UNMIXING CAUSAL DYNAMICS: LEARNING LATENT MECHANISMS FROM AGGREGATE TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Observed time series are often superpositions of multiple causal mechanisms that evolve concurrently over the same variables, so the causal relations of individual mechanisms are entangled in the aggregate signal. At each time step, each latent mechanism produces its own state vector, but we observe only the elementwise sum across mechanisms, together with sparse measurements that occasionally reveal mechanism states. We introduce UCD (**U**nmixing **C**ausal **D**ynamics) to recover mechanism-specific causal structure from such aggregated observations. UCD models each mechanism with a temporal structural equation model (SEM) with independent additive noise and performs particle-based inference over the unobserved mechanism trajectories. It maintains a set of particles, each representing a plausible joint history of all mechanisms. Between measurements, particles propose and reweight next-step mechanism states that are consistent with the observed aggregate; when sparse state measurements arrive, particles are corrected to match the revealed states and reweighted by their plausibility. We then learn each mechanism's causal graph and noise scales by maximizing the likelihood of the residuals between predicted and revealed mechanism states under the assumed noise model. Experiments show that UCD reliably recovers mechanism-specific causal structure from aggregated time series given sparse state measurements.

1 INTRODUCTION

Causal discovery is a key tool for understanding complex dynamical systems in time-series settings (Assaad et al., 2022), with applications ranging from climate science (Runge et al., 2023) and neuroscience to finance (Jiang & Shimizu, 2023). By characterizing how variables influence one another within and across time steps, temporal causal models support hypothesis generation and guide experimental follow-up (Camps-Valls et al., 2023). Achieving this from observational data is challenging because temporal signals reflect noisy, dependent dynamics and may mix multiple latent generating processes (Gerhardus & Runge, 2020; Runge et al., 2019; 2023).

Most temporal causal discovery methods infer a single causal system over the measured variables, e.g., PCMCI (Runge et al., 2019), DYNOTEARS (Pamfil et al., 2020), NTS-NOTEARS (Sun et al., 2023), CausalFormer (Kong et al., 2024), Rhino (Gong et al., 2023), and CUTS+ (Cheng et al., 2024). In many domains, however, each observed channel can be the superposition of several concurrently evolving mechanisms. Concretely, at each time step, each mechanism has its own latent state vector over the same variables, but the observation records only their elementwise sum. Different mechanisms may induce different directions, magnitudes, and delays of influence between the same variables, so fitting a single graph to the aggregate can conflate distinct pathways and mask mechanism-specific structure (D'Acunto et al., 2023a).

Recent work has begun to model such heterogeneity by introducing explicit combination structure. MS-CASTLE (D'Acunto et al., 2023a) recovers scale-specific graphs after wavelet decomposition, and REDCLIFF (Brown & Carlson, 2025) combines forecasting factors using history-dependent weights. These approaches organize multiple processes through a particular decomposition or weighting form. In contrast, we study the setting where mechanisms overlap in time and observations are an additive mixture of their states, aiming to recover each mechanism's causal dynamics without assuming spectral separation or a predefined weighting rule.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

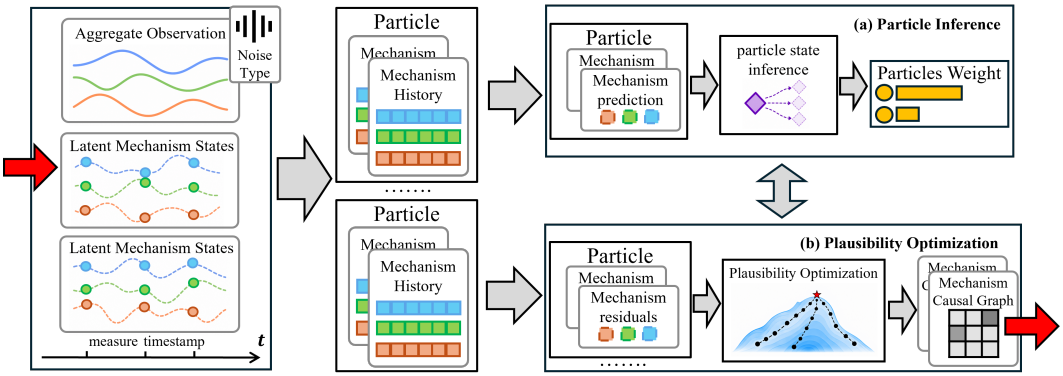

Figure 1: Overview of UCD. The observed time series is an additive sum of multiple latent mechanisms, with occasional direct measurements of individual mechanism states. We maintain a set of particles, each encoding a candidate joint history of all mechanisms consistent with the observations. (a) *Particle inference:* particles are propagated under the current temporal SEM, adjusted to match aggregate observations, and importance-weighted (with resampling) to track multiple plausible decompositions between measurements; at measurement times, the corresponding latent components are clamped and particles are reweighted. (b) *Plausibility optimization:* given the weighted particle histories, we score one-step prediction residuals between SEM-predicted and measured latent states under the assumed noise model, aggregate scores across particles to marginalize latent uncertainty, and update mechanism-specific causal graphs and noise scales.

We propose UCD to recover mechanism-specific temporal causal structure from an additive mixture time series, aided by sparse measurements of individual mechanism states (Figure 1). Each mechanism follows a temporal structural equation model, but the observation reveals only the sum across mechanisms, so the latent decomposition is generally not directly identifiable. To represent this ambiguity, UCD maintains a particle approximation to the posterior over joint latent histories. Between measurement times, each particle propagates all mechanisms forward under the current SEM, is adjusted to satisfy the aggregate observation constraint, and is importance-weighted and resampled. When a latent measurement arrives, particles are clamped and reweighted to match these measurements. Conditioned on the resulting weighted histories, we learn each mechanism's causal graph and noise scale by scoring one-step SEM prediction residuals under the noise model, averaging this plausibility score across particles, and alternating inference with parameter updates. The main contributions are:

(i) We introduce UCD, an overcomplete latent-mechanism formulation for temporal causal discovery from additive time-series observations with sparse mechanism-state supervision. The formulation supports multiple simultaneously active mechanisms, each with its own causal graph, without assuming a fixed decomposition (e.g., by scale or frequency).

(ii) We develop an observation-conditioned particle inference procedure that tracks a distribution over joint latent mechanism histories. Particles sample process noise and propagate mechanism states forward, are constrained and reweighted to match the aggregate observation, and are corrected via clamping and reweighting when sparse mechanism-state measurements are available.

(iii) We propose a measurement-time plausibility likelihood objective that marginalizes latent uncertainty by averaging conditional state densities over particles. Optimizing this objective while alternating between particle inference and parameter updates learns each mechanism's instantaneous and lagged causal coefficients, together with noise scales.

(iv) Experiments on synthetic and real-world benchmarks demonstrate that UCD more accurately recovers latent mechanism-specific causal structure from aggregated observations under sparse state measurements.

2 Related Works

Causal Discovery from Time Series Temporal causal discovery has been approached through conditional-independence testing, continuous optimization, and predictive modeling, including

PCMCI (Runge et al., 2019), DYNOTEARS (Pamfil et al., 2020), NTS-NOTEARS (Sun et al., 2023), Neural Granger causality (Tank et al., 2021), Rhino (Gong et al., 2023), CUTS (Cheng et al., 2023) and CUTS+ (Cheng et al., 2024). These methods recover instantaneous or lagged relations among measured variables, but do not explicitly separate multiple latent processes that jointly generate the same observed channels.

Causal Discovery with Multiple Mechanisms Prior work also considers causal heterogeneity across temporal regimes, variable-specific changes, or different sequences, as in CASTOR (Rahmani & Frossard, 2025), Causal-RuLSIF (Gao et al., 2025), and MCD (Varambally et al., 2024). These settings differ from ours, where multiple causal mechanisms evolve simultaneously and jointly generate each observation. Closer to our setting, MS-CASTLE (D'Acunto et al., 2023a) and MN-CASTLE (D'Acunto et al., 2023b) recover scale-specific causal structures through wavelet-based decomposition, while REDCLIFF (Brown & Carlson, 2025) combines forecasting factors through history-dependent weights. In contrast, UCD models simultaneously evolving mechanisms with separate latent histories and learns their causal structures from additive observations using sparse mechanism-state measurements.

3 PRELIMINARIES

Let d denote the number of observed variables, T the sequence length, K the number of latent mechanisms, and L the maximum lag. Indices $i, j \in \{1, \dots, d\}$ refer to variables, $k \in \{1, \dots, K\}$ to mechanisms, and $\ell \in \{1, \dots, L\}$ to lags. At each time t, we observe an aggregate signal $\mathbf{x}_t \in \mathbb{R}^d$ and posit latent mechanism states $\mathbf{z}_{k,t} \in \mathbb{R}^d$, with components x_t^j and $z_{k,t}^j$.

Definition 3.1 (Linear temporal causal mechanism)**.** Mechanism k evolves according to a time-invariant linear temporal structural equation model:

$$\mathbf{z}_{k,t}^\top = \mathbf{z}_{k,t}^\top \mathbf{W}_k + \sum_{\ell=1}^{L} \mathbf{z}_{k,t-\ell}^\top \mathbf{A}_k^{(\ell)} + \boldsymbol{\varepsilon}_{k,t}^\top. \tag{1}$$

$\mathbf{W}_k, \mathbf{A}_k^{(\ell)} \in \mathbb{R}^{d \times d}$ parameterize instantaneous and lagged causal effects, where $W_{k,ij}$ (resp., $A_{k,ij}^{(\ell)}$) encodes the effect of $z_{k,t}^i$ (resp., $z_{k,t-\ell}^i$) on $z_{k,t}^j$. The directed graph induced by $\mathbf{W}_k$ is assumed acyclic. Noise components $\varepsilon_{k,t}^j$ are mutually independent, and for each fixed (k, j) are i.i.d. over time. We denote all model parameters by θ.

Definition 3.2 (Additive observations with sparse state measurements)**.** The observed process is the coordinatewise sum of mechanism states, $\mathbf{x}_t = \sum_{k=1}^K \mathbf{z}_{k,t}$ (equivalently, $x_t^j = \sum_{k=1}^K z_{k,t}^j$). Let $\mathbf{z}_t \in \mathbb{R}^{Kd}$ be the vertical concatenation of $\mathbf{z}_{1,t}, \dots, \mathbf{z}_{K,t}$. In addition to aggregate observations $\mathbf{x}_{1:T}$, we observe the full concatenated latent state exactly at a sparse set of times $\mathcal{O} \subseteq \{1, \dots, T\}$, denoted by $\mathbf{z}_t^{\text{obs}}$ for $t \in \mathcal{O}$.

Given L, a specified structural-noise family, aggregate observations $\mathbf{x}_{1:T}$, and sparse complete-state measurements $\{\mathbf{z}_t^{\text{obs}} : t \in \mathcal{O}\}$, our goal is to estimate the mechanism-specific temporal graphs (instantaneous and lagged), their coefficients, and the noise scales.

Histories and Conditional Distributions Let $H_{t-1} = (\mathbf{z}_{t-1}, \dots, \mathbf{z}_{t-L})$ denote the length-L latent history, where each $\mathbf{z}_u$ is the concatenated state. Let $\mathcal{F}_{t-1}$ denote the information available strictly before time t, consisting of aggregate observations $\mathbf{x}_u$ for $u < t$ with $u \notin \mathcal{O}$, and state measurements $\mathbf{z}_u^{\text{obs}}$ for $u < t$ with $u \in \mathcal{O}$.

Because the additive observation model does not uniquely determine H_{t-1}, we represent uncertainty about the latent history by the filtering distribution $p_\theta(H_{t-1} \mid \mathcal{F}_{t-1})$. For a specified noise family, let $p_\theta(\mathbf{z}_t \mid H_{t-1})$ denote the conditional density of the next concatenated state given a candidate history. At a measurement time $t \in \mathcal{O}$, the one-step predictive density of the revealed state is obtained by marginalizing over possible histories:

$$p_\theta(\mathbf{z}_t^{\text{obs}} \mid \mathcal{F}_{t-1}) = \mathbb{E}_\theta\left[p_\theta(\mathbf{z}_t^{\text{obs}} \mid H_{t-1}) \,\middle|\, \mathcal{F}_{t-1}\right]. \tag{2}$$

The expectation is with respect to $H_{t-1} \sim p_\theta(H_{t-1} \mid \mathcal{F}_{t-1})$; θ and the evaluation point $\mathbf{z}_t^{\text{obs}}$ are fixed. Intuitively, each candidate history contributes its likelihood for the same measured state,

weighted by how plausible that history is under the past data in $\mathcal{F}_{t-1}$. Notably, $\mathcal{F}_{t-1}$ excludes both the current aggregate observation $\mathbf{x}_t$ and the current state measurement $\mathbf{z}_t^{\text{obs}}$.

4 METHOD

As illustrated in Figure 1, UCD alternates between (i) particle-based inference of latent mechanism histories from aggregate observations and (ii) plausibility optimization of causal parameters using sparse state measurements.

4.1 CAUSAL DYNAMICS

Centering. We subtract the full-sequence mean from each observed channel and, for each mechanism–variable pair, subtract the mean over available measurement times. This centering removes only the mean and does not rescale variances.

Reparameterized one-step dynamics. Let $\mathbf{S}_k = \operatorname{diag}(s_k^1, \ldots, s_k^d)$ collect learnable structural-noise scales, and write the structural noise as $\boldsymbol{\varepsilon}_{k,t}^\mathsf{T} = \boldsymbol{\eta}_{k,t}^\mathsf{T}\mathbf{S}_k$, where the components of $\boldsymbol{\eta}_{k,t}$ are independent and follow the specified zero-mean, unit-variance noise family. Since $\mathbf{I} - \mathbf{W}_k$ is invertible under acyclicity, Eq. (1) implies

$$\mathbf{z}_{k,t}^\mathsf{T} = \sum_{\ell=1}^{L} \mathbf{z}_{k,t-\ell}^\mathsf{T} \mathbf{A}_k^{(\ell)} (\mathbf{I} - \mathbf{W}_k)^{-1} + \boldsymbol{\eta}_{k,t}^\mathsf{T} \mathbf{S}_k (\mathbf{I} - \mathbf{W}_k)^{-1}. \tag{3}$$

Define the reduced lag coefficients $\mathbf{B}_k^{(\ell)} = \mathbf{A}_k^{(\ell)}(\mathbf{I} - \mathbf{W}_k)^{-1}$ and the noise map $\mathbf{G}_k = \mathbf{S}_k(\mathbf{I} - \mathbf{W}_k)^{-1}$. Then

$$\mathbf{z}_{k,t}^\mathsf{T} = \mathbf{m}_{\theta,k}(H_{t-1})^\mathsf{T} + \boldsymbol{\eta}_{k,t}^\mathsf{T}\mathbf{G}_k, \qquad \mathbf{m}_{\theta,k}(H_{t-1})^\mathsf{T} = \sum_{\ell=1}^{L} \mathbf{z}_{k,t-\ell}^\mathsf{T} \mathbf{B}_k^{(\ell)}. \tag{4}$$

Here the lagged states are entries of H_{t-1}. Thus $\mathbf{m}_{\theta,k}(H_{t-1})$ is the conditional mean, $\mathbf{B}_k^{(\ell)}$ propagates past states after accounting for instantaneous effects, and $\mathbf{G}_k$ maps standardized noise to the current state. The structural edges remain defined by $\mathbf{W}_k$ and $\mathbf{A}_k^{(\ell)}$. We parameterize $\theta = \{\mathbf{W}_k, \mathbf{B}_k^{(1)}, \ldots, \mathbf{B}_k^{(L)}, \mathbf{S}_k\}_{k=1}^K$.

To encourage stable temporal dynamics during estimation, we constrain the reduced lag coefficients via a companion-matrix rescaling; the full parameterization is given in Appendix A.2.

4.2 PARTICLE INFERENCE UNDER ADDITIVE OBSERVATIONS

For fixed parameters θ, we approximate the filtering distribution $p_\theta(H_{t-1} \mid \mathcal{F}_{t-1})$ using P particle histories (Arulampalam et al., 2002). For $b \in \{1, \ldots, P\}$, particle b stores a length-L history $H_{t-1}^{(b)}$ of all mechanisms and carries a weight; at the start of each step we work with equally weighted particles.

All inputs follow the centering convention in Section 4.1. To initialize, we simulate the latent mechanisms under θ with independent noise for a burn-in period and keep only the most recent L states. These histories approximate a stationary prior over H_0 and form the initial particle set.

We first describe the update for times $t \notin \mathcal{O}$, where only the aggregate observation $\mathbf{x}_t$ is available (measurement times are handled in Section 4.3). Given a particle history, the aggregate residual

$$\mathbf{r}_t^{(b)} = \mathbf{x}_t - \sum_{k=1}^{K} \mathbf{m}_{\theta,k}(H_{t-1}^{(b)}) \tag{5}$$

collects the part of $\mathbf{x}_t$ that must be explained by the current noises. Conditioning on $H_{t-1}^{(b)}$, the additive constraint requires $\big(\mathbf{r}_t^{(b)}\big)^\mathsf{T} = \sum_{k=1}^K \boldsymbol{\eta}_{k,t}^\mathsf{T}\mathbf{G}_k$.

Constrained noise sampling. We sample the standardized noises for mechanisms $1{:}(K-1)$ and determine mechanism K by the constraint above. Let $\mathbf{u} \in \mathbb{R}^{(K-1)d}$ concatenate the free noise vectors, with blocks $\mathbf{u}_k \in \mathbb{R}^d$ for $k = 1, \dots, K-1$. Define $\mathbf{G}_{\text{free}} = (\mathbf{G}_1^\mathsf{T}, \dots, \mathbf{G}_{K-1}^\mathsf{T})^\mathsf{T} \in \mathbb{R}^{(K-1)d \times d}$ and $\mathbf{C}_\theta = \sum_{k=1}^K \mathbf{G}_k^\mathsf{T} \mathbf{G}_k \in \mathbb{R}^{d \times d}$. Here $\mathbf{G}_{\text{free}}$ maps the free noises to their aggregate contribution, while $\mathbf{C}_\theta$ is the aggregate noise covariance conditional on a history.

To bias proposals toward values compatible with $\mathbf{r}_t^{(b)}$, we approximate the independent standardized-noise priors by standard Gaussians and condition this approximation on the residual, obtaining

$$\begin{aligned} q_G(\mathbf{u} \mid \mathbf{r}_t^{(b)}) &= \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}_t^{(b)}, \mathbf{V}_\theta), \\ \boldsymbol{\mu}_t^{(b)} &= \mathbf{G}_{\text{free}} \mathbf{C}_\theta^{-1} \mathbf{r}_t^{(b)}, \\ \mathbf{V}_\theta &= \mathbf{I}_{(K-1)d} - \mathbf{G}_{\text{free}} \mathbf{C}_\theta^{-1} \mathbf{G}_{\text{free}}^\mathsf{T}. \end{aligned} \tag{6}$$

The mean allocates the residual across free mechanisms, while the covariance captures remaining allocation ambiguity.

Let f_k denote the product density of independent standardized-noise components and define $f_{\text{free}}(\mathbf{u}) = \prod_{k=1}^{K-1} f_k(\mathbf{u}_k)$ (all learnable scales are contained in $\mathbf{G}_k$). To retain support in regions underrepresented by the Gaussian approximation, we use the mixture proposal

$$q_{\text{mix}}(\mathbf{u} \mid \mathbf{r}_t^{(b)}) = (1-\alpha) q_G(\mathbf{u} \mid \mathbf{r}_t^{(b)}) + \alpha f_{\text{free}}(\mathbf{u}), \qquad 0 < \alpha < 1. \tag{7}$$

State proposal and weighting. Draw $\mathbf{u}_t^{(b)} \sim q_{\text{mix}}(\cdot \mid \mathbf{r}_t^{(b)})$ and propose mechanism states

$$\begin{aligned} \big(\mathbf{z}_{k,t}^{(b)}\big)^\mathsf{T} &= \mathbf{m}_{\theta,k}(H_{t-1}^{(b)})^\mathsf{T} + \big(\mathbf{u}_{k,t}^{(b)}\big)^\mathsf{T} \mathbf{G}_k, \quad k = 1, \dots, K-1, \\ \mathbf{z}_{K,t}^{(b)} &= \mathbf{x}_t - \sum_{k=1}^{K-1} \mathbf{z}_{k,t}^{(b)}. \end{aligned} \tag{8}$$

By construction, the proposal satisfies the aggregate constraint exactly at times without state measurements. Since $\mathbf{G}_K$ is invertible, the implied noise for the last mechanism is

$$\big(\boldsymbol{\eta}_{K,t}^{(b)}\big)^\mathsf{T} = \Big(\big(\mathbf{r}_t^{(b)}\big)^\mathsf{T} - \big(\mathbf{u}_t^{(b)}\big)^\mathsf{T} \mathbf{G}_{\text{free}}\Big) \mathbf{G}_K^{-1}. \tag{9}$$

The observation provides d constraints, leaving $(K-1)d$ free noise coordinates.

For fixed $H_{t-1}^{(b)}$ and $\mathbf{x}_t$, the target density of the free noises is

$$\begin{aligned} p_\theta\Big(\mathbf{u}_t \mid \mathbf{x}_t, H_{t-1}^{(b)}\Big) &\propto f_{\text{free}}(\mathbf{u}_t)\, p_\theta\Big(\mathbf{x}_t \mid \mathbf{u}_t, H_{t-1}^{(b)}\Big) \\ &= \frac{f_{\text{free}}(\mathbf{u}_t)\, f_K\Big(\boldsymbol{\eta}_{K,t}(\mathbf{u}_t, \mathbf{r}_t^{(b)})\Big)}{|\det \mathbf{G}_K|}. \end{aligned} \tag{10}$$

The omitted normalizing constant is $p_\theta(\mathbf{x}_t \mid H_{t-1}^{(b)})$, which does not depend on $\mathbf{u}_t$. The factor $|\det \mathbf{G}_K|^{-1}$ follows from the change of variables that enforces the additive constraint.

Since $\mathbf{u}_t^{(b)}$ is sampled from $q_{\text{mix}}(\mathbf{u} \mid \mathbf{r}_t^{(b)})$, the unnormalized importance weight is

$$\widetilde{w}_t^{(b)} = \frac{f_{\text{free}}(\mathbf{u}_t^{(b)}) f_K(\boldsymbol{\eta}_{K,t}^{(b)})}{|\det \mathbf{G}_K|\, q_{\text{mix}}\Big(\mathbf{u}_t^{(b)} \mid \mathbf{r}_t^{(b)}\Big)}. \tag{11}$$

We normalize $w_t^{(b)} = \widetilde{w}_t^{(b)} / \sum_{b'} \widetilde{w}_t^{(b')}$. Each particle then appends $\mathbf{z}_t^{(b)}$ and discards the oldest state to form a candidate updated history $\widetilde{H}_t^{(b)}$. Resampling the candidates according to $w_t^{(b)}$ yields P equally weighted histories $H_t^{(b)}$ for the next time step.

4.3 Sparse State Measurements and Plausibility Optimization

At times $t \in \mathcal{O}$, the revealed mechanism states play two roles: (i) they *condition* the particle approximation by correcting the latent history distribution, and (ii) they provide a *learning signal* by scoring how plausible each particle's one-step prediction errors are under the structural-noise model. Both roles use the pre-assimilation histories $H_{t-1}^{(b)}$, i.e., before incorporating $\mathbf{z}_t^{\text{obs}}$.

Plausibility from standardized residuals. For a given history $H_{t-1}^{(b)}$, the discrepancy between the revealed mechanism state and the one-step prediction induces the standardized residual

$$\big(\mathbf{e}_{k,t}^{(b)}\big)^{\mathsf{T}} = \Big(\big(\mathbf{z}_{k,t}^{\mathrm{obs}}\big)^{\mathsf{T}} - \mathbf{m}_{\theta,k}(H_{t-1}^{(b)})^{\mathsf{T}}\Big)\,\mathbf{G}_k^{-1}. \tag{12}$$

Intuitively, $\mathbf{e}_{k,t}^{(b)}$ is the noise realization required to reconcile particle b with the measurement for mechanism k, expressed in standardized-noise coordinates. Its density under f_k therefore quantifies how plausible that reconciliation is. Using independence across mechanisms and the change of variables induced by $\mathbf{G}_k$, the conditional density of the revealed complete state under history b is

$$p_\theta\Big(\mathbf{z}_t^{\mathrm{obs}} \mid H_{t-1}^{(b)}\Big) = \prod_{k=1}^{K} \frac{f_k(\mathbf{e}_{k,t}^{(b)})}{|\det \mathbf{G}_k|}. \tag{13}$$

Histories that require more likely residuals receive higher plausibility. Averaging across particles yields the particle approximation to the predictive density in Eq. (2):

$$\widehat{p}_\theta\big(\mathbf{z}_t^{\mathrm{obs}} \mid \mathcal{F}_{t-1}\big) = \frac{1}{P}\sum_{b=1}^{P} p_\theta\Big(\mathbf{z}_t^{\mathrm{obs}} \mid H_{t-1}^{(b)}\Big)\,. \tag{14}$$

Learning objective. We optimize parameters by maximizing predictive likelihood at measurement times:

$$\mathcal{L}_{\mathrm{state}}(\theta) = -\frac{1}{|\mathcal{O}|}\sum_{t\in\mathcal{O}} \log \widehat{p}_\theta\big(\mathbf{z}_t^{\mathrm{obs}} \mid \mathcal{F}_{t-1}\big)\,. \tag{15}$$

This objective depends on aggregate-only observations only through the inferred histories $\{H_{t-1}^{(b)}\}$.

Measurement assimilation. After computing the predictive density, we condition the particle set on the measurement by clamping $\mathbf{z}_t^{(b)} = \mathbf{z}_t^{\mathrm{obs}}$ and reweighting particles by their conditional measurement likelihood, $w_t^{(b)} \propto p_\theta(\mathbf{z}_t^{\mathrm{obs}} \mid H_{t-1}^{(b)})$. Each particle appends the measured state and discards the oldest state to form candidate updated histories, and resampling produces P equally weighted histories $H_t^{(b)}$ for the next time step. This measurement update replaces the aggregate-only weighting in Eq. (11).

4.4 Alternating Optimization

We optimize θ by alternating between latent-history inference and parameter updates. Following Zheng et al. (2018), we impose a soft acyclicity penalty on the instantaneous graph, $h(\mathbf{W}_k) = \mathrm{tr}\big(\exp(\mathbf{W}_k \odot \mathbf{W}_k)\big) - d$. The complete training objective is

$$\begin{aligned}\mathcal{L}(\theta) = \lambda_{\mathrm{state}}\mathcal{L}_{\mathrm{state}}(\theta) + \frac{\lambda_W}{Kd^2}\sum_{k=1}^{K}\|\mathbf{W}_k\|_1 \\ + \frac{\lambda_A}{KLd^2}\sum_{k=1}^{K}\sum_{\ell=1}^{L}\|\mathbf{A}_k^{(\ell)}\|_1 + \lambda_{\mathrm{dag}}\sum_{k=1}^{K} h(\mathbf{W}_k).\end{aligned} \tag{16}$$

Here λ_{state} weights the predictive-likelihood term, while λ_W and λ_A encourage sparsity and λ_{dag} enforces acyclicity. Only $\mathbf{W}_k$ requires an acyclicity penalty because lagged edges are directed forward in time. We fix $\mathrm{diag}(\mathbf{W}_k) = 0$ and optimize the diagonal entries of $\mathbf{S}_k$.

Each iteration consists of two steps. *(Inference)* With θ fixed, we run the particle filter forward: at $t \notin \mathcal{O}$ we update particles using $\mathbf{x}_t$ as in Section 4.2; at $t \in \mathcal{O}$ we process $\mathbf{z}_t^{\mathrm{obs}}$ as in Section 4.3, while retaining the pre-assimilation histories $\{H_{t-1}^{(b)}\}$ to evaluate $\mathcal{L}_{\mathrm{state}}(\theta)$. *(Learning)* Holding the retained histories fixed, we take an optimization step on Eq. (16) by recomputing one-step predictions, measurement residuals, and their likelihood under the current noise model as functions of θ. The next iteration reruns inference with the updated parameters. A full algorithmic description is given in Appendix A.1.

	Method	$K=2$ SHD (↓)	$K=2$ F1 (↑)	$K=2$ AUPRC (↑)	$K=3$ SHD (↓)	$K=3$ F1 (↑)	$K=3$ AUPRC (↑)
match	DYNOTEARS	40.40±2.73	88.72±0.91	91.23±1.40	85.60±4.50	83.75±1.09	86.93±1.88
	NTS-NOTEARS	48.40±2.33	86.04±1.34	87.88±1.53	83.00±4.47	83.56±1.11	85.48±1.28
	MS-CASTLE	502.60±25.19	26.10±1.42	12.98±1.44	656.60±33.85	13.90±1.21	6.06±0.58
	UCD	**29.00±1.79**	**91.79±0.77**	**95.38±1.11**	**54.00±6.13**	**89.53±1.67**	**92.74±1.54**
aggregate	DYNOTEARS	38.20±2.14	89.15±0.76	91.83±1.33	76.40±3.07	84.95±1.05	88.69±1.62
	NTS-NOTEARS	45.20±3.31	86.71±1.56	89.08±1.58	75.00±5.55	84.63±1.23	87.68±0.92
	VAR-LiNGAM	74.40±11.60	80.70±1.73	85.46±1.87	139.20±7.28	71.14±1.96	72.32±2.56
	Rhino	118.00±11.88	65.70±1.74	66.95±2.48	201.20±14.22	55.14±2.21	54.84±2.37
	CASTOR	78.40±11.38	79.19±1.66	82.40±1.16	170.20±21.20	67.67±2.23	70.42±3.54
	UCD	**27.80±1.17**	**92.00±0.60**	**95.77±0.90**	**50.00±5.40**	**89.99±1.50**	**93.56±1.46**

	Method	$K=4$ SHD (↓)	$K=4$ F1 (↑)	$K=4$ AUPRC (↑)	$K=5$ SHD (↓)	$K=5$ F1 (↑)	$K=5$ AUPRC (↑)
match	DYNOTEARS	173.60±1.50	76.39±1.50	81.94±2.22	290.40±10.38	69.81±1.02	77.51±1.81
	NTS-NOTEARS	151.80±8.13	77.72±1.24	79.66±2.40	239.00±5.18	72.60±1.25	75.57±1.76
	MS-CASTLE	865.80±38.58	10.22±0.55	4.97±0.27	1036.60±46.18	9.28±0.30	4.63±0.13
	UCD	**94.00±10.49**	**86.35±1.83**	**90.04±1.80**	**127.00±11.70**	**85.12±1.25**	**88.97±1.37**
aggregate	DYNOTEARS	153.00±2.76	78.17±1.26	84.85±1.79	244.60±6.15	72.70±1.23	82.11±1.74
	NTS-NOTEARS	136.20±7.11	79.11±1.05	83.06±2.28	205.00±4.98	75.07±1.41	80.97±1.99
	VAR-LiNGAM	224.20±9.52	60.86±1.20	62.91±1.68	313.20±19.29	50.65±2.25	56.56±1.73
	Rhino	278.60±13.28	49.97±2.41	50.93±3.32	349.40±14.18	44.64±3.04	48.30±1.98
	CASTOR	248.20±20.01	61.93±1.86	62.41±3.24	310.80±26.48	58.71±1.62	60.69±1.80
	UCD	**84.40±10.21**	**87.25±1.59**	**92.05±1.39**	**111.60±12.63**	**86.15±1.32**	**91.40±0.76**

Table 1: Performance under different numbers of latent mechanisms

5 EXPERIMENT

5.1 EXPERIMENT SETUP

Baselines. We compare UCD with six temporal causal discovery baselines: DYNOTEARS (Pamfil et al., 2020), NTS-NOTEARS (Sun et al., 2023), MS-CASTLE (D'Acunto et al., 2023a), VAR-LiNGAM (Hyvärinen et al., 2010), Rhino (Gong et al., 2023), and CASTOR (Rahmani & Frossard, 2025). DYNOTEARS and NTS-NOTEARS are designed for a single observed process, so we extend them by fitting K mechanism-specific models with a joint loss that combines (i) mean-squared error on the revealed mechanism states at measurement times and (ii) mean-squared error between the observed aggregate series and the sum of inferred mechanism states at all times. MS-CASTLE natively outputs multiple scale-specific graphs via wavelet decomposition.

We report results in two settings. *Match:* methods that output K graphs are aligned to the ground-truth mechanisms and evaluated per mechanism. *Aggregate:* all mechanism graphs are merged into a single graph, reducing evaluation to standard single-graph discovery; VAR-LiNGAM, Rhino, and CASTOR are run directly on the aggregate series in this setting.

Synthetic Data. For each mechanism, we sample a contemporaneous Erdős–Rényi DAG (ERDdS & R&wi, 1959) and independently draw lagged edges up to lag L. We then simulate a linear additive-noise tSEM with independent Laplace noise. The aggregate series is observed for all t, while the full mechanism state is revealed at a random subset of times. Unless noted otherwise, we use $T = 4000$, $d = 30$, $K = 3$, and $L = 2$, with d contemporaneous edges per mechanism and lagged edges included with probability $1/d$ (so $\mathbb{E}[\#\text{lag edges at lag } \ell] = d$); measurements are available at 40% of time indices.

Real Data. We also evaluate on CausalRivers (Stein et al., 2025), a real-world benchmark of river-discharge time series recorded every 15 minutes over five years. We use the eastern Germany subset, follow the provided 10-node random-subgraph split, and truncate each sequence to the first 1000 time steps.

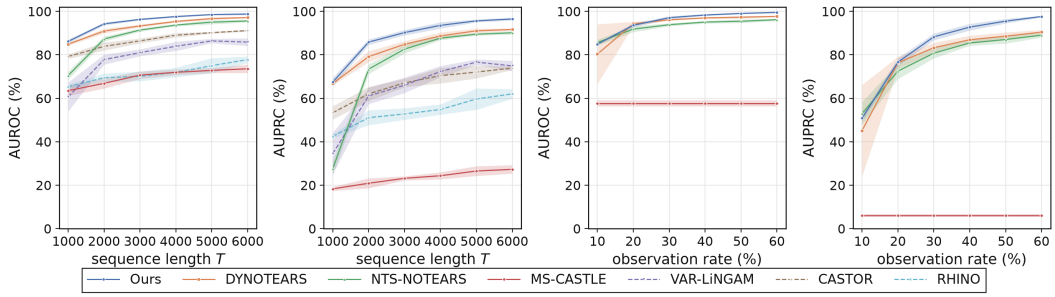

Figure 2: Performance under different sequence lengths and observation rates.

	Method	$d = 20$ F1 (↑)	$d = 20$ AUPRC (↑)	$d = 30$ F1 (↑)	$d = 30$ AUPRC (↑)	$d = 40$ F1 (↑)	$d = 40$ AUPRC (↑)
match	DYNOTEARS	84.10±1.95	86.95±1.69	83.75±1.09	86.93±1.88	83.25±1.22	85.66±1.72
	NTS-NOTEARS	83.16±2.29	84.61±2.47	83.56±1.11	85.48±1.28	82.99±1.64	82.37±1.75
	MS-CASTLE	15.64±0.57	8.04±0.32	13.90±1.21	6.06±0.58	13.63±1.45	4.96±0.21
	UCD	**90.16±2.33**	**94.54±1.45**	**89.53±1.67**	**92.74±1.54**	**86.53±0.98**	**90.62±0.80**
aggregate	DYNOTEARS	84.84±2.06	88.93±1.58	84.95±1.05	88.69±1.62	83.71±1.44	86.82±1.72
	NTS-NOTEARS	84.01±2.14	87.09±2.05	84.63±1.23	87.68±0.92	83.30±1.74	83.72±1.91
	VAR-LiNGAM	70.53±0.91	72.67±1.49	71.14±1.96	72.32±2.56	70.38±1.48	69.42±2.11
	Rhino	56.32±3.83	61.36±3.06	55.14±2.21	54.84±2.37	58.63±1.04	55.65±1.39
	CASTOR	67.70±2.88	69.06±2.17	67.67±2.23	70.42±3.54	68.07±1.21	67.93±2.36
	UCD	**90.65±2.34**	**95.34±1.46**	**89.99±1.50**	**93.56±1.46**	**86.99±1.15**	**91.62±1.01**

Table 2: Performance under different graph scales.

5.2 Experimental Results

Mechanism recovery under varying K. Table 1 reports causal recovery as the number of superposed mechanisms K increases. Under mechanism-level evaluation, UCD achieves the best F1 and AUPRC (and the lowest SHD) across all K. Because larger K implies more graphs and edges to recover, SHD is not directly comparable across different K; we therefore emphasize F1 and AUPRC. Importantly, UCD maintains a clear margin as the mixture becomes more ambiguous, suggesting that (i) particle inference effectively preserves multiple plausible latent decompositions consistent with the aggregate observations, and (ii) plausibility optimization uses sparse state measurements to identify causal structure by scoring the residuals implied by each candidate history under the noise model. In contrast, MS-CASTLE performs poorly because it ties mechanisms to predefined wavelet scales, whereas the latent mechanisms in our data need not align with fixed frequency bands.

Aggregate-level evaluation (merging mechanism graphs) shows the same trend. This indicates that the learned mechanisms are not only individually well recovered, but also collectively preserve the overall causal structure of the summed process. The agreement between mechanism-level and aggregate-level results supports the benefit of jointly inferring latent histories and learning causal parameters from residual plausibility.

Effect of sequence length and observation rate. Figure 2 studies sensitivity to sequence length T and the mechanism-state observation rate. As T increases, all methods generally improve with more temporal information, but UCD consistently attains the highest AUROC and AUPRC. This indicates that particle-based latent-state inference can accumulate constraints from aggregate observations over time, instead of depending primarily on the sparsely revealed mechanism states.

Varying the observation rate shows a complementary trend: UCD remains strong even with limited supervision and improves smoothly as more mechanism states are revealed. This matches the intended division of labor in UCD—aggregate observations shape the latent histories via particle inference, while each additional measurement strengthens the plausibility signal used to learn

Method	aggregate			match		
	SHD (↓)	F1 (↑)	AUPRC (↑)	SHD (↓)	F1 (↑)	AUPRC (↑)
UCD	**50.00±5.40**	**89.99±1.50**	**93.56±1.46**	**54.00±6.13**	**89.53±1.67**	**92.74±1.54**
UCD w/o particles	73.80±7.49	84.64±1.98	90.23±1.34	80.40±6.97	83.82±1.92	88.44±1.51
UCD w/o weight	74.80±8.06	86.06±1.84	92.35±1.42	81.20±8.70	85.40±1.96	90.98±1.63
UCD w/o supervision	262.20±12.27	12.20±1.44	18.84±1.38	293.40±12.55	5.05±1.42	5.56±0.42

Table 3: Performance of ablation study.

causal parameters. Overall, UCD effectively leverages dense aggregate data together with sparse mechanism-level measurements.

Scalability with graph size. Table 2 tests scalability by increasing the number of variables from $d = 20$ to $d = 40$. Across both mechanism-level and aggregate-level evaluation, UCD achieves the best AUPRC and F1. While performance decreases for all methods as the problem dimension grows, UCD degrades only moderately and retains a consistent margin over competing approaches. These results suggest that our particle-based latent-history inference and plausibility-driven parameter learning remain effective as the underlying temporal graphs become larger.

Ablation Study We compare UCD with three ablations that isolate its main design choices. UCD **w/o particles** keeps only a single latent-history hypothesis, removing posterior multi-modality in the particle representation. UCD **w/o weight** maintains multiple particles but disables importance weighting and resampling, so implausible histories are not pruned and particle degeneracy is not corrected. UCD **w/o supervision** removes sparse mechanism-state measurements and relies only on aggregate observations, using the aggregate-consistency signal at every time step for learning.

Table 3 shows that the full UCD performs best under both mechanism-level and aggregate-level evaluations. Using a single particle substantially degrades recovery, highlighting the need to track multiple candidate decompositions of the aggregate series. Disabling weighting also hurts performance, indicating that selectively amplifying high-plausibility histories is critical for stable inference over long horizons. The largest drop occurs without sparse state supervision, confirming that aggregate observations alone are typically insufficient to disentangle simultaneously active mechanisms. Overall, the ablations support the complementary roles of (i) multi-particle latent-history inference, (ii) importance-weighted resampling, and (iii) sparse state measurements that anchor learning through residual plausibility.

Result on Real Datasets On CausalRivers, we construct an additive-mixture task by treating two time series as latent mechanisms, observing only their elementwise sum, and revealing 30% of the mechanism states. UCD achieves the best performance under both mechanism-level and aggregate-level evaluations. While absolute gains are smaller than on synthetic data (reflecting the difficulty and noise of real measurements), the consistent improvement under both protocols indicates that UCD transfers beyond controlled simulations. Full dataset construction details and ROC curves are provided in Appendix A.3.

6 Conclusion

We study causal discovery from aggregate time series generated by multiple simultaneously evolving latent mechanisms, given sparse mechanism-state measurements and a specified structural-noise family. We propose UCD, which represents uncertainty over latent mechanism histories through particle-based inference and learns mechanism-specific causal structures by evaluating the plausibility of residuals at sparse state-measurement times. This design allows aggregate observations to continuously constrain latent-state inference while sparse mechanism measurements provide direct guidance for causal learning. Experiments demonstrate that UCD consistently recovers more accurate mechanism-specific causal structures than competing approaches and remains effective as the underlying recovery problem becomes more challenging.

AI USE STATEMENT

We used generative AI tools to assist with writing and editing, including grammar correction and paragraph restructuring, to improve clarity, conciseness, and academic style. The authors reviewed all content and verified that it reflects our intended technical contributions and does not introduce unsupported claims. We take full responsibility for the final content of this work.

ETHICS STATEMENT

We do not use sensitive personal data, conduct human-subject experiments, or introduce foreseeable ethical risks. We therefore identify no specific ethical concerns that require additional reporting.

REPRODUCIBILITY STATEMENT

We include the source code for UCD in the supplementary material. All experiments were implemented in Python 3.12 with PyTorch 2.11.0. Detailed computation and hyperparameters are provided in Appendix A.8.

REFERENCES

M Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.

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

Zachary C. Brown and David Carlson. Generating hypotheses of dynamic causal graphs in neuroscience: Leveraging generative factor models of observed time series. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *International Conference on Machine Learning*, 2025.

Gustau Camps-Valls, Andreas Gerhardus, Urmi Ninad, Gherardo Varando, Georg Martius, Emili Balaguer-Ballester, Ricardo Vinuesa, Emiliano Diaz, Laure Zanna, and Jakob Runge. Discovering causal relations and equations from data. *Physics Reports*, 1044:1–68, 2023.

Yuxiao Cheng, Runzhao Yang, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. CUTS: neural causal discovery from irregular time-series data. In *International Conference on Learning Representations*, 2023.

Yuxiao Cheng, Lianglong Li, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. Cuts+: High-dimensional causal discovery from irregular time-series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11525–11533, 2024.

Gabriele D'Acunto, Paolo Di Lorenzo, and Sergio Barbarossa. Multiscale causal structure learning. *Trans. Mach. Learn. Res.*, 2023, 2023a.

Gabriele D'Acunto, Gianmarco De Francisci Morales, Paolo Bajardi, and Francesco Bonchi. Learning multiscale non-stationary causal structures. *Trans. Mach. Learn. Res.*, 2023, 2023b.

PARW ERDdS and A R&wi. On random graphs i. *Publ. Math. Debrecen*, 6(290-297):18, 1959.

Jan Eriksson and Visa Koivunen. Identifiability, separability, and uniqueness of linear ica models. *IEEE Signal Processing Letters*, 11(7):601–604, 2004.

Shanyun Gao, Raghavendra Addanki, Tong Yu, Ryan A. Rossi, and Murat Kocaoglu. Causal discovery-driven change point detection in time series. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Mohammad Emtiyaz Khan (eds.), *International Conference on Artificial Intelligence and Statistics*, pp. 3754–3762, 2025.

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

Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. In *International Conference on Learning Representations*, 2023.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11 (5):1709–1731, 2010.

Yi Jiang and Shohei Shimizu. Linkages among the foreign exchange, stock, and bond markets in japan and the united states. In *Causal Analysis Workshop Series*, pp. 1–19. PMLR, 2023.

Lingbai Kong, Wengen Li, Hanchen Yang, Yichao Zhang, Jihong Guan, and Shuigeng Zhou. Causalformer: An interpretable transformer for temporal causal discovery. *IEEE Transactions on Knowledge and Data Engineering*, 37(1):102–115, 2024.

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

Abdellah Rahmani and Pascal Frossard. Causal temporal regime structure learning. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Mohammad Emtiyaz Khan (eds.), *International Conference on Artificial Intelligence and Statistics*, pp. 4546–4554, 2025.

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

Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, 2023.

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

Gideon Stein, Maha Shadaydeh, Jan Blunk, Niklas Penzel, and Joachim Denzler. Causalrivers-scaling up benchmarking of causal discovery for real-world time-series. In *International Conference on Learning Representations*, pp. 64516–64531, 2025.

Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart. NTS-NOTEARS: learning nonparametric dbns with prior knowledge. In Francisco J. R. Ruiz, Jennifer G. Dy, and Jan-Willem van de Meent (eds.), *International Conference on Artificial Intelligence and Statistics*, pp. 1942–1964, 2023.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2021.

Sumanth Varambally, Yian Ma, and Rose Yu. Discovering mixtures of structural causal models from time series data. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *International Conference on Machine Learning*, pp. 49171–49202, 2024.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, 2018.

A Appendix

A.1 Learning Scheme

Algorithm 1 UCD: Learning Latent Mechanisms from Aggregate Time Series

Input: aggregate series $\mathbf{x}_{1:T}$; sparse complete-state measurements $\{\mathbf{z}_t^{\text{obs}} : t \in \mathcal{O}\}$; maximum lag L; a specified noise family; number of training iterations $Iter_{\text{total}}$; and hyperparameters (particle count P, mixture weight α, regularization strengths).
Notation: $k \in \{1, \dots, K\}$ (mechanism), $t \in \{1, \dots, T\}$ (time), $b \in \{1, \dots, P\}$ (particle).
Output: mechanism-specific graphs $\{\hat{\mathbf{W}}_k, \hat{\mathbf{A}}_k^{(1:L)}\}_{k=1}^K$.

```
Initialize parameters θ.
for r = 1 to Iter_total do
    Initialize particle histories {H_0^(b)}_{b=1}^P via burn-in.
    for t = 1 to T do
        if t ∉ O then
            ▷ Aggregate-only update
            Propose {z_{k,t}^(b)} and compute importance weights via Eqs. (8–11).
        else
            ▷ Measurement update + learning signal
            Compute conditional measurement likelihoods via Eqs. (13–14).
            Reweight particles by these likelihoods and clamp z_t^(b) ← z_t^obs.
            Retain pre-assimilation histories for optimizing L_state(θ).
        end if
        Resample particle histories according to their weights.
    end for
    ▷ Parameter update
    Take one optimization step on Eq. (16).
end for
Return {Ŵ_k, Â_k^(1:L)}_{k=1}^K.
```

Algorithm 1 summarizes UCD: we run a particle filter to infer latent histories from aggregate observations, use sparse complete-state measurements to clamp particles and to compute the plausibility objective, and update θ by alternating inference with optimization.

A.2 Stable Parameterization of Temporal Dynamics

To encourage stable temporal dynamics during estimation, we constrain the reduced lag coefficients via their companion matrix. Let $\mathbf{C}_k \in \mathbb{R}^{Ld \times Ld}$ act on the row-wise concatenation of the most recent L mechanism states. Its first block column contains $\mathbf{B}_k^{(1)}, \dots, \mathbf{B}_k^{(L)}$, and the remaining nonzero blocks shift stored states forward by one step:

$$\mathbf{C}_k = \begin{pmatrix} \mathbf{B}_k^{(1)} & \mathbf{I} & \cdots & \mathbf{0} \\ \mathbf{B}_k^{(2)} & \mathbf{0} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \mathbf{I} \\ \mathbf{B}_k^{(L)} & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix}. \tag{17}$$

Writing $\rho(\cdot)$ for spectral radius, we fix an upper bound $\rho_{\max} = 0.95$ and compute a contraction factor

$$\gamma_k = \frac{\rho_{\max}}{\max\{\rho(\mathbf{C}_k), \rho_{\max}\}}. \tag{18}$$

We then rescale the reduced lag coefficients and map back to the original lag matrices via

$$\mathbf{B}_k^{(\ell)} \leftarrow \gamma_k^\ell \mathbf{B}_k^{(\ell)}, \qquad \mathbf{A}_k^{(\ell)} = \mathbf{B}_k^{(\ell)}(\mathbf{I} - \mathbf{W}_k). \tag{19}$$

This construction leaves the dynamics unchanged when $\rho(\mathbf{C}_k) \le \rho_{\max}$ and otherwise contracts them. The lag-dependent scaling γ_k^ℓ multiplies every companion eigenvalue by γ_k, ensuring $\rho(\mathbf{C}_k) \le \rho_{\max}$.

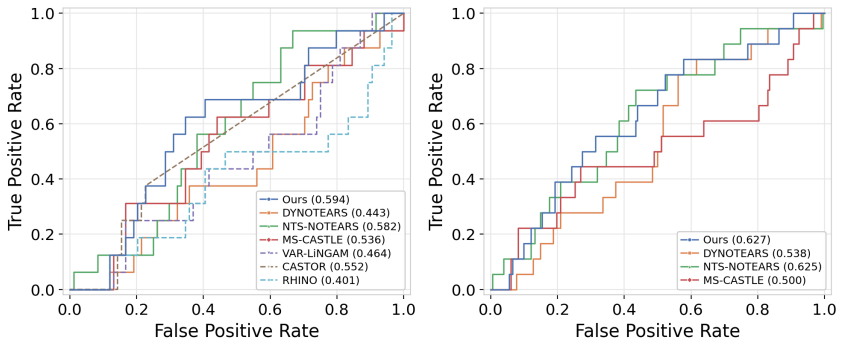

Figure 3: ROC curves on the constructed additive-mixture task from CausalRivers (*left*: match evaluation; *right*: aggregate evaluation).

Method	Particles	AUROC (↑)	AUPRC (↑)	F1 (↑)	SHD (↓)	Time (s) (↓)
UCD	16	98.07±0.41	91.31±1.18	87.90±1.27	62.00±5.06	54.0±0.0
	32	98.27±0.37	91.89±1.84	87.85±1.82	62.20±7.05	72.4±0.5
	64	98.07±0.27	92.13±1.26	89.07±1.21	56.00±4.82	106.4±0.5
	128	98.19±0.27	92.32±1.02	89.12±2.18	56.00±9.82	184.8±1.3
	256	98.33±0.37	92.74±1.54	89.53±1.67	54.00±6.13	361.8±1.0
	512	98.55±0.44	93.27±1.44	89.98±0.95	51.80±3.71	660.2±2.6
DYNOTEARS	–	96.98±0.36	86.93±1.88	83.75±1.09	85.60±4.50	110.0±7.3
NTS-NOTEARS	–	95.06±0.59	85.48±1.28	83.56±1.11	83.00±4.47	150.8±8.6
MS-CASTLE	–	57.60±1.42	6.06±0.58	13.90±1.21	656.60±33.85	15.4±1.7

Table 4: Performance and computational cost under different numbers of particles.

A.3 Result on CausalRivers

We evaluate UCD on CausalRivers using the eastern Germany subset, the provided 10-node random-subgraph split, and the first 1000 time steps. To construct a two-mechanism additive-mixture setting, we standardize each sequence per variable and then treat two sequences as latent mechanisms, observe only their elementwise sum, and randomly reveal 30% of the mechanism states.

Because CausalRivers does not provide lag-specific ground truth, we set the maximum lag to $L = 1$ for all methods, collapse recovered relations across lags into a single summary graph and evaluate against the provided causal graph. Figure 3 reports ROC curves under both aggregate-level and mechanism-level evaluation. Overall performance is modest for all methods due to the difficulty of real-world time-series discovery and the additional ambiguity introduced by additive superposition. Nevertheless, UCD achieves the strongest results under both protocols. While gains are smaller than on synthetic data (and NTS-NOTEARS is competitive under matching), the consistent improvement suggests that UCD remains effective in noisy real-world settings.

A.4 Effect of Particle Number and Computational Cost

Table 4 evaluates how the particle count P affects causal recovery and runtime. As P increases, UCD generally improves (higher AUROC/AUPRC/F1 and lower SHD), consistent with particles providing a richer approximation to the posterior over latent mechanism histories and thus better coverage of plausible decompositions of the aggregate series.

This accuracy gain comes at a clear computational cost: runtime grows roughly linearly with P. While $P = 512$ yields the best scores, the improvement over $P = 256$ is modest relative to the added runtime. We therefore use $P = 256$ by default as a practical accuracy–cost trade-off. Even with smaller P, UCD remains competitive with or better than the baselines, while avoiding the cost of the largest-particle configuration.

Noise scale	AUROC (↑)	AUPRC (↑)	F1 (↑)	SHD (↓)
0.5	97.89±0.26	91.46±0.88	87.92±1.86	61.80±8.08
1	98.33±0.37	92.74±1.54	89.53±1.67	54.00±6.13
2	98.32±0.39	92.71±1.52	88.89±1.94	58.80±7.49
4	98.24±0.17	92.02±1.44	85.67±2.05	80.60±9.39

Table 5: Sensitivity of UCD to the structural-noise scale.

A.5 Results of Noise Scale

Table 5 varies the structural-noise standard deviation in the data-generating process to assess robustness. Because UCD learns mechanism-specific noise scales jointly with the causal parameters, it can adapt to different stochasticity levels rather than assuming a fixed noise magnitude. Overall, causal recovery remains stable over a broad range of noise scales.

As noise increases, F1 decreases and SHD increases more noticeably, while AUROC and AUPRC remain comparatively stable. Since AUROC and AUPRC summarize performance across thresholds, this suggests that UCD largely preserves the *ranking* of candidate edges, whereas the optimal operating threshold shifts with the noise level.

A.6 Generic Identifiability of Latent Temporal Mechanisms

Setting and propagation representation. We study identifiability of the *generative model* itself, separate from the specific inference and optimization procedure used by UCD. Section 3 assumes mutual independence of structural-noise components. For the analysis in this appendix, we strengthen this to assume that $\epsilon^j_{k,t}$ are not only mutually independent, but also non-Gaussian. These conditions enable the Overcomplete Independent Component Analysis (OICA) argument used below.

Using the reduced coefficients $\mathbf{B}_k^{(\ell)}$ from Section 4.1, the mechanism dynamics can be written in column form as

$$\mathbf{z}_{k,t} = \sum_{\ell=1}^{L} \big(\mathbf{B}_k^{(\ell)}\big)^{\mathsf{T}} \mathbf{z}_{k,t-\ell} + (\mathbf{I} - \mathbf{W}_k)^{-\mathsf{T}} \boldsymbol{\varepsilon}_{k,t}. \tag{20}$$

For each mechanism k, define propagation matrices $\boldsymbol{\Psi}_{k,h} \in \mathbb{R}^{d\times d}$ by the recurrence

$$\begin{aligned} \boldsymbol{\Psi}_{k,0} &= (\mathbf{I} - \mathbf{W}_k)^{-\mathsf{T}}, \\ \boldsymbol{\Psi}_{k,h} &= \sum_{\ell=1}^{L} \big(\mathbf{B}_k^{(\ell)}\big)^{\mathsf{T}} \boldsymbol{\Psi}_{k,h-\ell}, \qquad h \geq 1, \end{aligned} \tag{21}$$

We adopt the zero-padding convention $\boldsymbol{\Psi}_{k,h} = \mathbf{0}$ for $h < 0$. The propagation vector associated with source (k, j) at horizon $h \geq 0$ is $\boldsymbol{\psi}_{k,j,h} = \boldsymbol{\Psi}_{k,h}[:, j]$, yielding the trajectory $\boldsymbol{\psi}_{k,j,0}, \boldsymbol{\psi}_{k,j,1}, \ldots, \boldsymbol{\psi}_{k,j,\tau}$.

Iterating Eq. (20) over a finite horizon τ yields the truncated OICA-style representation

$$\mathbf{x}_t = \mathcal{R}_t^{(\tau)} + \sum_{k=1}^{K} \sum_{j=1}^{d} \sum_{h=0}^{\tau} \boldsymbol{\psi}_{k,j,h}\, \epsilon^j_{k,t-h}, \tag{22}$$

where $\mathcal{R}_t^{(\tau)}$ collects contributions originating before the window. Our arguments below are conditional on either conditioning on $\mathcal{R}_t^{(\tau)}$ or choosing τ large enough that this term is negligible; under stable dynamics, contributions from sufficiently remote states decay. We therefore do not interpret Eq. (22) without $\mathcal{R}_t^{(\tau)}$ as an exact, unconditional truncation of the infinite-history process.

Under independent non-Gaussian sources, linear ICA implies that the mixing structure in Eq. (22) is identifiable up to permutation and scaling (Eriksson & Koivunen, 2004). We focus on a *complete, non-degenerate* setting in which all $Kd(\tau + 1)$ finite-window propagation columns are present, nonzero, and pairwise non-collinear.

The permutation analysis below additionally assumes an OICA normalization that resolves the scaling ambiguity needed to compare propagation columns. We take this normalization as given, and focus on the remaining temporal and mechanism-level permutation ambiguities.

Single-source recurrence and horizon ordering. Fix a single structural-noise source (k, j) and suppose that the corresponding $\tau + 1$ *normalized* propagation columns have been isolated from the OICA output. The remaining ambiguity is their *ordering* over horizons. Suppressing (k, j) in the notation, write the (unknownly ordered) trajectory as $\boldsymbol{\psi}_0, \ldots, \boldsymbol{\psi}_\tau$. For $h = 1, \ldots, \tau$, define the stacked lag vector $\mathbf{u}_h = (\boldsymbol{\psi}_{h-1}^\mathsf{T}, \ldots, \boldsymbol{\psi}_{h-L}^\mathsf{T})^\mathsf{T}$, and collect $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_\tau]$ and $\mathbf{V} = [\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_\tau]$. The propagation recurrence can then be written as

$$\mathbf{V} = \mathbf{M}_k \mathbf{U}. \tag{23}$$

Theorem A.1 (Uniqueness of the single-source recurrence)**.** *Given an ordered source trajectory satisfying Eq. (23), the recurrence matrix $\mathbf{M}_k$ is unique if and only if*

$$\operatorname{rank}(\mathbf{U}) = dL. \tag{24}$$

Proof. If $\mathbf{M}_1\mathbf{U} = \mathbf{M}_2\mathbf{U}$ and $\operatorname{rank}(\mathbf{U}) = dL$, then $\mathbf{U}$ has a right inverse, hence

$$(\mathbf{M}_1 - \mathbf{M}_2)\mathbf{U} = \mathbf{0} \quad \Longrightarrow \quad \mathbf{M}_1 = \mathbf{M}_2. \tag{25}$$

Conversely, if $\operatorname{rank}(\mathbf{U}) < dL$, there exists a nonzero matrix $\mathbf{D}$ with $\mathbf{D}\mathbf{U} = \mathbf{0}$, so $(\mathbf{M} + \mathbf{D})\mathbf{U} = \mathbf{M}\mathbf{U}$ and the recurrence is not unique. □

Since $\mathbf{U} \in \mathbb{R}^{dL \times \tau}$, full row rank requires $\tau \geq dL$. We impose the stronger condition $\tau > dL$ so that additional horizons remain after the recurrence has been identified.

Let $\pi : \{0, \ldots, \tau\} \to \{0, \ldots, \tau\}$ be a candidate permutation of horizon indices, where $\pi(h)$ is the original horizon assigned to position h in the candidate ordering. The reordered trajectory is

$$\bar{\boldsymbol{\psi}}_h = \boldsymbol{\psi}_{\pi(h)}, \qquad h = 0, \ldots, \tau. \tag{26}$$

Construct $\mathbf{U}_\pi$ and $\mathbf{V}_\pi$ from $\{\bar{\boldsymbol{\psi}}_h\}$ as above. We call π *admissible* if it yields a uniquely determined recurrence, i.e.,

$$\operatorname{rank}(\mathbf{U}_\pi) = dL. \tag{27}$$

For any admissible ordering, dL independent propagation columns uniquely determine the recurrence matrix. When $\tau > dL$, the remaining horizons impose additional recurrence constraints. These constraints are algebraic in the causal parameters: under an acyclic instantaneous model, $(\mathbf{I} - \mathbf{W}_k)^{-1}$ admits a finite polynomial expansion, and therefore each finite-horizon propagation vector is an algebraic function of the model parameters. An incorrect ordering can satisfy all constraints generically only if the constraints hold identically over the admissible parameter space; otherwise, the exceptional parameter values lie in a Lebesgue measure-zero algebraic set. It therefore suffices to consider only structural horizon symmetries.

Structural horizon symmetries. Consider the formal rescaling

$$\mathbf{B}_k^{(\ell)} \longmapsto t^\ell \mathbf{B}_k^{(\ell)}, \qquad \ell = 1, \ldots, L, \tag{28}$$

with $\mathbf{W}_k$ fixed. From Eq. (21), an induction on h gives the induced horizon-wise scaling

$$\boldsymbol{\psi}_{k,j,h} \longmapsto t^h \boldsymbol{\psi}_{k,j,h}. \tag{29}$$

The auxiliary variable t serves only as a bookkeeping device for algebraic degree; it is not a model parameter. Under this rescaling, the horizon index is encoded by the power of t carried by each propagation vector. Consequently, any horizon permutation that preserves the recurrence *identically* over the admissible parameter space must preserve these degree relations. In particular, the coefficient playing the lag-one role must have the same scaling degree at every step.

In the complete non-degenerate setting, $\boldsymbol{\Psi}_{k,1} \neq \mathbf{0}$. Since $\boldsymbol{\Psi}_{k,0} = (\mathbf{I} - \mathbf{W}_k)^{-\mathsf{T}}$ is invertible, the identity $\boldsymbol{\Psi}_{k,1} = (\mathbf{B}_k^{(1)})^\mathsf{T} \boldsymbol{\Psi}_{k,0}$ implies $\mathbf{B}_k^{(1)} \neq \mathbf{0}$. A structural horizon symmetry must therefore reuse the same lag-one recurrence coefficient at every applicable step, which forces a constant first difference

$$\pi(h) - \pi(h-1) = a, \tag{30}$$

for some integer a independent of h.

Theorem A.2 (Structural horizon symmetries)**.** *In the complete non-degenerate setting, any structural horizon permutation must be either*

$$\pi(h) = h \qquad \textit{or} \qquad \pi(h) = \tau - h. \tag{31}$$

Proof. Eq. (30) implies $\pi(h) = ah+b$. Since π is a bijection from $\{0, \dots, \tau\}$ onto itself, necessarily $|a| = 1$, giving the identity and reversal. □

Ruling out reversal via startup constraints ($L > 1$). The remaining question is whether reversal is compatible with the causal startup boundary. For compactness, define $\mathbf{D}_\ell = (\mathbf{B}_k^{(\ell)})^\mathsf{T}$. If $\mathbf{D}_L$ is invertible, then the internally reversed recurrence has coefficients

$$\bar{\mathbf{D}}_r = -\mathbf{D}_L^{-1}\mathbf{D}_{L-r}, \quad r = 1, \dots, L-1, \qquad \bar{\mathbf{D}}_L = \mathbf{D}_L^{-1}. \tag{32}$$

Thus, the internal recurrence alone does not always distinguish forward and reversed trajectories.

A valid causal propagation trajectory must also satisfy the startup equations at horizons $r = 1, \dots, L-1$, where only lags $\ell \le r$ can contribute:

$$\boldsymbol{\psi}_{k,j,r} = \sum_{\ell=1}^{r} \big(\mathbf{B}_k^{(\ell)}\big)^\mathsf{T} \boldsymbol{\psi}_{k,j,r-\ell}, \qquad r = 1, \dots, L-1. \tag{33}$$

For $L = 1$ there are no startup constraints, so a reverse ambiguity may remain.

For $L > 1$, reversal is *not* a structural symmetry of the full causal trajectory. To see this, we exhibit a parameter point where the reversed startup constraint fails, and then appeal to a small-perturbation argument to reach the complete non-degenerate setting.

Take $\mathbf{W}_k = \mathbf{0}$ and $\mathbf{B}_k^{(\ell)} = a_\ell \mathbf{I}$ with $a_\ell > 0$ for $\ell = 1, \dots, L$, choosing a_ℓ sufficiently small to ensure stability. For each source j, let $\mathbf{e}_j \in \mathbb{R}^d$ be the jth standard basis vector. Then

$$\boldsymbol{\psi}_{k,j,h} = p_h \mathbf{e}_j, \qquad h \ge 0, \tag{34}$$

where $p_0 = 1$ and, for $h \ge 1$,

$$p_h = \sum_{\ell=1}^{\min\{L,h\}} a_\ell p_{h-\ell}. \tag{35}$$

Since $a_1 > 0$, induction gives $p_h > 0$ for all $h \ge 0$.

Here $\mathbf{D}_\ell = a_\ell \mathbf{I}$, so the lag-one coefficient of the reversed recurrence is

$$\bar{\mathbf{D}}_1 = -\mathbf{D}_L^{-1}\mathbf{D}_{L-1} = -\frac{a_{L-1}}{a_L}\mathbf{I}. \tag{36}$$

For the reversed trajectory, $\bar{\boldsymbol{\psi}}_0 = \boldsymbol{\psi}_{k,j,\tau}$ and $\bar{\boldsymbol{\psi}}_1 = \boldsymbol{\psi}_{k,j,\tau-1}$. The first startup equation for the reversed trajectory would require

$$\boldsymbol{\psi}_{k,j,\tau-1} = \bar{\mathbf{D}}_1 \boldsymbol{\psi}_{k,j,\tau} = -\frac{a_{L-1}}{a_L}\boldsymbol{\psi}_{k,j,\tau}, \tag{37}$$

which is impossible because $p_{\tau-1}$, p_τ, and a_{L-1}/a_L are positive.

This construction shows that the reversed startup constraint is not an algebraic identity; hence, by the same algebraic measure-zero argument, it can hold only on a Lebesgue measure-zero subset of the causal parameter space. Moreover, the startup failure persists under sufficiently small perturbations of the coefficients. A generic small perturbation can be chosen within the complete non-degenerate parameter set so that (i) finite-horizon propagation columns are nonzero and pairwise non-collinear and (ii) stability and invertibility of $\mathbf{D}_L$ are preserved. Therefore, reversal is not a structural horizon symmetry when $L > 1$.

Theorem A.3 (Generic identifiability of a single-source temporal trajectory)**.** *Consider a normalized set of propagation columns known to belong to one structural-noise source in the complete non-degenerate setting. Suppose*

$$\operatorname{rank}(\mathbf{U}) = dL, \qquad \tau > dL, \qquad L > 1. \tag{38}$$

Then, outside a Lebesgue measure-zero subset of the continuous parameter space, the true horizon ordering is the unique admissible ordering satisfying both the temporal recurrence and the startup equations. Consequently, the ordered source trajectory generically determines a unique dynamic signature $\mathbf{M}_k$.

Proof. By Theorem A.1, the correctly ordered trajectory uniquely determines $\mathbf{M}_k$. Since $\tau > dL$, any incorrect admissible ordering must also satisfy additional recurrence constraints; unless these constraints arise from a structural symmetry, they hold only on a measure-zero subset of the parameter space. By Theorem A.2, the only nontrivial structural candidate is reversal, and the argument above rules it out when $L > 1$. Hence the true ordering is generically unique. □

Mechanism grouping and causal recovery. All d structural-noise sources belonging to mechanism k share the same dynamic signature $\mathbf{M}_k$. To make this signature sufficient for grouping, assume distinct mechanisms have distinct signatures:

$$\mathbf{M}_k \neq \mathbf{M}_q, \qquad k \neq q. \tag{39}$$

Then, once single-source trajectories are identified, equality of recovered signatures yields a unique partition into K mechanism groups, up to permutation of mechanism labels.

Within each recovered mechanism, the d sources may still be permuted across observed variables. This within-mechanism permutation is resolved from the horizon-zero propagation columns using the standard ICA-LiNGAM permutation argument (Shimizu et al., 2006). After this alignment,

$$\mathbf{\Psi}_{k,0} = [\boldsymbol{\psi}_{k,1,0} \quad \cdots \quad \boldsymbol{\psi}_{k,d,0}] = (\mathbf{I} - \mathbf{W}_k)^{-\mathsf{T}}. \tag{40}$$

Therefore,

$$\mathbf{W}_k = \mathbf{I} - \mathbf{\Psi}_{k,0}^{-\mathsf{T}}. \tag{41}$$

The recovered signature yields $\mathbf{B}_k^{(1)}, \ldots, \mathbf{B}_k^{(L)}$, and the structural lag coefficients follow from

$$\mathbf{A}_k^{(\ell)} = \mathbf{B}_k^{(\ell)}(\mathbf{I} - \mathbf{W}_k), \qquad \ell = 1, \ldots, L. \tag{42}$$

Theorem A.4 (Generic recovery from identified source trajectories)**.** *Assume the complete non-degenerate OICA setting, the OICA scaling normalization specified above, and the conditions of Theorem A.3 for every identified source trajectory. If Eq. (39) holds, then conditional on identifying the single-source trajectories, the mechanism-specific causal parameters*

$$\left\{\mathbf{W}_k, \mathbf{A}_k^{(1)}, \ldots, \mathbf{A}_k^{(L)}\right\}_{k=1}^{K} \tag{43}$$

are generically unique up to permutation of the K latent mechanisms.

Proof. Each identified source trajectory generically determines a unique dynamic signature. Condition equation 39 then yields a unique grouping of sources by mechanism, up to relabeling. Within each recovered mechanism, the remaining permutation of its d sources is uniquely resolved from the horizon-zero propagation columns via the ICA-LiNGAM permutation argument, and the same ordering applies to all horizons. Eq. (41) then determines the instantaneous matrix, and Eq. (42) determines all lagged matrices. Hence the collection of mechanism-specific temporal causal models is generically unique up to permutation of mechanism labels. □

A.7 Limitation to nonlinear mechanisms

Our identifiability arguments rely on the linear propagation structure and do not directly extend to nonlinear mechanisms. The core difficulty is a reparameterization ambiguity between the latent decomposition and the nonlinear structural functions. For example, suppose two latent mechanisms both contribute to an observed coordinate Y:

$$x_t^Y = z_{1,t}^Y + z_{2,t}^Y. \tag{44}$$

If Y has no instantaneous or lagged children in either mechanism, then for any transfer term δ_t we can define an alternative decomposition

$$\tilde{z}_{1,t}^Y = z_{1,t}^Y - \delta_t, \qquad \tilde{z}_{2,t}^Y = z_{2,t}^Y + \delta_t, \tag{45}$$

which leaves the aggregate observation x_t^Y unchanged and does not affect any other variable. With unrestricted nonlinear structural functions, this transferred mass can be absorbed by changing parent–child functional relationships. As a result, different mechanism-specific graphs (e.g., $X \to Y$ versus $Z \to Y$) can yield observationally equivalent aggregate processes. Establishing nonlinear identifiability therefore requires additional restrictions (e.g., on the function class, interventions, or constraints on the latent decomposition); we leave such extensions to future work and focus on the linear setting.

A.8 Computation and Hyperparameters

Experiments were run on Ubuntu 22.04 LTS servers. CPU experiments used Intel Platinum 8468V processors and GPU experiments used NVIDIA RTX 3090 GPUs. For each method, we performed a parameter search and report the best-performing configuration below; unspecified hyperparameters follow the original implementations.

- UCD
  - Adam learning rate: 0.015.
  - Acyclicity penalty: $\lambda_{\mathrm{dag}} = 1.0$.
  - Sparsity penalties: $\lambda_W = 0.6$ (instantaneous) and $\lambda_A = 0.3$ (lagged).
  - Plausibility weight: $\lambda_{\mathrm{state}} = 0.01$.
  - Particle inference: $P = 256$, $\alpha = 0.1$, burn-in 300 steps.
  - Alternating optimization: 40 iterations.
  - Stability bound: $\rho_{\max} = 0.95$.
  - Pruning threshold: 0.08.
- DYNOTEARS
  - Sparsity penalties: $\lambda_w = \lambda_a = 0.03$.
  - Augmented Lagrangian: up to 100 outer rounds; inner iterations start at 100 and increase by 25 per round up to 3000.
  - Pruning threshold: 0.06.
- NTS-NOTEARS
  - Sparsity penalties: $\lambda_1 = 0.001$ and $\lambda_w = 0.02$.
  - Augmented Lagrangian: up to 100 outer rounds; inner iterations start at 100 and increase by 25 per round up to 3000.
  - Pruning threshold: 0.28.
- MS-CASTLE
  - ℓ_1 coefficient: 0.03.
  - Solver iterations: 100.
  - Pruning threshold: 0.02.
- VAR-LiNGAM
  - Lag order: selected by Akaike information criterion.
  - Pruning threshold: 0.05.
- CASTOR
  - Sparsity penalties: $\lambda_w = \lambda_a = 0.05$.
  - Pruning threshold: 0.03.
- Rhino
  - Training: 300 epochs, mini-batch size 128 time fragments.
  - Sparsity prior: set $\gamma/(T-L) = 6.5 \times 10^{-5}$ for all sequence lengths.
  - Pruning threshold: 0.55.