Adversarial Causal Tuning for Realistic Time-series Generation

Nikolaos Gkorgkolis* GKORGKOLIS@CSD.UOC.GR
Nikolaos Kougioulis* KOUGIOULIS@CSD.UOC.GR
Ioannis Tsamardinos TSAMARD@CSD.UOC.GR
University of Crete & FORTH, Heraklion, Greece

Mingxue Wang WANGMINGXUE1@HUAWEI.COM
Bora Caglayan BORA.CAGLAYAN@HUAWEI.COM
Andrea Tonon ANDREA.TONON1@HUAWEI-PARTNERS.COM
Dario Simionato DARIO.SIMIONATO1@H-PARTNERS.COM
Huawei Ireland Research Centre, Dublin, Ireland

arXiv:2506.02084v2 [cs.LG] 12 May 2026

Abstract

We address the problem of generating simulated, yet realistic, time-series data from a causal model with the same observational and interventional distributions as a given real dataset (probabilistic causal digital twin). While non-causal models (e.g., GANs) also strive to simulate realistic data, causal models are fundamentally more powerful, able to simulate the effect of interventions (what-if scenarios), optimize decisions, perform root-cause analysis, and counterfactual causal reasoning. We introduce the Adversarial Causal Tuning (ACT) methodology, which outputs the optimal causal model that fits the data, along with a quantification of the goodness-of-fit. The returned causal model can then be employed to simulate new data or to perform other causal reasoning tasks. ACT adopts ideas from Generative Adversarial Network training and AutoML to search for optimal causal pipelines and discriminators that detect deviations between the distributions of real and simulated data. It also adapts a permutation testing procedure from established causal tuning methods to penalize models for complexity. Through extensive experiments on real, semi-synthetic, and synthetic datasets, we show that (a) employing multiple optimized discriminators is paramount for selecting the optimal causal models and quantifying goodness-of-fit, (b) ACT selects the optimal causal model in synthetic datasets while avoiding overfitting, generating data indistinguishable from the true data distribution (c) all state-of-the-art generative and causal simulation methods, exhibit room for improvement in reproducing real data distributions; generating realistic temporal data is still an open research challenge.

Keywords: Causal Discovery, Causal model selection, Generative Models, Time-series

1. Introduction

Causal models encode the generative mechanisms underlying the data. Unlike predictive or generative models, they support the simulation of interventions to explore what-if scenarios, decision optimization, root-cause analysis, and counterfactual reasoning (Pearl, 2009). Such models are rarely fully specified based on prior knowledge alone. Hence, causal discovery algorithms have been developed to induce them from observational data, or a limited number of interventional data (Spirtes et al., 2001; Runge, 2018; Runge et al., 2019). Numerous such algorithms have been proposed in the literature (Zanga et al., 2022). However, no universally optimal algorithm exists: the efficacy of an algorithm depends on the input dataset. Therefore, one should search the space of causal modeling

*. Authors contributed equally to this work.

© N. Gkorgkolis*, N. Kougioulis*, I. Tsamardinos, M. Wang, B. Caglayan, A. Tonon & D. Simionato.

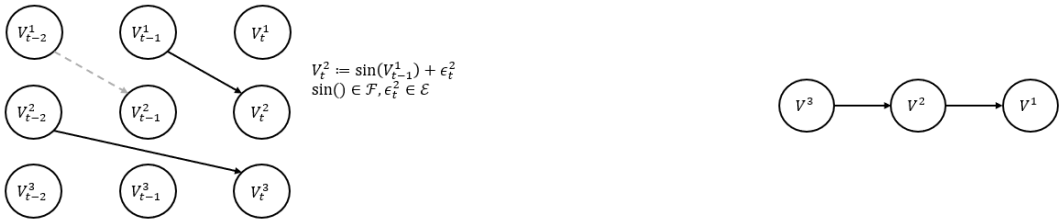
pipelines to tune algorithmic choices and their hyperparameters to identify the best-fitting causal model, giving rise to the **causal tuning** problem. The latter is conceptually equivalent to tuning predictive models in automated machine learning (AutoML) systems. However, causal tuning is an *unsupervised task* which requires the development of special methodologies.

In this work, we present a general principle for causal tuning called **Adversarial Causal Tuning** or **ACT**. Given as input a multiple time-series dataset, it returns an optimal temporal causal model and a measure of goodness-of-fit. Our main motivation for ACT in this paper is *realistic data generation*. Specifically, it is to identify a causal model that generates data indistinguishable from the real ones. Crucially, it should also be able to simulate what-if scenarios where one intervenes to set the value of a variable and produce data from the interventional distribution $P(X|\mathrm{do}(C = c))$. Such *a causal model can thus serve as a probabilistic Causal Digital Twin and a foundation for privacy-preserving data sharing* (Homaei et al., 2025). Notice that non-causal generative models, such as conditional GANs (Mirza and Osindero, 2014), can generate data from the conditional distribution $P(X|C = c)$ which is not always equal to $P(X|\mathrm{do}(C = c))$.

ACT also aims to solve another fundamental problem in causal discovery, namely, realistic benchmark creation for causal algorithms. Specifically, to benchmark a causal algorithm, one needs to have available a dataset as well as its corresponding ground-truth, data-generating causal model. Existing benchmarks are almost exclusively comprised of **synthetic causal models**. These are produced by randomly sampling the causal graph from a graph distribution (e.g., Erdős–Rényi or Barabási–Albert distributions (Brouillard et al., 2020)). The causal functional dependencies are sampled from small hand-crafted families (e.g., linear Gaussian or simple nonlinear additive noise models), while the noise distributions are typically parametric. *It is highly questionable whether the community corpus of benchmark causal models resembles real causal models encountered in practice*. ACT can be employed to fit real datasets; whenever an acceptable goodness-of-fit has been achieved by the winning causal model, it can then be added to a benchmark corpus of realistic causal models. Finally, ACT can serve as the heart of automated causal discovery platforms (**AutoCD**) (Biza et al., 2024) that try to automate the induction of causal models and causal inference tasks.

ACT frames causal model selection as a minimax problem: a search in the space of causal pipelines (*causal configurations*) tries to identify a causal model that *minimizes* the ability of discriminators to distinguish real from simulated data; simultaneously, a search in the space of predictive pipelines (*discrimination configurations*) tries to identify the discriminator that *maximizes* the identification of the source of a sample: real or simulated. If the best discriminator performs no better than random guessing for a causal model, the latter achieves goodness-of-fit. ACT also includes a sparsity penalty that prevents overfitting, based on permutation testing (Biza et al., 2022). ACT ensures distributional fit rather than causal fit; the true causal graph may be different, but it should at least be a Markov equivalent one (Spirtes et al., 2001). Causal fit can only be validated with experiments. ACT is comparable to a model-agnostic and discrete version of the GAN minimax procedure. The searches in the causal and discrimination configuration spaces can employ AutoML techniques (Hutter et al., 2019) such as black-box optimization and meta-level learning. The current implementation of ACT addresses the tuning of Temporal Structural Causal Models, assuming causal sufficiency, no contemporaneous causal effects, and additive noise.

Our motivational experiments show that one should optimize over discriminators, or risk underfitting. Using a single discriminator is common in the literature and has led prior works to *falsely* conclude that their methodologies can fit models that generate realistic data (Cheng et al., 2024). Our experiments show that this conclusion could be an artifact of the lack of optimization.



(a) Lagged Causal Graph

(b) Summary Causal Graph

Figure 1: (a) The lagged graph of a TSCM with max lag $\ell_{\text{max}} = 2$ where V^1 causes V^2 with lag 1 and V^2 causes V^3 with lag 2. The functional dependency of V^2_t on its direct causes (parents) is shown. (b) The corresponding summary causal graph (as in (Cheng et al., 2024)) discards time lag information, representing only the aggregated causal influences.

Our contributions (a) include the formalization of ACT and an open-source implementation `(url withheld to maintain anonymity)`. (b) We show that penalizing for sparsity is essential for preventing overfitting of the selected model. (c) We demonstrate that ACT outperforms state-of-the-art methodologies in fitting datasets from synthetic causal models. In addition, (d) ACT can most often successfully generate realistic data under interventions. Finally, (e) we show that on real datasets, all state-of-the-art methods fail to successfully fit a generative model. Realistic time-series simulation remains an open challenge, motivating the need for improved causal discovery algorithms. ACT can be used for generating realistic data, for causal model selection within automated causal discovery pipelines, and for creating realistic causal models for benchmarking algorithms.

2. Preliminaries & Problem Formulation

Temporal structural causal models (TSCM) We consider a discrete multivariate process of a vector of variables $V \in \mathbb{R}^N$ over time points $1, \ldots, T$. V^j_t denotes the value of variable $V^j \in \mathbf{V}$ at time t, while $\mathbf{V}^j_{<t}$ denotes the past values of V^j, i.e., the set $\{V^j_1, \ldots, ...V^j_{t-1}\}$. We assume direct causal relations are given by structural equations of the form: $V^j_t := f^j(\text{Pa}(V^j_t)) + \epsilon^j_t$, where $\text{Pa}(V^j_t) \subseteq \cup_j \{V^j_{<t}\}$ and is called the *parent set* of V^j. Thus, V^j_t is causally determined by a set of past values of $\mathbf{V}$, but not contemporaneous values of $\mathbf{V}$. The noise terms ϵ^j_t are independent of each other and all variables, hence we assume no latent confounding factors (causal sufficiency). We assume that the causal functional relationships f^j and the distribution of the noise terms do not depend on time. The causal equations can be qualitatively represented with a graph G where the vertices correspond to each V^j_t and there is an edge $V^j_{t'} \to V^k_t$, whenever $V^j_{t'} \in \text{Pa}(V^k_t)$. We assume the graph structure does not depend on absolute time, hence we consider only the portion of the graph with variables $V_{t-\ell_{\text{max}}}, \ldots, V_t$, where ℓ_{max} is the maximum lag for which an edge $V^j_{t-\ell_{\text{max}}} \to V^k_t$ exists. We refer to this subgraph as the **Lagged Causal Graph** of the process. The lagged graph captures the dynamics of the process. A **TSCM** is a tuple $\langle G, F, E, C \rangle$, where G is a lagged causal graph, $\mathcal{F} = \{f^i, i = 1, \ldots, N\}$ the set of functional dependencies, $\mathcal{E}$ the set of noise distributions, and C the probability distribution of the initial conditions $V_1, \ldots, V_{\ell_{\text{max}+1}}$. The **Summary Causal Graph** is defined as a graph containing a single vertex for every variable and

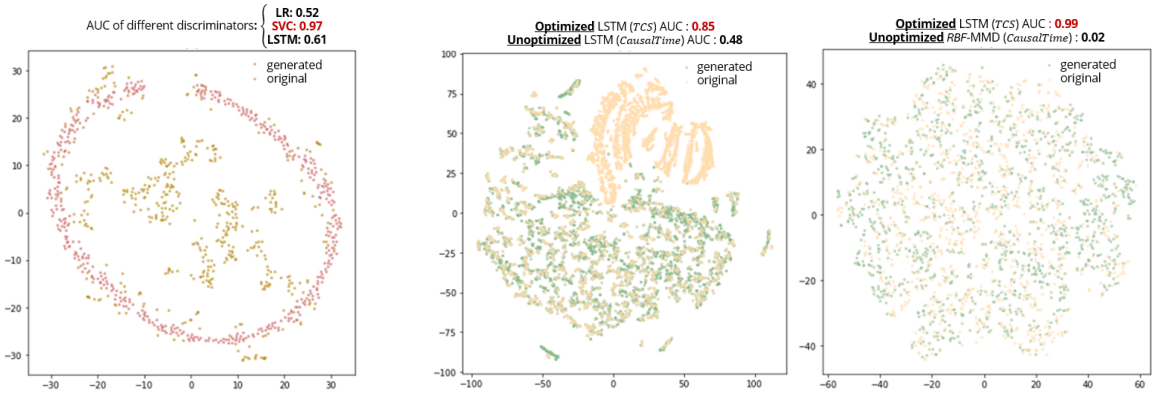


Figure 2: t-SNE (Van der Maaten and Hinton, 2008) 2D projections of original vs. simulated data for a synthetic, real, and semi-synthetic dataset. (Left) The data distributions are visually different, but one needs to optimize over several discrimination algorithms to identify the one that achieves high AUC. (Middle) The same algorithm (LSTM) is able to distinguish the data distributions when its hyperparameters are tuned properly. (Right) t-SNE is not able to visually depict the differences between the data distributions, but an optimized LSTM achieves discrimination $\text{AUC} = 0.99$. An unoptimized kernel MMD is not able to detect distributional differences. Prior works that rely on unoptimized discriminators and metrics may overclaim successful fitting of real datasets with generative models.

an edge $V^j \to V^k$ if $V^j_{t-\ell} \to V^k_t$ for some lag ℓ. Given that we are mostly interested in capturing the dynamics of the process, we will drop from the model the initial distribution C and represent TSCMs with the tuple $\langle G, F, E \rangle$.

Data generation (simulation) from a TSCM. Given a TSCM data can be generated by *ancestral sampling* (Murphy, 2023): parent values are retrieved from previous lags, noise terms are drawn, and structural equations are evaluated at each time step. Initial values V^j_t for $t \leq \ell_{\max}$ are sampled from the distribution $\epsilon^j \in \mathcal{E}$. Initial warm-up steps are discarded to reduce any noisy dependence on initial conditions.

Problem statement. Let $\{\mathbf{x}_t\}_{t=1}^T$, $\mathbf{x}_t \in \mathbb{R}^N$, denote an observed, real multivariate time-series of N quantities observed over T timesteps. **The objective is to learn a TSCM that generates simulated data whose distribution is indistinguishable from the real data**. Achieving this requires estimating a lagged causal graph, functional causal (structural) relationships, and noise distributions.

3. Related Work on Realistic Temporal Data Generation

We focus on prior works, causal or non-causal, that, given an input real dataset, generate simulated data whose distributions should be indistinguishable from the real dataset. Regarding non-causal methods, the work of Zhang et al. (2022) concerns the CPAR (Conditional Probabilistic Auto-regressive) model for generating multivariate time-series samples, while TimeVAE (Desai et al., 2021) follows a variational autoencoder (Murphy, 2023) approach. Both these methods are non-causal and hence cannot simulate interventional data. CausalTime (Cheng et al., 2024) fits a

residual-based neural network to model a VAR process and then extracts a *hypothetical causal graph* (using the authors' terminology) using feature-importance methods such as DeepSHAP (Lundberg and Lee, 2017). However, the resulting graph is a **summary graph** without lag information (Figure 1) and does not correspond to a temporal causal model. As emphasized in their work, *feature-importance scores do not constitute causal discovery, and the generative mechanism learned by their network does not approximate a TSCM*. As the method does not model an explicit TSCM, interventional simulations or counterfactual reasoning are not supported.

While our focus is the generation of realistic time series data from a causal model, ACT is a general methodology for causal model selection. Prior work on this problem is the Out-of-sample Causal Tuning (OCT) algorithm (Biza et al., 2022). However, OCT does not focus on the quality of the data generation and is implemented only for atemporal (cross-sectional, i.i.d.) data. Significant adaptations are required to enable the comparison of OCT and ACT as causal tuning methods for temporal data which are left for future work.

4. Motivation for optimizing over discriminators

Prior works use different methodologies to assess the successful fitting of a model and its ability to generate realistic data. These methodologies typically rely on using a single discrepancy measure (e.g., MMD) (Gretton et al., 2012) or discriminator (e.g. SVM, LSTM, LR) (Lopez-Paz and Oquab, 2017). However, different discriminators vary dramatically in their statistical power depending on the dataset and hyperparameter configuration. As a result, poorly tuned discrimination can incorrectly suggest that simulated and real data come from the same distribution, even when their discrepancy is visually or statistically evident. For example, CausalTime uses a single, fixed discriminator, leading the authors to believe successful fitting in several real datasets; our experiments in Section 6 below disprove these claims. Figure 2 illustrates this issue across a synthetic (based on trigonometric functions (left)), real (AirQualityUCI (Ye et al., 2024) (middle)), and semi-synthetic dataset (fMRI from Smith et al. (2011) (right)) (see Section 6 for a definition). Generative models were fit using CausalTime on each corresponding dataset. This demonstrates the need to optimize over discrimination algorithms as well as hyperparameters. In addition, visualizations cannot always be trusted to spot distributional differences.

5. Adversarial Causal Tuning

Adversarial Causal Tuning or ACT seeks to **identify an optimal TSCM that fits the input dataset** and estimate its goodness-of-fit. The returned TSCM can be employed to generate simulated data under observational and/or interventional conditions. Informally, we define ACT as the following selection principle:

> **ACT principle**: Given an input dataset T, select as best-performing causal model, the one that, despite our best discrimination efforts, generates data as indistinguishable as possible from the real dataset. Select the simplest causal model whose discrimination performance is statistically equivalent to the best-performing one.

We now discuss how to implement this principle for the specific context of TSCMs and multivariate time series data. First, we focus on *measuring and estimating discrimination* between two multivariate time series datasets (empirical distributions) T and T' on the same variables. As our metric

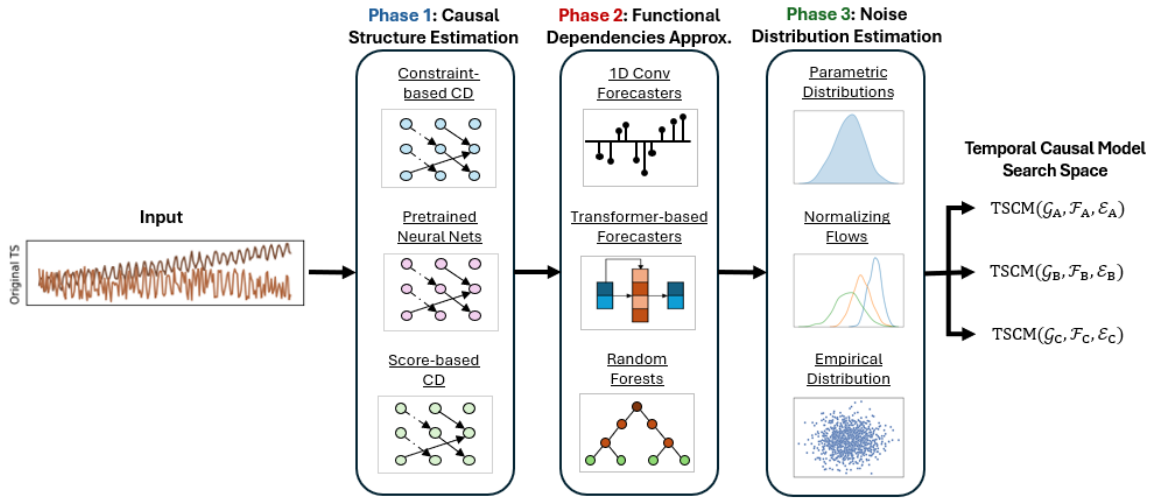


Figure 3: ACT's production of TSCM space: Given a multivariate time-series sample of real data, ACT induces lagged causal graphs (*Phase 1*), functional dependencies (*Phase 2*), and noise distributions (*Phase 3*) using all combinations of available algorithms and their hyperparameters. Each combination of Phases 1–3 produces a temporal causal model.

of choice, we use the Area Under the Receiver's Characteristic Curve or **AUC** achieved by a discriminator D when presented with the binary classification problem of distinguishing the source T or T' of a sample, denoted as $d_C(T, T')$. This is equivalent to a Classifier 2-Sample Test (C2ST) (Lopez-Paz and Oquab, 2017). To account for the temporal information of data on a generic classifier, a time series dataset is converted to time-lagged format. Specifically, a sample $[V_{t-k}, \ldots, V_t]$ is created for every $t = \ell_{\text{max}} + 1, \ldots, T$ from the previous k time steps. For time-series-specific classifiers like LSTM, temporal information is determined by the sequence length parameter, as in CausalTime (Cheng et al., 2024). To estimate $d_C(T, T')$ we hold out 25% of the data as a test set and train the discriminator on the remaining 75%.

Let us now define the TSCM space $\mathcal{C}$ as the set of TSCMs under consideration. This set can be produced by pipelines of algorithms and their hyperparameters that accept as input a multivariate timeseries T and output a TSCM. Any such a pipeline should include algorithms for learning the causal structure, the functional dependencies, and the noise distribution, but could also include steps for data transformation or imputation of missing values. We also define $\mathcal{D}$ as the space of all discrimination models that can be produced from an input dataset. $\mathcal{D}$ can be produced by any pipeline that accepts time series data and produces a binary classifier. We denote with S_C a simulated dataset that can be produced by TSCM C. The **discrimination performance** of a TSCM C is defined as $d_C = \max_{D \in \mathcal{D}} d_D(T, S_C)$. We can then define the **best-performing** TSCM C^* as:

$$C_* = \arg_C \min_{C \in \mathcal{C}} d_C = \arg_C \min_{C \in \mathcal{C}} \max_{D \in \mathcal{D}} d_D(T, S_C) \tag{1}$$

This minimax procedure is conceptually equivalent to the search implemented by Generative Adversarial Networks (GAN) (Goodfellow et al., 2014): the search in TSCM space aims to fit a causal

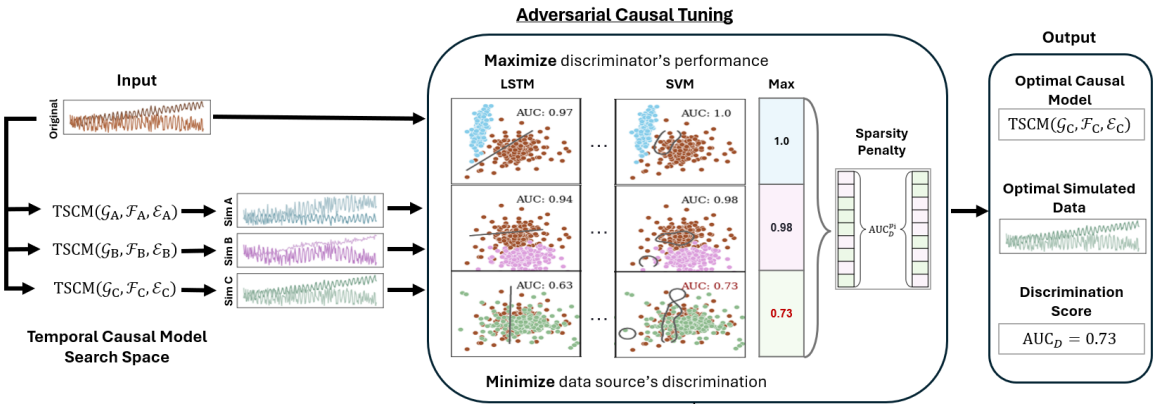


Figure 4: ACT production of discriminator space and selection of the final TSCM: For each TSCM, the discriminator with maximal power is selected. The best performing TSCM is the one with the minimum maximum discrimination performance (TSCM C, third row). The sparsity penalty is applied to select the final model. Its discrimination score is also returned as an estimate of goodness-of-fit.

model that generates data that decrease discrimination, while the search in the discriminator space (adversary) aims to increase it. In GANs, both spaces are defined by Deep Neural Networks (DNN) and are differentiable; ACT as presented, implements discrete searches and is algorithm-agnostic, not tied to DNNs. The size of the simulated data S_C is set to the size of the input dataset T to avoid imbalanced classification problems.

Sparsity Penalty: The best performing TSCM C_* may be overfitting the data, be overly complex, or uninformative. For example, a TSCM with a full causal graph can fit any data distribution. To avoid this situation, it is important to impose a complexity penalty on the selected TSCM. We adopt the same penalty scheme developed for optimizing causal models as in Biza et al. (2022). ACT selects as optimal the simplest TSCM, denoted by C_{ACT}, whose discrimination performance is statistically indistinguishable from d_{C_*}. As a simplicity metric for a TSCM $\langle G, \mathcal{F}, \mathcal{E} \rangle$, we use the number of edges (direct causal relations) in the lagged causal graph, denoted as $|G|$. To determine statistical equivalence of the performance of a TSCM C with C_*, we test the null hypothesis $H : d_{C_*} = d_C$ using a permutation-based test (Biza et al., 2022). Specifically, we consider the vectors of predictions v and v' of the best discriminator for C_* and C respectively, on the test set. Subsequently, the test returns a p-value p_C for the hypothesis that the two AUCs d_{C_*} and d_C are equal. If $p_C > a$ for some significance level a, the performance of C is deemed statistically indistinguishable from the best performance achieved. In the implementation, we employ the standard level $a = 0.05$. The TSCM selected by ACT is:

$$C_{ACT} = \arg_C \min_{C=\langle G,\mathcal{F},\mathcal{E}\rangle \in \mathcal{C}} |G|, \text{ s.t. } p_C > a \tag{2}$$

Goodness-of-fit: ACT returns as a metric of the successful fitting of the input dataset T with a causal model, the discrimination AUC performance of C_{ACT} on the test set, namely $d_{C_{ACT}}$. The

closer it is to 0.5, corresponding to random guessing, the better the fit. Values significantly lower than 0.5 indicate good discrimination with the classification labels inverted and should not occur.

Search algorithm: ACT's uses a simple grid search algorithm over discrete TSCM and discriminator spaces. Figures 3 and 4 graphically depict the ACT procedure. Table 2 in appendix A summarizes the search spaces.

Generation of the TSCM Space: The causal pipelines in $\mathcal{C}$ include three phases producing a causal graph, functional dependencies, and noise distributions, respectively. In phase I, ACT runs all combinations of available causal discovery algorithms and their hyperparameter values, within prespecified sets of possible values. The current implementation supports methods spanning constraint-based, continuous-optimization, and neural approaches (e.g., PCMCI (Runge, 2018), DYNOTEARS (Pamfil et al., 2020), and CausalPretraining - CP (Stein et al., 2024)). In phase II, given a lagged causal graph G, functions $f^j(Pa(V_t^j)$ are fit using regression and time-series forecasting models, such as Random Forests (Breiman, 2001) and time-series forecasting foundational models (Das et al., 2024). In phase III, for each graph G and functions $\mathcal{F} = \{f^j\}$ the residuals r_i^j are computed for each variable and time-point. The distribution ϵ^j of residuals for each V^j is modeled using density estimators (e.g., flow-based models (Rezende and Mohamed, 2015; Dinh et al., 2017)) and parametric distributions. At the end of all phases, TSCMs $\langle G, \mathcal{F}, \mathcal{E} \rangle$ are produced for every combination of algorithm and hyperparameter value for all phases (≈**700 TSCMs**).

Generation of the Discriminator space $\mathcal{D}$: The discriminator space $\mathcal{D}$ can be produced by any pipeline that accepts real and simulated data T and S_c, respectively, and outputs a discrimination (binary classification) model. The current implementation supports pipelines with single algorithms, namely Support Vector Machines (Hearst et al., 1998) and LSTMs (Hochreiter, 1997) producing ≈**150 discriminators**.

6. Experimental Comparative Evaluation of ACT

Datasets : We evaluate ACT on corpora of **real, semi-synthetic, and synthetic** datasets. **Seven real-world data** are drawn from commonly used benchmarks in the forecasting literature (Hahn et al., 2023), spanning energy, weather, transportation, environmental sensing, and air-quality monitoring. These include the ETTh1/ETTm1 electricity transformer datasets (Wu et al., 2021), the WTH weather dataset (Jiang et al., 2023), AirQualityUCI (AQUCI), hourly bike–pedestrian counts from the Burke–Gilman Trail (BU), outdoor sensor data from Nanning (OD), and the multi-station Air_Quality dataset used in Cheng et al. (2024) (AQ). Dataset descriptions are provided in Appendix B. We impute missing values through interpolation as in *CausalTime*. Each time-series stationarity has been confirmed by the Augmented Dickey-Fuller (Dickey and Fuller, 1979) unit root test. To provide standard errors for the computed metrics for each dataset, we sample 10 times 2000 timepoints, similar to Block Bootstrap (Gonçalves and Politis, 2011; Härdle et al., 2003). **Semi-synthetic systems** are mathematical models (ordinary or partial differential equations) with known causal structure of real physical systems or processes, which are then employed to simulate data. These consist of financial time-series generated from the Fama–French model (Huang and Kleinberg, 2015; Nauta et al., 2019) (5 datasets), and 6 nonlinear fMRI BOLD simulations constructed via balloon models (Smith et al., 2011; Buxton et al., 1998). Finally, we produce **20 synthetic causal models** and corresponding datasets by varying the following parameters. A random graph

Table 1: Comparison of unoptimized MMD vs AUC_D. There are instances (annotated with bold **red** color) where *MMD* values are low but AUC_D scores indicate high discrimination. Unoptimized MMD, commonly used in the literature, could lead to falsely conclude successful generation of realistic time series data.

Metric	Real							Semi-synthetic		Synthetic
	WTH	AQUCI	AQ	ETTh1	ETTm1	BU	OD	*f*MRI	Finance	
MMD	**0.28**$_{\pm.07}$	0.46$_{\pm.02}$	1.30$_{\pm.74}$	0.75$_{\pm.16}$	**0.27**$_{\pm.02}$	**0.23**$_{\pm.02}$	2.36$_{\pm.00}$	0.08$_{\pm.01}$	**0.26**$_{\pm.02}$	0.09$_{\pm.11}$
AUC_D	**0.85**$_{\pm.01}$	0.99$_{\pm.00}$	0.97$_{\pm.00}$	0.92$_{\pm.02}$	**0.93**$_{\pm.00}$	**0.88**$_{\pm.01}$	0.99$_{\pm.00}$	0.71$_{\pm.01}$	**0.99**$_{\pm.00}$	0.57$_{\pm.00}$

is sampled from either Erdős-Rényi (Brouillard et al., 2020) or Barabási-Albert (Barabási and Albert, 1999) scale-free models. Functional relationships are linear combinations of functions g of the form $V_t^j = \sum \alpha_i g(X_i)$, where the coefficients α_i are randomly sampled, and g belongs to the set of identity, power, exponential, sigmoid, ReLU, and trigonometric functions, leading to **non-linear** functional dependencies (except when g is selected to be the identity). Noise distributions are either uniform or Gaussian. The maximum lag $\ell_{\text{max}} \in [1, 3]$, the number of variables in $[3, 12]$, and datasets are simulated for 1000 timepoints (excluding the warmup timepoints). Finally, if a model simulates data with a trend, we resample the functional dependencies until their mean is stationary, similar to Stein et al. (2024). More details on the characteristics of the considered datasets can be found at Appendix B.

Metrics : We use two metrics for comparing distributions: (a) the discrimination performance on the test set, measured by the AUC; in figures and tables, denoted as AUC_D to explicitly show that it measures the AUC. (b) The *Maximum Mean Discrepancy – MMD* (Gretton et al., 2012) is a metric of distributional discrimination. In both cases, *lower is better* (AUC_D's lowest value is 0.5). For measuring causal graph similarity, we will use (i) the *Structural Hamming Distance (SHD)* (Tsamardinos et al., 2006) that counts the number of add or remove an edge or reverse orientation operations to transform the PDAG of the estimated graph to the PDAG of the true graph (*lower is better*); (ii) *Adjacency AUC* ($\mathrm{AUC}_{\text{adj}}$), following Cheng et al. (2024); Stein et al. (2024): we rank edges in a graph according to a causal model's confidence and compute the AUC. Higher values are better; they indicate that the causal model about the edges present in the true graph.

Unoptimized MMD Can be a Misleading Metric of Discrimination: Table 1 shows the unoptimized MMD metric using the default parameters (Gaussian kernel) against the AUC_D for the corpus datasets, including 10 synthetic ones. Red fonts indicate situations where MMD is relatively low (< 0.3) but AUC_D significantly higher than 0.5. Standard deviation on the real datasets has been produced by repeating the experiment 10 times. Such low MMDs (0.21 to 0.24) led Cheng et al. (2024) to conclude that CausalTime has been validated "to generate realistic timeseries".

Evaluation of the Estimated Causal Structure: We evaluate ACT as a causal model selection method. We compare the SHD of ACT without the sparsity penalty (i.e., selecting the best performing model C_*, Eq. 1)), with the penalty (C_{ACT}, Eq. 2), and without using ACT at all, but choosing a TSCM at random out of all TSCMs considered. Results are shown on 19 synthetic datasets with 10 variables and a varying number of edges in their lagged graph. In Figure 5(Left) we show the

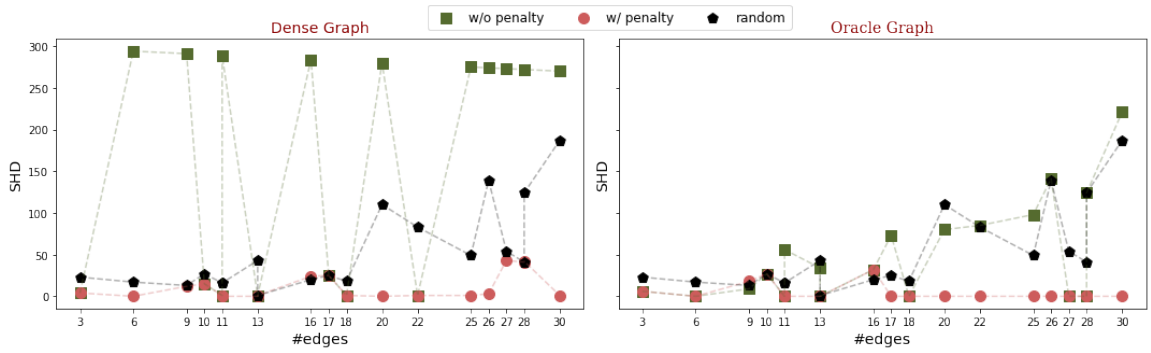


Figure 5: SHD of ACT with/without penalty and random selection of causal model on 19 synthetic datasets. **(a)** TSCM search space includes the fully connected graph. ACT with sparsity penalty avoids selecting the dense model in all cases. **(b)** TSCM search space includes the true causal graph. ACT with sparsity penalty selects it in 16/19 cases, compared to 7/19 without the penalty. ACT's selections are always better than random guessing.

SHD when we add to the TSCM search space $\mathcal{C}$ the fully connected graph (Dense Graph). Without the sparsity penalty, ACT overfits the graph selection (high SHD, green line) selecting the full graph 8 times, while ACT with the penalty avoids it in all cases. In Figure 5(Right) we add to the TSCM search space the true causal graph. We observe that ACT with the sparsity penalty selects it 16/19 cases, compared to 7/19 cases without the penalty. Overall, we conclude that *ACT (i) consistently avoids the trivial fully connected solution, (ii) selects the true graph causal graph if it is included in its search space, (iii) is always better than selecting a TSCM randomly within its search space*.

Comparison to Temporal Data Generation Methods for Observational Distributions We compare the efficacy of the methodologies in Section 3 against ACT in generating realistic datasets. At present, we consider datasets generated from the observational distribution, i.e., without simulating the effect of interventions. We compare against the authors' implementation of *CausalTime*, *CPAR* as implemented by *SDV* (Patki et al., 2016), and TimeVAE (Desai et al., 2021) available through the Synthcity (Qian et al., 2023) framework. Figure 6 shows the results. We observe that *ACT outperforms all other methodologies on the synthetic datasets, even though it attempts to solve a harder problem, i.e., fitting a generative causal model instead of a non-causal generative model*. On the corpus of real datasets, all methodologies fail to generate realistic datasets, with CausalTime obtaining better performance. Detailed results on the semi-synthetic datasets are omited for brevity; all methodologies are on par on the fMRI datasets (AUC_D in [0.71, 0.78]) and fail on the finance datasets (AUC_D in [0.78, 0.99]), with ACT and CausalTime being the winners in each corpus of semi-synthetic datasets, respectively. We note again, that despite its name, CausalTime does not produce a lagged causal graph, but a summary graph, thus solving an easier problem.

ML Efficacy : We evaluate the ML Efficacy of ACT and CausalTime. ML efficacy is measured with the AUC_{adj}, comparing the ranking (according to confidence) of discovered causal edges using the generated data against the one using the original data as the ground truth. A high ML efficacy score implies that one could share the generated data instead of the real ones, and still obtain similar results in terms of causal discovery. To learn the causal models, we use PCMCI with the default

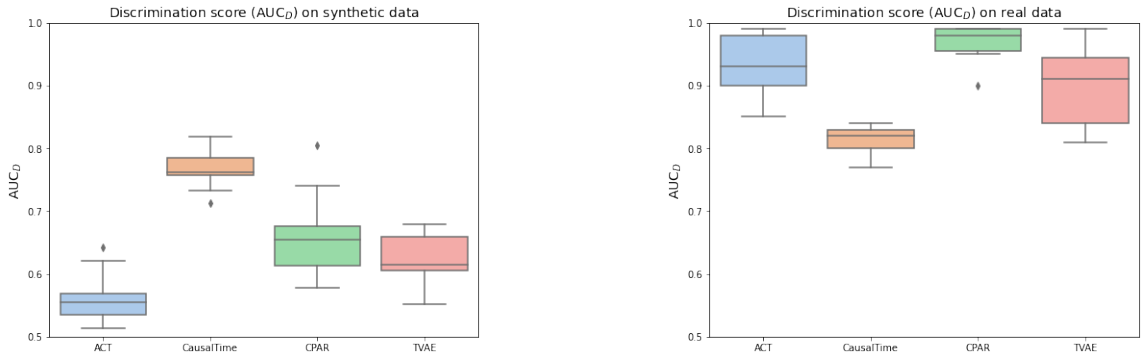


Figure 6: Discrimination performances of state-of-the-art on the corpora or synthetic (left) and real (right) datasets. ACT outperforms all benchmark methods on the synthetic datasets. All methods fail to capture the input data distribution on the real data. ACT is the only methodology that captures the data distribution by fitting a causal model, thus solving a harder problem.

parameters. The average AUC_{adj} on the real, semi-synthetic, and synthetic corpora is 0.66, 0.74, and 0.98 for ACT and 0.60, 0.72, and 0.88 for CausalTime, respectively. Overall, results indicate that when fitting a causal model is successful (low AUC_D), such as in the case of synthetic data, ML efficacy is achieved (high AUC_{adj}). Detailed results can be seen at appendix C.

Temporal Data Generation for Interventional Distributions: A key advantage of fitting a causal model is the ability to generate data under interventional conditions. We evaluate realistic interventional data generation on 29 synthetic causal models with 5-10 variables and up to 3 lags. ACT is applied to learn an optimal causal model from 1000 observational time-points. Interventional datasets of 1000 time-points are generated by using the true causal model and the estimated one by ACT. To simulate interventions, a time point within [1, 750] is selected, and an intervention is simulated by setting the value of one of the variables to its minimum or maximum in the observational data set. Hence, we simulate a hard intervention at a single time point that resembles a fault-injection. Figure 7 shows the optimized observational and interventional discrimination AUCs. In $19/29$ cases, the observational and interventional AUC_Ds differ by less than 0.1, indicating that interventional data under the estimated causal model resemble the corresponding intervention by the true model. In $4/29$ the AUC_D difference is in $[0.1, 0.2]$, and in $6/29$ larger than 0.2. The results indicate that in most cases, the TSCM learned by ACT could serve as a causal digital twin that correctly predicts the effect of experiments. We have explored the cases where the interventional discrimination is large and noticed that it could happen even when the causal structure is learned perfectly from the observational data. One explanation for the occasional large differences is that they unavoidable; specifically, interventional datasets may have different distributions even for the same model and intervention, depending on the initial conditions due to non-linear dynamics. We intend to further explore this issue. Finally, notice that a comparison with the other generative methodologies is impossible since none of them can simulate interventions; non-causal models could potentially simulate conditional distributions, but that would require non-trivial adaptations of related work.

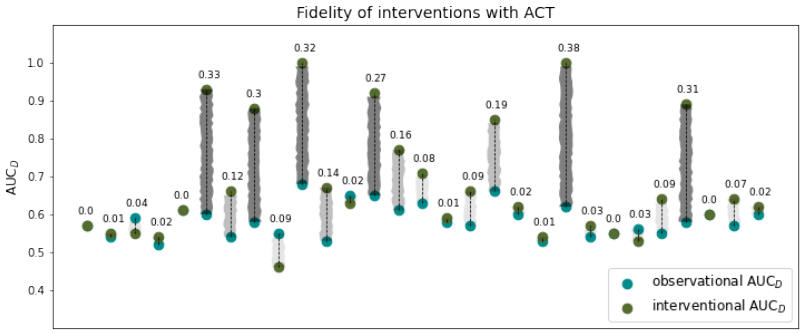


Figure 7: Comparison between the optimized AUC_D of the model selected by ACT on observational and interventional datasets, for 29 synthetic TSCMs. The vertical lines are labeled with the AUC_D differences. In 19/29 experiments, the differences are smaller than 0.1 (light gray shade). This indicates the TSCM learned from observational data, generates interventional data similar to the data produced by the true model.

7. Conclusions, Limitations, and Discussion

We introduce Adversarial Causal Tuning (ACT), which can be used for timeseries generation under either observational or interventional settings, creation of benchmark causal models, and causal model selection. Experimental results on real, synthetic, and semi-synthetic datasets show that penalizing for complexity is paramount in causal model selection. ACT outperforms the state-of-the-art of non-causal time series data generation on synthetic models, despite solving a harder problem. However, all proposed algorithms fail to successfully fit distributions of real time series data; this still remains an open research problem. ACT returns an optimized metric of goodness-of-fit that warns regarding unsuccessful fitting. In contrast, some prior works do not optimize the discrimination metric and are led to falsely conclude that successful fitting has been achieved.

ACT is currently limited to time series data under the assumptions of additive noise, causal sufficiency, and no contemporaneous causalities. To improve performance on real datasets, it may be necessary to remove these assumptions. This is not trivial: for example, fitting functional relationships in the presence of latent confounding variables is challenging. ACT can benefit from including better causal discovery algorithms in its search space, as it is shown that when it is presented with the true causal graph, it manages to successfully identify it among dozens of other alternatives. It can also benefit from employing better search techniques as Bayesian Optimization and meta-level learning adapted for minimax problems. Other improvements are possible: estimation of discrimination performance can be improved by using (repeated) cross-validation for small sample sizes. One could generate multiple simulated datasets S_C and take the average discrimination over all of them for better estimation, sacrificing computational efficiency. We note that the discrimination performance reported stems from a minimax procedure, which may introduce estimation bias. Techniques such as Bootstrap Bias Correction (BBC) (Tsamardinos et al., 2018) should be adopted to correct the returned AUCs for this bias. Discrimination pipelines could be extended by including other types of algorithms than classifiers, such as time series selection (Vareille et al., 2024), providing additional information regarding the variables that create the distributional differences.

8. Reproducibility Statement

The complete source code, including documentation, implementation, and benchmarking scripts, is publicly available at [removed for anonymous submission].

Acknowledgments

We would publicly like to thank Konstantina Biza and Étienne Vareille for their fruitful comments throughout this work.

References

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286 (5439), 1999.

Konstantina Biza, Ioannis Tsamardinos, and Sofia Triantafillou. Out-of-sample tuning for causal discovery. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4), 2022.

Konstantina Biza, Antonios Ntroumpogiannis, Sofia Triantafillou, and Ioannis Tsamardinos. Towards automated causal discovery: a case study on 5g telecommunication data. *arXiv preprint arXiv:2402.14481*, 2024.

Leo Breiman. Random forests. *Machine Learning*, 45, 2001.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 2020.

Richard B Buxton, Eric C Wong, and Lawrence R Frank. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magnetic Resonance in Medicine*, 39(6), 1998.

Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. Causaltime: Realistically generated time-series for benchmarking of causal discovery. In *International Conference on Learning Representations*, 2024.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2024.

Abhyuday Desai, Cynthia Freeman, Zuhui Wang, and Ian Beaver. Timevae: A variational autoencoder for multivariate time series generation. *arXiv preprint arXiv:2111.08095*, 2021.

David A Dickey and Wayne A Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a), 1979.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using RealNVP. *International Conference on Learning Representations*, 2017.

Sílvia Gonçalves and Dimitris Politis. Discussion: Bootstrap methods for dependent data: A review. *Journal of the Korean Statistical Society*, 40(4), 2011.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Advances in Neural Information Processing Systems*, 2014.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13, 2012.

Yannik Hahn, Tristan Langer, Richard Meyes, and Tobias Meisen. Time series dataset survey for forecasting with deep learning. *Forecasting*, 5(1), 2023.

Wolfgang Härdle, Joel Horowitz, and Jens-Peter Kreiss. Bootstrap methods for time series. *International Statistical Review*, 71(2), 2003.

M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Schölkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4), 1998.

S Hochreiter. Long short-term memory. *Neural Computation*, 1997.

Mohammadhossein Homaei, Mehran Tarif, Pablo Garcia Rodriguez, Andres Caro, and Mar Avila. Causal digital twins for cyber-physical security: A framework for robust anomaly detection in industrial control systems. *arXiv preprint arXiv:2510.09616*, 2025.

Yuxiao Huang and Samantha Kleinberg. Fast and accurate causal inference from time series data. In *The Twenty-Eighth International Flairs Conference*, 2015.

Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated Machine Learning: aMethods, Systems, Challenges*. Springer Nature, 2019.

Maowei Jiang, Pengyu Zeng, Kai Wang, Huan Liu, Wenbo Chen, and Haoran Liu. Fecam: Frequency enhanced channel attention mechanism for time series forecasting. *Advanced Engineering Informatics*, 58, 2023.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Kevin P Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

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

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. DYNOTEARS: Structure learning from time-series data. In *AISTATS*. PMLR, 2020.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *IEEE DSAA*. IEEE, 2016.

Judea Pearl. *Causality*. Cambridge University Press, 2nd edition, 2009.

Zhaozhi Qian, Bogdan-Constantin Cebere, and Mihaela van der Schaar. Synthcity: Facilitating innovative use cases of synthetic data in different data modalities, 2023.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*. PMLR, 2015.

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

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

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2), 2011.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2001.

Gideon Stein, Maha Shadaydeh, and Joachim Denzler. Embracing the black box: Heading towards foundation models for causal discovery from time series data. *AAAI AI4TS Workshop*, 2024.

Ioannis Tsamardinos, Laura Brown, and Constantin Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65:31–78, 10 2006. doi: 10.1007/s10994-006-6889-7.

Ioannis Tsamardinos, Elissavet Greasidou, and Giorgos Borboudakis. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Machine learning*, 107(12):1895–1922, 2018.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

Etienne Vareille, Michele Linardi, Ioannis Tsamardinos, and Vassilis Christophides. Chronoepilogi: Scalable time series selection with multiple solutions. *Advances in Neural Information Processing Systems*, 37:134287–134316, 2024.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 2021.

Junchen Ye, Weimiao Li, Zhixin Zhang, Tongyu Zhu, Leilei Sun, and Bowen Du. Mvts-library: An open library for deep multivariate time series forecasting. *Knowledge-Based Systems*, 283: 111170, 2024.

Alessio Zanga, Elif Ozkirimli, and Fabio Stella. A survey on causal discovery: Theory and practice. *International Journal of Approximate Reasoning*, 151:101–129, 2022. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2022.09.004.

Kevin Zhang, Neha Patki, and Kalyan Veeramachaneni. Sequential models in the synthetic data vault. *arXiv preprint arXiv:2207.14406*, 2022.

Appendix A. Search Spaces

Table 2: Search spaces for Temporal Causal Modeling (TSCM, left) and Discriminator (C2ST, right). Overall approximately **700 TSCMs** and **150** discriminators are searched over.

(a) TSCM Search Space

Phase	Method	Param.	Values
Phase 1	CP	Model size	`398M` (2 models)
	PCMCI	ℓ_{max}/iters	2, 3 / 10
	DYNOTEARS	ℓ_{max}	1, 3
Phase 2	RF	$n_{\text{estimators}}$	100, 500, 1000
	TCDF	Levels / ks / lr	0, 2 / 2, 3 / 1e–2, 1e–3
	TimesFM	Model	`timesfm-1.0-200m`
Phase 3	Gaussian	μ / σ	empirical
	Uniform	a / b	empirical
	Residuals	-	empirical
	Splines	lr / bins	1e-2 / 16
	RealNVP	lr	1e-4

(b) Discriminator Search Space

Method	Param.	Values
SVM	C	1.0, 0.75, 0.5, 0.25
	Kernel	linear, poly, rbf
	Degree	3
	Gamma	auto, scale
LSTM	Batch size	32, 64
	Hidden dim	128, 256
	Layers	2, 3
	Dropout	0.05, 0.1
	SeqLen	10, 20
	Epochs	10, 50
	LR	1e–4, 1e–3

Appendix B. Datasets

Table 3: Overview of Real Time-Series Datasets.

Dataset	Variables	Timesteps	Granularity	Start Date	Domain
ETTh1	7	17420	1 hour	07/01/2016	Power
ETTm1	7	69680	15 min	07/01/2016	Power
WTH	12	35064	1 hour	01/01/2020	Weather
Air_Quality	36	8760	1 hour	-	Weather
AirQualityUCI	12	9357	1 hour	03/10/2024	Weather
Bike-usage	5	552584	1 hour	01/01/2014	Transportation
Outdoors	3	1440	1 sec	21/02/2016	Environmental

Table 4: Overview of the Synthetic Time-Series Dataset Collection.

Collection	Num. of Datasets	Variables	Timesteps	Max Lag	Functional Relationships
C1	10	3-12	1000	1-3	Linear & Non-linear

Appendix C. ML Efficacy

Table 5: Overview of Semi-synthetic Time-Series Dataset Collections.

Collection	**Num. of Datasets**	**Variables**	**Timesteps**	**Max Lag**	**Functional Relationships**
fMRI	6	5	1200 - 5000	1	Non-linear
Finance	5	25	4000	1	Non-linear

Table 6: ML Efficacy: AUC_{adj} of the causal graphs derived from simulated data against the ground truth. A high AUC_{adj} score indicates the simulated data act similar to the original on CD tasks. In most cases *ACT* clearly outperforms *CausalTime* (see *ETTh1*, *ETTm1*, *BU*, *OD*, *fMRI* and *Synthetic*). There are 3 cases where *CausalTime* scores higher than *ACT* (*WTH*, *AQUCI* and *Finance*), but the difference is to the second decimal. Another observation is that both methodologies seem incapable of capturing the behavior of the original data for certain cases (*WTH*, *AQ*, *OD* and *Finance*), a realization that aligns with our remarks about the strain of current the current state-of-the-art to causally simulate real time-series.

Method	Real							Semi-synthetic		Synthetic
	WTH	AQUCI	AQ	ETTh1	ETTm1	BU	OD	fMRI	Finance	
ACT	0.58	0.72	0.56	**0.76**	**0.77**	**0.72**	**0.50**	**0.86**	0.61	**0.98**
CausalTime	**0.59**	**0.73**	0.56	0.59	0.68	0.58	0.48	0.77	**0.67**	0.88