Causal Structure Learning for Dynamical Systems with Theoretical Score Analysis

Nicholas Tagliapietra[1,2], **Katharina Ensinger**[1], **Christoph Zimmer**[3], **Osman Mian**[4]
[1]Bosch Center for Artificial Intelligence, Renningen, Germany
[2]Computer Science Department, TU Darmstadt, Germany
[3]Baden-Wuerttemberg Cooperative State University Mannheim, Germany
[4]Institute for AI in medicine IKIM, Germany
tagliapietra.nicholas@gmail.com

Abstract

Real world systems evolve in continuous-time according to their underlying causal relationships, yet their dynamics are often unknown. Existing approaches to learning such dynamics typically either discretize time —leading to poor performance on irregularly sampled data— or ignore the underlying causality. We propose CADYT , a novel method for causal discovery on dynamical systems addressing both these challenges. In contrast to state-of-the-art causal discovery methods that model the problem using discrete-time Dynamic Bayesian networks, our formulation is grounded in Difference-based causal models, which allow milder assumptions for modeling the continuous nature of the system. CADYT leverages exact Gaussian Process inference for modeling the continuous-time dynamics which is more aligned with the underlying dynamical process. We propose a practical instantiation that identifies the causal structure via a greedy search guided by the Algorithmic Markov Condition and Minimum Description Length principle. Our experiments show that CADYT outperforms state-of-the-art methods on both regularly and irregularly-sampled data, discovering causal networks closer to the true underlying dynamics.

1 Introduction

Real-world physical systems are fundamentally governed by continuous-time dynamics (Strogatz 2000) with intrinsic causal mechanisms. For instance, in a mass-spring system as shown in Figure n:1, the position of each mass (S_i) induces a force influencing its own velocity (V_i) and the velocities of other masses connected to it with a spring. The position of each mass, however, depends only on its own velocity. This example remarks the importance of incorporating the directionality of causal relationships to achieve physical plausiblity in learned models. While differential equations are the de facto standard for modeling dynamical systems, inferring them from data remains challenging. Moreover, when leveraging data-driven approaches such as neural networks or Gaussian processes to learn these models, there are rarely guarantees that the true underlying dynamics will include causality, and spurious correlations might be incorporated by accident. These challenges highlight the need for a causal discovery framework for learning systems in continuous time.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

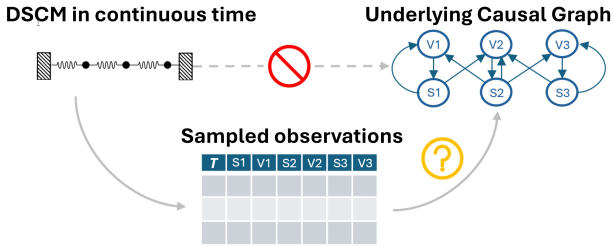

Figure 1: CADYT discovers the unknown causal structure (Top-right) using trajectories sampled from a continuous-time dynamical system (e.g. n-mass spring system). Our learned continuous-time model adapts to arbitrary timelines, including irregularly sampled ones.

Existing state-of-the-art for causal discovery on time-series focuses on learning the causal structure, but rarely captures the underlying continuous dynamics. Indeed, most methods learn a discretized version of the true dynamics (Hyvärinen et al. 2010; Peters, Janzing, and Schölkopf 2013; Runge 2020; Pamfil et al. 2020) while assuming regular sampling, and are not designed for irregularly-sampled data. Dynamic-systems modeling, on the other hand, solve this issue, and current approaches adapt to irregularly-sampled data by learning a continuous-time model (Chen et al. 2019; Hedge et al. 2022). These methods, however, ignore causality and do not guarantee the generalization capabilities of causal models.

In this work, we tackle both challenges: we propose a novel approach capable of performing causal discovery on dynamical systems in a continuous-time fashion, relaxing the regular sampling assumption. We review the conditions under which dynamics can be modeled with Dynamic Structural Causal Models (Mooij, Janzing, and Schölkopf 2013), and build our method around those conditions. Our approach, CADYT, leverages the Gaussian Process-based framework developed by Ensinger et al. (2024) to learn a continuous-time model of the dynamics. We incorporate those in our novel score, that leverages the Algorithmic Markov Condition (AMC) postulate (Janzing and Schölkopf 2010). AMC allows us to identify the causes of a target variable as the ones providing the *simplest* description, in terms of Kolmogorov complexity. Our score upper bounds the,

otherwise uncomputable, Kolmogorov complexity via the Minimum Description Length principle (Grünwald 2007). We then minimize this score via structure search for the underlying causal structure. Our contributions are as follows:

- We propose a novel approach for causal discovery in continuous-time dynamical systems that addresses both irregular sampling and causal structure identification—two challenges previously tackled in isolation.
- We leverage the Gaussian Process framework for exact inference in continuous time, enabling nonparametric modeling of system dynamics.
- We develop an end-to-end algorithm, **Ca**usal Discovery for **Dy**namic **T**imeseries (CADYT), combining Gaussian Process dynamical system modeling methods for continuous-time inference with structure search, enabling exact evaluation of dynamics while recovering true causal mechanisms.

2 Background

In this section we introduce the basic formalism for dynamical systems described by differential equations, and how they are typically learned. Next, we provide a causal interpretation of dynamical systems through dynamic structural causal models (DSCM). We end this section by explaining how we can use the Algorithmic framework of Janzing and Schölkopf (2010) to perform causal discovery over such dynamical systems defined using DSCMs.

Dynamical Systems

We consider a multivariate real-valued stochastic process $\mathbf{X}^{(t)} = \{X_1^{(t)}, \dots, X_D^{(t)}\}$ on a compact time interval, i.e. $t \in [0, T]$ and each $X_i^{(\cdot)} \in \mathbb{R}$. We characterize the evolution of the system's states through the framework of dynamical systems (Strogatz 2000), where its dynamics are governed by a defined set of rules that control state transitions. In the continuous-time setting relevant to our work, the dynamics are typically formalized using systems of *autonomous Ordinary Differential Equations* (ODE) of the form

$$\dot{\mathbf{X}}^{(t)} = F(\mathbf{X}^{(t)}) \quad \text{with } F : \mathbb{R}^D \to \mathbb{R}^D, \tag{1}$$

where $\dot{\mathbf{X}}^{(t)}$ is the time derivative of $\mathbf{X}^{(t)}$ representing its rate of change at time t, and $F = (F_0, \dots, F_D)$ is the vector field describing the system's dynamics. A trajectory of a dynamical system is the path traced starting from an initial condition $\mathbf{X}^{(0)}$ for which Eq.(1) holds. The system of ODEs induces causal dependencies between components of the dynamics (and/or trajectories) describing if and how each one influences another, called local dependency. Bellot, Branson, and van der Schaar (2022) provide the following formalization

Definition 1 (Local dependency)**.** *Two components X_i and X_j are locally dependent given any other processes iff X_i appears in the differential equation of X_j i.e. $|\partial_i F_j| \neq 0$.*

Essentially, a component X_j is locally dependent on component X_i if X_i directly influences X_j in Eq.(1), and independent otherwise. These dependencies entail an associated directed (potentially cyclic) graph, $\mathcal{G}$, where the components $X_i \in \mathbf{X}$ are the nodes and there is a directed edge from $X_i \to X_j$ if and only if X_j is locally dependent on X_i.

Dynamics Model Learning: In the field of Dynamics Model Learning, we aim at learning an approximation of the dynamics F from discrete-time trajectory data. Since observations are discrete while the underlying dynamics are continuous, we must first discretize the ODE using a numerical integrator. This discretization enables matching continuous dynamics to discrete observations. In this work, we use multi-step integrators (Hairer, Nørsett, and Wanner 2008) since they allow for exact GP inference. They approximate a point in the trajectory $X^{(n+s)}$ by using a weighted sum of the last $s-1$ points $\bar{\mathbf{X}}^{[n:n+s-1]} = \{\bar{\mathbf{X}}^{(n)}, \dots, \bar{\mathbf{X}}^{(n+s-1)}\}$ and are of the form

$$\sum_{j=0}^{s} a_{jn} \bar{X}^{(n+j)} = \sum_{j=0}^{s} b_{jn} F(\bar{X}^{(n+j)}), \tag{2}$$

where a_{jn} and b_{jn} are integrator-specific coefficients (e.g., Adams-Bashforth or Adams-Moulton methods). The number of stages is determined by s and often corresponds to the order of convergence (Hairer, Nørsett, and Wanner 2008).

Gaussian Processes Gaussian Processes (GPs) (Rasmussen and Williams 2005) are a class of probabilistic models which describe a random function $F : \mathbb{R}^D \to \mathbb{R}$. They are a generalization of multivariate normal distributions, and analogously, are fully described by their mean function $m(x)$ and covariance function $k_\theta(x, y)$, where θ denotes the trainable parameters such as lengthscales. Further, the covariance function $k(\cdot, \cdot) : \mathbb{R}^D \to \mathbb{R}^D$ can be a kernel such as the *Radial Basis Function* (RBF) kernel, or the *Polynomial* kernel. Here, we assume a zero-mean Gaussian process prior $F \sim \mathcal{N}(0, k_\theta(x, y))$. Further, a GP conditioned on a number of test points $\mathbf{X} = \{\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(N)}\}$ and their associated observations $\mathbf{Y} = \{Y^{(1)}, \dots, Y^{(N)}\}$ still follows (by definition) a multivariate Gaussian distribution, with mean $\mu(X^*)$ and variance $\Sigma(X^*)$ derived as

$$\mu(X^*) = k(X^*)^T (K + \lambda I)^{-1} Y, \tag{3}$$

$$\Sigma(X^*) = k(X^*, X^*) - k(X^*)^T (K + \lambda I)^{-1} k(X^*), \tag{4}$$

where $K \in \mathbb{R}^{N \times N}$ is the covariance matrix evaluated at points $X^{[1:N]}$ i.e. $K_{ij} = k(X^{(i)}, X^{(j)})$. This conditioning operation is called *Gaussian Process Regression* (GPR).

When applied to Dynamics Model Learning, different variants of GPR are used for approximating the dynamics function F in Eq.1 by conditioning the GP on measured trajectory points. Following Ensinger et al. (2024), multi-step integrators (Hairer, Nørsett, and Wanner 2008) enable exact GP inference and make it possible to evaluate the learned F conditioned on observations. To do so, Ensinger et al. (2024) derived kernels leveraging multi-step integrators as,

$$\begin{aligned} K(\bar{X}^{(n)}, \bar{X}^{(m)}) &= \mathbf{b}_n^\top \, k(\bar{X}^{[n:n+s]}, \bar{X}^{[m:m+s]}) \, \mathbf{b}_m, \\ k(\bar{X}^*) &= \mathbf{b}_n^\top \, k(\bar{X}^*, \bar{X}^{[n:n+s]}), \end{aligned} \tag{5}$$

where $\mathbf{b}_n \in \mathbb{R}^s$ is the n-th column of the matrix $B \in \mathbb{R}^{s \times N}$ containing the integration coefficients, and $k(\bar{X}^{[i:j]}, \bar{X}^{[k:l]})$

is a block matrix obtained after evaluating the chosen kernel between $\bar{X}^{[i:j]}$ and $\bar{X}^{[k:l]}$. In essence, given N (potentially irregularly sampled) trajectory points, $\mathbf{X}^{[t_1,...,t_N]}$ we can evaluate the dynamics at any point $\bar{X}^*$ by performing the GPR in Eq.(3),(4) using the kernels in Eq. (5)

$$F_D(\mathbf{X}^\star|\mathbf{X}^{t_1:t_N}) \sim \mathcal{N}(\mu_{post}(\mathbf{X}^\star), \Sigma_{post}(\mathbf{X}^\star)), \quad (6)$$

where the equations for mean and covariance depend on the chosen multi-step integrators. The advantage of using multi-step integrators within GPR is that, depending on their order, they permit a better representation of the continuous dynamics F and allow for evaluations of F at any test point $\mathbf{X}^*$.

Most dynamics model learning methods discussed above ignore the causal structure and fail to preserve local dependencies. This could lead to inaccurate predictions of a component X_j when an independent, unrelated component X_i undergoes a distribution shift. In the following we show how to equip these methods to take causal relations into account.

Dynamic Structural Causal Models

The system of ODEs shown in Eq.(1) induces different local dependencies (Def. 1) between stochastic processes. This dependency can be thought of as modularization of the system. The set of individual component trajectories is called modular if the admitted trajectories of the entire system can be detached into trajectories of individual components (Mooij, Janzing, and Schölkopf 2013). Under the dynamical stability assumption, which states that asymptotic dynamics of the system of ODE's converge to a unique element irrespective of the initial conditions, a system of ODE can be converted to a *Dynamic Structural Causal Model* (DSCM) (Rubenstein et al. 2016), defined as follows.

Definition 2 (Rubenstein et al. 2016)**.** *Let* DYN_i *be the trajectory for component* X_i*, and let* $\text{DYN} = \bigcup_{X_i \in \mathbf{X}} \text{DYN}_i$ *be a modular set of trajectories. A deterministic Dynamic Structural Causal Model (DSCM) on time indexed variables* $\mathbf{X}$ *taking values in* DYN *is a collection of equations.*

$$\mathcal{S} : \{X_i = F_i(Pa_i), \;\; X_i \in \mathbf{X}\} \,, \quad (7)$$

where $Pa_i \subseteq \mathbf{X} \backslash \{X_i\}$ *and each* F_i *is a map that outputs the trajectory of the effect variable* X_i *in terms of the trajectory of its direct causes* Pa_i*.*

In the above definition, the parents Pa_i should be interpreted as *direct causes* of X_i, and the function F_i as *causal mechanism* that maps the direct causes to the effect. Further, self-loops in the causal graph are allowed, although they do not explicitly appear in Def.2, which describes instead the stationary asymptotic behavior of the system. In essence in a DSCM, we explicitly decompose Eq. (1) into multiple simpler sub-equations, one for each component, as a direct consequence of local consistency. Rubenstein et al. (2016) proves that it is possible to derive a DSCM that allows us to reason about the asymptotic dynamics of the underlying ODE if the dynamic stability assumption holds, and impossible otherwise. Hence going forward, we assume dynamic stability. We adopt the instantaneous gradient assumption, meaning that causal relationships are encoded in the time derivatives of $\mathbf{X}$. This assumption is more aligned with the ODE structure and milder than the instantaneous-effect assumption, which requires instantaneous dependencies between variables. Consequently, our proposal is more aligned with Difference-based Causal Models (DBCM) (Voortman, Dash, and Druzdzel 2010) as opposed to the Dynamic Bayesian Networks.

Let trajectory $\mathcal{T} \in \mathbb{R}^{N \times D}$ be a sequence of observations of $\mathbf{X}$ sampled from a DSCM over time-steps $\{t_1, ..., t_N\}$ and denoted by $\mathcal{T} = \{\mathbf{X}^{(1)}, ..., \mathbf{X}^{(N)}\}$, we aim to find the underlying directed graph $\mathcal{G}$ of process interactions entailed by the DSCM and learn the dynamics of each variable using only its causal parents. Doing so is impossible unless we make assumptions on how $\mathcal{T}$ was generated (Pearl 2009). To that end, we assume causal sufficiency, i.e. for each component X_i there are no unobserved components that influence X_i. In addition we assume $\mathcal{G}$ is μ-Markovian with respect to $\mathcal{T}$ which implies that local independencies of the underlying DSCM are reflected in $\mathcal{T}$. Conversly we assume $\mathcal{T}$ is μ-faithful with respect to $\mathcal{G}$ which implies that all local independencies that we find in $\mathcal{T}$ also hold in $\mathcal{G}$. Together these assumptions ensure that independence statements derived from $\mathcal{T}$ can be interpreted as absence of edges in $\mathcal{G}$, thereby letting us deduce local independencies (Mogensen and Hansen 2020). Let ϕ_{max} be the highest frequency present among all the components within $\mathcal{S}$ and define $\Delta_{max} = \max(t_{i+1} - t_i)\; for\; 1 \leq i < N$, we assume that $\Delta_{max} \in (0, \frac{1}{2\phi_{max}})$, meaning that $\mathcal{T}$ has been sampled at a rate finer than its critical frequency. Next, we explain how to learn the underlying causal structure entailed by a DSCM.

Information Theoretic Causal Discovery Information theoretic causal discovery relies on the algorithmic Markov condition (AMC) (Janzing and Schölkopf 2010), and is grounded in Kolmogorov complexity. The Kolmogorov complexity of a binary string x is the length of the shortest binary program p^* that outputs x and halts on a universal Turing machine U (Kolmogorov 1965; Vereshchagin and Vitányi 2004). For a probability distribution P, this complexity $K(P)$ is the length of the shortest program that outputs $P(x)$ to within precision q on input $\langle x, q \rangle$. Formally,

$$K(P) = \min_{p \in \{0,1\}^*} \left\{ ||p|| : |U(p, x, q) - P(x)| \leq \frac{1}{q} \right\}$$

The AMC states that a graph $\mathcal{G}$ over $\mathbf{X}$, with joint distribution P, is an admissible causal graph only if the shortest description of P factorizes as $K(P(X_1, \ldots, X_D)) = \sum_{j=1}^{D} K(P(X_j \mid Pa_j)) + \mathcal{O}(1)$. Thus, the true causal graph minimizes Kolmogorov complexity i.e. each Pa_j provides the tersest description of its causal child.

While Kolmogorov complexity is not computable due to the halting problem, it can be bounded from above in a statistically well-founded way using the Minimum Description Length (MDL) principle (Grunwald 2004). Marx and Vreeken (2021) show that for sample sizes approaching infinity, the true causal graph can be identified by minimizing an appropriate lossless MDL-score.

Given a model class $\mathcal{M}$, the MDL principle selects the optimal model $M \in \mathcal{M}$ for data D by minimizing the total description length: $L(D, M) = L(M) + L(D \mid M)$, where $L(M)$ represents the number of bits required to describe the model M, and $L(D \mid M)$ is the number of bits needed to describe the data D once the model M is known. Existing methods have already used this algorithmic model to discover causal graphs for non-timeseries data with reasonable success (Kaltenpoth and Vreeken 2019; Mian, Marx, and Vreeken 2021; Mameche, Kaltenpoth, and Vreeken 2023).

3 MDL for Dynamical Systems

To use information theoretic causal discovery with dynamical systems, we need to build a suitable MDL score for time-series trajectories sampled from a DSCM. We will do so for the well known case of Additive Noise Models (Hoyer et al. 2009) and assume that we have

$$X_i^{(t)} = \hat{X}_i^{(t)} + \nu_i^{(t)} \ , \tag{8}$$

with $\hat{X}_i^{(t)}$ following the dynamical system defined in Eq. (1) for $t \in [0, T]$ and $\nu_i^{(t)} \sim \mathcal{N}(0, \sigma_i^2)$ being independent gaussian noise terms such that $\nu_i \perp\!\!\!\perp X_i \ \forall i$ and $\nu_i \perp\!\!\!\perp \nu_j \ \forall i, j$. This setup can be interpreted as a noisy observation model in dynamical systems, where a deterministic process is perturbed by independent additive noise. Given a trajectory $\mathcal{T}$ of size N for (possibly irregular) time steps $\{t_1, ..., t_N\}$, we want to find the model such that

$$M^* = \operatorname*{argmin}_{M \in \mathcal{M}} \ L(\mathcal{T}, M), \tag{9}$$

$$= \operatorname*{argmin}_{M \in \mathcal{M}} \left(L(M) + \sum_{i=1}^{D} L(X_i^{[t_1:t_N]} \,|\, Pa_i, F_i) \right), \tag{10}$$

$$= \operatorname*{argmin}_{M \in \mathcal{M}} \left(L(M) + \sum_{i=1}^{D} L(\nu_i) \right) \ , \tag{11}$$

where we use the notation $X_i^{[a:b]}$ to denote the values for X_i from time-steps a to b included. The summation term in Eq.(10) follows from Eq.(7) and measures compression of each component given its causal parents and the parametrization enforced by M. We simplify this in Eq.(11) to show that encoding a component given the model reduces to encoding the noise terms. To be able to use MDL as a practical stand-in for Kolmogorov complexity inside AMC, we need to define a model class and an encoding scheme measuring the complexity of the class resp. data under that model class. This we do next.

Encoding the Model The model cost $L(M)$ consists of a global cost $L_{global}(M)$ plus the sum of the local model costs for each individual variable X_i, i.e. $\sum_{i=1}^{D} L(M_i)$. For a pre-specified integrator of step-size s, the global cost measures the complexity of storing the initial s samples of a given trajectory $\mathcal{T}$. Formally,

$$L_{global}(M) = \log N + r_d \cdot D \cdot s \ , \tag{12}$$

where we encode the integrator stepsize s using $\log N$ bits and first s samples of the trajectory. We assign a fixed cost of r_d bits to each component value $X_i^{(t)} \in \mathcal{T}^{[t_1:t_s]}$.

For each X_i, we define a local model M_i where we store its causal parents and the parameters of the structural equation F_i in Eq. (7). We encode M_i as

$$L(M_i) = L_{\mathbb{N}}(||Pa_i||) + ||Pa_i|| \log D + L_F(F_i) \ , \tag{13}$$

where $L_{\mathbb{N}}$ encodes the number of parents using the MDL-optimal encoding for integers $z \geq 0$ (Rissanen 1983). It is defined as $L_{\mathbb{N}}(z) = \log^* z + \log c_0$, where $\log^* z = \log z + \log\log z + \ldots$ and only positive terms are considered. Further, c_0 is a normalization constant to ensure the Kraft-inequality holds (Kraft 1949). Next, we identify those $||Pa_i||$ variables and encode the function F_i over them.

Encoding the Functions To compute each local model M_i, we learn a continuous dynamics function F_i for each X_i by using the GPR scheme developed by Ensinger et al. (2024). We do so due to two main reasons namely 1) they offer a natural fit for modeling systems of ODEs by learning a continuous model and not a discrete one and are therefore capable of handling irregularly sampled trajectories and, 2) their non-parametric modeling-nature saves us from imposing parametric assumptions on F_i. Each model M_i regresses the dynamics related to variable X_i from its causal parents Pa_i. In essence, we learn a GP as defined in Eq.(6) using the kernels in Eq.(5), yielding an estimator in the form

$$F_i(X_i^{t_\star} | Pa_i^{[t_1:t_\star)}) \sim \mathcal{N}\big(\mu_{post}(X_i^{t_\star}), \Sigma_{post}(X_i^{t_\star})\big), \tag{14}$$

where we can estimate the dynamics of a X_i at an arbitrary point $t_\star$ using the history of $Pa_i^{[t_1:t_\star)}$ up to this point. Once a GP is trained, $X_i^{(t)}$ can be estimated by numerically integrating the dynamics $F_i(\cdot)$ learned by the GP over an arbitrary timeline. We formally score $F_i(\cdot)$ as,

$$L_F(F_i) = \log\left(\frac{1}{r_\lambda}\right)\frac{N(N-1)}{2} + L_\phi([\boldsymbol{\alpha_i}, \boldsymbol{\beta_i}, \boldsymbol{\Lambda_i}]) \ . \tag{15}$$

The components of F_i comprise the kernel matrix K_i and the corresponding length-scale and noise-variance parameters $\boldsymbol{\alpha_i}$ and $\boldsymbol{\beta_i}$, computed from Pa_i. Since the integrator coefficients are deterministically obtained from $\boldsymbol{\alpha_i}$, $\boldsymbol{\beta_i}$, K_i, and the initial trajectory in Eq. (12), they need not be stored. To store K_i efficiently, we apply Singular Value Decomposition $K_i = V_i \Lambda_i V_i^\top$, where V_i and Λ_i denote the orthonormal eigenvector and diagonal eigenvalue matrices, respectively. The orthonormal matrix $V_i \in \mathbb{R}^{N \times N}$ can be represented using at most $N(N-1)/2$ rotation angles at a predefined precision r_λ. Finally, the length scales $\boldsymbol{\alpha_i}$, variances $\boldsymbol{\beta_i}$, and eigenvalues in $\boldsymbol{\Lambda_i}$ are encoded using L_p.

Encoding the Parameters To encode the length-scale resp. eigenvalues obtained from the SVD, we use the score proposed by Marx and Vreeken (2019) for encoding parameters up to a user-specified precision p. We have

$$L_p(\theta) = 2||\theta|| + \sum_{i=1}^{||\theta||} L_{\mathbb{N}}(|\rho_i|) + L_{\mathbb{N}}(\lceil \theta_i \cdot 10^{\rho_i} \rceil) \ , \tag{16}$$

with ρ_i being the smallest integer such that $|\theta_i| \cdot 10^{\rho_i} \geq 10^p$. To simplify, $p = 2$ implies that we consider first two digits of the parameter. We need two bits to store the signs of ρ_i and the parameter, then we encode the shift ρ_i and the shifted parameter θ_i.

Encoding Data Given Model As a final step in constructing a lossless score, we encode the residual noise that remains after our the model has encoded the underlying causal structure and data-generating process. As our goal is to minimize the variance of the residuals across the timeseries trajectory, we encode each ν_i as zero mean Gaussian noise using the encoding provided by Grunwald (2004), formally

$$L(\nu_i) = \frac{N}{2}\left(\frac{1}{\ln 2} + \log 2\pi\hat{\sigma}_i^2\right) , \tag{17}$$

where $\hat{\sigma}_i^2$ is the empirical estimate of residual noise variance. Combining the above, we obtain a lossless MDL score for a timeseries trajectory modeled using a causal graph.

Theoretical Analysis

While lack of computability of Kolmogorov complexity impedes us from directly providing identifiability guarantees using the Algorithmic Markov Condition (AMC), we can still independently prove that our score acts as a valid regularized log-likelihood score with an upperbound asymptotically similar to the BIC score (Schwarz 1978). Doing so however necessitates two additional assumptions.

Assumption 1 (Finite dimensions)**.** *K is finite-dimensional.*

Assumption 2 (Bounded hyperparameters and precision)**.** *The length scale parameters α_i, the variance parameters β_i are upper bounded. All precisions $|\rho_i|$ are upper bounded.*

Kernel classes satisfying Assm. (1) include, Polynomial kernels, Wendland kernels (Wendland 1995), Buhmann kernels (Martin, Buhmann, and Ablowitz 2003), Truncated resp. Random Fourier kernels (Rahimi and Recht 2007), and Nyström Approximated kernels (Williams and Seeger 2000). Intuitively, Assms. 1 and 2 ensure that the cost of storing the eigenvalues in Eq. (15) scales sub-linearly with the number of trajectory points, which is necessary to provide theoretical guarantees. Let $C = \log\left(\frac{1}{r_\lambda}\right) \cdot \frac{N(N-1)}{2} + 2||\theta|| + \frac{N}{2}\left(\frac{1}{\ln 2} + \log(2\pi)\right)$, we show the following.

Lemma 1. *Given a DSCM $\mathcal{S}$, let $\mathcal{T}$ be a trajectory generated from $\mathcal{S}$ and let $\bar{L}(\mathcal{T}, M_j) = L(M_j) + L(\nu_j) - C$. If Assms. 1,2 hold, it holds asymptotically*

$$\bar{L}(\mathcal{T}, M_j) \leq c_0 \cdot N \cdot log(\hat{\sigma}_j^2) + c_1^{(j)} log(N) + c_2^{(j)}.$$

with constants $c_0, c_1^{(j)}, c_2^{(j)}$ independent of N.

Theorem 2. *Given a DSCM $\mathcal{S}$, let $\mathcal{T}$ be a trajectory generated from $\mathcal{S}$ and let $\bar{L}(\mathcal{T}, M) = \sum_{i=1}^{D} \bar{L}(\mathcal{T}, M_i)$. If Assms. 1,2 hold, it holds asymptotically.*

$$\bar{L}(\mathcal{T}, M) \leq c_0 \cdot N \cdot log(\hat{\sigma}^2) + c_1 log(N) + c_2.$$

with constants c_0, c_1, and c_2 independent of N and $\bar{L}(\mathcal{T}, M)$ asymptotically is a valid regularized log-likelihood score according to Definition A.2.

Corollary 3. *Model selection using $L(\mathcal{T}, M)$ is equivalent to model selection using $\bar{L}(\mathcal{T}, M)$.*

We provide the proof in Appendix A. Intuitively, our proof shows that the upper bound of $L(\mathcal{T}, M)$ behaves like sum of component-wise regularized log-likelihood scores (such as the BIC). These guarantees, however, only hold if we score all graphs over $\mathcal{T}$. This is an intractable brute-force approach because the search space grows super-exponentially in the number of variables. To nevertheless have a practical instantiation for minimizing $L(\mathcal{T}, M)$, we use a general greedy structure search algorithm which we describe next.

4 The CADYT Algorithm

We present the score-based method **Ca**usal Discovery for **Dy**namic **T**imeseries (CADYT) for discovering causal graphs of multivariate continuous-valued dynamical systems. We incorporate our proposed score into a common three-step search procedure (Mian, Marx, and Vreeken 2021; Mameche, Kaltenpoth, and Vreeken 2023) namely, edge scoring, forward and backward search. This is the next best alternative to exhaustive search for our case. The well known Greedy Equivalence Search (Chickering 2002) is not built for timeseries and methods for topological search (Wang et al. 2017; Xu, Mameche, and Vreeken 2025) do not directly apply to cyclic systems. We provide full pseudocode in Appendix E.

We start with the edge-ranking phase that computes the gain for all pairwise causal connections. The gain Γ_{ij}, between each pair X_i and X_j, is given by

$$\Gamma_{ij} = L(\mathcal{T}, M) - L(\mathcal{T}, M_{\oplus ij}) , \tag{18}$$

where $M_{\oplus ij}$ implies model M with edge $X_i \to X_j$ included. Intuitively, the higher the Γ_{ij}, the more confident we are that this is the correct causal edge. The edge scoring phase returns a priority queue of tuples $(\Gamma_{ij}, (X_i, X_j))$ ordered by decreasing gain.

The forward search iteratively adds the highest-scoring edge from the priority queue. After adding an edge $X_i \to X_j$, all incoming edges to X_j are re-evaluated using Eq. (18) and updated in the queue. Before inclusion, each edge is tested for statistical significance using the no-hypercompression inequality (Grünwald 2007). The search terminates when no further additions improve the score. The subsequent backward phase prunes redundant edges by removing any whose deletion increases $L(\mathcal{T}, M)$, continuing until no such edges remain. We then return the final graph.

CADYT has overall computational complexity of $\mathcal{O}(N^3 D^3 \log D)$ where N^3 derives from our choice of non-parametric regression functions (GPs) for L_F. In Appendix C we give a detailed derivation on the computational complexity and show how it is at least on-par with existing methods. CADYT , moreover, can be inherently parallelized, and we implement it as such, resulting in a fast runtime.

5 Related Work

GPs for time-series modeling Most works in this field address discrete-time dynamics (Deisenroth and Rasmussen 2011; Wang, Fleet, and Hertzmann 2005) while we aim to learn GP dynamics models in continuous time. However,

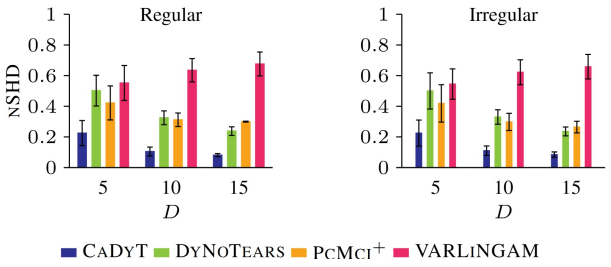

Figure 2: [Lower is better, NSHD] for random graphs of sizes $D \in \{5, 10, 15\}$ for regularly sampled data (left) and irregularly sampled data (right). CaDyT finds graphs closer to the ground truth resulting in a lower NSHD.

there is also work that addresses continuous-time modeling (Heinonen et al. 2018; Hedge et al. 2022; Ridderbusch, Ober-Blöbaum, and Goulart 2023). Glass, Ensinger, and Zimmer (2024) apply the inference scheme (Hedge et al. 2022) to active learning. We leverage the method proposed by Ensinger et al. (2024) due to the beneficial properties of exact inference even under irregular sampling. None of the mentioned approaches, as opposed to our work, can discover the underlying causality in dynamical systems.

Causal models for time-series Causal discovery from time-series data has received active research interest over the past decade. Early methods focused on Granger-causality (G-causality) (Granger 1969), which implies that X_i G-causes X_j if including past of X_i helps in predicting the present of X_j. Methods have been designed for both linear (Geweke 1982; Barrett, Barnett, and Seth 2010) and non-linear (Nauta, Bucur, and Seifert 2019) causal models. Methods that are not explicitly based on G-causality lie in the category of constraint-based methods (Chu, Glymour, and Ridgeway 2008; Sun, Taylor, and Bollt 2015; Runge 2020), score-based methods (Pamfil et al. 2020) or noise-based methods (Hyvärinen et al. 2010; Peters, Janzing, and Schölkopf 2013). These methods assume regularly sampled discrete trajectories and overlook the continuous-time structure of time series and the conditions required to reliably learn the underlying dynamics.

Voortman, Dash, and Druzdzel (2010) were among the first ones to study under which conditions dynamical systems allow for a causal interpretation, and proposed Difference-based causal models (DBCM). DBCM differ from well known dynamic Bayesian networks in that they force all causation to go through derivatives. This idea was extended by Mooij, Janzing, and Schölkopf (2013) to show that (D)SCMs can model the asymptotic behavior of systems of ODEs with limitations on the possible interventions under DBCM. Those limitations are addressed in recent works (Blom, Bongers, and Mooij 2020; Cinquini et al. 2025).

Our proposal aims to model a DBCM rather than a dynamic Bayesian network. In contrast to the existing work, our work is uniquely positioned at the intersection of dynamical system learning and causal discovery as it provides a theory-backed approach to modeling the former, while allowing for learning the laws of the process via the latter.

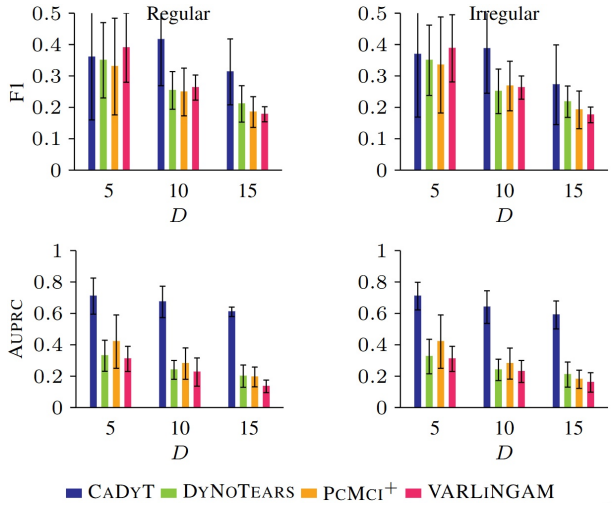

Figure 3: [Higher is better, F1 score (top) and AUPRC (bottom)] for random graphs of sizes $D \in \{5, 10, 15\}$ for regularly (left) resp. irregularly sampled data (right). CaDyT improves over baselines in terms of F1. CaDyT having high AUPRC indicates higher confidence about the correct causal edges as opposed to spurious ones.

6 Experiments

Setup We instantiate CaDyT using parallelized greedy search. We perform our experiments with GPs with RBF-Kernel leveraging explicit Adams-Bashforth (AB) integrators of order $s \in \{1, 2, 3\}$. Even though Thm. 1 applies to finite-dimensional Kernels, our choice is motivated by RBF-Kernel's exact-inference capabilities. Even with over-regularization that could result due to the use of RBF-Kernel, we find that we still outperform the competition. We include the results for CaDyT using Polynomial kernels in the supplementary material. We compare CaDyT with a variety of baselines: The constraint-based PCMCI$^+$ (Runge 2020) using non-parametric Kernelized independence test, the score-based DyNoTears (Pamfil et al. 2020), and the noise-based VARLiNGAM (Hyvärinen et al. 2010).

We generate synthetic data using Diamond structure (4 variables) and Erdős–Rényi random graphs with and without cycles for $D \in \{5, 10, 15\}$ for both regular and irregular timelines. To evaluate the predicted structures we measure the *Structural Hamming Distance* (SHD) (Tsamardinos, Brown, and Aliferis 2006) which counts the edge mismatch in true and predicted structures. For comparability across structures of different sizes, we normalize SHD between 0 and 1 by dividing with D^2 and call this NSHD. To evaluate precision and recall over predicted edges we use the F1 metric, and use *Area Under Precision Recall Curve* (AUPRC) to assess how correct each method is on the edges it is most confident about. We repeat all experiments 20 times with different seeds and report the mean. We report results for the more challenging setting of Erdős–Rényi graphs with both regular and irregular sampling in the manuscript and postpone full experimental details to Appendix B.

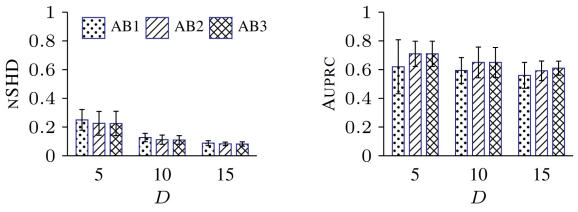

Figure 4: [NSHD (left) Lower is better, AUPRC (right) Higher is better] for Adams-Bashforth integrator of order $s \in \{1, 2, 3\}$ for irregularly sampled data. CADYT shows gradual improvement with higher order integrators.

Results We perform a sanity check and assess the robustness of CADYT to false-positives. We generated 10 graphs made of 4 independent ODEs. CADYT with AB3 never discovered a single spurious edge. Surprisingly, baselines report causal edges for independent data. VARLINGAM found spurious edges 30% of times, whereas PCMCI$^+$ and DYNOTEARS 60% and 100% respectively.

Next we sample random dynamical systems and evaluate how well the methods can discover the underlying causal structure. We sample cyclic and acyclic dynamical systems with equal probability and report the results. We report NSHD in Figure n: 2 where we observe that CADYT consistently outperforms all baselines by a clear margin. Both DYNOTEARS and PCMCI$^+$ demonstrate a high false positive rate which worsens the NSHD. Overall, all methods worsen slightly on irregularly-sampled data, but CADYT still remains best by a visible margin.

To evaluate how well we perform in-terms of precision and recall, we report the F1 score in Figure n: 3. We see that CADYT is on-par with the baselines for variable sizes 5, and continues to be robust to false-positives by maintaining a high precision as network size increases. Competing methods on the other hand have very low precision as they tend to predict spurious edges quite frequently.

While NSHD and F1 score could give us a summarized picture of how well the methods perform, we are also interested in how correct are the methods on the causal relationships that they are most confident about. To that end, we calculate the AUPRC metric by ordering the predicted edges of each algorithm in decreasing confidence. For CADYT this confidence is computed using Eq. (18), For PCMCI$^+$ we use the p-values associated with each edge whereas for DYNOTEARS resp. VARLINGAM we use the strength of the causal edge as present in the predicted adjacency matrix. Looking at the results in Figure n: 3 we see that CADYT again performs well across benchmarks. A high AUPRC indicates that CADYT is mostly correct about high-confidence edges.

Effect of Integration Order To study the effect of the integration scheme, we conduct an ablation study whose result we report in Figure n:4. We compare integrators on irregular timelines and see that higher-order integrators (AB2/AB3) generally outperform AB1. This is consistent with the domain knowledge that higher-order integration schemes approximate the underlying continuous-time dynamics better.

Method	DblMass	DblLinear	Rössler
DYNOTEARS	0.22	0.59	0.34
PCMCI$^+$	0.24	0.30	0.22
VARLINGAM	0.39	0.44	0.30
CADYT (ours)	**0.79**	**0.79**	**0.55**

Table 1: [Higher is better, AUPRC] for the simulated 2-mass spring system (DblMass), for the real double-linear system (DblLinear), and for the Rössler Oscillator (Rössler).

We observed that addition of cycles into the data-generating process affects the methods differently. We find CADYT tends to improve with the order of integration, where versions leveraging higher-order schemes (AB2 and AB3) stay robust or even improve, whereas the lower-order AB1 variant deteriorates.

Oscillators and Chaotic Systems We further test on the simulated 2-mass spring system, its analogous real-world counterpart (Schmidt and Lipson 2009), and the hyper-chaotic Rössler Oscillator. To stay fair to the baselines, we use regular sampling. We report results averaged over 10 runs in Table n: 1 which shows that CADYT outperforms the competing methods and finds causal structures closer to the underlying dynamics (lower SHD), while denoting high confidence about true edges (high AUPRC).

7 Discussion and Conclusions

We proposed CADYT , a method for uncovering causal structure in continuous-time dynamics from discrete trajectory data. Our approach combines the exact inference framework of Ensinger et al. (2024) with the Algorithmic Markov Condition (Janzing and Schölkopf 2010). We proved that our score is a valid regularized log-likelihood score (Def.A.2) with an upper-bound asymptotically similar to the BIC, and demonstrated empirically that our score outperforms existing methods. Going forward we see potential lines of improvements as future work.

First, we used greedy graph search and the Adams-Bashforth integrators as the two main components for CADYT. We do not, however, claim that these to be optimal choices. It is possible that alternate search strategies resp. integrators yield better performance and we aim to investigate such alternative approaches. Second, while using the proposal of Ensinger et al. (2024) allowed us to naturally work with both regular and irregular-sampled trajectories, it has been known to be sensitive to high amount of noise. We conjecture that we could further improve the results by evolving the machinery to be robust to noise. Third, we could extend CADYT to relax the causal sufficiency assumption by searching for instantaneous edges such that both edge directions give a high score, post backward search. This could potentially point to a hidden confounder. Last, currently CADYT assumes ODEs can be well-modeled by a GP regression under noisy observation model. A recently proposed extension of DSCMS to chaotic models (Boeken and Mooij 2024) could allow us to relax this assumption further and is a potentially rewarding line of future work.

Acknowledgments

Nicholas Tagliapietra and Katharina Ensinger are supported by Robert Bosch GmbH. Osman Mian is supported by the German Federal Ministry of Research, Technology and Space (DECIPHER-M, 01KD2420C).

References

Barrett, A. B.; Barnett, L.; and Seth, A. K. 2010. Multivariate Granger causality and generalized variance. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 81(4).

Bellot, A.; Branson, K.; and van der Schaar, M. 2022. Neural graphical modelling in continuous-time: consistency guarantees and algorithms. arXiv:2105.02522.

Blom, T.; Bongers, S.; and Mooij, J. M. 2020. Beyond Structural Causal Models: Causal Constraints Models. In Adams, R. P.; and Gogate, V., eds., *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *PMLR*, 585–594. PMLR.

Boeken, P.; and Mooij, J. M. 2024. Dynamic Structural Causal Models. arXiv:2406.01161.

Chen, R. T. Q.; Rubanova, Y.; Bettencourt, J.; and Duvenaud, D. 2019. Neural Ordinary Differential Equations. arXiv:1806.07366.

Chickering, D. M. 2002. Optimal structure identification with greedy search. *JMLR*, 3.

Chu, T.; Glymour, C.; and Ridgeway, G. 2008. Search for Additive Nonlinear Time Series Causal Models. *Journal of Machine Learning Research*, 9(5).

Cinquini, M.; Beretta, I.; Ruggieri, S.; and Valera, I. 2025. A Practical Approach to Causal Inference over Time. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39: 14832–14839.

Deisenroth, M. P.; and Rasmussen, C. E. 2011. PILCO: a model-based and data-efficient approach to policy search. In *Proceedings of the 28th ICML*, ICML'11. Omnipress.

Ensinger, K.; Tagliapietra, N.; Ziesche, S.; and Trimpe, S. 2024. Exact inference for continuous-time Gaussian process dynamics. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Geweke, J. 1982. Measurement of linear dependence and feedback between multiple time series. *Journal of the American statistical association*, 77(378).

Glass, L.; Ensinger, K.; and Zimmer, C. 2024. Safe Active Learning for Gaussian Differential Equations. arXiv:2412.09053.

Granger, C. W. J. 1969. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3): 424–438.

Grunwald, P. 2004. A tutorial introduction to the minimum description length principle. arXiv:math/0406077.

Grünwald, P. D. 2007. *The minimum description length principle*. MIT press.

Hairer, E.; Nørsett, S.; and Wanner, G. 2008. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer Series in Computational Mathematics. Springer Berlin Heidelberg. ISBN 9783540566700.

Hedge, P.; Yildiz, C.; Lähdesmäki, H.; Kaski, S.; and Heinonen, M. 2022. Variational multiple shooting for Bayesian ODEs with Gaussian processes. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022), PMLR*, Proceedings of Machine Learning Research, 790–799. United States: JMLR. Conference on Uncertainty in Artificial Intelligence, UAI ; Conference date: 01-08-2022 Through 05-08-2022.

Heinonen, M.; Yildiz, C.; Mannerström, H.; Intosalmi, J.; and Lähdesmäki, H. 2018. Learning unknown ODE models with Gaussian processes. In *Proceedings of the 35th ICML, ICML 2018*, volume 5 of *PMLR*, 3120–3132. United States: International Machine Learning Society.

Hoyer, P.; Janzing, D.; Mooij, J. M.; Peters, J.; and Schölkopf, B. 2009. Nonlinear causal discovery with additive noise models. In *NeurIPS*, volume 21. Curran.

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).

Janzing, D.; and Schölkopf, B. 2010. Causal inference using the Algorithmic Markov Condition. *IEEETPAMI*, 56: 5168–5194.

Kaltenpoth, D.; and Vreeken, J. 2019. We Are Not Your Real Parents: Telling Causal from Confounded by MDL. In *SDM*. SIAM.

Kolmogorov, A. N. 1965. Three approaches to the quantitative definition ofinformation'. *Problems of information transmission*, 1(1).

Kraft, L. G. 1949. *A device for quantizing, grouping, and coding amplitude-modulated pulses*. Ph.D. thesis, Massachusetts Institute of Technology.

Mameche, S.; Kaltenpoth, D.; and Vreeken, J. 2023. Learning Causal Models under Independent Changes. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 75595–75622. Curran Associates, Inc.

Martin, B.; Buhmann, M.; and Ablowitz, J. 2003. Radial basis functions: theory and implementations. *Cambridge University (ISBN: 0-521-63338-9.)*.

Marx, A.; and Vreeken, J. 2019. Telling cause from effect by local and global regression. *KAIS*, 60(3): 1277–1305.

Marx, A.; and Vreeken, J. 2021. Formally Justifying MDL-based Inference of Cause and Effect. arXiv:2105.01902.

Mian, O.; Marx, A.; and Vreeken, J. 2021. Discovering Fully Oriented Causal Networks. In *AAAI*.

Mogensen, S. W.; and Hansen, N. R. 2020. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, 48(1).

Mooij, J. M.; Janzing, D.; and Schölkopf, B. 2013. From ordinary differential equations to structural causal models: the deterministic case. *arXiv preprint arXiv:1304.7920*.

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

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*. Pmlr.

Pearl, J. 2009. *Causality*. Cambridge university press.

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.

Rahimi, A.; and Recht, B. 2007. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20.

Rasmussen, C. E.; and Williams, C. K. I. 2005. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Ridderbusch, S.; Ober-Blöbaum, S.; and Goulart, P. 2023. The past does matter: correlation of subsequent states in trajectory predictions of Gaussian Process models. In *Uncertainty in Artificial Intelligence, UAI 2023, July 31 - 4 August 2023, Pittsburgh, PA, USA*, volume 216, 1752–1761. PMLR.

Rissanen, J. 1983. A Universal Prior for Integers and Estimation by Minimum Description Length. *AnnalsStatistics*, 11(2): 416–431.

Rubenstein, P. K.; Bongers, S.; Schölkopf, B.; and Mooij, J. M. 2016. From deterministic ODEs to dynamic structural causal models. *arXiv preprint arXiv:1608.08028*.

Runge, J. 2020. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on uncertainty in artificial intelligence*. Pmlr.

Schmidt, M.; and Lipson, H. 2009. Distilling Free-Form Natural Laws from Experimental Data. *Science*, 324(5923).

Schwarz, G. 1978. Estimating the dimension of a model. *The annals of statistics*.

Strogatz, S. H. 2000. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*. Westview Press.

Sun, J.; Taylor, D.; and Bollt, E. M. 2015. Causal network inference by optimal causation entropy. *SIAM Journal on Applied Dynamical Systems*, 14(1).

Tsamardinos, I.; Brown, L. E.; and Aliferis, C. F. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65.

Vereshchagin, N. K.; and Vitányi, P. M. 2004. Kolmogorov's structure functions and model selection. *IEEETIT*, 50(12).

Voortman, M.; Dash, D.; and Druzdzel, M. J. 2010. Learning causal models that make correct manipulation predictions with time series data. In *Causality: Objectives and Assessment*. PMLR.

Wang, J. M.; Fleet, D. J.; and Hertzmann, A. 2005. Gaussian Process Dynamical Models. In *NeurIPS 2005*. Cambridge, MA, USA: MIT Press.

Wang, Y.; Solus, L.; Yang, K. D.; and Uhler, C. 2017. Permutation-based Causal Inference Algorithms with Interventions. In *NIPS*.

Wendland, H. 1995. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in computational Mathematics*, 4(1).

Williams, C.; and Seeger, M. 2000. Using the Nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13.

Xu, S.; Mameche, S.; and Vreeken, J. 2025. Information-Theoretic Causal Discovery in Topological Order. In *AISTATS 2025*.