multivariate identifiability for any models that belongs to IFMOC. It is easy to see that if the assumptions in Theorem 4 are met, the history-dependent PNL belongs to IFMOC *w.r.t. the instantaneous parents*. It should be noted that the entire history-dependent PNL *DOES NOT belong to IFMOC*, but this does not affect our results since we only care about the instantaneous parents under this case. □

E Other theoretical considerations

E.1 Soundness of variational objective

Here, we show the validity of the variational objective (Eq. (8)) in the sense that optimizing it can lead to the ground truth graph. Remarkably, Theorem 1 in Geffner et al. [8] justifies the validity of the variational objective under the same set of assumptions as Rhino. Although Geffner et al. [8] focused on static data, the generality of the theorem is not limited to the static case.

Theorem 5 (Validity of variational objective [8])**.** Assuming the conditions in Theorem 1 are satisfied, and we further assume that there is no model misspecification, then the solution $(\theta', q'_\phi(\boldsymbol{G}))$ from optimizing Eq. (8) with infinite data satisfies $q'_\phi(\boldsymbol{G}) = \delta(\boldsymbol{G} = \boldsymbol{G}')$, where $\boldsymbol{G}'$ is a unique graph. In particular, $\boldsymbol{G}' = \boldsymbol{G}^*$ and $p_{\theta'}(\boldsymbol{X}_{0:T}; \boldsymbol{G}') = p(\boldsymbol{X}_{0:T}; \boldsymbol{G}^*)$, where $\boldsymbol{G}^*$ is the ground truth graph and $p(\boldsymbol{X}_{0:T}; \boldsymbol{G}^*)$ is the true data generating distribution.

E.2 Relation to other methods

VARLiNGaM [14] VARLiNGaM [14] is a causal discovery method for time series data based on the linear vector auto-regression, which can model both lagged and instantaneous effects. Its SEM is defined as Eq. (10), where the noise ϵ_t^i is an independent non-Gaussian noise. It is easy to observe that this is a special case of Rhino (Eq. (3)) by setting f_i as the matrix multiplication of the weighted adjacency $\boldsymbol{G}_{0:K}$ with the nodes, and g_i as the identity mapping. For the training objective,

VARLiNGaM adopted a two stage training to sidestep the difficulty of directly optimizing the log likelihood. From the Theorem 5 for Rhino, we note that the solution from optimizing the variational objective is equivalent to maximizing the log likelihood under infinite data limit. Therefore, by setting large enough DAGness penalty coefficient α, ρ, the inferred graph from both methods should be equivalent.

DYNOTEARS [24] The formulation of DYNOTEARS is the same as VARLiNGaM, which is based on linear vector auto-regression. The main novelty is the usage of the DAGness penalty $h(\boldsymbol{G})$, which continuously relaxes the DAG constraint. The training objective is the mean square error with augmented Lagrange scheme for DAGness penalty. Thus, it is obvious that DYNOTEARS is a special case of Rhino with linear transformations and identity g_i. Similarly, Theorem 5 shows the connections between the variational objective and maximum likelihood, which is equivalent to mean square error if the noise distribution is *Gaussian with equal variances*.

cMLP cMLP [37] combines Granger causality with deep neural networks. The model formulation is

$$X_t^i = f_i(\boldsymbol{X}_{0:t-1}^1, \dots, \boldsymbol{X}_{0:t-1}^D) + \epsilon_t^i$$

where f_i is a function based on MLP. Although the input is the entire history, the one that matters is the node that has the connection to X_t^i (i.e. lagged parents). Therefore, it is easy to see they are closely related to Rhino without *instantaneous parents* $\mathbf{Pa}_G^i(t)$ and history-dependent noise. Since the training objective of cMLP is based on the mean square error with sparseness constraint, by the same argument as before, the variational objective is equivalent to mean square error with equal variance Gaussian noise and large training data.

TiMINo [26] TiMINo is most similar to our work among all the aforementioned methods in terms of model formulation. TiMINo proposed a very general formulation based on IFMOC [28] and showed the conditions for structural identifiability. Rhino generalizes the TiMINO in a way such that noise history dependency can be incorporated. Thus, Rhino only belongs to IFMOC w.r.t. the instantaneous parents. Therefore, Rhino without the history-dependent noise is a TiMINo model. The training objective of TiMINo is based on the dependence minimization between the noise residuals and causes, and can only infer summary graph instead of temporal causal graph. Zhang et al. [42] proved the equivalence of the mutual information minimization to maximum likelihood, which is equivalent to our variational objective under infinite data.

F Treatment Effect Estimation

We now show how to leverage the fitted Rhino for estimating the *conditional average treatment effect* (CATE). For simplicity, we only consider a special case of CATE defined as

$$\mathrm{CATE}(a,b) = \mathbb{E}_{q_\phi(\boldsymbol{G})}\left[\mathbb{E}_{p(\boldsymbol{X}_{t+\tau}^Y|\boldsymbol{X}_{<t},\mathrm{do}(X_t^I=a),\boldsymbol{G})}[X_{t+\tau}^Y] - \mathbb{E}_{p(\boldsymbol{X}_{t+\tau}^Y|\boldsymbol{X}_{<t},\mathrm{do}(X_t^I=b),\boldsymbol{G})}[X_{t+\tau}^Y]\right] \tag{23}$$

We assume the conditioning variable can only be $\boldsymbol{X}_{<t}$ (i.e. the entire history before t), and the intervention and target variable can only be either at current time t or sometime in the future $t+\tau$. We emphasize that this formulation is for simplicity, and Rhino can be easily generalized to more cases as Geffner et al. [8]. Once fitted, the idea is to draw target samples $X_{t+\tau}^Y$ from the interventional distribution $p(\boldsymbol{X}_{t+\tau}^Y|\boldsymbol{X}_{<t}, \mathrm{do}(X_t^I), \boldsymbol{G})$ for each graph sample $\boldsymbol{G} \sim q_\phi(\boldsymbol{G})$. Then, unbiased Monte Carlo estimation can be used to compute CATE. For sampling from the interventional distribution, we can use the "multilated" graph $\boldsymbol{G}_{\mathrm{do}(X_t^I)}$ to replace $\boldsymbol{G}$, where all incoming edges to X_t^I are removed. The intervention samples can be obtained by simulating the Rhino with history $\boldsymbol{X}_{<t}$, $X_t^I = a$ or b and $\boldsymbol{G}_{\mathrm{do}(X_t^I)}$.

F.1 Causal Inference Results

Here, we provide the preliminary results for CATE performance of Rhino by calculating the RMSEs of the estimated CATEs comparing to the true CATE from the interventional samples (lower is better). The ground truth intervention samples are generated according to Appendix H.1. We present boxplots of the performance in Fig. 2. All Rhino-based method perform similarly. Surprisingly, the CATE

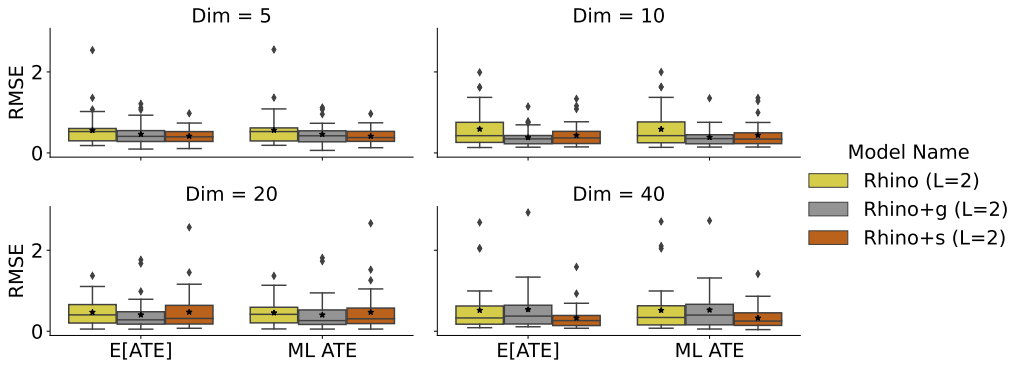

Figure 2: Comparison of the RMSE of the average treatment effects (CATEs) of the different instantiations of Rhino depending on the dimensionality. $\mathbb{E}[\text{CATE}]$ refers to RMSE of the expected CATE over the posterior graph distribution (i.e. $\boldsymbol{G} \sim q_\phi(\boldsymbol{G})$). ML ATE uses the most likely graph to calculate the ATE. These results are obtained by averaging 160 datasets, similar to the discovery setup.

performance seems to have little correlation to the causal discovery performance and warrants further study in the future.

G Variational distribution formulation

Here we provide the detailed formulation of the independent Bernoulli distribution $q_\phi(\boldsymbol{G})$. Since this distribution is responsible for modelling the temporal adjacency matrix $\boldsymbol{G}_{0:K}$, we use Σ_k to represents the edge probability in $\boldsymbol{G}_k$. We further split the edge probability matrices into the instantaneous part Σ_0 and lagged parts $\Sigma_{1:K}$.

To avoid the constrained optimization of $\Sigma_{1:K}$ (i.e. the value needs to be within $[0, 1]$), we adopt the following formulation:

$$\sigma_{k,ij} = \frac{\exp(u_{k,ij})}{\exp(u_{k,ij}) + \exp(v_{k,ij})} \tag{24}$$

where $u_{k,ij} \in \boldsymbol{U}_k$, $v_{k,ij} \in \boldsymbol{V}_k$ and $\boldsymbol{U}_k, \boldsymbol{V}_k \in \mathbb{R}^{D\times D}$ for all $k = 1, \dots, K$. Since we do not require lagged adjacency matrix to be a DAG, $\boldsymbol{U}_k, \boldsymbol{V}_k$ has no constraints during optimization.

On the other hand, $\boldsymbol{G}_0$ needs to be a DAG for instantaneous effect. By smart formulation, we can get rid of the length-1 cycles. The intuition is that for a pair of node i, j, only three mutually exclusive possibilities can exist: (1) $i \to j$; (2) $j \to i$; (3) no edge between them. Thus, instead of using a full probability matrix Σ_0, we use three lower triangular matrices $\boldsymbol{U}_0$, $\boldsymbol{V}_0$ and $\boldsymbol{E}_0$ to characterise the above three scenarios. For node $i > j$,

$$\begin{aligned}
p(i \to j) &= \frac{\exp(u_{ij})}{\exp(u_{ij}) + \exp(v_{ij}) + \exp(e_{ij})} \\
p(j \to i) &= \frac{\exp(v_{ij})}{\exp(u_{ij}) + \exp(v_{ij}) + \exp(e_{ij})} \\
p(\text{no edge}) &= \frac{\exp(e_{ij})}{\exp(u_{ij}) + \exp(v_{ij}) + \exp(e_{ij})}.
\end{aligned}$$

Thus, by this formulation, the corresponding instantaneous adjacency matrix will not contain length-1 cycles.

H Synthetic Experiments

H.1 Data generation

We create the synthetic datasets in a four step process: 1) generate random Erdös–Rényi (ER) or scale-free (SF) graphs that specify the lagged and instantaneous causal relationships; 2) drawing random MLPs for the functional relationships as well as a random *conditional* spline transformation to modulate the scale of the Gaussian noise variables ϵ; 3) sample initial starting conditions and follow Eq. (1) with the additive noise to simulate the temporal progression; 4) removing the burn-in period and return stable timeseries. We consider four different axes of variation for the data generation: number of nodes $N_{nodes} \in [5, 10, 20, 40]$; ER or SF graphs; instantaneous or no instantaneous effects; and history-dependent or history-independent noise (i.e. Gaussian noise). All combinations are generated with 5 different seeds, yielding 160 different datasets. Datasets with instantaneous effects have $4 \times N_{nodes}$ edges in the instantaneous adjacency matrix. All datasets have $2 \times N_{nodes}$ connections in the lagged adjacency matrices. The MLPs for the functional relationships are fully-connected with two hidden layers,64 units and ReLU activation. In case of history-independent noise, we are using Gaussian as the base distribution. The history dependency is modelled as a product of a scale variable obtained by the transformation of the averaged lagged parental values through a random-sampled quadratic spline, and Gaussian noise variable.

The datasets with 40 nodes are generated with a series length of 400 steps, a burn-in period of 100 steps, and 100 training series. All other datasets are generated with a time-series length of 200, burn-in period of 50 steps and 50 training series. We generate random interventions for all the datasets by setting the treatment variable to 10 for intervention and -10 for reference. 5000 ground-truth intervention samples are used to estimate the true treatment effect.

H.2 Methods

All benchmarks for the synthetic experiments are run by using publicly available libraries: VAR-LiNGaM [14] is implemenented in the `lingam`[1] python package. PCMCI$^+$[31] is implemented in `Tigramite`[2]. We use the implementation in `causalnex`[3] to run DYNOTEARS[24]. We use the default parameters for all these baselines. For PCMCI$^+$, we enumerate all graphs in the Markov equivalence class to evaluate the causal discovery performance (see Appendix I.1 for details).

For Rhino and its variants, we use the same set of hyper-parameters for all 160 datasets to demonstrates our robustness. By default, we allow Rhino and its variants to model instantaneous effect; set the model lag to be the ground truth 2 except for ablation study; the $q_\phi(\boldsymbol{G})$ is initialized to favour sparse graphs (edge probability< 0.5); quadratic spline flow is used to for history-dependent noise. For the model formulation, we use 2 layer fully connected MLPs with 64 (5 and 10 nodes), 80 (10 nodes) and 160 (40 nodes) for all neural networks in Rhino-based methods. We also apply layer normalization and residual connections to each layer of the MLPs. For the gradient estimator, we use the Gumbel softmax method with a hard forward pass and a soft backward pass with temperature of 0.25. All spline flows uses 8 bins. The embedding sizes for transformation (i.e. Eq. (4) and conditional spline flow) is equal to the node number.

For the sparseness penalty λ_s in Eq. (7), we use 9 for Rhino and Rhino+s, and 5 for Rhino+g. We set $\rho = 1$ and $\alpha = 0$ for all Rhino-based methods. For optimization, we use Adam [17] with learning rate 0.01. The training procedure follows from Appendix B.1 in Geffner et al. [8].

H.3 Additional Causal Discovery Results

Ablation: different type of graphs The first study is to test our model robustness to different types of graphs. Fig. 3 shows the discovery performance over ER or SF graph averaged over all other possible data setting combinations. Most methods perform better on ER graphs than on SF graphs, with only DYNOTEARS [24] as an exception. We note that the PCMCI$^+$ runs on SF graphs with 40 nodes exceed our maximum run time of **1 week**, showing its computational limitation in high dimensions. Nevertheless, Rhino achieves consistent performance throughout all graph settings.

[1] see `https://lingam.readthedocs.io`

[2] see `https://jakobrunge.github.io/tigramite/`

[3] see `https://causalnex.readthedocs.io/en/latest/`

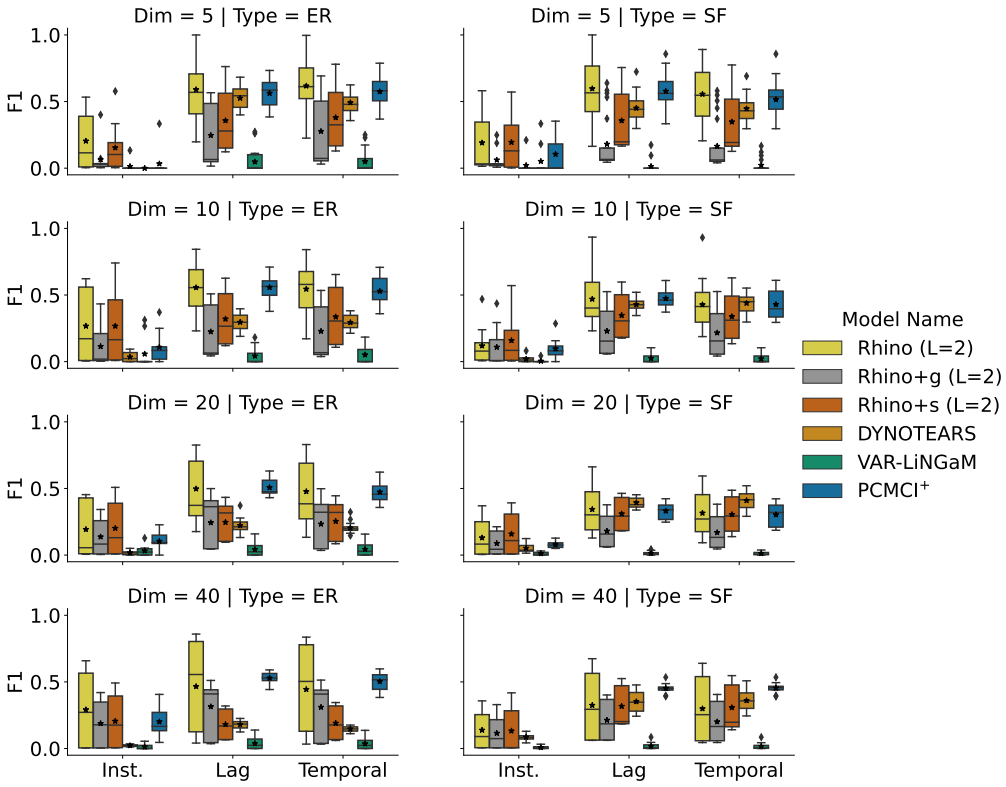

Figure 3: Comparison of the F_1 score of the different baseline methods as well as Rhino (light yellow) depending on the dimensionality and the graph type. Inst. refers to the performance on the instantaneous adjacency matrix, Lag refers to the lagged adjancency matrices and temporal considers the full temporal matrix.

Ablation: history dependency Figure 4 explores the performance difference of all methods on data generated with/without history-dependent noise. Interestingly, most methods perform better on the history-dependent datasets than the history-independent ones. The possible reasons are (1) the difficulty of the discovery also depends on the randomly sampled functions; (2) the default hyperparameters of all methods are initially chosen to favor the datasets with history-dependent noise and instantaneous effects. We find that PCMCI$^+$ is the most robust across both settings, followed by Rhino and DYNOTEARS. On the other hand, the two variants of Rhino seems to be less robust. When the Rhino is correctly specified, it achieves the best performance. In summary, Rhino demonstrates reasonable robustness to history-dependency mismatch and achieves the best when correctly specified.

Ablation: instantaneous effect We investigate the impact of instantaneous effects in the data. Figure 5 shows the F_1 score averaged over all possible setting combinations other than instantaneous effect. All methods seem to be robust across both settings with PCMCI$^+$ and Rhino performing the best. The score of the instantaneous adjacency matrix when instantaneous effects are disabled is not defined and therefore not plotted.

I Real-world Experiment Details

I.1 Post-processing temporal adjacency matrix

The ground truth graphs for Netsim and DREAM3 datasets are summary graph, which is essentially the temporal graph aggregated over time. We provide a formal definition of summary graph:

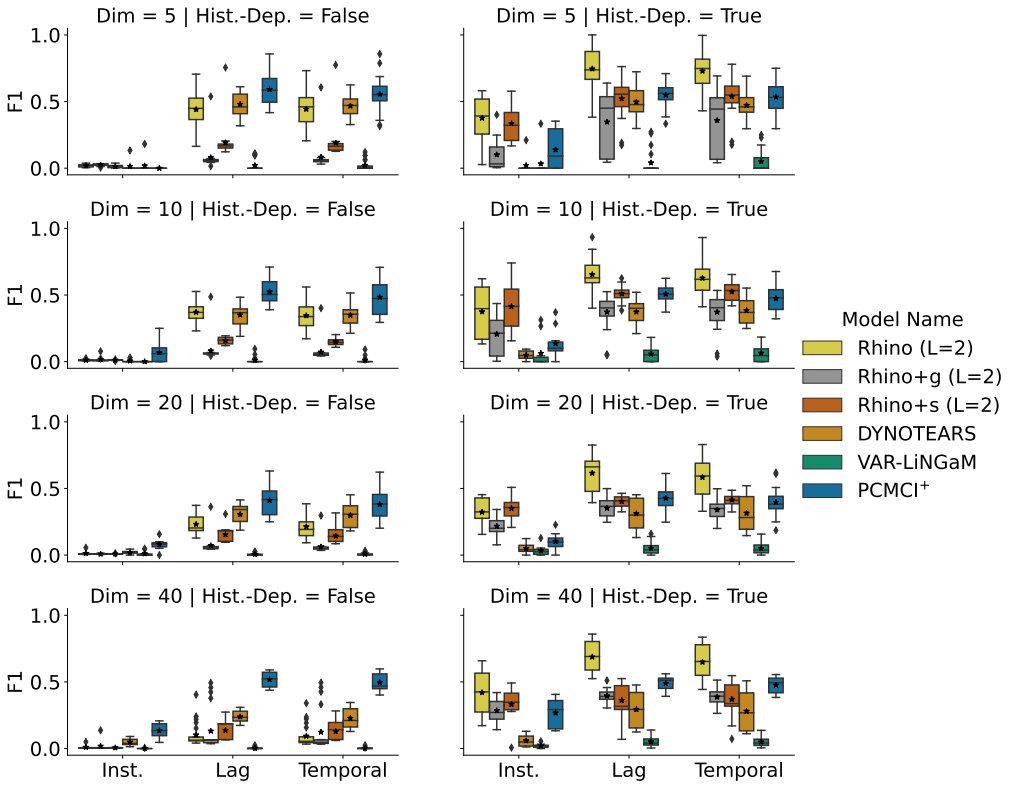

Figure 4: Comparison of the F_1 score of the different baseline methods as well as Rhino (light yellow) depending on the dimensionality and whether the data is generated with history-depence or not. Inst. refers to the performance on the instantaneous adjacency matrix, Lag refers to the lagged adjancency matrices and temporal considers the full temporal matrix.

Definition I.1 (Causal summary graph [3])**.** Let $\boldsymbol{X}_t$ be a multivariate temporal process, and $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$ be a summary graph. The edge $p \to q$ exists if and only if there exists some time t and some lag τ such that $\boldsymbol{X}^p_{t-\tau}$ causes $\boldsymbol{X}^q_t$ at time t with a lag $0 \leq i$ for $p \neq q$ and with a time lag of $0 < i$ for $p = q$.

Unlike the some of the Granger causality baselines, Rhino (and its variants), DYNOTEARS, VAR-LiNGaM produces the temporal adjacency matrix after training. For DREAM3 and Netsim datasets, this creates the incompatibility during evaluation. Thus, we need to aggregate the temporal graph into a summary graph before comparing to the ground truth. For binary adjacency matrix, we sum over the time steps followed by a step function, i.e. $\text{step}(\sum_k \boldsymbol{G}_k)$. Thus, there will be an edge $i \to j$ in summary graph as long as there is a connection from i to j at any timestamp. For the Bernoulli probability matrix from Rhino and its variants, we take a $\max(\cdot)$ over the timestamp to generate the probability matrix for the summary graph.

An exception is PCMCI$^+$, which can only produce MECs for the instantaneous adjacency matrix. In such case, we will enumerate up to 10000 possible instantaneous DAGs from the MECs. Together with the lagged adjacency matrix, we will perform the above post-processing step to generate the corresponding aggregated adjacency matrix. We also estimate the corresponding edge probabilities by taking the average over all possible DAGs.

For DREAM3 experiments, we ignore the self-connections by setting the diagonal of the aggregated adjacency matrix to be 0.

For Netsim, self-connections are not ignored, following the same settings as Khanna and Tan [16].

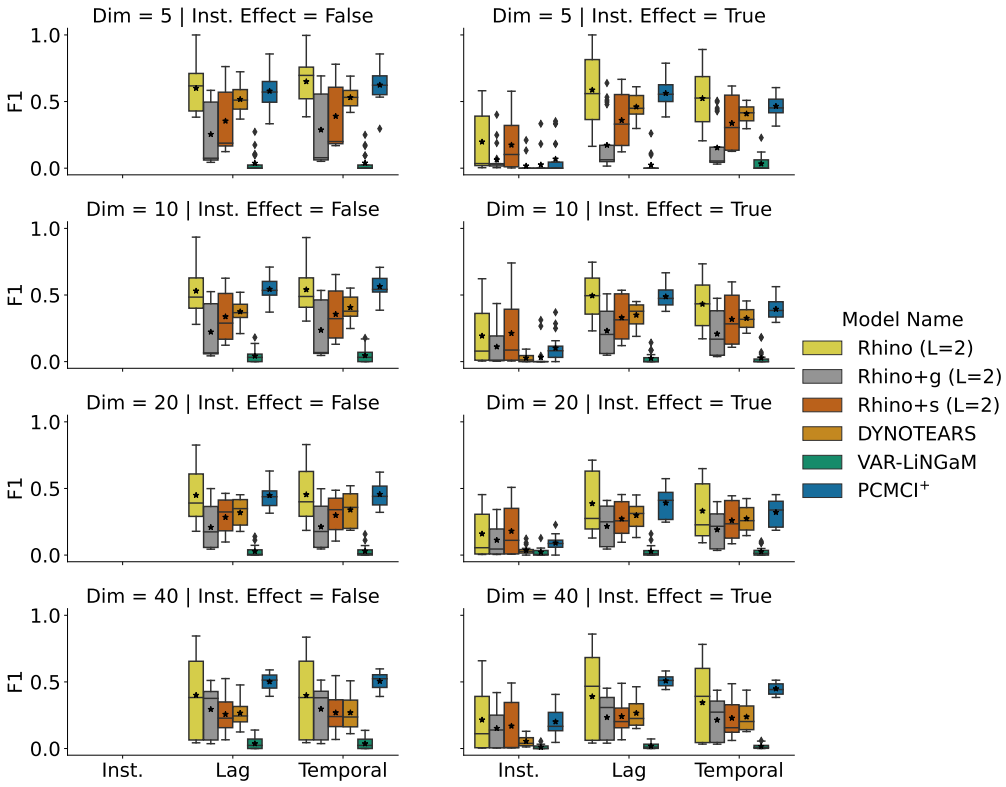

Figure 5: Comparison of the F_1 score of the different baseline methods as well as Rhino (light yellow) depending on the dimensionality and whether the data is generated with instantaneous effects or not. Inst. refers to the performance on the instantaneous adjacency matrix, Lag refers to the lagged adjancency matrices and temporal considers the full temporal matrix.

I.2 Netsim Hyperparameter setting

For the Netsim experiment, we extract subject 2-6 in *Sim-3.mat* to form the training data and use subject 7-8 as validation dataset. We tune the hyperparameters of Rhino and its variants based on the validation log likelihood; DYNOTEARS with MSE on validation dataset; and use default settings of PCMCI$^+$ from Tigramite package.

It is worth noting that our setup of Netsim experiment is different from Khanna and Tan [16], where they train the baselines using a **single subject** and compute the corresponding AUROC, followed by averaging over subjects 2-6. Our setup is to train all methods using the entire data from subject 2-6 before computing AUROC. Thus, the hyperparameters for Granger causality are slightly different, and the AUROC increases for the baselines compared to those reported in Khanna and Tan [16].

Rhino we use 2-layer MLPs with 64 hidden units for both $\ell_{\tau,j}, \zeta_i$ in Eq. (4) and the hyper-network for conditional spline flow (8 bins). The node embedding dimension is 15. All the MLPs use residual connections and layer-norm at every hidden layer. We use linear conditional spline flow [6] for better training stability. We also initialise the Bernoulli probability $q_\phi(\boldsymbol{G})$ to have no preference (i.e. edge probability $= 0.5$). For prior $p(\boldsymbol{G})$, we set the initial value $\rho = 1$, $\alpha = 0$ and $\lambda_s = 25$. For the gradient estimator, we use the Gumbel softmax method with a hard forward pass and a soft backward pass with temperature of 0.25. We use batch size 64, learning rate 0.001 with Adam optimizer [17]. The training procedure follows from Appendix B.1 in Geffner et al. [8]. The above is also used for Rhino variants.

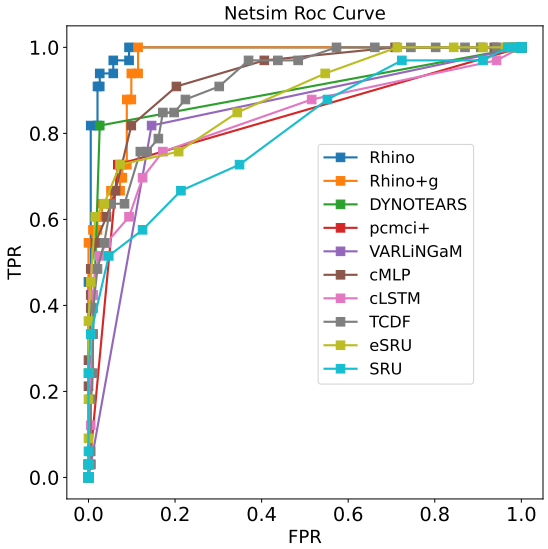

Figure 6: The ROC curve plots of Rhino and other baselines for Netsim dataset. Similar to Fig. 7, we only select 1 run out of 5 for Rhino, Rhino+g, DYNOTEARS, PCMCI$^+$ for illustration purpose.

DYNOTEARS, PCMCI$^+$ and VARLiNGaM For DYNOTEARS, we set lag to be 2, $\lambda_a = 0.5$ and $\lambda_w = 0.5$. For PCMCI$^+$, we use `parcorr` independence test with lag 3. For VARLiNGaM, we use lag 2 with default settings as `https://lingam.readthedocs.io/en/latest/`.

Granger Causality For computing AUROC, we follow the same method as Khanna and Tan [16], Tank et al. [37] by sweeping through a range of hyperparameters. Specifically, we use the same hyperparameters for SRU and eSRU as [16]. For cMLP, we choose the ridge penalty as 0.43 and sweep through the group sparse penalty in range $[0.1, 1]$. For cLSTM, we set the ridge penalty to be 0.045, and sweep the group sparse penalty in range $[0.1, 1]$.For TCDF, we sweep through the threshold in range $[-1, 2]$ for the attention scores to extract corresponding summary graphs. Other than the above hyperparameters, everything else follows the setup as in Khanna and Tan [16].

I.3 Additional Netsim Results

Figure 6 shows the ROC curve plot for Rhino and other baselines. It is clear that Rhino achieves significantly better TPR-FPR trade-offs compared to others.

I.4 Additional real-world benchmark: DREAM3

We evaluate Rhino performance with an additional real-world biology benchmark called *DREAM3* [29, 21]. These datasets are also used to evaluate Granger causality [16, 37, 23, 4] but recently adopted for SEM-based method [24]. The dataset consists *in silico* measurements of gene expression levels for 5 different networks. Each network contains $d = 100$ genes. Each time series represents a perturbation trajectory with time length $T = 21$. For each network, 46 perturbation trajectories are recorded. The goal is to infer the causal structure of each network. AUROC metric is also used here. We consider the same baselines as in the synthetic experiments (i.e. DYNOTEARS and PCMCI$^+$) without *VARLiNGaM* since its default implementation fails when the number of variables ($d = 100$) is greater than the series length ($T = 21$). Additionally, we also consider relevant Granger causality methods as Netsim experiments. Their corresponding results are directly cited from Khanna and Tan [16] due to the identical setup. Appendix I.4.1 specifies Rhino hyperparameters.

Table 3 demonstrates the AUROC of the summary graph inferred after training. It is clear that Rhino and its variant outperform all other methods. Although Rhino is not formulated to solve the summary graph discovery, this result together with Netsim experiment confirm a clear advantage compared to the state-of-the-art Granger causality. Thus, Rhino can be used to infer either temporal or summary graph depending on users' needs.

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 ± 0.002	0.519 ± 0.002	0.530 ± 0.003	0.510 ± 0.001	0.512 ± 0
Rhino+g	**0.673±0.013**	0.665**±0.009**	**0.659±0.005**	**0.598±0.004**	**0.588±0.005**
Rhino	**0.685±0.003**	**0.680±0.007**	**0.664±0.006**	0.585±0.004	0.567±0.003

Table 3: The AUROC of the summary graph adjacency matrix for 5 datasets in DREAM3, where self-connections are ignored. DYNO. means DYNOTEARS. For Rhino and PCMCI$^+$, the mean results with standard error are reported by averaging over 5 runs. Khanna and Tan [16] only reported the single-run results for baselines.

By inspecting the hyperparameters of Rhino in Appendix I.4.1, instantaneous effects seem to provide no help for discovery in these datasets. It suggests the recording intervals are fast enough to avoid any aggregation effect. This explains why the Granger causality can also perform reasonably well.

Unlike the strong performances of DYNOTEARS and PCMCI$^+$ in synthetic experiments, they perform poorly in DREAM3. The linear nature of DYNOTEARS seems to harm its performance drastically. On the other hand, PCMCI$^+$ suffers from the low independence test power under small training data.

Another interesting ablation is to compare with Rhino+g, which performs on par with Rhino and achieves better scores on 2 out of 5 datasets. Although we have no access to the true noise mechanism, we suspect that the added noise is not history-dependent and highly likely to be Gaussian. Despite the model mismatch, Rhino is still one of the best methods for this problem. This further strengthens our belief in the robustness of our model under misspecification.

I.4.1 DREAM3 Hyperparameter setting

For tuning the hyper-parameters of Rhino, its variants and DYNOTEARS, we split each of the 5 datasets into $80\%/20\%$ training/validation. We tune baselines in the same way as in Netsim experiment (Appendix I.2). For other Granger causality baselines, refer to Table 7-11 in Khanna and Tan [16].

Hyperparams	Node Embedding	Instantaneous eff.	Node Embed. (flow)	lag	λ_s	Auglag
Rhino (Ecoli1)	16	False	16	2	19	30
Rhino (Ecoli2)	16	False	100	2	25	80
Rhino (Yeast1)	32	False	100	2	25	10
Rhino (Yeast2)	32	False	100	2	25	80
Rhino (Yeast3)	32	False	16	2	25	5
Rhino+g (Ecoli1)	100	False	N/A	2	15	60
Rhino+g (Ecoli2)	100	False	N/A	2	25	25
Rhino+g (Yeast1)	100	False	N/A	2	15	5
Rhino+g (Yeast2)	100	False	N/A	2	19	125
Rhino+g (Yeast3)	100	False	N/A	2	9	10

Table 4: The hyperparameter setup for Rhino. `Node embedding` is the dimensionality of $\boldsymbol{u}_{\tau,i}$ below Eq. (4); `Instantaneous eff.` specifies whether it models the instantaneous effect or not; `Node Embed. (flow)` represents the dimensionality of the node embedding for the hyper-network used for conditional spline flow g_i since the hyper-network shares the similar structure as Eq. (4); `lag` defines the model lag order; and λ_s is the sparseness penalty in the prior (Eq. (7)); `Auglag` is the number of augmented Lagrangian steps, each step consists of 2000 training iterations.

Other than the hyper-parameters reported in Table 4, we use 1-layer MLPs with 10 hidden units for neural networks in function f_i and hyper-network. Other settings are the same as Netsim (Appendix I.2).

Hyperparams	lag	λ_a	λ_w
Ecoli1	2	0.01	0.5
Ecoli2	2	0.1	0.01
Yeast1	2	0.005	0.1
Yeast2	3	0.01	0.01
Yeast3	2	0.01	0.005

Table 5: The hyperparameter setup for DYNOTEARS.

Table 5 contains the hyper-parameters setup for DYNOTEARS. We set the maximum training iterations to be 1000 with DAGness tolerance 10^{-8}. The threshold value for the weighted adjacency matrix is 0.05. For PCMCI$^+$, the maximum lag is set to 2. The conditional independence test is set to `parcorr`, which is based on linear ordinary least square (OLS). A more powerful choice can be a nonlinear independence test based on GP, called `GPDC`. However, PCMCI$^+$ with $GPDC$ is too slow to finish the training.

I.4.2 Additional DREAM3 Results

Here, Fig. 7 shows the additional ROC curve plots for all 5 datasets in DREAM3. For the visualization purpose, we only select a single run for Rhino and this will not affect the curve much due to small standard error in Table 3.

J AUROC Metric

AUROC metric is a one of the standard metrics for evaluating the causal discovery, which measures the trade-off between the *true positive rate* (TPR) and *false positive rate* (FPR). However, during the experiments, we found out that AUROC does not necessarily correlate well with other discovery metrics. From Fig. 8, it is clear that the F_1 score continues to increase whereas AUROC and validation likelihood starts to decrease after few steps. Since the dataset of Netsim is relatively small, this indicates the possible overfitting. This disagreement originates from the different aspects these metrics care about. For AUROC, it cares about the trade-off between TPR and FPR with various decision thresholds, and it penalizes the wrong decisions with certainty harshly. On the other hand, F_1 score cares about the final inferred binary adjacency matrix with a fixed decision threshold. For example, if we multiply the Bernoulli probability matrix by a small factor (e.g. 10^{-5}), the AUROC score will remain the same but the F_1 score will tends to 0 with the default decision threshold 0.5.

Thus, model overfitting tends to drive the edge probabilities towards 1 or 0, which may help the F_1 score but these extreme decisions can result in a large decrease in the AUROC score. Thus, for small dataset, we believe AUROC is a better metric than F_1, which also agrees with validation likelihood.

In addition, the Bayesian setup of Rhino may also help with better AUROC for small dataset. From the same figure, even the large decrease of validation likelihood suggests potential model overfitting, the AUROC still maintains a reasonable value. This may be due to the Bayesian view of the causal graph, where the posterior edge probability does not converge to extreme values.

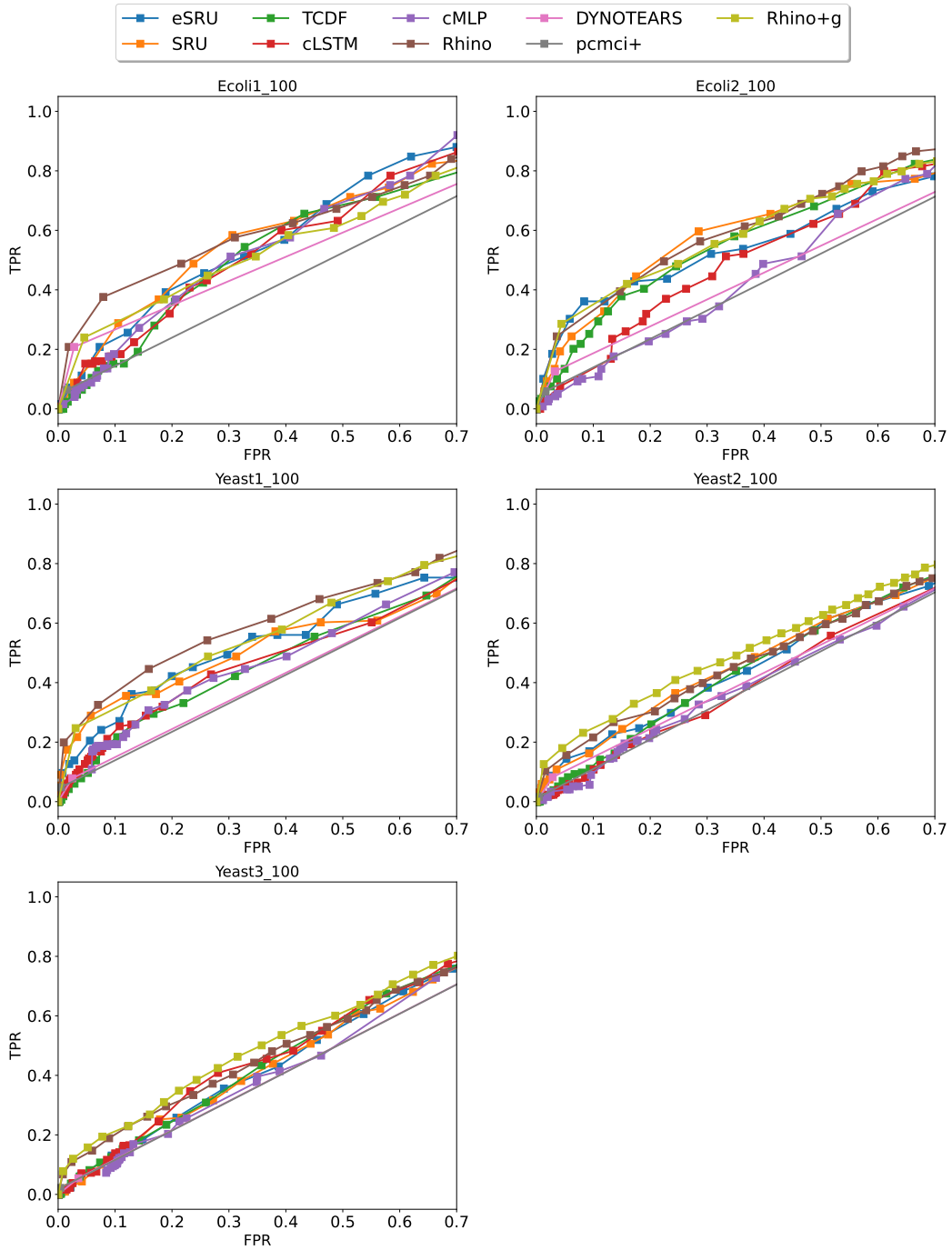

Figure 7: The ROC curve plots of Rhino and other baselines for DREAM3 datasets. For illustration purpose, we only select a single run of Rhino, Rhino+g, DYNOTEARS and PCMCI$^+$ to plot ROC curve. Since the standard error reported in Table 3 is relatively small, the plot should not vary much for other runs. The ROC curve of other baselines are directly taken from figure 7 in Khanna and Tan [16].

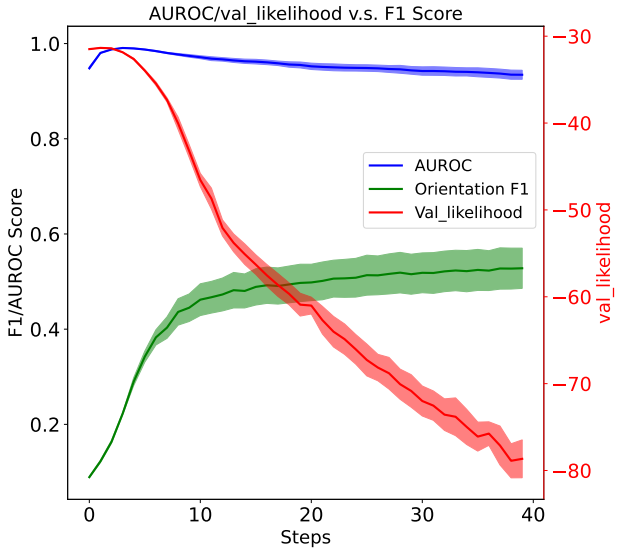

Figure 8: The curves of orientation F_1, AUROC and validation likelihood during training. Each curve is obtained by averaging over 5 random seeds. The validation curve agrees well with the AUROC curve, but shows an opposite trends as F_1 curve. This potentially indicates model overfitting in the later stage of training.