SEQ2CAUSE: A UNIFIED FRAMEWORK FOR CAUSAL DISCOVERY IN EVENT SEQUENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Complex systems such as vehicles, patients and genomes emit discrete event sequences whose operative question is causal, not predictive: which events cause which other events, and which cause higher-level outcomes such as failures or diseases? This question decomposes along two axes — dependency type (event $\rightarrow$ event vs. event $\rightarrow$ outcome) and causal scope (single sequence vs. population) — yielding four structurally distinct regimes with different identifiability conditions. No existing method addresses more than one, because all assume multi-stream structure with low vocabulary, and none scales beyond a few hundred event types. We present SEQ2CAUSE, a unified framework that resolves all four regimes through a single shared primitive: a pretrained autoregressive model repurposed as an amortized conditional independence testing engine requiring no task-specific retraining. We establish a prediction–causality duality: the model's excess cross-entropy simultaneously bounds causal identification error across all four regimes, so that every improvement in next-token prediction tightens causal guarantees for free. On nonlinear SCMs (vocabularies up to $8,000$ types) and real-world vehicle diagnostic logs (29K event types, 474 failure outcomes), SEQ2CAUSE is the first method to populate all four regimes at scale with a single frozen backbone. Existing methods are either inapplicable, inaccurate, or computationally intractable in this setting.

1 INTRODUCTION

Discrete event sequences are the canonical data modality of modern complex systems. A vehicle emits hundreds of diagnostic trouble codes before a failure Pirasteh et al. (2019); Math et al. (2025a); Math & Lienhart (2026). A patient's trajectory through a hospital unfolds as a sequence of interventions, diagnoses, and observations Li et al. (2021); He et al. (2022). A genome expresses itself as an ordered chain of regulatory events Pratapa et al. (2020); Shen et al. (2023); Theodoris et al. (2023); Cui et al. (2024). In each of these domains, the operative scientific question is not predictive but causal: which events cause which other events, and which events cause higher-level outcomes such as failures, diseases, or phenotypes?

This question decomposes into four structurally distinct regimes (Tab. 1), defined by two orthogonal axes. The first is *dependency type*: one may seek event-to-event structure ($\mathcal{X} \rightarrow \mathcal{X}$), how events cause

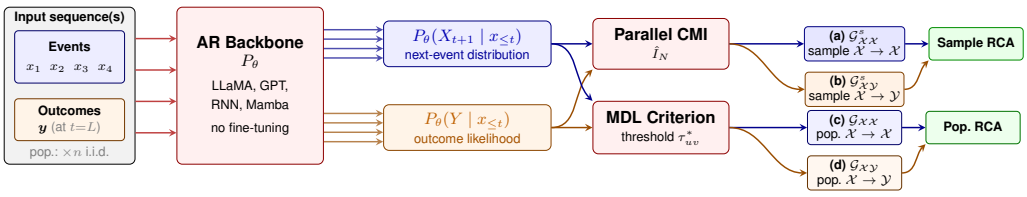
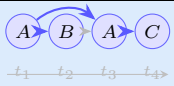
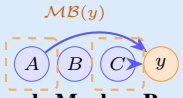
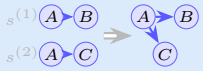
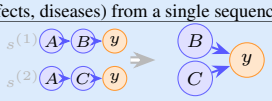


Figure 1: **SEQ2CAUSE: one frozen backbone, four causal regimes.** Any pretrained AR model (LLaMA, GPT, Mamba) is repurposed via teacher forcing as a parallelised causal discovery engine on GPUs, with no fine-tuning. Two heads share the same backbone: the event conditional (blue) and the outcome likelihood (orange). Using a single sequence, the sample-level graphs (a, b) follow from CMI testing; for multiple sequences, population-level graphs (c, d) are obtained from Minimum Description Length (MDL) Grunwald (2004). Their union enables root-cause identification.

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.

Table 1: **The four regimes of causal discovery in discrete event sequences.** Rows separate causal scope: a single observed sequence (sample-level) vs. an entire dataset (population-level). Columns separate dependency type — event-to-event ($\mathcal{X} \to \mathcal{X}$, blue) vs. event-to-outcome ($\mathcal{X} \to \mathcal{Y}$, orange). These four cells have different identifiability conditions, different estimands, and different failure modes. No existing method addresses more than one at scale.

	Event-to-Event: $\mathcal{X} \to \mathcal{X}$	**Events-to-Outcomes: $\mathcal{X} \to \mathcal{Y}$**
Sample-Level (Single Sequence)	 **(a) Sample Time Graph Recovery** Recover the time causal graph specific to a single sequence of events with lagged effects.	 **(b) Sample Markov Boundary** Identify the direct causes of each outcome y (e.g., defects, diseases) from a single sequence.
Population-Level (Full Dataset)	 **(c) Population Summary Graph** Summary causal graph over event types $\mathcal{X}$: *which event caused other events across samples?*	 **(d) Population-Markov Boundary** Identify the Causes of outcomes aggregated across sequences: *What events lead to this product defect?*

other events, or event-to-outcome structure ($\mathcal{X} \to \mathcal{Y}$), how events cause final outcomes. The second is *causal scope*: inference may target a single observed sequence (sample-level), where the causal object is specific to one sample, or an entire dataset (population-level), where the goal is a type-level summary graph generalizable across sequences. These four regimes are not interchangeable: the sample-level question "did 'Fire@t1' cause 'Smoke@t2'?" and the population-level question "does 'Fire' cause 'Smoke' in general?" have different identifiability conditions, different estimands, and different failure modes. Surprisingly, no existing framework addresses more than one Assaad et al. (2022); Gong et al. (2024).

The dominant paradigm operates on $|\mathcal{X}|$ parallel streams, each assigned to a distinct entity, a design structurally mismatched with the single-stream, high-vocabulary setting we target (Sec. 2 details why Granger-, Hawkes-, and constraint-based methods, as well as classical Markov boundary algorithms, all fail to scale). The single-stream, high-vocabulary regime is thus left entirely unaddressed.

The key observation underlying this work is that any pretrained autoregressive (AR) model Sutskever et al. (2014); Touvron et al. (2023) for event sequences, whether a lightweight RNN or a frontier foundation model, already computes everything causal discovery requires. Trained by next-token prediction, it implicitly estimates the full joint distribution via the chain rule $\prod_t P(x_t|x_{<t})$; these conditionals are *precisely* the quantities required to evaluate conditional mutual information (CMI) and perform conditional independence (CI) testing. No additional causal-specific model or retraining is required: causal discovery is read directly off the next-token probabilities the AR model already produces. Given a pretrained AR model with excess cross-entropy ϵ, causal discovery reduces to a batched forward pass spanning all four regimes with identification error vanishing at rate $\epsilon \to 0$. **Every frontier-scale next-token checkpoint is, for free, a causal discovery engine**. We introduce SEQ2CAUSE[1], the first method to populate all four cells simultaneously at scale. SEQ2CAUSE repurposes any pretrained AR model, without fine-tuning for parallelized causal discovery on GPUs. It recovers the MB of each outcome ($\mathcal{X} \to \mathcal{Y}$) and the causal graph over event types ($\mathcal{X} \to X$), from a single sequence at inference time (*sample-level*) or aggregated across a dataset (*population-level*), with guarantees derived directly from the model's pretraining quality ϵ. Our contributions are:

- **A four-regime taxonomy and its unification**. We formalize causal discovery in discrete event sequences with large vocabularies as a 2×2 problem defined by dependency type ($\mathcal{X} \to \mathcal{X}$ vs. $\mathcal{X} \to \mathcal{Y}$) and causal scope (sample-level vs. population-level). We establish that a single pretrained backbone serves all four cells without task-specific adaptation.
- **A prediction–causality duality**. We prove that the AR model's excess cross-entropy ϵ simultaneously controls (i) CMI estimation error (Theorem 1), (ii) sample-time graph soundness (Theorem 2), and (iii) MDL population edge-inclusion error (Lem. 1). Causal

[1]Code and unified evaluation will be released as the *seq2cause* Python package upon acceptance.

identification across all four regimes is therefore a function of pretraining quality, with no regime-specific calibration.

- **Empirical validation across all four regimes at scale.** On nonlinear SCMs ($|\mathcal{X}| \in [100, 8000]$) and real-world vehicle diagnostic sequences ($|\mathcal{X}| = 29,100, |\mathcal{Y}| = 474$), SEQ2CAUSE is the first method to populate all four cells of Tab 1 with a single frozen backbone where all existing methods are either inaccurate, or computationally intractable.

2 Related Work

Causal discovery in sequence of discrete events splits into two structurally distinct regimes that are not interchangeable. For a full overview, see Appendix E.

Multi-stream (Standard). The dominant paradigm operates on $|\mathcal{X}|$ parallel streams, each assigned to a distinct entity (sensor, user). The causal target is a population-level summary graph $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ estimated from $T \to \infty$ observations per stream. Methods in this family, such as Granger-based Granger (1969); Nauta et al. (2019); Zhang et al. (2020), constraint-based Runge et al. (2019), functional Hyvärinen et al. (2010), optimization-based Pamfil et al. (2020), and neural point processes Zuo et al. (2020); Qiao et al. (2023); Cüppers et al. (2024) scale quadratically or worse with the number of events $|\mathcal{X}|$. For Hawkes-based methods specifically, the excitation matrix $\Phi \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ carries $|\mathcal{X}|^2$ parameters, and its maximum likelihood estimation scales quadratically with the event types or worse Laub et al. (2025), rendering the entire family computationally intractable at $|\mathcal{X}| > 10^3$. We explain this theoretically in Appendix B.1.

Single-stream (Sample- and Population-level). The general setting is fundamentally different. By analogy with NLP, a single sentence over a large vocabulary is the sample; a corpus is the population. We observe n i.i.d sequences $\{s^{(k)}\}_{k=1}^{n}$ over a shared massive vocabulary $|\mathcal{X}| > 1000$, this is the population-level regime. If during inference only one realization $s = (x_0, \cdots, x_L)$ is available, this is the sample-level regime. The causal target decomposes into two subproblems: recovering event-to-event structure and event-to-outcomes structure (Tab. 1). This regime is strictly harder than multi-stream for two compounding reasons. First, the co-occurrence frequency of any pair $(u, v) \in \mathcal{X}^2$ within a single sequence of moderate length L satisfies:

$$\mathbb{E}\left[\#(u, v \text{ co-occur in } s)\right] = \mathcal{O}\left(\frac{L^2}{|\mathcal{X}|^2}\right) \xrightarrow{|\mathcal{X}|\to\infty} 0 \tag{1}$$

making empirical joint distributions degenerate. Second, there is no long stationary trajectory to extract a population from: one vehicle trip, one patient visit, one genome sequence is the observation of interest. Sliding-window arguments that justify multi-stream methods do not apply here.

Multi-label Causal Discovery. Multi-label causal discovery seeks to identify the Markov Boundary $\mathcal{MB}(Y_j)$ of each outcome label: its minimal conditioning set rendering Y_j independent of all other variables Tsamardinos & Aliferis (2003). Classical local structure learning (LSL) algorithms estimate these boundaries directly: IAMB Tsamardinos et al. (2003), CMB Gao & Ji (2015), MB-by-MB Wang et al. (2014), PCDbyPCD Yin et al. (2008) and MI-MCF Ma et al. (2025) all recover the $\mathcal{MB}$ via iterative CI testing. Their extension to event sequences is obstructed by dimensionality ($|\mathcal{X}|^2$ candidate pairs) and distributional assumptions that rarely hold. Critically, none of these methods addresses the $\mathcal{X} \to \mathcal{Y}$ dependencies in the single-stream, high-vocabulary setting, nor do they handle $\mathcal{X} \to \mathcal{X}$ discovery jointly (see Appendix, Tab 4). Similarly, existing RCA methods for event sequences (identifying primary causes of an outcome) either require a known system topology Yao et al. (2024), operate on continuous multivariate streams Han et al. (2025), or reduce root causes to anomaly scores without a causal graph Assaad et al. (2023), crucially no work at the sample-level contrary to SEQ2CAUSE with the Cor. 1.

3 Background and Problem Formulation

Data-Generating Process. We model the joint generation of events and outcomes as a stochastic process over a shared time index $t \in \mathbb{N}$ (Appendix, Fig. 4). Notation is described in Appendix A.

Let $\{X_t\}_{t\geq 0}$ be a stochastic process taking values in a finite discrete alphabet $\mathcal{X}$. We make no finite-order Markov assumption; the full conditional $P(X_t|X_{<t})$ may depend on the entire history — a great departure from standard methods Gong et al. (2024). A realization of length $L+1$ is denoted $s = (x_0, \ldots, x_L) \in \mathcal{X}^{L+1}$. Let $\mathcal{Y} = \{Y_1, \ldots, Y_c\}$ be a set of discrete outcome variables. Each binary outcome Y_j is generated conditionally on the event history as $P(Y_j \mid X_{<t})$ for c outcomes. Critically, the event chain Eq. equation 2 is autonomous: it does not depend on $\mathcal{Y}$. We assume that outcomes are solely driven by the events (no inter-outcome effects Wu et al. (2020b)). The joint distribution therefore factorizes as:

$$P(X_{0:L}, Y_{1:n}) \;=\; \prod_{t=0}^{L} P(X_t \mid X_{<t}) \prod_{j=1}^{c} P(Y_j \mid X_{<L})\,, \tag{2}$$

Remark 1 (Two-chain DGP and Markovianity)**.** *The two-chain structure defines a natural edge partition. Intra-chain edges $X_{t-\ell} \to X_t$ describe the event-to-event time causal graph $\mathcal{G}^s_{\mathcal{X}\mathcal{X}}$; cross-chain edges $X_t \to Y_j$ form the event-to-outcome subgraph $\mathcal{G}^s_{\mathcal{X}\mathcal{Y}}$. No edges of the form $Y_j \to X_t$ or $Y_j \to Y_k$ exist, thus the joint sample-level graph $\mathcal{G}^s$ is a DAG by construction. The sample-level regime of Fig. 4 corresponds exactly to recovering subgraphs at different levels of aggregation.*

Sample-Level Causal Graphs. In several domains, a practitioner has access to a single sequence at inference time. Population-level aggregation is either unavailable or scientifically undesirable, as the causal structure of this sequence not the average across a cohort, is the object of interest. We define the causal graph types following the literature Assaad et al. (2022):

Definition 1 (Sample-Time Causal Graph)**.** *Let s be a realization of the event chain in Eq. equation 2. The Sample-Time Causal Graph $\mathcal{G}^s_{\mathcal{X}\mathcal{X}} = (\mathcal{T}, \mathcal{E}_t)$ is a DAG where nodes $\mathcal{T} = \{0, \ldots, L\}$ are time steps. A directed edge $(t-\ell) \to t$ exists in $\mathcal{E}_t$ if and only if the realization of the event at $t-\ell$ is a direct cause of the event at t. $\mathcal{G}^s_{\mathcal{X}\mathcal{X}}$ answers: did $Fire@t_1 \to Smoke@t_2$ at the sample level.*

Definition 2 (Sample-Markov Boundary Graph)**.** *The Sample Markov Boundary Graph $\mathcal{G}^s_{\mathcal{X}\mathcal{Y}} = (\mathcal{T} \cup \mathcal{Y}, \mathcal{E}_y)$ is a bipartite DAG. Edge $t \to Y_j$ exists in $\mathcal{E}_y$ if and only if the event realization x_t causes Y_j. Under Eq. equation 2, Y_j has neither children nor spouses, so $\mathcal{MB}(Y_j) = \mathrm{Pa}_{\mathcal{G}^s_{\mathcal{X}\mathcal{Y}}}(Y_j)$.*

4 The seq2cause Framework

4.1 Neural Autoregressive Density Estimation

To learn the described DGP we employ two AR models Math et al. (2025a) namely the event and outcome models $\mathrm{f}_{\theta_x}, \mathrm{f}_{\theta_y}$ respectively with parameters θ_x, θ_y[2]. They are trained once via cross-entropy minimization. They both map the trajectory $x_{<t}$ to a distribution over the next event type and outcomes:

$$P_{\theta_x}(X_t \mid X_{<t}) = \; \mathrm{Softmax}(\mathrm{f}_{\theta_x}(x_{<t})) \tag{3}$$

$$P_{\theta_y}(Y_j \mid X_{<t}) = [\mathrm{Sigmoid}\big(\mathrm{f}_{\theta_y}(\mathrm{f}_{\theta_x}(x_{<t}))\big)]_j, \quad j = 1, \ldots, c. \tag{4}$$

Crucially, the likelihood of Y_j is estimated at every step t even though it is only realized at $t = L$; this forces the model to recover the outcome likelihood from partial histories Math et al. (2025a), which is precisely the sequential conditional probabilities required by the described CI-test in §4.3.

4.2 Assumptions and ϵ-Oracle Model

seq2cause assumes no parametric form for the DGP, above the expressivity of the AR estimator. Beyond temporal precedence (A2) and causal sufficiency (A3), two assumptions distinguish our framework from prior work Assaad et al. (2022):

Assumption 1 (ϵ-Oracle Model)**.** *The AR model f_θ trained by maximum likelihood on data from P satisfies:*

$$D_{\mathrm{KL}}\big(P(X_t \mid X_{<t}) \,\|\, P_\theta(X_t \mid X_{<t})\big) \leq \epsilon \quad \forall\, t. \tag{5}$$

[2] f_{θ_y} is stacked on top of the event model which serves as a backbone containing the majority of the parameters.

Assumption 1 does not presuppose a perfect model: it states that whatever quality the backbone achieves, the causal identification error is bounded by that same quantity (Thm. 1, Cor. 2). Empirically, Fig. 2a confirms that structural recovery succeeds well before full convergence ($\epsilon \approx 0.1$ suffices). No oracle is assumed; only that the practitioner can monitor the gap (Appendix H.2). Secondly, standard faithfulness Spirtes & Glymour (1991) requires CI-tests to return exact zeros — unrealistic under imperfect density estimation. We instead adopt strong faithfulness Uhler et al. (2013): valid causal associations must exceed a significance threshold τ. We provide a robustness analysis to assumption violation in Appendix C.2.

4.3 Sample-Level: The Universal CMI Primitive

Both the $\mathcal{X} \to \mathcal{X}$ and $\mathcal{X} \to \mathcal{Y}$ regimes reduce to a single statistical primitive: testing whether a past event carries information about a target variable (event or outcome), given the observed trajectory. Event types and outcome labels both belong to a shared universe $\mathcal{X} \cup \mathcal{Y}$, and the target is simply an indicator over an element of this universe: let $T \in \{E_{t+1}\} \cup \{Y_1, \ldots, Y_c\}$, where $E_{t+1} \triangleq \mathbf{1}[X_{t+1} = x_{t+1}]$ is the binary indicator that the next event equals a specific element of $\mathcal{X}$, and $Y_j \in \mathcal{Y}$ is the j-th outcome variable. Nodes in $\mathcal{G}^s$ are indexed by these binary random variables; an edge $E_{t-\ell} \to E_t$ denotes a direct causal effect between the corresponding event realizations[3]. T is thus always binary, regardless of whether it indexes an event in $\mathcal{X}$ or an outcome in $\mathcal{Y}$.

Conditional Mutual Information. In a sequence, we measure the remaining uncertainty of a target T when observing event E_t given $X_{<t} \triangleq \{X_0, \cdots, X_{t-1}\}$:

$$I(T,\, E_t \mid X_{<t}) \;\triangleq\; H(T \mid X_{<t}) \;-\; H(T \mid E_t,\, X_{<t}) \;=\; \mathbb{E}_{e_t, x_{<t}}[I_G(T,\, e_t \mid x_{<t})]\,, \tag{6}$$

where $I_G(T, e_t \mid x_{<t}) = D_{\mathrm{KL}}\big(P(T \mid e_t, x_{<t}) \,\|\, P(T \mid x_{<t})\big)$ is the information gain Quinlan (1986) and I is the conditional mutual information (CMI) Cover & Thomas (2006). The CI-test follows:

$$T \not\perp E_t \mid X_{<t} \iff I(T,\, E_t \mid X_{<t}) > 0. \tag{7}$$

Crucially, Eq. equation 6 is identical for events or outcomes. The only regime-specific element is which conditional distribution $P_\theta(T \mid \cdot)$ is queried by the event or outcome head.

Remark 2 (Full-History Conditioning and Memory)**.** *We condition on the full trajectory $X_{<t}$ rather than the coarsened binary history $E_{<t}$. Under causal sufficiency (A3), this blocks a potential backdoor path that would otherwise remain hidden in the binary projection. The sparse variant (Appendix G.3) truncates $X_{<t}$ to an horizon m to reduce memory consumption in long sequences for $\mathcal{X} \to \mathcal{X}$. Empirically, the F1 score > 80% persists up to $m = 20$ (see Appendix, Fig. 9(c)).*

Lagged Effects $\mathcal{X} \to \mathcal{X}$. Notably, for lagged effects $(t - \ell) \to t$ between events, the mediators $M = X_{t-\ell+1:t-1}$ open an indirect path from $E_{t-\ell}$ to E_t; we randomize M to close it, which combined with temporal precedence (A2) and causal sufficiency (A3) making $X_{<t-\ell}$ a valid backdoor adjustment set. We thus identify the existence of a direct edge $E_{t-\ell} \to E_t$ as a randomized Controlled Direct Effect Pearl (2009). For the $\mathcal{X} \to \mathcal{Y}$ regime this issue does not arise: by the two-chain DGP (Eq. 2), Y_j has no children or spouses, so $X_{<t}$ is already a valid adjustment set and no intervention is needed[4].

Estimation and Error Bound. Since the true distribution P is unknown, we use an ϵ-oracle AR model (A1) to simulate N i.i.d. history particles $\{x_{<t}^{(l)}\}_{l=1}^{N} \sim P_{\theta_x}(X_{<t})$ autoregressively via next-token prediction. Hence, the Monte Carlo estimator of Eq. equation 6 is given by:

$$\hat{I}_N(T,\, E_t \mid X_{<t}) \;=\; \frac{1}{N} \sum_{l=1}^{N} \mathbb{E}_{e_t \sim P_\theta}\Big[I_G\Big(T,\, e_t \;\Big|\; x_{<t}^{(l)}\Big)\Big]\,. \tag{8}$$

[3] E_t projects the multinomial sequence onto a binary sequence in which causal dependencies between events realization are well-defined in the sample-level graph $\mathcal{G}^s$: $E_0 \to E_4$ means that the occurrence of x_0 at t_0 caused event x_4 at t_4.

[4] We provide the full derivation, implementation, and parallelization in Appendix G, and use T throughout for clarity.

This estimator is consistent (Prop. 1) and requires only two queries to the model per particle: $P_\theta(T \mid x^{(l)}_{<t})$ and $P_\theta(T \mid e_t, x^{(l)}_{<t})$ (i.e., *without* and *with* cause E_t). Importantly, it is fully parallelizable on GPUs across positions t and particles l. Since $\hat{I}_N$ is an approximation and estimation of the true CMI I, we now show that it is bounded by the model approximation error ϵ (Thm. 1).

4.4 Sample-Level Soundness

Theorem 1 (CMI Error Bound)**.** *Let $\hat{I}_N$ be the Monte Carlo estimator (Eq. 8) and I the true CMI for the binary target T. Under the ϵ-oracle assumption (A1) the error is bounded by:*

$$\limsup_{N\to\infty} \left|\hat{I}_N - I\right| \;\le\; 2\,\Phi\Big(\sqrt{\epsilon_T\,/\,2}\Big) \;\le\; 2\,\Phi\Big(\sqrt{\epsilon/2}\Big), \tag{9}$$

where $\Phi(\delta) := \delta \ln 2 + (1{+}\delta)\, h_b\big(\frac{\delta}{1+\delta}\big)$, h_b is the binary entropy, and ϵ_T is the binary approximation error for binary target T. The right inequality is the computable worst-case bound, monitored via the cross entropy loss. The left inequality is the operative bound: since each CI test queries a binary projection of the full categorical distribution, ϵ_T absorbs only the fraction of ϵ relevant to target T.

(Proof Sketch) We use the Alicki-Fannes-Winter Winter (2016) inequality for the difference between conditional entropies of two distributions with a small total variation distance $TV(P, P_\theta) \le 1/2$.

Using Thm. 1, as $\epsilon \to 0$, the CMI error vanishes and structural recovery becomes exact using standard faithfulness or must exceed the noise floor induced by ϵ_T. Even for finite N, it is empirically dominated by the approximation bias induced by ϵ_T for $N > 64$ (Appendix, Fig. 11).

Theorem 2 (Sample-Level Soundness)**.** *Under ϵ-strong faithfulness (Def. 6), causal sufficiency (A3), and temporal precedence (A2),* SEQ2CAUSE *recovers the correct Sample-Time Causal Graph $\mathcal{G}^s$ as $N \to \infty$: an edge $(t{-}\ell) \to t$ is included iff $\hat{I}_N > 2\Phi(\sqrt{\epsilon_T/2})$ and excluded iff $\hat{I}_N \le 2\Phi(\sqrt{\epsilon_T/2})$, with no false positives in the limit $N \to \infty$ and $\epsilon \to 0$.*

(Proof Sketch) By induction, we show that for each sequential step t, we can recover the potential causes $E_{<t}$ of the effect event E_t or Y_j. By temporal precedence (A2) and causal sufficiency (A3), we can conclude that no other events affect the effect and thus verify the heredity.

Corollary 1 (Root-Cause Analysis as Graph Traversal)**.** *Under the conditions of Thm. 2,* SEQ2CAUSE *recovers the complete joint graph $\mathcal{G}^s = \mathcal{G}^s_{\mathcal{X}\mathcal{X}} \cup \mathcal{G}^s_{\mathcal{X}\mathcal{Y}}$ from a single sequence s. The root cause of outcome Y_j is then identifiable as the event realization x_{t^*} at the earliest time index admitting a directed path to the Markov boundary of Y_j:*

$$\mathrm{RC}(Y_j,\, s) \;=\; x_{t^*}, \qquad t^* = \arg\min_t \Big\{\, t \;\Big|\; \exists\, t' \in \mathrm{MB}(Y_j),\;\; t \to t' \text{ in } \mathcal{G}^s_{\mathcal{X}\mathcal{X}} \Big\}, \tag{10}$$

This is the first identifiability result for sample-level root-cause analysis in discrete event sequences without topology assumptions or interventional data Assaad et al. (2023); Wang et al. (2023); Yao et al. (2024); Han et al. (2025) under the listed assumptions.

4.5 Population-Level Discovery via Minimum Description Length

Population-level causal discovery seeks the global population summary graph $\mathcal{G} = (\mathcal{X} \cup \mathcal{Y},\, \mathcal{E})$ given n i.i.d. sequences $\{s^{(k)}\}_{k=1}^n$. A directed edge $u \to v$ encodes the type-level causal claim: "`Fire` causes `Smoke` across the population" Assaad et al. (2022), as opposed to the sample-level claim "`Fire`@t_1 caused `Smoke`@t_2" recovered by the single sample time graph Math et al. (2025b).

Two-Part Code. To aggregate causal evidences, we formulate a minimum description length (MDL) Grunwald (2004) criterion that operates directly on the summary graph $\mathcal{G}$ over event types and outcomes. Formally, the best graph is the one that minimized the cost $\mathcal{L}(\mathcal{D}, \mathcal{G})$ Janzing & Schölkopf (2010) as:

$$\mathcal{L}(\mathcal{D}, \mathcal{G}) = \underbrace{\mathcal{L}(\mathcal{D}|\mathcal{G})}_{\text{data encoding cost}} + \underbrace{\mathcal{L}(\mathcal{G})}_{\text{model cost}} \tag{11}$$

Each type-level edge $u \to v$ encodes a source $u \in \mathcal{X}$ and target $v \in \mathcal{X} \cup \mathcal{Y}$ at a cost of $c_{\text{type}} = \log|\mathcal{X}| + \log(|\mathcal{X}| + |\mathcal{Y}|)$ nats by Rissanen's prefix-free integer code Rissanen (1983). Then, the data

encoding cost is the negative log likelihood (NLL) of the dataset under $\mathcal{G}$ Cover & Thomas (2006), well-defined through the acyclicity of the sample-level factorization (Eq. 2). Therefore, adding edge $u \to v$ changes $\mathcal{L}(\mathcal{D}, \mathcal{G})$:

$$\Delta\mathcal{L}(\mathcal{D}, \mathcal{G}) = \sum_{(k,t,\ell) \in \mathcal{M}_{uv}} \log \frac{P\big(T^{(k)} = 1 \mid \mathrm{Pa}_{\mathcal{G}}(T^{(k)}) \cup \{E^{(k)}_{t-\ell}\}\big)}{P\big(T^{(k)} = 1 \mid \mathrm{Pa}_{\mathcal{G}}(T^{(k)})\big)} - c_{\mathrm{type}} \tag{12}$$

where $\mathcal{M}_{uv} := \{(k, t, \ell) : x^{(k)}_{t-\ell} = u, x^{(k)}_t \text{ or } y^{(k)}_j = v\}$ are the matching positions of source type u at lag ℓ and target v at t in sequence k, $\mathrm{Pa}_{\mathcal{G}}(T^{(k)})$ the current parent set of T in $\mathcal{G}$, and $E_{t-\ell} = \mathbf{1}[X_{t-\ell} = u]$ the candidate cause at lag ℓ. Crucially, the edge $u \to v$ is included if $\Delta\mathcal{L} > 0$ (the accumulated likelihood gain exceeds the model cost).

Generative Error Correction for MDL Inclusion. Since P is replaced by P_θ (A1), the approximated log-ratio in Eq. 12 deviates at every matched position. It therefore accumulates across sequences and degrades precision as the number of samples n grows. The following Lem. 1 characterizes this deviation in function of ϵ and yields the self-corrected inclusion criterion of Cor. 2.

Lemma 1 (Data Encoding Cost Deviation under the ϵ-Oracle)**.** *Let ΔL^P and ΔL^θ denote the true and approximated MDL gains for edge $u \to v$ (Eq. 12) under P and P_θ respectively. Under A1, the accumulated deviation satisfies:*

$$\big|\Delta\mathcal{L}^\theta - \Delta\mathcal{L}\big| \;\le\; \sqrt{\epsilon_T/2} \;\cdot\; \eta_{uv} \tag{13}$$

where $\epsilon_T \le \epsilon$ is the local binary approximation error (Prop. 2), and η_{uv} is fully computable from P_θ and the data (Appendix D.5, Eq. 37). Concretely, $\eta_{uv} = \sum_{(k,t)\in\mathcal{M}_{uv}} \kappa^{(k)}_t$ accumulates per-position inverse-margin terms $\kappa^{(k)}_t \propto 1/(q_\pm - \sqrt{\epsilon_T/2})$, where $q_\pm = P_\theta(T \mid Z_\pm)$ are the same two AR forward passes already evaluated for $\Delta\mathcal{L}^\theta$, so η_{uv} requires no additional model queries.

Corollary 2 (Adaptive Population Threshold)**.** *Using Lem. 1, edge $u \to v$ is included in the population summary graph $\mathcal{G}$ if and only if:*

$$\boxed{u \to v \in \mathcal{G} \iff \Delta\mathcal{L}^\theta > \tau^*_{uv}, \qquad \tau^*_{uv} = \underbrace{c_{\mathrm{type}}}_{\textit{MDL penalty}} + \underbrace{\sqrt{\epsilon_T/2} \cdot \eta_{uv}}_{\textit{error correction}}} \tag{14}$$

The threshold τ^*_{uv} is adaptive in two compounding senses. The MDL penalty c_{type} is a fixed one-time encoding cost, penalizing weak evidence for adding $u \to v$. And the error correction $\sqrt{\epsilon_T/2} \cdot \eta_{uv}$ is pair-specific, vanishing as $\epsilon_T \to 0$ such that a better oracle requires no manual threshold recalibration.

Remark 3 (Computation of τ^*_{uv})**.** *Importantly, τ^*_{uv} requires no Monte-Carlo estimation, yielding memory complexity $\mathcal{O}(B \times M \times L^2)$ for the GPU staircase implementation (Appendix, Fig. 6) for $\mathcal{X} \to \mathcal{X}$ and just $\mathcal{O}(B \times L)$ for $\mathcal{X} \to \mathcal{Y}$, where B is the batch size and M mediators to sample.*

5 Experiments

We evaluate SEQ2CAUSE across all four regimes of Tab. 1, unifying results from synthetic non-linear SCMs and real-world vehicle diagnostic logs. Evaluation always proceeds in two phases: (1) reuse or train a pretrained AR model, (2) apply SEQ2CAUSE. Full protocol, robustness analysis and ablations are in Appendix H. Sample-Level results are averaged over 10 cross-validation runs; population-level results report the mean directly.

5.1 Setup

Nonlinear-SCMs ($\mathcal{X} \to \mathcal{X}$, *sample and population*). We validate SEQ2CAUSE on sequences generated by nonlinear-SCMs with controllable memory m, sequence length L, and vocabulary size $|\mathcal{X}|$, training a LLaMa model Touvron et al. (2023) and monitoring the normalized oracle score $\hat{\epsilon}$ (Eq. 42) before applying SEQ2CAUSE to recover the summary graphs. At the sample level we compare against Neural Granger (same AR backbone, probability difference instead of CMI), Attention-BERT Devlin et al. (2019), Saliency Shrikumar et al. (2017), Shapley Value Sampling Lundberg & Lee (2017), and Random/Frequency baselines; at the population level, against Hawkes Processes (SHTP) Qiao et al. (2023), CASCADE Cüppers et al. (2024), PMI Church & Hanks (1990), and TF-IDF Sparck Jones (1988) with individual significance thresholds.

Table 2: **Unified evaluation of SEQ2CAUSE across all four causal regimes. (a)** $|\mathcal{X}| = 1{,}000$, $L = 64$, ϵ=0.05, τ=3×10^{-5}, N=128, $H(P) = 2.4$ nats. **(b, d)** Diagnostic vehicle sequences leading to critical failures as outcomes, $|\mathcal{X}|$=29,100, $|\mathcal{Y}|$=474, L≈150, $n = 50,000$ sequences, weighted avg, †: timeout >3 days **(c)** $|\mathcal{X}| = 1{,}000$, $L = 64$, ϵ=0.05, n=50,000 sequences, $H(P) = 2.4$ nats.

	Event-to-Event $\mathcal{X} \to \mathcal{X}$				**Events-to-Outcomes** $\mathcal{X} \to \mathcal{Y}$			
	(a) *Sample-Summary Graph* $\mathcal{G}^s_{\mathcal{X}\mathcal{X}}$ (nonlinear-SCM)				**(b)** *Sample-Markov Boundary* $\mathcal{G}^s_{\mathcal{X}\mathcal{Y}}$ (Vehicle Diagnostics)			
	Method	SHD↓	F1↑	Prec.↑	*Method*	Prec.↑	Rec.↑	F1↑
Sample-Level	Attention (BERT)	321.0±15	50.1±.01	35.0±.01	IAMB Tsamardinos et al. (2003)	–†	–†	–†
	Saliency (IG, LLaMA) Shrikumar et al. (2017)	160.2±6.6	67.3±.01	51.4±.01	CMB Gao & Ji (2015)	–†	–†	–†
	Shapley (LLaMA) Lundberg & Lee (2017)	148.0±5.1	60.0±.01	55.5±.01	MB-by-MB Xie et al. (2024)	–†	–†	–†
	Neural Granger (LLaMA) Granger (1969)	100.2±14.6	69.2±.04	71.6±.04	PCDbyPCDYin et al. (2008)	–†	–†	–†
	seq2cause	**28.6**±2.8	**91.3**±.01	**89.7**±.01	**seq2cause**	**58.2**±1.4	**35.5**±0.8	**46.1**±1.0
	(c) *Population-Summary Graph* $\mathcal{G}_{\mathcal{X}\mathcal{X}}$ (nonlinear-SCM)				**(d)** *Population-Markov Boundary* $\mathcal{G}_{\mathcal{X}\mathcal{Y}}$ (Vehicle Diagnostics)			
	Method	Prec.↑	Rec.↑	F1↑	*Method*	Prec.↑	Rec.↑	F1↑
Population-Level	PMI Church & Hanks (1990) ($\tau = 1$)	5.6	91.6	10.5	IAMB Tsamardinos et al. (2003)	–†	–†	–†
	TF-IDF Sparck Jones (1988) ($\tau = 0.05$)	17.4	89.6	29.1	CMB Gao & Ji (2015)	–†	–†	–†
	SHP Qiao et al. (2023) (Hawkes)	–†	–†	–†	PCDbyPCD Yin et al. (2008)	–†	–†	–†
	CASCADE Cüppers et al. (2024)	–†	–†	–†	MI-MCF Ma et al. (2025)	–†	–†	–†
	SEQ2CAUSE *w/o corr.*	4.2	99.7	**8.0**	TF-IDF Sparck Jones (1988)	41.3±2.1	18.7±1.4	25.8±1.6
	seq2cause	**81.5**	**54.6**	**65.4**	**seq2cause**	**68.6**±1.4	**48.0**±1.6	**56.8**±1.2

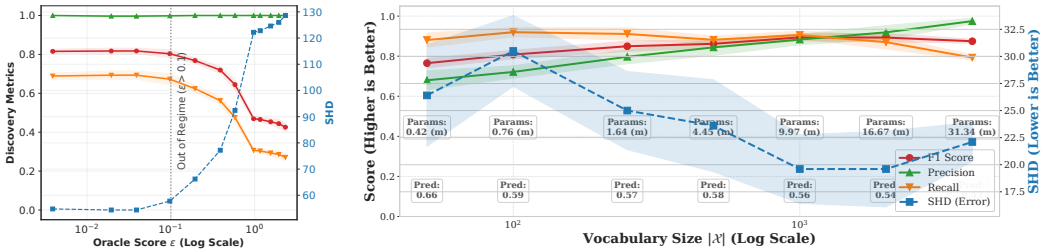
Vehicle Diagnostic Sequences ($\mathcal{X} \to \mathcal{Y}$, *sample and population*). We adopt a real-world vehicle dataset Math et al. (2025a) of m=300,000 error-code sequences over $|\mathcal{X}|$=29,100 event types and $|\mathcal{Y}|$=474 failure outcomes (150 events/sequence on average). The two frozen AR backbones from Math et al. (2025a) (105M parameters, 1.91 nats test cross-entropy) serve as f_{θ_x} and f_{θ_y}, neither having seen the test set. Ground truth MB were domain-expert annotated (partially for some outcomes, a conservative evaluation). We compare against IAMB Tsamardinos et al. (2003), CMB Gao & Ji (2015), MB-by-MB Xie et al. (2024), PCDbyPCD Yin et al. (2008), and MI-MCF Ma et al. (2025) (via *PyCausalFS* Yu et al. (2020)) plus TF-IDF Sparck Jones (1988), reporting weighted Precision, Recall, and F1 (example graph in Fig. 5).

5.2 Unified Four-Regime Evaluation

Cell (a) — Sample-level $\mathcal{X} \to \mathcal{X}$. SEQ2CAUSE outperforms the strongest baseline (Neural Granger) by over 20 F1 points, achieving 91% F1 and 89% Precision at $|\mathcal{X}| = 1{,}000$. Attention-based methods fail to separate correlation from causation (50% F1 score); saliency and Shapley methods offer partial signal (67% and 60% F1 respectively) but lack the interventional grounding of CMI testing. This confirms that **measuring the CMI is a fundamentally more reliable causal signal** than monitoring single-token probability fluctuations or gradient-based attributions.

Cell (c) — Population-level $\mathcal{X} \to \mathcal{X}$. Hawkes-based methods (SHP) and CASCADE are architecturally excluded: co-occurrence degeneracy (Eq. 1) renders their excitation matrices unidentifiable at $|\mathcal{X}|$=1,000, and both time out before N=100 sequences. Correlation baselines (PMI, TF-IDF) terminate in time but conflate causation with association: PMI achieves Recall 91.6% at Precision 5.6%, effectively predicting all edges while TF-IDF's heuristic threshold yields F1 29.1%, with no mechanism to distinguish direct causes from spurious co-occurrences. SEQ2CAUSE *without* the MDL error correction (Cor. 2) mirrors this failure: Precision collapses to 4.2% as oracle approximation errors accumulate unchecked beyond N≈1,000 sequences, providing direct empirical validation of Lem. 1. With correction, SEQ2CAUSE achieves F1 65.4% (Prec. 81.5%, Rec. 54.6%): the only method to maintain both high precision and moderate recall simultaneously, and **the only causally grounded solution viable**. Association baselines are acutely sensitive to τ: a shift of 0.02 in TF-IDF's threshold alone can collapse F1 by over 20 points, with no principled criterion to guide that choice (Appendix H.3). SEQ2CAUSE's decision boundary is **derived analytically from the error bound and requires no calibration**. Full evolution across samples is shown in Appendix Fig. 8. We provide a deeper discussion regarding threshold calibration for the baselines in Appendix H.3.

Cells (b, d) — Sample and Population $\mathcal{X} \to \mathcal{Y}$. Classical algorithms are architecturally excluded from this regime by two compounding degeneracies. *Degeneracy 1* (vocabulary-driven intractability):



(a) Sample-Level: Robustness to ϵ

(b) Sample-Level: Scalability to Vocabulary Size

Figure 2: **Robustness and scalability of SEQ2CAUSE. (a)** Phase transition at $\epsilon \approx 0.1$: Precision stays near 0.95 as ϵ increases (conservative failure mode). **(b)** Stable performance across vocabularies from $|\mathcal{X}|=10^2$ to 8×10^3; Shannon Redundancy (*Pred* = $1 - H(P)/H_{max}$) accounts for modeling difficulty ($\hat{\epsilon} = 0.01, L = 64, N = 64, \tau = 10^{-4}$).

at $|\mathcal{X}| = 29,100$, the candidate conditioning-set space is $2^{|\mathcal{X}|}$, and even the greedy IAMB shell requires $\mathcal{O}(|\mathcal{X}|)$ CI tests per outcome whose joint support is never populated; all five baselines exceed a 3-day budget on $n = 50,000$ sequences across 4 GPUs. *Degeneracy 2* (sample-driven impossibility): restricting to the top-K event types and $n \in \{500, 5000\}$ (Appendix Tab. 6) rules out "more compute would help", even where classical MB algorithms run to completion (K=500, n=500), the expected co-occurrence per (u, Y_j) pair tends to 0, so every asymptotic CI test rejects independence; Precision and Recall collapse to 0 across all five baselines, with CI tests returning "independent" on essentially all pairs. Scaling up either axis trades one degeneracy for the other: more samples at fixed K eases Deg. 2 but worsens Deg. 1; more vocabulary at fixed n re-opens Deg. 2. SEQ2CAUSE escapes this trap because the AR backbone amortizes density estimation across all (u, Y_j) pairs simultaneously, reusing the same representations for every CI test rather than re-estimating joint supports per pair.

5.3 Generative Error and Scalability

Identifiability precedes convergence. Figure 2(a) shows that exact model convergence ($\epsilon \to 0$) is not necessary for structural recovery: a phase transition at $\epsilon \approx 0.1$ recovers the coarse-grained causal graph before the model masters fine-grained transition probabilities. SEQ2CAUSE exhibits a **conservative failure mode**: as ϵ increases, Recall degrades while Precision stays near 0.95, so errors manifest as missed rather than hallucinated edges, desirable when false positives are costlier than false negatives.

Breaking the curse of dimensionality. Figure 2(b) evaluates SEQ2CAUSE across vocabularies from $|\mathcal{X}|=10$ to 8,000. F1 remains stable at $\approx 81\%$ despite the exponential growth of the candidate edge space, confirming that the pretrained AR backbone effectively amortizes the density estimation cost across all event types. No existing constraint-based or Hawkes-based method can operate at $|\mathcal{X}| > 100$ in this regime.

Self-correcting MDL threshold τ^*_{uv} prevents precision collapse. Without correction, Precision collapses to near zero beyond $N \approx 1{,}000$ sequences as approximation errors accumulate (Lem. 1); τ^*_{uv} instead maintains Precision above 80% across $n \in [10,\ 20 \cdot 10^4]$ (Tab. 2, Appendix Fig. 10).

6 Conclusion

Empirically, SEQ2CAUSE is the first method to populate all four regimes of Tab. 1 with a single frozen backbone: 91% F1 on sample-level event graphs versus 69% for the strongest baseline, and tractable Markov boundary recovery at $|\mathcal{X}| = 29,100$ where classical algorithms are intractable, with population-level boundaries recovered in minutes via MDL aggregation across thousands of sequences (Lem. 1). Thms. 1, 2 explain why: minimizing next-token cross-entropy simultaneously bounds CMI error, guarantees soundness, and self-calibrates the population threshold through the single monitorable quantity ϵ, so every advance in sequence modeling tightens causal identification guarantees without retraining.

AI USE STATEMENT

In this work, we used generative AI tools for editing and formatting latex. We have not used generative AI tools for general paper idea, theorems, proofs, experiments, ablations, or bibliography.

REPRODUCIBILITY STATEMENT

All theoretical claims are proved with full assumptions in the appendix: Theorem 1 and Proposition 2 in Appendix A, Lemma 1 in Appendix D.5, and Corollary 2 follows directly. The ϵ-oracle assumption (Assumption 1) and its practical estimation from training diagnostics are detailed in Appendix H.2, with the empirical certificate (Fig. 2a) that recovery precedes full convergence. The synthetic SCM (sparse random-weight MLP mechanism, ground truth verified by counterfactual intervention rather than likelihood) is fully specified in Appendix H.1, including sampling, sparsity/lag-decay parameters, and every swept vocabulary size (Appendix H). All hyperparameters (backbone size, particle counts, context length, thresholding scheme) are listed with the corresponding experiment in Section 5.1 and Appendix H. The proprietary vehicle diagnostic corpus cannot be released; we instead release the full synthetic data generator, reproducing every reported trend (Figs. 2–12), together with an anonymized code repository implementing the estimator, do-intervention construction, MDL threshold, and every baseline in Table 2. All numbers are averaged over multiple seeds with the per-seed spread reported alongside each figure/table.

REFERENCES

Charles K. Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. J. Artif. Int. Res., 73, May 2022. ISSN 1076-9757. doi: 10.1613/jair.1.13428. URL https://doi.org/10.1613/jair.1.13428.

Charles K. Assaad, Imad Ez-Zejjari, and Lei Zan. Root cause identification for collective anomalies in time series given an acyclic summary causal graph with loops. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, volume 206 of Proceedings of Machine Learning Research, pp. 8395–8404. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/assaad23a.html.

Emmanuel Bacry, Martin Bompaire, Stéphane Gaffas, and Jean-Francois Muzy. Sparse and low-rank multivariate hawkes processes. Journal of Machine Learning Research, 21(50):1–32, 2020. URL http://jmlr.org/papers/v21/15-114.html.

Alexis Bellot and Mihaela van der Schaar. Conditional independence testing using generative adversarial networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/dc87c13749315c7217cdc4ac692e704c-Paper.pdf.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1):22–29, 1990. URL https://aclanthology.org/J90-1003/.

Thomas M. Cover and Joy A. Thomas. Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience, July 2006. ISBN 0471241954.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. Nature Methods, 21(8):1470–1480, 8 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

Joscha Cüppers, Sascha Xu, Musa Ahmed, and Jilles Vreeken. Causal discovery from event sequences by local cause-effect attribution. Advances in Neural Information Processing Systems, 37, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In North American Chapter of the

Association for Computational Linguistics, 2019. URL https://api.semanticscholar.org/CorpusID:52967399.

Arnaud Doucet, Nando de Freitas, and Neil Gordon. An Introduction to Sequential Monte Carlo Methods, pp. 3–14. Springer New York, New York, NY, 2001. ISBN 978-1-4757-3437-9. doi: 10.1007/978-1-4757-3437-9_1. URL https://doi.org/10.1007/978-1-4757-3437-9_1.

Bao Duong and Thin Nguyen. Normalizing flows for conditional independence testing. Knowledge and Information Systems, 66, 08 2023. doi: 10.1007/s10115-023-01964-w.

Tian Gao and Qiang Ji. Local causal discovery of direct causes and effects. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), Advances in Neural Information Processing Systems, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/fcdf25d6e191893e705819b177cddea0-Paper.pdf.

Chang Gong, Chuzhe Zhang, Di Yao, Jingping Bi, Wenbin Li, and YongJun Xu. Causal discovery from temporal data: An overview and new perspectives. ACM Comput. Surv., 57(4), December 2024. ISSN 0360-0300. doi: 10.1145/3705297. URL https://doi.org/10.1145/3705297.

C W J Granger. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. Econometrica, 37(3):424–438, 7 1969. URL https://ideas.repec.org/a/ecm/emetrp/v37y1969i3p424-38.html.

Peter Grunwald. A tutorial introduction to the minimum description length principle, 2004. URL https://arxiv.org/abs/math/0406077.

Xiao Han, Saima Absar, Lu Zhang, and Shuhan Yuan. Root cause analysis of anomalies in multivariate time series through granger causal discovery. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=k38Th3x4d9.

Weijie He, Xiaohao Mao, Chao Ma, Yu Huang, José Miguel Hernàndez-Lobato, and Ting Chen. Bsoda: A bipartite scalable framework for online disease diagnosis. In Proceedings of the ACM Web Conference 2022, WWW '22, pp. 2511–2521, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3512123. URL https://doi.org/10.1145/3485447.3512123.

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 (56):1709–1731, 2010. URL http://jmlr.org/papers/v11/hyvarinen10a.html.

Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. IEEE Trans. Inf. Theor., 56(10):5168–5194, October 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2060095. URL https://doi.org/10.1109/TIT.2010.2060095.

Patrick J. Laub, Young Lee, Philip K. Pollett, and Thomas Taimre. Hawkes models and their applications. Annual Review of Statistics and Its Application, 12 (Volume 12, 2025):233–258, 2025. ISSN 2326-831X. doi: https://doi.org/10.1146/annurev-statistics-112723-034304. URL https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-112723-034304.

Jing Li, Botong Wu, Xinwei Sun, and Yizhou Wang. Causal hidden markov model for time series disease forecasting. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12100–12109, 2021. URL https://api.semanticscholar.org/CorpusID:232417085.

Zhuoqun Li and Mingxuan Sun. Sparse transformer hawkes process for long event sequences. In Danai Koutra, Claudia Plant, Manuel Gomez Rodriguez, Elena Baralis, and Francesco Bonchi (eds.), Machine Learning and Knowledge Discovery in Databases: Research Track, pp. 172–188, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43424-2.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, pp. 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Lin Ma, Liang Hu, Yonghao Li, Weiping Ding, and Wanfu Gao. Mi-mcf: A mutual information-based multilabel causal feature selection. IEEE Transactions on Neural Networks and Learning Systems, pp. 1–15, 2025. doi: 10.1109/TNNLS.2025.3556128.

Hugo Math and Rainer Lienhart. Context-informed sequence classification: A multimodal approach to vehicle diagnostics. In 1st ICLR Workshop on Time Series in the Age of Large Models, 2026. URL https://openreview.net/forum?id=G4iAE9xOpb.

Hugo Math, Rainer Lienhart, and Robin Schön. Harnessing event sensory data for error pattern prediction in vehicles: A language model approach. Proceedings of the AAAI Conference on Artificial Intelligence, 39(18):19423–19431, 4 2025a. doi: 10.1609/aaai.v39i18.34138. URL https://ojs.aaai.org/index.php/AAAI/article/view/34138.

Hugo Math, Robin Schön, and Rainer Lienhart. One-shot multi-label causal discovery in high-dimensional event sequences. In NeurIPS 2025 Workshop on CauScien: Uncovering Causality in Science, 2025b. URL https://openreview.net/forum?id=z7NT8vGWC2.

Hugo Math, Julian Lorenz, and Rainer Lienhart. Neuro-symbolic rule discovery: Empowering LLMs with causality for vehicle diagnostics. In ICLR 2026 Workshop on Logical Reasoning of Large Language Models, 2026. URL https://openreview.net/forum?id=M5ZszfsJxm.

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=AjXkRZIvjB.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. Machine Learning and Knowledge Extraction, 1(1):312–340, January 2019. ISSN 2504-4990. doi: 10.3390/make1010019.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In Silvia Chiappa and Roberto Calandra (eds.), Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, volume 108 of Proceedings of Machine Learning Research, pp. 1595–1605. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/pamfil20a.html.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: an imperative style, high-performance deep learning library. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2019. Curran Associates Inc.

Judea Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 1558604790.

Judea Pearl. Causality: Models, Reasoning and Inference. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Parivash Pirasteh, Slawomir Nowaczyk, Sepideh Pashami, Magnus Löwenadler, Klas Thunberg, Henrik Ydreskog, and Peter Berck. Interactive feature extraction for diagnostic trouble codes in predictive maintenance: A case study from automotive domain. In Proceedings of the Workshop on Interactive Data Mining, WIDM'19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362962. doi: 10.1145/3304079.3310288. URL https://doi.org/10.1145/3304079.3310288.

Aditya Pratapa, Amogh Jalihal, Jeffrey Law, Aditya Bharadwaj, and T. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nature Methods, 17:1–8, 02 2020. doi: 10.1038/s41592-019-0690-6.

Jie Qiao, Ruichu Cai, Siyu Wu, Yu Xiang, Keli Zhang, and Zhifeng Hao. Structural hawkes processes for learning causal structure from discrete-time event sequences. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI '23, 2023. ISBN 978-1-956792-03-4. doi: 10.24963/ijcai.2023/633. URL https://doi.org/10.24963/ijcai.2023/633.

J. R. Quinlan. Induction of decision trees. Machine Learning, 1:81–106, 1986.

Jorma Rissanen. A universal prior for integers and estimation by minimum description length. The Annals of Statistics, 11(2):416–431, 1983. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2240558.

Raanan Yehezkel Rohekar, Yaniv Gurwicz, and Shami Nisimov. Causal interpretation of self-attention in pre-trained transformers. In Thirty-seventh Conference on Neural Information Processing Systems, 2023. URL https://openreview.net/forum?id=DS4rKySlYC.

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. doi: 10.1126/sciadv.aau4996. URL https://www.science.org/doi/abs/10.1126/sciadv.aau4996.

Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/02f039058bd48307e6f653a2005c9dd2-Paper.pdf.

Claude Elwood Shannon. Prediction and entropy of printed english. Bell System Technical Journal, 30:50–64, January 1951. URL http://languagelog.ldc.upenn.edu/myl/Shannon1950.pdf.

Hongru Shen, Jilei Liu, Jiani Hu, Xilin Shen, Chao Zhang, Dan Wu, Mengyao Feng, Meng Yang, Yang Li, Yichen Yang, Wei Wang, Qiang Zhang, Jilong Yang, Kexin Chen, and Xiangchun Li. Generative pretraining from large-scale transcriptomes for single-cell deciphering. iScience, 26(5): 106536, 2023. ISSN 2589-0042. doi: https://doi.org/10.1016/j.isci.2023.106536. URL https://www.sciencedirect.com/science/article/pii/S2589004223006132.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Doina Precup and Yee Whye Teh (eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pp. 3145–3153. PMLR, 2017. URL https://proceedings.mlr.press/v70/shrikumar17a.html.

Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval, pp. 132–142. Taylor Graham Publishing, GBR, 1988. ISBN 0947568212.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. Social Science Computer Review, 9(1):62–72, 1991. doi: 10.1177/089443939100900106. URL https://doi.org/10.1177/089443939100900106.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14, pp. 3104–3112. MIT Press, 2014.

MASAYUKI TAKAYAMA, Tadahisa OKUDA, Thong Pham, Tatsuyoshi Ikenoue, Shingo Fukuma, Shohei Shimizu, and Akiyoshi Sannai. Integrating large language models in causal discovery: A statistical causal approach. Transactions on Machine Learning Research, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Reh1S8rxfh.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. Nature, 618:616–624, 2023. URL https://api.semanticscholar.org/CorpusID:259002047.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Ioannis Tsamardinos and Constantin F. Aliferis. Towards principled feature selection: Relevancy, filters and wrappers. In Christopher M. Bishop and Brendan J. Frey (eds.), Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, volume R4 of Proceedings of Machine Learning Research, pp. 300–307. PMLR, 1 2003. URL https://proceedings.mlr.press/r4/tsamardinos03a.html. Reissued by PMLR on 01 April 2021.

Ioannis Tsamardinos, Constantin Aliferis, and Alexander Statnikov. Algorithms for large scale markov blanket discovery. In Proceedings of the sixteenth international Florida artificial intelligence research society conference, pp. 376–381, 01 2003.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, Bin Yu, et al. Geometry of the faithfulness assumption in causal inference. The Annals of Statistics, 41(2):436–463, 2013.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N. Balasubramanian, and Amit Sharma. Causal order: The key to leveraging imperfect experts in causal inference. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=9juyeCqL0u.

Changzhang Wang, You Zhou, Qiang Zhao, and Zhi Geng. Discovering and orienting the edges connected to a target variable in a dag via a sequential local learning approach. Computational Statistics and Data Analysis, 77:252–266, 2014. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2014.03.003. URL https://www.sciencedirect.com/science/article/pii/S0167947314000802.

Dongjie Wang, Zhengzhang Chen, Yanjie Fu, Yanchi Liu, and Haifeng Chen. Incremental causal graph learning for online root cause analysis. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '23, pp. 2269–2278, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599392. URL https://doi.org/10.1145/3580305.3599392.

Andreas Winter. Tight uniform continuity bounds for quantum entropies: Conditional entropy, relative entropy distance and energy constraints. Communications in Mathematical Physics, 347 (1):291–313, March 2016. ISSN 1432-0916. doi: 10.1007/s00220-016-2609-8. URL http://dx.doi.org/10.1007/s00220-016-2609-8.

Xingyu Wu, Bingbing Jiang, Kui Yu, Huanhuan Chen, and Chunyan Miao. Multi-label causal feature selection. Proceedings of the AAAI Conference on Artificial Intelligence, 34(04):6430–6437, Apr. 2020a. doi: 10.1609/aaai.v34i04.6114. URL https://ojs.aaai.org/index.php/AAAI/article/view/6114.

Xingyu Wu, Bingbing Jiang, Yan Zhong, and Huanhuan Chen. Multi-label causal variable discovery: Learning common causal variables and label-specific causal variables. CoRR, abs/2011.04176, 2020b. URL https://arxiv.org/abs/2011.04176.

Feng Xie, Zheng Li, Peng Wu, Yan Zeng, Chunchen Liu, and Zhi Geng. Local causal structure learning in the presence of latent variables. In Proceedings of the 41st International Conference on Machine Learning, ICML'24. JMLR.org, 2024.

Yanfeng Yang, Shuai Li, Yingjie Zhang, Zhuoran Sun, Hai Shu, Ziqi Chen, and Renming Zhang. Conditional diffusion models based conditional independence testing. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pp. 22020–22028, 2025. URL https://ojs.aaai.org/index.php/AAAI/article/view/34354.

Zhenhe Yao, Changhua Pei, Wenxiao Chen, Hanzhang Wang, Liangfei Su, Huai Jiang, Zhe Xie, Xiaohui Nie, and Dan Pei. Chain-of-event: Interpretable root cause analysis for microservices through automatically learning weighted event causal graph. In Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering, FSE 2024, pp. 50–61, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706585. doi: 10.1145/3663529.3663827. URL https://doi.org/10.1145/3663529.3663827.

Jianxin Yin, You Zhou, Changzhang Wang, Ping He, Cheng Zheng, and Zhi Geng. Partial orientation and local structural learning of causal networks for prediction. In Isabelle Guyon, Constantin Aliferis, Greg Cooper, André Elisseeff, Jean-Philippe Pellet, Peter Spirtes, and Alexander Statnikov (eds.), Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008, volume 3 of Proceedings of Machine Learning Research, pp. 93–105, Hong Kong, 6 2008. PMLR. URL http://proceedings.mlr.press/v3/yin08a.html.

Kui Yu, Xianjie Guo, Lin Liu, Jiuyong Li, Hao Wang, Zhaolong Ling, and Xindong Wu. Causality-based feature selection: Methods and evaluations. ACM Comput. Surv., 53(5), September 2020. ISSN 0360-0300. doi: 10.1145/3409382. URL https://doi.org/10.1145/3409382.

Wei Zhang, Thomas Kobber Panum, Somesh Jha, Prasad Chalasani, and David Page. Cause: learning granger causality from event sequences using attribution methods. In Proceedings of the 37th International Conference on Machine Learning, ICML'20. JMLR.org, 2020.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer Hawkes process. In Hal Daumé III and Aarti Singh (eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pp. 11692–11702. PMLR, 7 2020. URL https://proceedings.mlr.press/v119/zuo20a.html.

A Notations and Mathematical Setup

Notations We use capital letters (e.g., X) to denote random variables, $P(X)$ the probability distribution of X, $P(X = x) = p(x)$ the probability of the realisation x for the random variable X and bold capital letters (e.g., $\boldsymbol{X}$) for sets of random variables.We only work with discrete distributions. Let $\mathcal{X}$ denote an alphabet and D_{KL} the *Kullback-Leibler divergence* Cover & Thomas (2006).

Probability space. Let $\Omega = \mathcal{X}^{\mathbb{N}}$ be the space of infinite sequences $\omega = (x_0, x_1, \ldots)$ over the finite alphabet $\mathcal{X}$. Let $\mathcal{F} = \mathcal{B}(\mathcal{X}^{\mathbb{N}})$ be the product σ-algebra generated by cylinder sets, and let P be the law of the event chain Eq. 2 on this space.

Labels $Y_1, \ldots, Y_c$ are not part of Ω: by the joint DGP equation 2, each Y_j is a function $Y_j : \Omega \to \{0, 1\}$ defined by $Y_j(\omega) = \mathbf{1}[\omega \in A_j]$ for some measurable set $A_j \in \mathcal{F}$ depending on the full trajectory. The label head P_{θ_y} estimates $P(Y_j = 1 \mid X_{<t})$ at each step t from the event prefixes.

The target variable $T : \Omega \to \{0, 1\}$ is therefore, always a measurable function on $(\Omega, \mathcal{F}, P)$, whether $T = E_t \triangleq \mathbf{1}_{X_t = x_t}$ (event indicator) or $T = Y_j$ (outcome label). This unified representation is what allows a single probability space to cover all four regimes.

Graph Types

A.1 Definitions

Definition 3 (Markov Boundary)**.** *Tsamardinos & Aliferis (2003). In a faithful BN $< \boldsymbol{U}, \mathcal{G}, P >$, for a set of variables $\boldsymbol{Z} \subset \boldsymbol{U}$ and label $Y \in \boldsymbol{U}$, if all other variables $X \in \{\boldsymbol{X} - \boldsymbol{Z}\}$ are independent of Y conditioned on $\boldsymbol{Z}$ and any proper subset of $\boldsymbol{Z}$ do not satisfy the condition, then $\boldsymbol{Z}$ is the Markov Boundary of Y: $\boldsymbol{MB}(Y)$.*

A.2 Summary Graph

Operators are often confronted with observing a single sequence during inference to understand generalizable rules (e.g., "Fire causes Smoke") rather than specific timestamps ($\mathcal{G}^s$). We adapt the

Table 3: **Graph objects in SEQ2CAUSE.** All time graphs are DAGs by temporal precedence. Summary graphs may contain cycles Assaad et al. (2022). The sample-time graph decomposes into two subgraphs ($\mathcal{G}^s_{\mathcal{X}\mathcal{X}}$ and $\mathcal{G}^s_{\mathcal{X}\mathcal{Y}}$) by dependency type; their union is the complete sample-level causal graph $\mathcal{G}^s$.

Symbol	Name	Nodes	Edge meaning
$\mathcal{G}^s_{\mathcal{X}\mathcal{X}}$	Sample-time event-to-event graph	Time indices $\{0,\dots,L-1\}$	$(t-\ell) \to t$: realization $x_{t-\ell}$ caused x_t in s
$\mathcal{G}^s_{\mathcal{X}\mathcal{Y}}$	Sample-time event-to-outcome graph	Time indices $\{0,\dots,L-1\}$ $\cup$ outcomes $\mathcal{Y}_s$	$t \to Y_j$: realization x_t caused Y_j in s
$\mathcal{G}^s = \mathcal{G}^s_{\mathcal{X}\mathcal{X}} \cup \mathcal{G}^s_{\mathcal{X}\mathcal{Y}}$	Sample-time causal graph	Time indices $\{0,\dots,L-1\}$ $\cup$ outcomes $\mathcal{Y}_s$	Joint DAG over all events and outcomes in s
$\bar{\mathcal{G}} = \pi(\mathcal{G}^s)$	Sample-summary graph	Event types $\mathcal{X}_s \subseteq \mathcal{X}$, outcomes $\mathcal{Y}_s \subseteq \mathcal{Y}$	$u \to v$: type u caused type v in s
$\mathcal{G}$	Population summary graph	Event types $\mathcal{X}$, outcomes $\mathcal{Y}$	$u \to v$: type u causes type v across $\mathcal{D}$

terminology from Assaad et al. (2022) regarding summary causal graph (SCG), which treats the unique event types in $\mathcal{X}$ as the variables of interest. Therefore, for a single sequence s, we aim to recover its Summary Causal Graph.

Definition 4 (Sample-Summary Graph)**.** *Let $\mathcal{X}_s \subseteq \mathcal{X}$ be the set of unique event types in s. The Sample-Summary Graph $\bar{\mathcal{G}} = (\mathcal{X}_s, \mathcal{E}_s)$ is the surjective projection of the Sample-Time Causal Graph $\mathcal{G}^s$ onto $\mathcal{X}_s$. Specifically, a type-level edge $u \to v$ exists in $\bar{\mathcal{G}}$ if and only if it appears at least once in the time graph:*

$$u \to v \in \mathcal{E}_s \iff \exists t, k \text{ s.t. } ((t-k) \to t) \in \mathcal{E}_t \wedge x_{t-k} = u \wedge x_t = v$$

B On the Limitation of Multivariate Representation for Event Sequences

We detail the limitation of representing discrete event sequences as multivariate time series in the high-dimensional regime.

B.1 Multivariate Hawkes Process Log-likelihood Derivation

Definition 5 (Mutually-Exciting Hawkes processes)**.** *Consider $\mathbf{N}(t) = (N^1(t), \cdots, N^{|\mathcal{X}|}(t))$ as a collection of $|\mathcal{X}|$ counting processes with $N^k(t)$'s occurrence times denoted as t_1^k, t_2^k etc. They are mutually-exciting Hawkes processes if $N^k(t)$'s conditional intensity follows:*

$$\lambda_k^*(t) = \lambda + \sum_{j=1}^{|\mathcal{X}|} \sum_{t_i^j < t} \mu_{j,k}(t - t_i^j) \text{ for } k = 1, \cdots, |\mathcal{X}|$$

where $\mu_{j,k}(s) \geq 0$

The likelihood for any point process parametrized by θ with observations $\{t_1, \cdots, t_L\}$ within a time horizon $0 \leq t \leq T$ can be computed as the sum of the log-likelihood for each process Laub et al. (2025):

$$\ln \mathcal{L}(\theta | t_1, \cdots, t_L, T) = \sum_{k=1}^{|\mathcal{X}|} \ln \mathcal{L}_k(\theta | t_1, \cdots t_L)$$

where each term is defined by:

$$\ln \mathcal{L}_k(t_1, \cdots, t_L) = \ln \left(\left[\prod_{i=1}^{L} \lambda_k^*(t_i) \right] \exp\left(-\int_0^T \lambda_k^*(t)dt\right) \right) \tag{15}$$

This expression reduces to:

$$\begin{aligned} &= \sum_{i=1}^{L} \ln \lambda_k^*(t_i) + \ln \left(\exp\left(-\int_0^T \lambda_k^*(t)dt\right) \right) \\ &= \sum_{i=1}^{L} \ln \lambda_k^*(t_i) - \int_0^T \lambda_k^*(t)dt \end{aligned} \tag{16}$$

Finally, we have for the total log-likelihood:

$$\ln \mathcal{L}(\theta | t_1, \cdots, t_L, T) = \sum_{k=1}^{|\mathcal{X}|} \sum_{i=1}^{L} \ln \lambda_k^*(t_i) - \sum_{k=1}^{|\mathcal{X}|} \int_0^T \lambda_k^*(t)dt \tag{17}$$

B.2 Limitations

Number of parameters. The N_t^j process has either an excitatory effect on $N_t^k(\phi_{j,k} > 0)$ or no effect on $N_t^k(\phi_{j,k} = 0)$ for $j = k$ or $j \neq k$. Importantly, these processes have $\mathcal{O}(|\mathcal{X}|^2)$ parameters to fit, which is often intractable in practice for many applications Laub et al. (2025). For $|\mathcal{X}| = 10^3$ the number of parameters is already 1 million. Moreover, the actual number of scalars being optimized in practice is more than $|\mathcal{X}|^2$, since there is the base intensity λ, the parameters of the excitation functions μ (decay rates) and interaction magnitude. The literature often assumes $|\mathcal{X}|^2 + |\mathcal{X}|$ Laub et al. (2025).

Optimization. To understand why the Hawkes process is insufficient for our setting, we first show that the MLE of its parameters is computationally intractable at scale. We therefore seek the parameters (base rate α, decay β) that make the observed sequence of events most probable. The likelihood for any point process parametrized by θ with observations $\{t_1, \cdots, t_L\}$ within a time horizon $0 \leq t \leq T$ can be computed as the sum of the log-likelihood for each process:

$$\ln \mathcal{L}(\theta | t_1, \cdots, t_L, T) = \sum_{k=1}^{|\mathcal{X}|} \sum_{i=1}^{L} \ln \lambda_k^*(t_i) - \sum_{k=1}^{|\mathcal{X}|} \int_0^T \lambda_k^*(t)dt \tag{18}$$

Importantly, Eq. 18 cannot be computed easily since the sum over all event types $|\mathcal{X}|$ and the double summation over all pairs of event t_i and t_j, leading to $\mathcal{O}(|\mathcal{X}|^2 \cdot N(T)^2)$ at worst and $\mathcal{O}(|\mathcal{X}|^2 \cdot N(T))$ if using exponential decay Laub et al. (2025). This is a regime where classical MLE is computationally prohibitive and statistically prone to overfitting Bacry et al. (2020); Li & Sun (2023).

C Assumption

C.1 Definitions

Assumption 2 (Temporal Precedence)**.** *Given a recorded sequence with monotonically increasing timestamps $0 \leq t_1 \leq \cdots \leq t_L$, an event x_t may influence any subsequent event $x_{t'}$ only if $t < t'$. Edge orientation in $\mathcal{G}$ is therefore determined by time order.*

Assumption 3 (Causal Sufficiency)**.** *All causally relevant variables are observed. There are no latent confounders affecting the events or the outcomes.*

Definition 6 (ϵ-Strong Faithfulness)**.** *Let $\tau_{\epsilon_T} = 2\Phi(\sqrt{\epsilon_T/2})$ be the asymptotic approximation error bound of the model (Thm. 1). A distribution P is ϵ-**Strong Faithful** to a causal graph $\mathcal{G}$ with respect to the estimator P_θ if, for every active edge $E_{t-\ell} \to T$ the **true** CMI satisfies:*

$$I(T; E_{t-\ell} \mid X_{<t}) > 2 \cdot \tau_{\epsilon_T} \tag{19}$$

Example 1 (ϵ-Strong Faithful CI-test at two marginal frequencies)**.** *Fix $|\mathcal{X}| = 1{,}000$, $\epsilon = 0.05$ (a well-trained backbone, cf. Table 9), and consider a candidate cause $E_{t-\ell}$ whose target $T = E_{t+1}$ has marginal frequency $p_v = \mathbb{P}(X_{t+1} = v)$. By Prop. 2, the binary approximation error for target T satisfies $\epsilon_T \approx \epsilon \cdot p_v \cdot |\mathcal{X}|^{-1} \cdot |\mathcal{X}| = \epsilon \cdot p_v$; more precisely, for the two regimes below:*

Typical event *($p_v = 1/|\mathcal{X}| = 10^{-3}$). Prop. 2 gives $\epsilon_T \approx \epsilon/|\mathcal{X}| = 5 \times 10^{-5}$. The noise floor evaluates to*

$$\tau_{\epsilon_T} = 2\Phi\left(\sqrt{\epsilon_T/2}\right) = 0.053 \text{ nats}, \tag{20}$$

Strong faithfulness (Def. 6) then requires any true causal association to satisfy $I(T; E_{t-\ell} \mid X_{<t}) > 2\tau_{\epsilon_T} \approx 0.01$ nats. Two consequences follow directly. First, the noise floor decreases as p_v decreases: rarer events admit a tighter decision boundary, because the binary projection of a low-probability target retains only a small slice of the total approximation error ϵ. Second, both floors vanish as $\epsilon \to 0$, confirming that an exact oracle recovers every causal edge regardless of event frequency. This counterintuitive scaling (larger vocabularies yield lower per-test noise floors) follows directly from the binary projection inequality (Prop. 2).

C.2 Limitations

We now include a discussion regarding the main assumptions taken in this paper and the one that we exclude.

C.2.1 Causal Sufficiency

A fundamental assumption in causal discovery is *Causal Sufficiency* (Assumption 3)—the premise that no unobserved confounders influence the system. Since SEQ2CAUSE relies on pre-trained backbones which may have learned from noisy or incomplete data, we empirically evaluate its robustness under controlled violations of causal sufficiency, focusing on two realistic forms of hidden confounding.

Measurement Error (Noise Injection). In Fig. 3(a), we simulate measurement error by randomly replacing valid tokens in the history with noise (P_{noise}). While Recall naturally degrades as the true causal parents are obscured, **Precision remains high** (> 0.8) even when 40% of the context is corrupted. This confirms that SEQ2CAUSE does not "rationalize" noise; if the causal signal $X \to Y$ is destroyed by measurement error, the model assigns $CMI \approx 0$ rather than hallucinating a spurious link.

Missing Intermediaries (Temporal Drops). In the same Fig. 3(b), we simulate missing data by randomly dropping time steps, effectively hiding intermediate nodes in the causal chain ($X \to Z_{hidden} \to Y$). This is a more critical scenario where the conditioning sets are broken. We observe that SEQ2CAUSE is robust to moderate data loss ($P_{drop} < 0.2$).

C.2.2 Temporal Precedence & Instantaneous Effects

We assumes that causal influence respects temporal precedence, which means observation are perfectly recorded over time and therefore does not model instantaneous causal effects between events occurring at the same time index. This assumption is standard in sequential causal discovery and ensures that the recovered causal graph is acyclic and identifiable in the single observed sequence setting.

From a theoretical standpoint, instantaneous effects are not identifiable from a single observed trajectory without additional parametric assumptions, repeated samples, or access to interventions. In practice, apparent simultaneity often arises from time discretization, logging resolution, or batching effects. SEQ2CAUSE interprets such cases through the earliest observable temporal ordering, yielding a conservative but identifiable causal structure. As a result, the recovered summary graph captures directed causal influence with positive temporal delay, rather than true simultaneity.

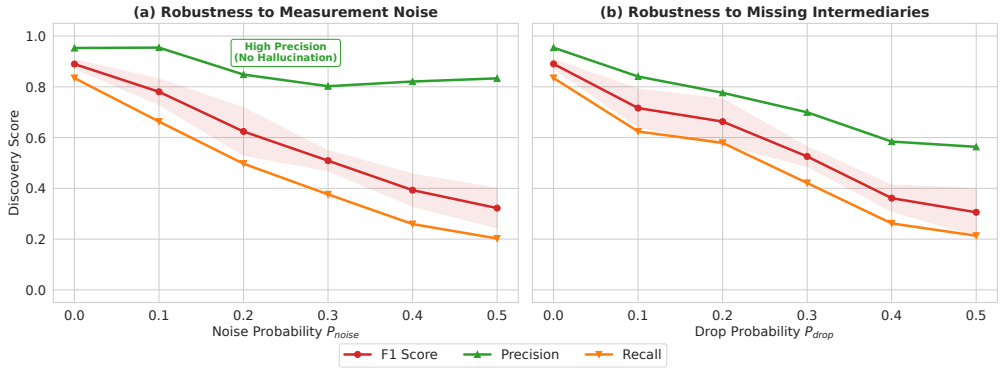


Figure 3: **Robustness to Hidden Confounding.** Evaluation of SEQ2CAUSE under violations of causal sufficiency. **(a)** Measurement Error: Random noise is injected into the context. Precision stays high, indicating resistance to hallucination. **(b)** Temporal Drops: Time steps are randomly deleted, thus conditioning sets are broken. SEQ2CAUSE still recovers the correct sample-summary causal graph structure despite missing intermediaries but the discovery scores quickly decrease.

C.2.3 CONSISTENCY THROUGH TIME

We assume consistency through time: the causal mechanisms governing the generation of events are invariant across time indices. In other words, causal directionality does not reverse over time. For instance, if $A \to B$, then latter it is assume that $B \not\to A$

This assumption is strictly weaker than stationarity Assaad et al. (2022). While the marginal distribution of $\{X_t\}$ may vary over time, the underlying causal dependencies—encoded by the directed edges of the instance time causal graph remain stable. This form of causal invariance is standard in sequential settings and underlies the validity of summary causal graphs that collapse time-indexed relations into event-to-event dependencies.

C.2.4 NO INTER-LABEL EFFECTS

We assume $Y_j \perp Y_k \mid X_{<L}$ for all $j \neq k$: labels do not cause one another, they are jointly caused by the event history. This is not a restrictive assumption in our setting — it is structurally implied by the two-chain DGP (Eq. 2). Since outcomes are realized at the end of the sequence and the event chain is autonomous (no $Y \to X$ edges exist), any apparent statistical dependence between Y_j and Y_k is fully mediated by their shared event causes in $X_{<L}$. Conditioning on $X_{<L}$ therefore blocks every path between labels, making inter-label CI-tests redundant rather than informative. This assumption is standard in multi-label causal discovery where the label space is defined exogenously (Wu et al., 2020a;b), and matches the real-world structure of our target domains: a vehicle failure Y_j and a separate failure Y_k are both downstream of the same diagnostic event stream, not causally related to each other.

Nevertheless, we can consider for the $\mathcal{X} \to \mathcal{Y}$ a second light phase to check inter-label dependencies. Let's consider the Markov Boundaries $MB_1 = [X_1, X_3], MB_2 = [X1, X_2]$ of two labels Y_1, Y_2. We propose to investigate a 'Phase 2', focusing on inter-label dependencies through simulated interventions. For instance, if we consider a sequence S_1 of two labels Y_1, Y_2 with the MB above, we can perform counterfactual interventions by applying $do(X_1 = 0), do(X_3 = 0)$ to S_1 (under A3). Then we would observe the average change in the likelihood of Y_1 which if it is non-zero, would indicate a dependence between Y_1 and Y_2.

C.2.5 NOT-REQUIRED FOR SEQ2CAUSE

We now list the notable assumption that we **don't take** in this paper.

Stationarity The stationarity assumption states that the generative process $\{X_t\}$ does not change with respect to time. Stationarity of the underlying process is a common simplifying assumption in time series causal discovery Assaad et al. (2022), but causal structure itself—defined in terms of

temporal precedence and directed edges—is a property of the generative mechanisms and does not by itself imply stationarity of the observed sequence.

Importantly, SEQ2CAUSE does not require stationarity of the observed sequence. In practice, if using autoregressive Transformers as the AR Model, they can represent non-stationary distributions through contextualized representations (e.g., positional or time embeddings Zuo et al. (2020)), allowing the model to adapt its predictions to evolving regimes without assuming time-invariant marginals. This enables modeling evolving regimes common in real-world logs or patients trajectories.

Parametric Assumption No parametric form is assumed for the transition dynamics beyond the expressivity of the autoregressive model

C.2.6 BROADER IMPACT LIMITATIONS

Risks under degraded models and imbalanced data. SEQ2CAUSE inherits the biases of its autoregressive backbone: if the AR model is trained on imbalanced event data—where certain event types are overrepresented due to systematic logging biases, sensor placement, or population skew—the estimated conditionals $P_\theta(X_t \mid X_{<t})$ will be poorly calibrated, inflating ϵ non-uniformly across the vocabulary. By Thm. 1, this inflates the CMI error bound precisely for the under-represented pairs, causing the framework to systematically miss causal edges involving minority event types while confidently recovering edges among frequent ones. In safety-critical deployments (e.g., vehicle diagnostics, clinical pathways), this failure mode is insidious: the returned causal graph appears high-precision but silently omits rare-but-critical failure chains. Practitioners should monitor the excess cross-entropy loss and treat any causal graph produced with $\epsilon > 0.1$ as unreliable (Fig. 2(a). More broadly, deploying SEQ2CAUSE on biased event logs without human oversight risks reinforcing existing system biases—for instance, concluding that a fault code is non-causal simply because it was underlogged in the training fleet.

D PROOFS

D.1 PROOF OF PROPOSITION 1 (CMI ESTIMATOR CONSISTENCY)

Proposition 1 (Convergence to the ϵ-Proxy)**.** *The estimator $\hat{I}_N$ is a consistent estimator of the ϵ-oracle induced CMI denoted as I_θ. By the Strong Law of Large Numbers, as $N \to \infty$:*

$$\hat{I}_N \xrightarrow[N\to+\infty]{a.s.} I_\theta(E_{t+1}; E_t \mid X_{<t})$$

Proof. The particles $x_{<t}^{(l)}$ are sampled directly from the model f_θ. Let $f_\theta(x_{<t}^{(l)})$ represents the estimation CMI for a fixed history $x_{<t}$ and I_θ the CMI with the approximated distribution P_θ. Expressing this as a difference of conditional entropies:

$$\begin{aligned} 0 &\le \mathbb{E}_{e_t\sim P_\theta} I_G\Big(E_{t+1}, e_t | x_{<t}^{(l)}\Big) \\ &= H_\theta\Big(E_{t+1}|x_{<t}^{(l)}\Big) - H_\theta\Big(E_{t+1}|E_t, x_{<t}^{(l)}\Big) \\ &\le H_\theta(E_{t+1}) \le \log 2. \end{aligned} \tag{21}$$

Thus the posterior variance of $f_\theta(x_{<t}^{(l)})$ satisfies $\sigma_f^2 \triangleq \mathbb{E}_{x_{<t}}[f_\theta^2(x_{<t})] - I_\theta^2(f_\theta) < +\infty$ Doucet et al. (2001) then the variance of $\hat{I}_N(f)$ is equal $\frac{\sigma_f^2}{N}$ and from the strong law of large numbers:

$$\hat{I}_N \xrightarrow[N\to+\infty]{\text{a.s.}} \mathbb{E}_{e_t\sim P_\theta, x_{<t}\sim P_\theta}[I_G(E_{t+1}, e_t \mid x_{<t})] \triangleq I_\theta(f_\theta) \tag{22}$$

□

D.2 PROOF OF PROPOSITION 2 (BINARY PROJECTION OF ϵ)

Proposition 2 (Binary Projection Inequality)**.** *Let P, Q be distributions over $\mathcal{X}$ with $D_{\mathrm{KL}}(P\|Q) = \epsilon$. For any event type $v \in \mathcal{X}$, let $p_v = P(X{=}v)$, $q_v = Q(X{=}v)$, and $T = \mathbf{1}[X{=}v]$. Then the model*

approximation error ϵ_T for the binary target T is:

$$\epsilon_T \;=\; D_{\mathrm{KL}}\big(\mathrm{Bern}(p_v) \,\big\|\, \mathrm{Bern}(q_v)\big) \;=\; \epsilon \;-\; (1-p_v)\cdot D_{\mathrm{KL}}\big(P_{\neg v} \,\big\|\, Q_{\neg v}\big), \tag{23}$$

where $P_{\neg v}(u) := p_u \,/\, (1-p_v)$ for $u \neq v$ is the conditional distribution over $\mathcal{X} \setminus \{v\}$, and similarly for $Q_{\neg v}$.

Proof. We apply the chain rule for relative entropy (Cover & Thomas (2006), Thm. 2.5.3). Any distribution over $\mathcal{X}$ can be viewed as a two-stage draw: first sample the binary indicator $B = \mathbf{1}[X{=}v] \sim \mathrm{Bern}(p_v)$, then, if $B = 0$, draw from the conditional $P_{\neg v}$ over the remaining $|\mathcal{X}| - 1$ types; if $B = 1$, output v deterministically. The chain rule gives:

$$D_{\mathrm{KL}}(P\|Q) = D_{\mathrm{KL}}\big(\mathrm{Bern}(p_v)\|\mathrm{Bern}(q_v)\big) + p_v \cdot D_{\mathrm{KL}}\big(P(\cdot \mid B{=}1) \,\|\, Q(\cdot \mid B{=}1)\big) + \tag{24}$$

$$(1-p_v)\cdot D_{\mathrm{KL}}\big(P_{\neg v} \,\|\, Q_{\neg v}\big). \tag{25}$$

Conditioned on $B = 1$, both distributions place all mass on v, so $D_{\mathrm{KL}}(P(\cdot \mid B{=}1)\|Q(\cdot \mid B{=}1)) = D_{\mathrm{KL}}(\delta_v\|\delta_v) = 0$. Substituting $D_{\mathrm{KL}}(P\|Q) = \epsilon$ and rearranging yields Eq. (23).

Since $D_{\mathrm{KL}}(P_{\neg v}\|Q_{\neg v}) \geq 0$, the immediate consequence is $\epsilon_T \leq \epsilon$. The gap $\epsilon - \epsilon_T = (1-p_v)\cdot D_{\mathrm{KL}}(P_{\neg v}\|Q_{\neg v})$ quantifies the fraction of the total categorical error that is internal to $\mathcal{X} \setminus \{v\}$ and irrelevant to the binary CI test for target v.

□

D.3 Proof of Thm. 1 (CMI Bounded by Model Approximation Error)

Proof. By definition, the Conditional Mutual Information is the difference of two conditional entropies Cover & Thomas (2006):

$$I(T; E_{t-\ell}|X_{<t}) = H(T|X_{<t}) - H(T|E_{t-\ell}, X_{<t})$$

With $T \in \{E_t\} \cup \{Y_1, \ldots, Y_c\}$ is the target (effect) and $E_{t-\ell}$ is the cause. Our framework operates in the Teacher Forcing regime. We do not sample the history $X_{<t}$ from the model's joint distribution. Instead, we estimate the CMI conditioned on the observed history $x^{(l)}_{<t}$. Consequently, the relevant error metric is the per-step conditional divergence at time t, given the fixed history.

Let $\Delta = |I - I_\theta|$ be the CMI estimation error. By the triangle inequality:

$$\Delta \leq \underbrace{|H_P(T \mid X_{<t}) - H_\theta(T \mid X_{<t})|}_{\text{Term A}} + \underbrace{|H_P(T \mid E_{t-\ell}, X_{<t}) - H_\theta(T \mid E_{t-\ell}, X_{<t})|}_{\text{Term B}}$$

We apply the sharp continuity bound for conditional entropy in classical systems (Winter (2016), Lem. 2) with a small total variation distance $\mathrm{TV}(P, P_\theta)) \leq 1/2$. Let ρ and σ be the true and model distributions respectively. Let $\delta = |(P(E_t|\cdot) - P_\theta(E_t|\cdot)|_1$ be the total variation distance. Since the target variable E_t is binary ($d_A = 2$), the bound is:

$$|H_\rho - H_\sigma| \leq \delta \ln(d_A) + (1+\delta) h_b\left(\frac{\delta}{1+\delta}\right)$$

Substituting $d_A = 2$ (so $\ln 2$ in nats since we are using cross entropy loss in PyTorch Paszke et al. (2019)):

$$|H_P(T \mid X_{<t}) - H_\theta(T \mid X_{<t})| \leq \delta \ln(2) + (1+\delta) h_b\left(\frac{\delta}{1+\delta}\right)$$

Term B represents the same entropy difference conditioned on an augmented set $\{E_{t-\ell}, X_{<t}\}$. Since the target dimension d_A remains 2, the same bound applies. Summing the terms:

$$|I - I_\theta| \leq 2\delta \ln(2) + 2(1+\delta) h_b\left(\frac{\delta}{1+\delta}\right)$$

Substituting $\delta = \sqrt{\epsilon/2}$ from Pinsker's inequality, and $h_b(x)$ monotonically increasing for small x, we obtain the final bound in terms of the oracle score ϵ as:

$$|I - I_\theta| \leq 2\sqrt{\epsilon/2}\ln(2) + 2(1+\sqrt{\epsilon/2})h_b\left(\frac{\sqrt{\epsilon/2}}{1+\sqrt{\epsilon/2}}\right) \tag{26}$$

Finally, we decompose the total error into estimation variance and approximation bias using the triangle inequality:

$$|\hat{I}_N - I| = |(\hat{I}_N - I_\theta) + (I_\theta - I)| \leq \underbrace{|\hat{I}_N - I_\theta|}_{\text{Estimation Error}} + \underbrace{|I_\theta - I|}_{\text{Approximation Bias}} \tag{27}$$

Given that $\hat{I}_N$ is a consistent estimator of the model's internal CMI, I_θ (Prop. 1). By the Strong Law of Large Numbers, $\hat{I}_N \xrightarrow{a.s.} I_\theta$ as $N \to \infty$ the stochastic estimation error vanishes, leaving only the irreducible approximation bias:

$$\limsup_{N\to\infty} |\hat{I}_N - I| \leq 0 + |I - I_\theta|$$

We thus obtain the final bound in terms of the oracle score as:

$$\limsup_{N\to\infty} |\hat{I}_N - I| \leq 2\sqrt{\epsilon/2}\ln(2) + 2(1+\sqrt{\epsilon/2})h_b\left(\frac{\sqrt{\epsilon/2}}{1+\sqrt{\epsilon/2}}\right)$$

Using the model approximation error ϵ_T for the binary target T which is lower than the global model approximation error ϵ, using Prop. 2 we have:

$$\limsup_{N\to\infty} \left|\hat{I}_N - I\right| \;\leq\; 2\,\Phi\Big(\sqrt{\epsilon_T\,/\,2}\Big) \;\leq\; 2\,\Phi\Big(\sqrt{\epsilon/2}\Big), \tag{28}$$

where $\Phi(\delta) := \delta\ln 2 + (1+\delta)\,h_b\big(\frac{\delta}{1+\delta}\big)$, h_b is the binary entropy.

This confirms that minimizing the cross-entropy loss (ϵ) directly minimizes the upper bound on structural causal error using the CMI as causal strength. □

D.4 Proof of Thm. 2 (Soundness of the Sample-Level)

Proof. We proceed by induction on the time index $t \in \{1, \ldots, L\}$, exploiting temporal precedence (A2) to reduce the parent search at each step to the strict past $\{E_{t-\ell}\}_{\ell\geq 1}$. The goal is to show that the CMI-based inclusion rule recovers $\widehat{\mathrm{Pa}}(T) = \mathrm{Pa}_{\mathcal{G}^s}(T)$ for every target $T \in \mathcal{T}_t = \{E_{t+1}\} \cup \{Y_1, \ldots, Y_c\}$, and hence the complete joint graph $\mathcal{G}^s = \mathcal{G}^s_{\mathcal{X}\mathcal{X}} \cup \mathcal{G}^s_{\mathcal{X}\mathcal{Y}}$ (Fig. 4).

Base case ($t = 1$). Consider any target $T \in \mathcal{T}_1$. By temporal precedence (A2), no event precedes $t = 1$; the candidate set $\{E_{1-\ell}\}_{\ell\geq 1}$ is empty. SEQ2CAUSE therefore returns $\widehat{\mathrm{Pa}}(T) = \emptyset = \mathrm{Pa}_{\mathcal{G}^s}(T)$.

Inductive step. Fix $t \geq 2$ and assume that for all $j < t$ the parent sets have been correctly identified. Let $T \in \mathcal{T}_t$ be an arbitrary target and let $E_{t-\ell}$, $\ell \in \{1, \ldots, t-1\}$, be a candidate cause.

For $T = E_{t+1}$ with $\ell > 1$, the CMI is evaluated after do-intervening on the intermediate events (Def. 7); this blocks the front-door $X_{t-\ell+1:t-1}$ by randomizing the mediators. Using the lagged information gain I^L_G (Eq. 39), the CMI is replaced and we condition on $X_{<t-\ell}$ while controlling for the mediators. The bound of Eq. 9 applies.

For $T = Y_j$ or $\ell = 1$, the standard CMI of Eq. equation 7 applies without modification, no intervention needed. In both cases the inclusion rule is identical and the two structural cases below apply.

$E_{t-\ell}$ is a true parent ($E_{t-\ell} \in \mathrm{Pa}_{\mathcal{G}^s}(T)$). By ϵ-strong faithfulness (Def. 6), every true causal association satisfies

$$I(T;\, E_{t-\ell} \mid X_{<t}) > 4\Phi\left(\sqrt{\epsilon_T/2}\right). \tag{29}$$

By Thm. 1, the Monte Carlo estimator satisfies $\limsup_{N\to\infty}\left|\hat{I}_N - I\right| \le 2\Phi\Big(\sqrt{\epsilon_T/2}\Big)$, so asymptotically $\hat{I}_N > 2\Phi\Big(\sqrt{\epsilon_T/2}\Big)$ and the edge is accepted.

$E_{t-\ell}$ is not a parent ($E_{t-\ell} \notin \mathrm{Pa}_{\mathcal{G}^s}(T)$). By the two-chain factorization (Eq. 2) and causal sufficiency (A3), conditioning on the full history $X_{<t}$ blocks every active path from $E_{t-\ell}$ to T via the Causal Markov Condition. Hence the true CMI vanishes:

$$I(T;\, E_{t-\ell} \mid X_{<t}) = 0. \tag{30}$$

By Remark 2, full-history conditioning also closes any backdoor path that would remain open under the coarsened binary history $E_{<t}$. Thm. 1 then gives $\hat{I}_N \le 2\Phi\Big(\sqrt{\epsilon_T/2}\Big)$ asymptotically, so the edge is rejected.

Since the accept/reject decision is correct for every candidate $E_{t-\ell}$, $\ell = 1, \dots, t-1$, independently of whether $T = E_{t+1}$ (the $\mathcal{X} \to \mathcal{X}$ regime) or $T = Y_j$ (the $\mathcal{X} \to \mathcal{Y}$ regime), we obtain $\widehat{\mathrm{Pa}}(T) = \mathrm{Pa}_{\mathcal{G}^s}(T)$.

By induction, $\widehat{\mathrm{Pa}}(T) = \mathrm{Pa}_{\mathcal{G}^s}(T)$ holds for all $t = 1, \dots, L$ and all $T \in \mathcal{T}_t$. The sample-time causal graph $\mathcal{G}^s_{\mathcal{X}\mathcal{X}}$ is the union of the intra-chain parent sets; the sample Markov boundary graph $\mathcal{G}^s_{\mathcal{X}\mathcal{Y}}$ is the union of the cross-chain parent sets. Their union $\mathcal{G}^s = \mathcal{G}^s_{\mathcal{X}\mathcal{X}} \cup \mathcal{G}^s_{\mathcal{X}\mathcal{Y}}$ is therefore recovered. *Note:* As $\epsilon \to 0$ since $\Phi\Big(\sqrt{\epsilon_T/2}\Big) \to 0$, under standard faithfulness, we recover the joint graph for perfect density estimation. □

D.5 Proof of Lem. 1 (Data Encoding Cost Deviation under the ϵ-Oracle)

Proof. Fix a candidate edge $u \to v$ and a matched position $(k, t) \in \mathcal{M}_{uv}$. Let $Z_+^{(k,t)}$ and $Z_-^{(k,t)}$ denote the conditioning sets with and without the candidate cause $E^{(k)}_{t-\ell} = u$ (randomized mediators for $\mathcal{X} \to \mathcal{X}$; trivial for $\mathcal{X} \to \mathcal{Y}$ by the two-chain DGP). To lighten notation, write

$$p_\pm \;=\; P(T^{(k)} = t^{(k)}_{t,v} \mid Z_\pm^{(k,t)}), \qquad q_\pm \;=\; P_\theta(T^{(k)} = t^{(k)}_{t,v} \mid Z_\pm^{(k,t)}),$$

so $\ell_t^{(k)} = \log p_+ - \log p_-$ and $\ell_t^{(k),\theta} = \log q_+ - \log q_-$.

By the triangle inequality,

$$\left|\ell_t^{(k),\theta} - \ell_t^{(k)}\right| \;\le\; |\log q_+ - \log p_+| \;+\; |\log q_- - \log p_-|. \tag{31}$$

Each summand is handled identically; we bound the generic term $|\log q - \log p|$ for $p, q \in (0, 1)$.

For any $a, b \in (0, 1)$, $|\log a - \log b| \le |a - b| / \min(a, b)$ Cover & Thomas (2006). Hence

$$|\log q - \log p| \;\le\; \frac{|q - p|}{\min(p, q)}. \tag{32}$$

The binary target T_v is a deterministic function of the full categorical outcome X_t, so by the data processing inequality for KL divergence,

$$D_{\mathrm{KL}}\big(\mathrm{Bern}(p) \,\|\, \mathrm{Bern}(q)\big) \;\le\; D_{\mathrm{KL}}\big(P(\cdot \mid Z) \,\|\, P_\theta(\cdot \mid Z)\big) \;\le\; \epsilon_T^{(k,t)} \;\le\; \epsilon,$$

where $\epsilon_T^{(k,t)}$ is the local binary approximation error (Prop. 2). Pinsker's inequality then gives

$$|q - p| \;\le\; \mathrm{TV}\big(\mathrm{Bern}(p), \mathrm{Bern}(q)\big) \;\le\; \sqrt{\epsilon_T^{(k,t)}/2}. \tag{33}$$

The denominator $\min(p, q)$ involves p, which we cannot access. However, q is computed directly from the frozen AR model, and combining equation 33 with the triangle inequality yields

$$p \;\ge\; q - |q - p| \;\ge\; q - \sqrt{\epsilon_T^{(k,t)}/2}. \tag{34}$$

Two cases arise. Case A ($q > \sqrt{\epsilon_T^{(k,t)}/2}$): equation 34 certifies $p > 0$, and

$$\min(p, q) \;\geq\; q - \sqrt{\epsilon_T^{(k,t)}/2} \;>\; 0. \tag{35}$$

Case B ($q \leq \sqrt{\epsilon_T^{(k,t)}/2}$): The Pinsker bound cannot certify $p > 0$; the log-ratio may be unbounded. We declare this position under-resolved at the current oracle quality and exclude it from the MDL accumulation. At oracle-optimal quality ($\epsilon \to 0$) Case B never triggers: any position with $q = P_\theta(T_v \mid Z) > 0$ is eventually resolved as $\epsilon_T^{(k,t)}$ shrinks below $2q^2$.

Combining equation 32–equation 35 over both conditioning sets and both non-excluded cases,

$$\left|\ell_t^{(k),\theta} - \ell_t^{(k)}\right| \;\leq\; \sqrt{\epsilon_T^{(k,t)}/2} \cdot \underbrace{\left[\frac{\mathbf{1}\{q_+ > \sqrt{\epsilon_T^{(k,t)}/2}\}}{q_+ - \sqrt{\epsilon_T^{(k,t)}/2}} + \frac{\mathbf{1}\{q_- > \sqrt{\epsilon_T^{(k,t)}/2}\}}{q_- - \sqrt{\epsilon_T^{(k,t)}/2}}\right]}_{\kappa_t^{(k)}(u,v)}. \tag{36}$$

Both factors of $\kappa_t^{(k)}(u, v)$ are computable from the two forward passes through P_θ already required to evaluate $\ell_t^{(k),\theta}$; no additional model queries are needed.

Summing equation 36 over matched positions and using $\epsilon_T^{(k,t)} \leq \epsilon_T$,

$$\left|\Delta L^\theta - \Delta L\right| \;\leq\; \sqrt{\epsilon_T/2} \cdot \underbrace{\sum_{(k,t)\in\mathcal{M}_{uv}} \kappa_t^{(k)}(u, v)}_{\eta_{uv}}. \tag{37}$$

This yields the claim of Lem. 1. □

Remark 4 (η_{uv} is a well-defined function of the model and data)**.** *Every term in η_{uv} is observable without access to the true distribution P: $q_\pm$ is computed from the frozen AR forward pass, and ϵ_T is read off the training diagnostics (Appendix H.2). The indicator on each term makes η_{uv} robust to under-resolved positions — excluding them rather than letting a spurious $1/0$ dominate the sum.*

E Related Work

E.1 Comparison

E.2 Generative Model for CI-testing

Our work can be viewed as the discrete sequential analog of the generative-CI-test lineage initiated by Sen et al. (2017) and extended via Bellot & van der Schaar (2019); Duong & Nguyen (2023); Yang et al. (2025) and score-based variants. These methods learn a fresh conditional generator per CI query over continuous i.i.d. data; SEQ2CAUSE instead repurposes a single, pretrained, autoregressive model whose conditional is native to the data modality (discrete sequences), amortizing density estimation across all candidate edges. We provide theoretical bound analysis (Thm. 1) for the sample-level and extend it to the population-level with the MDL error-correction term (Lem. 1) to handle population aggregation.

E.3 LLM as Causal Reasoner

A parallel line of work treats LLMs themselves as causal oracles, either answering pairwise causal-direction queries from variable metadata Vashishtha et al. (2025), serving as imperfect experts TAKAYAMA et al. (2025). These methods rely on the semantic knowledge encoded in LLM pretraining and are critically limited by hallucination, prompt sensitivity, and the absence of statistical guarantees — as recently formalized by Mirzadeh et al. (2025). SEQ2CAUSE uses the AR model in a fundamentally different way: not as a textual oracle queried in natural language, but as a calibrated

Table 4: **Comparison of causal discovery methods for event sequences**. **Discrete**: Operates on symbolic sequences. **Non-Param.**: No assumed functional form. **Single Stream**: Inference from one observed sequence. $\mathcal{X}\to\mathcal{X}$: Event-to-event dependencies. $\mathcal{X}\to\mathcal{Y}$: Event-to-outcome. **Scalable**: Tractable at $|\mathcal{X}|>10^3$ (linear complexity in $|\mathcal{X}|$). **RCA**: Sample-level root-cause identification from a single sequence without topology assumptions or interventional data.

Method	**Discrete**	**Non-Param.**	**Single Stream**	$\mathcal{X}\to\mathcal{X}$	$\mathcal{X}\to\mathcal{Y}$	**Scalable**	**RCA**
Constraint-based (PCMCI, FCI)	×	✓	×	✓	×	×	×
Score-based (DYNOTEARS)	×	×	×	✓	×	×	×
Granger (TCDF, CAUSE)	×	×	×	✓	×	×	×
Noise-based (VarLiNGAM)	×	×	×	✓	×	×	×
Hawkes / TPP (THP, SHTP)	✓	×	×	✓	×	×	×
Info-Theoretic (NPHC, CASCADE)	✓	✓	×	✓	×	×	×
Multi-label (IAMB, CMB)	✓	×	×	×	✓	×	×
RCA methods (AERCA Han et al. (2025), CoE Yao et al. (2024))	×	×	○	○	×	×	○†
SEQ2CAUSE (Ours)	✓	✓	✓	✓	✓	✓	✓

† AERCA performs RCA on continuous multivariate streams with known dimensionality $|\mathcal{X}| \leq 100$; CoE requires a known service topology and labeled incident history.

conditional density estimator over the event alphabet, with CMI tests grounded in the model's explicit output distribution and requires no assumptions regarding the data generating process Rohekar et al. (2023). The guarantee travels through ϵ, not through the prompt. Therefore, SEQ2CAUSE aim to enhance an LLM's reasoning capability via a causal graph, similarly to Math et al. (2026).

E.4 Data Settings

Table 5: **Data Paradigm Comparison.** Contrast between traditional event sequence causal discovery (multivariate) and our single sequence setting (Session-based/NLP-like). Standard methods require the structure on the left and scale poorly to the structure on the right.

Feature	**Traditional Approach** (e.g., Hawkes, Granger, PCMCI)	**Ours (seq2cause)** (Autoregressive Model)
Input Format	**Long / Vertical Stream**	**Wide / Horizontal Batches**

Structure	`Time  Event_Type`
	`0.0  E_A`
	`0.4  E_B`
	`1.2  E_C`
	`...  ...`
		`Seq_ID  Sequence (Tokens)  Time`
		`0  [E_A, E_B, E_A, ...]  [0.0, 1.3, 1.4, ...]`
		`1  [E_C, E_D, E_G, ...]  [0.0, 1.2, 5.9, ...]`
		`...  ...`

Dimensionality	Low $\|\mathcal{X}\| \leq 100$	Massive $\|\mathcal{X}\| > 1,000$
Discovery Scope	Global (Graph of all processes)	Local / Sample (Summary Graph of specific trace)
Vocabulary Complexity	Often $O(\|\mathcal{X}\|^2)$ or $O(\|\mathcal{X}\|^3)$	$O(L \cdot \|\mathcal{X}\|)$ (Inference)

F Additional Visualization

F.1 Causal Graph

G Algorithm & Parallelization

G.1 Lagged Effects via Simulated Interventions

To evaluate the lagged effects of an event $E_{t-\ell}$ on E_t, we control for the intermediate events, so-called *mediators* $\boldsymbol{M} = X_{t-\ell+1:t-1}$ by simulating a Controlled Direct Effect (CDE) Pearl (2009) of $E_{t-\ell}$ on E_t.

Definition 7 (Randomized Interventional Do-Operator)**.** *Let $\boldsymbol{M} = X_{t-\ell+1:t-1} \subset \mathcal{X}$ be the set of intermediate events between cause $E_{t-\ell}$ and effect E_t. We define the intervention $do(\boldsymbol{M} \sim Q)$ as*

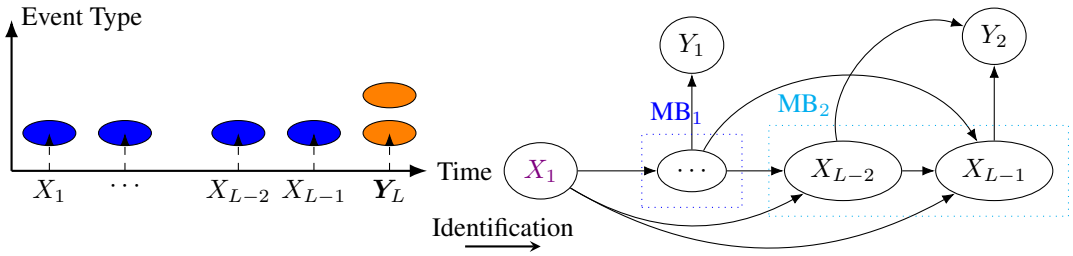


Figure 4: **An example of a sample-level causal graph** event sequence leading to outcomes where MB_1 represents the Markov Boundary of Y_1 and MB_2 the Markov Boundary of Y_2. X_1 is the root-cause of Y_2

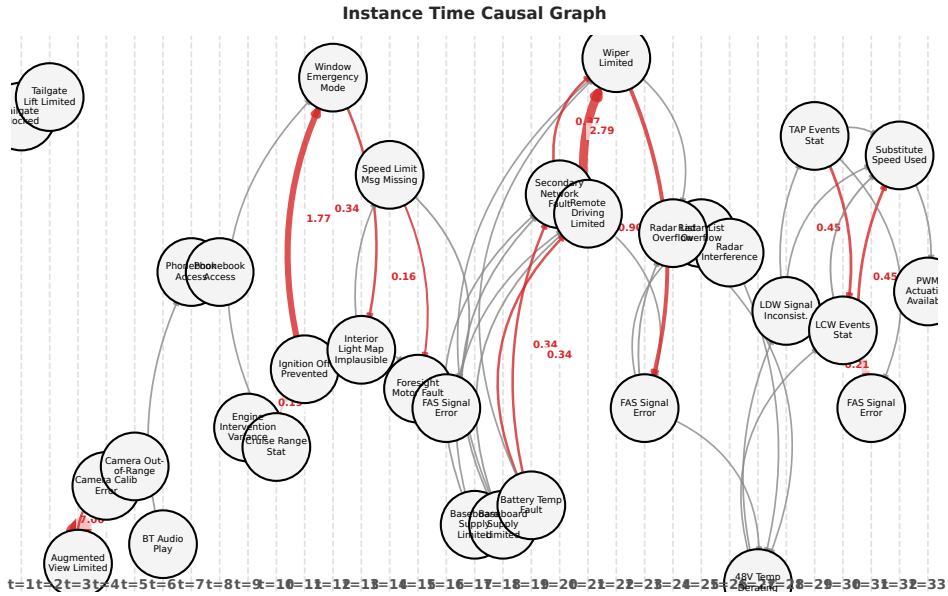


Figure 5: **Sample Time Causal Graph $\mathcal{G}^s_{\mathcal{X}\mathcal{X}}$.** Temporal evolution of a diagnostic defect cascade in a vehicle ($|\mathcal{X}| \approx 29,100$). SEQ2CAUSE effectively captures causal relationships, revealing distinct **error clusters** at different time steps (e.g., initial sensor failures at $t = 3$ triggering mechanical faults at $t = 12$, battery at issue $t = 17$). This enables actionable root-cause analysis by isolating the specific onset of a failure mechanism and their strength using The CMI $\hat{I}_N$

the expectation over counterfactual realizations sampled from a proposal Q (e.g., Uniform over $|\mathcal{X}|$) and average this effect for M counterfactual such as:

$$P(E_t|do(\boldsymbol{M} \sim Q), x_{<t-\ell}) \triangleq \mathbb{E}_{\boldsymbol{M}\sim Q} P(E_t|\boldsymbol{M}, x_{<t-\ell}) \approx \frac{1}{M}\sum_{l=1}^{N} P(E_t \mid \mathbf{m}^{(l)}, x_{<t-\ell}) \tag{38}$$

Remark 5. *This effectively marginalizes out the intermediate causal mechanisms only if we assume that there are no hidden confounders (Assumption 3).*

Using the previous definition, we modify the standard information gain to detect lagged information gain from $E_{t-\ell}$ to E_t, namely $I^{\mathcal{L}}_G$:

Definition 8 (Lagged Information Gain)**.** *Let $E_{t-\ell}$ be the cause, E_t be the effect and $\boldsymbol{M} = X_{t-\ell+1:t-1}$ the set of intermediate events. The Lagged Information Gain $I^{\mathcal{L}}_G$ is defined:*

$$I^{\mathcal{L}}_G(E_t; e_{t-\ell} \mid x_{<t-\ell}) \triangleq D_{\mathrm{KL}}\Big(P\big(E_t \mid do(\boldsymbol{M} \sim Q), e_{t-\ell}, x_{<t-\ell}\big) \,\Big\|\, P(E_t \mid do(\boldsymbol{M} \sim Q), x_{<t-\ell})\Big) \tag{39}$$

G.2 Parallelization of seq2cause via Teacher Forcing

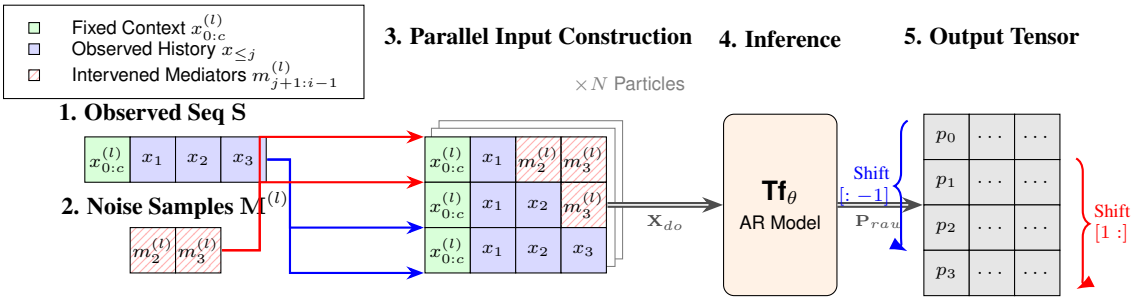


Figure 6: **Overview of SEQ2CAUSE Parallel CI-tests.** We construct a single broadcasted tensor $\mathbf{X}_{do}$ where each row j incrementally fixes the history $x_{\leq j}$ while randomizing the future (staircase pattern). The model processes this tensor in parallel to produce raw probabilities $\mathbf{P}_{raw}$ (grey). We then compute the CMI by comparing adjacent rows: the distribution at row $j-1$ serves as the baseline ($\mathbf{P}_{base}$, blue) for the intervention at row j ($\mathbf{P}_{do}$, red).

G.3 Scalability of seq2cause

Sample-Level To parallelize the Monte-Carlo estimation (Eq. equation 8) and counterfactual sampling (Eq. 39) we avoid computing the CMI for the full trajectory $x_{0:t-1}$ but a truncated version, called context as

$$x_{<t} \approx x_{0:c} \text{ for } c < t \text{ and } 0 < c \ll L$$

We argue that it makes it possible to parallelize the CI-tests by having one common dimension for the sampling and renders the problem feasible on GPUs. We usually take $c = \max(0.1L, 20)$ in our experiments. Although it might break Markovianity for long sequence, empirical results show robustness to this truncation. We provide ablation to unseen sequence lengths during training and show our method to be robust to high delayed effects in Fig 9.

Sparse Approximation for lagged effect in $\mathcal{X} \to \mathcal{X}$. For each time step t, we must perform t CI-tests (one for every potential lag per step). Thus, for a sequence of length L, the total number of CI-tests is given by $\sum_{t=1}^{L} t = \frac{L(L+1)}{2}$ which grows quadratically with the sequence length. As a result, even on multiple GPUs, it becomes tricky to infer for long sequences $L > 100$.

To solve this, we propose a sparse variant for which we bound the lagged effects of previous events on future events up to a memory m. Thus the DGP $\{X_t\}$ becomes an m-order Markov chain. With $m \ll L$, SEQ2CAUSE scales linearly with the sequence length L. The memory complexity transitions from:

$$\mathcal{O}(N \cdot (L-c) \cdot L \cdot |\mathcal{X}|) \xrightarrow{\text{Bounded Memory}} \mathcal{O}(N \cdot m \cdot L \cdot |\mathcal{X}|)$$

Population-Level. Population-level inference via the MDL criterion (Eq. 12) is strictly cheaper than sample-level inference: it requires no Monte-Carlo estimation. For each matched position $(k,t) \in \mathcal{M}_{uv}$, the per-step log-ratio $\ell_t^{(k),\theta}(u,v)$ reduces to two forward passes through the frozen AR model — one at $Z_u^{(k,t)}$ (cause present) and one at $Z_\varnothing^{(k,t)}$ (cause absent) — whose outputs are directly compared as a log-ratio. No particle sampling, no history simulation, no stochastic approximation: the total accumulated evidence $\hat{\Delta}_m(u,v)$ is obtained by summing n_{uv} (number of event pairs $u \to v$) such log-ratios deterministically across the dataset, at a memory complexity of:

$$\mathcal{O}(L^2 \cdot M \cdot |\mathcal{X}|)$$

where M is the number of sampled mediators required to evaluate the do-operator (empirically M=4 suffices, Appendix 12). For the self-corrected threshold τ_{uv}^* (Cor. 2), the error correction term $\sqrt{\epsilon_T/2} \cdot \eta_{uv}$ requires the local binary approximation error ϵ_T at each matched position. This quantity is obtained at zero additional cost: recalling that the cross-entropy loss decomposes as $\mathcal{L}_{\text{AR}}(\theta) = H(P) + D_{\text{KL}}(P \| P_\theta)$, the generative error ϵ_T is read off directly as the binary cross-entropy minus the entropy floor $H(P)$. No additional model queries are required. This inference

is fully parallelizable on GPUs because it uses the parallel lagged-effect implementation shown in Fig. 6.

H EVALUATION

Synthetic experiments settings. We used an $ml.g5.4xlarge$ instance from AWS Sagemaker, which contains 8 vCPUs and 1 NVIDIA A10G as GPU with 24GiB for training and inference.

Vehicle diagnostics experiments settings. We used an $ml.g5.12xlarge$ instance from AWS Sagemaker, which contains 8 vCPUs and 4 NVIDIA A10G as GPU with 24GiB for inference.

H.1 NONLINEAR SCMS

As we saw, standard causal discovery algorithms for multivariate time series (e.g., PCMCI, Neural Hawkes Processes, CASCADE) are not applicable in our setting. Moreover, evaluating causal discovery in high-dimensional event sequences is notoriously difficult due to the lack of ground-truth annotations in real-world traces (e.g., server logs, medical records). Furthermore, generic high-order MC are computationally intractable to materialize due to the state space $|\mathcal{X}|^m$ of $\{X_t\}$ and unlearnable in high dimensions without structural assumptions.

Synthetic Data. We introduce a synthetic benchmark based on a nonlinear Structural Causal Model (SCM). Let $\mathcal{X}$ be the set of event types, $s = (x_1, \ldots, x_L) \in \mathcal{X}^L$ a discrete sequence, and h the history window. The SCM generative process is

$$P(X_t \mid X_{t-h:t-1}) = \mathrm{softmax}\left(\mathbf{b} + \sum_{k=1}^{h} e^{-(k-1)} \, \mathbf{W}[x_{t-k}] + \mathrm{ReLU}\big(\big[E_{x_{t-h}}, \ldots, E_{x_{t-1}} \big] \, \mathbf{W}_1 \big) \, \mathbf{W}_2 \right), \tag{40}$$

where $\mathbf{W} \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ is a sparse interaction matrix (sparsity $= 0.9$, mixed-sign), $E \in \mathbb{R}^{|\mathcal{X}| \times 16}$ are fixed random embeddings, and $\mathbf{W}_1 \in \mathbb{R}^{h \cdot 16 \times 64}$, $\mathbf{W}_2 \in \mathbb{R}^{64 \times |\mathcal{X}|}$ form a one-hidden-layer MLP that captures non-linear interactions. The ground-truth adjacency is determined by counterfactual intervention (Pearl, 2009): $i \to j$ if $\mathbb{E}\big[\mathrm{KL}(P_{\mathrm{orig}} \,\|\, P_{\mathrm{do}(x_{t-k}=\epsilon)})\big] > \delta$. We tune the sparsity of the weight matrix $\mathbf{W}$ to obtain the Shannon redundancy Shannon (1951) as $1 - \frac{H(P)}{H_{max}} = 1 - \frac{H(P)}{\log(|\mathcal{X}|)}$ superior or equal to 58% across the benchmarked SCMs.

Evaluation via intervention. We always evaluate SEQ2CAUSE on the summary graph. To measure performance, we perform atomic interventions by measuring the average KL divergence over 10 by uniformly randomizing $E_{t-\ell}$ and measuring the average KL divergence over 10 counterfactual between post-intervention and observational distributions of E_t. If the divergence is above $\tau > 0.05$, an edge $E_t \to E_{t-\ell}$ exists in the summary graph.

H.2 ϵ IN PRACTICE

What ϵ stands for. ϵ replaces parametric structural assumptions (linearity, Gaussianity, additive noise) with a single monitorable quantity: the excess cross-entropy

$$\epsilon \;=\; \mathcal{L}_{\mathrm{AR}}(\theta) - H(P), \tag{41}$$

measurable during AR pretraining without ground-truth causal labels.

Practical estimation. Two quantities in equation 41 need handling:

(i) The cross-entropy $\mathcal{L}_{\mathrm{AR}}(\theta)$ is the standard held-out validation loss, directly available from training diagnostics.

(ii) The entropy floor $H(P)$ is inaccessible in practice. We approximate it by the minimum held-out validation loss achievable as training converges, under the standard assumption that $\mathcal{L}_{\mathrm{AR}}(\theta^*) \to H(P)$ as capacity and data grow. Empirically we find $\epsilon < 0.05$ is sufficient for structural recovery (Fig. 2(a), phase transition at $\epsilon \approx 0.1$).

Comparing ϵ across vocabularies: the Oracle Score. Different DGPs have different entropy floors $H(P)$, making raw ϵ incomparable across settings. For the nonlinear-SCM experiments (Fig. 2(b)) we use the Oracle Score $\hat{\epsilon}$ as a normalized estimator of the excess entropy:

$$\hat{\epsilon}(P_\theta) \;=\; \frac{\mathcal{L}_{\mathrm{AR}}(\theta) - H(P)}{H_{\max} - H(P)}, \tag{42}$$

where $H_{\max} = \log |\mathcal{X}|$ is the maximum entropy (uniform noise). $\hat{\epsilon} \to 0$ implies convergence to the theoretical limit ($P_\theta \to P$); $\hat{\epsilon} = 1$ is the uniform prior.

H.3 Population-Level $\mathcal{X} \to \mathcal{X}$ baselines

We compare SEQ2CAUSE against several families of baselines for the $\mathcal{X} \to \mathcal{X}$ regime that span association-based, information-theoretic, cascade-based, point-process, and neural attribution methods. Like for SEQ2CAUSE we apply an individual significance threshold τ to the estimated adjacency matrix to binarize it. Importantly, even when we grant PMI and TF-IDF oracle access to the threshold that maximises F1 on the test set — an advantage unavailable in practice — SEQ2CAUSE still outperforms them (Appendix. 7). In contrast, SEQ2CAUSE requires no threshold calibration: its MDL criterion derives a principled decision boundary directly from the density error bound.

Pointwise Mutual Information (PMI). For each ordered pair (i, j), we count co-occurrences within a lag window of size h:

$$\widehat{\mathrm{PMI}}(i \to j) = \log \frac{\hat{P}(x_t = j,\, x_{t-k} = i \text{ for some } k \le h)}{\hat{P}(x_{t-k} = i)\, \hat{P}(x_t = j)}, \tag{43}$$

clipped at zero (positive PMI). PMI measures symmetric association; it is included as a simple count-based reference and does not exploit temporal order beyond the lag constraint (Church & Hanks, 1990).

TF-IDF. We adapt the classical retrieval score (Sparck Jones, 1988) for identifying causal pairs over a discrete event vocabulary $\mathcal{V}$. Treating each source type i as a document whose terms are the event types that follow it within h steps, we define

$$\mathrm{Score}(i \to j) \;=\; \underbrace{\frac{\mathrm{count}(i \rightsquigarrow j)}{\mathrm{count}(i)}}_{\mathrm{f}(i \to j)} \;\times\; \underbrace{\log\!\left(\frac{|\mathcal{V}| + 1}{|\{i' : \mathrm{count}(i' \rightsquigarrow j) > 0\}| + 1}\right)}_{\mathrm{IDF}(j)}, \tag{44}$$

where $\mathrm{count}(i \rightsquigarrow j)$ is the number of times type j appears within h steps after type i across all observed sequences. The TF term measures how reliably i is followed by j; the IDF term down-weights effect types j that are triggered by many distinct sources, which in this setting correspond to high-frequency background events with no specific cause. Compared to PMI, which penalises both rare causes and rare effects, IDF operates in log-count space and only suppresses effects that lack causal specificity, making it more robust when many event types are infrequent. An edge $i \to j$ is predicted whenever $\mathrm{Score}(i \to j) > \tau$.

Cascade (MDL-based). The Cascade method (Cüppers et al., 2024) frames causal discovery as a minimum description length (MDL) problem: it selects the directed graph G that minimises the two-part code $L(G) + L(\mathbf{x} \mid G)$, where the second term encodes the observed event sequences given G. An event i is declared a cause of j if adding the directed edge $i \to j$ compresses the data by more than the cost of encoding the edge. We use the reference implementation of Cüppers et al. (2024)[5] with `max_delay`$= h$ and `allow_instant`$=$ False, treating each sequence as an independent stream separated by a temporal gap $> h$ to prevent spurious cross-sequence edges. Cascade returns a binary adjacency matrix directly; no threshold is applied.

SHTP; Structural Hawke Processes. We use the implementation of Qiao et al. (2023) for SHTP[6] that model the event apparition as a Hawkes Process. We treat each sequence as an independent stream exactly like Cascade.

[5] `https://github.com/joschac/cascade`

[6] `https://github.com/DMIRLAB-Group/SHP`

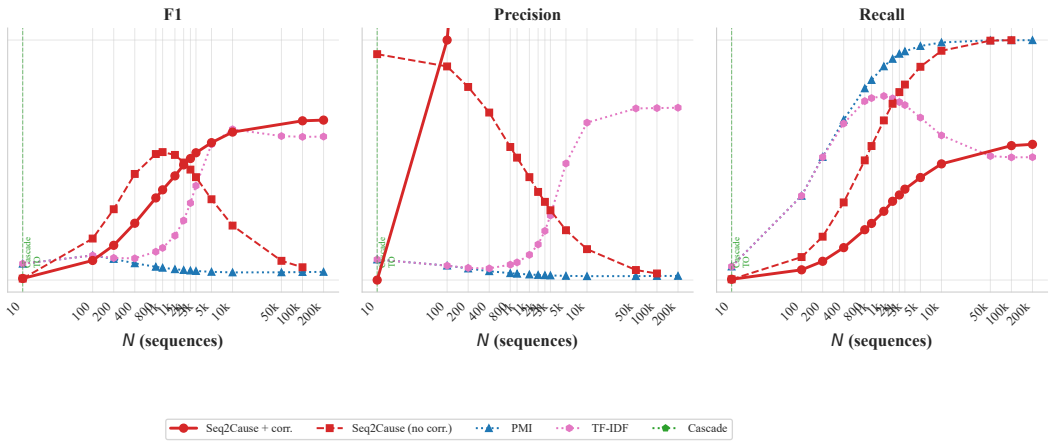


Figure 7: **Population-level $\mathcal{X} \to \mathcal{X}$ baseline comparison with oracle threshold selection.** F1, Precision, and Recall as a function of the number of observed sequences N for each method on the nonlinear SCM benchmark ($|\mathcal{V}| = 1000$, $m = 6$). For PMI and TF-IDF, scores are binarised at the threshold τ^* that maximises F1 on the held-out ground truth; this constitutes an upper bound on their achievable performance, since such oracle calibration is unavailable at test time. SEQ2CAUSE uses its MDL-derived threshold (Corollary 2) with no access to the ground truth. Despite this advantage, PMI and TF-IDF remain below SEQ2CAUSE across all sample sizes, confirming that the performance gap is not an artefact of threshold miscalibration but reflects a fundamental limitation of association-based scores for causal discovery.

H.3.1 Differentiation with Granger

SEQ2CAUSE works like a constraint based algorithm Spirtes & Glymour (1991) whereas Granger causality's core logic remains on predictability.

First, Granger causality assumes a fixed parametric form (linear VAR) for the DGP and requires stationarity; SEQ2CAUSE assumes neither, replacing structural assumptions with the single monitorable quantity ϵ (Assumption 1). Second, for lagged effects ($\ell > 1$) in the $\mathcal{X} \to \mathcal{X}$ regime, SEQ2CAUSE performs a do-intervention on intermediate mediators rather than conditioning on them, which distinguishes direct from total effects — a distinction Granger methods cannot make without additional structural assumptions Pearl (1988). Third, under causal sufficiency (A3), the full-history conditioning set $X_{<t}$ satisfies the back-door criterion for every candidate cause $E_{t-\ell}$, so the CMI test identifies the direct causal effect rather than a regression residual. The empirical superiority over Neural Granger (same backbone, probability difference instead of CMI, ΔF1 >20 points, Tab. 2) isolates the contribution of the interventional CMI signal over regressing on predictive residuals — confirming that the difference is not merely architectural.

H.4 Ablations

H.4.1 Robustness analysis

Scalability to sequence length. Figure 9(b) evaluates performance and GPU memory on sequences beyond the AR model's training window. The sparse variant (m-bounded lags) achieves $\mathcal{O}(m \cdot L \cdot |\mathcal{X}|)$ memory versus $\mathcal{O}(L^2 \cdot |\mathcal{X}|)$ for the full variant, enabling inference on sequences far exceeding the training length. Precision remains stable ($\approx$0.95) while Recall degrades gracefully, again confirming the conservative failure mode.

Deep temporal dependencies. Figure 9(c) stress-tests the parallelised intervention mechanism by extending the SCM memory horizon to $m = 48$. SEQ2CAUSE maintains F1 > 0.80 up to $m = 20$ (one third of the sequence length), confirming effective capture of long-range causal signals without exponential degradation.

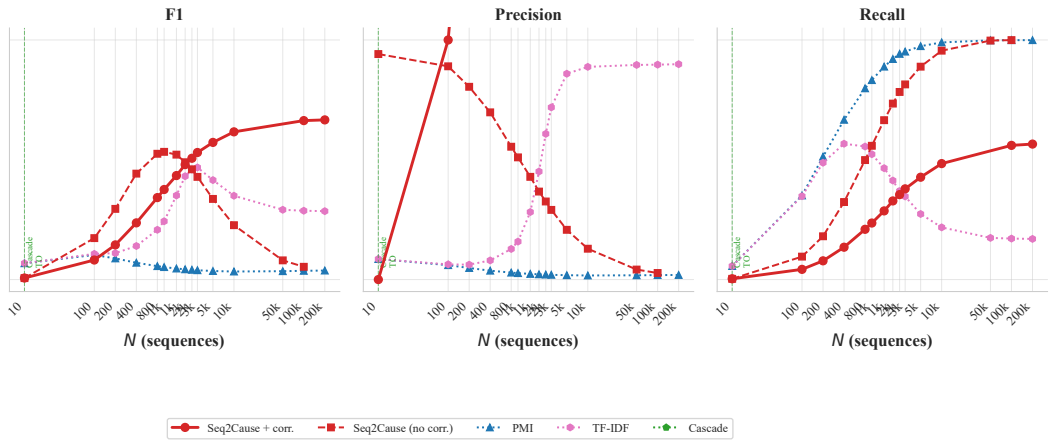


Figure 8: **Population-level $\mathcal{X} \to \mathcal{X}$ discovery as a function of number of sequences** n ($|\mathcal{X}|$=1,000, nonlinear-SCM, ϵ=0.05). The green dashed line marks the CASCADE timeout (N=100); SHP times out at the same scale and is omitted beyond that point. *Correlation baselines* (PMI, TF-IDF): Recall saturates near 1.0 early as heuristic thresholds admit all co-occurring pairs, while Precision remains near zero — confirming that frequency-based methods cannot distinguish causal structure from spurious association regardless of evidence volume. SEQ2CAUSE *w/o correction* (dashed red): Precision peaks near N=200 then collapses to 4.2% as oracle approximation errors $\sqrt{\epsilon_T/2} \cdot \eta_{uv}$ accumulate without correction, empirically validating Lem. 1 — the uncorrected threshold becomes equivalent to no threshold at scale. SEQ2CAUSE *w/ correction* (solid red): the adaptive τ^*_{uv} absorbs error accumulation, maintaining Precision above 0.80 across the full range while Recall grows steadily as genuine weak causal signals clear the corrected threshold — the only method achieving balanced F1 at any scale. (L=64, m=6, τ^*_{uv} from Cor. 2.)

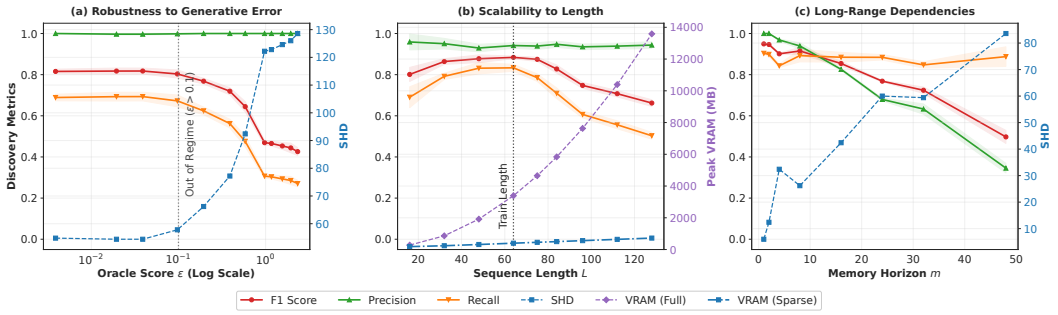


Figure 9: **Sample-Level ($\mathcal{X} \to \mathcal{X}$) Robustness and Scalability Analysis** ($|\mathcal{X}| = 1000, N = 128, \tau = 10^{-4}, L = 64$). **(a) Robustness to Generative Error:** Performance as a function of the model's oracle score ϵ. SEQ2CAUSE exhibits a phase transition, recovering structure even for imperfect models ($\epsilon < 0.1$) and maintaining high Precision even as fidelity degrades. **(b) Scalability to Length:** Performance and GPU memory usage vs. sequence length L. The **Sparse** variant (Appendix. G.3) demonstrates linear memory scaling ($O(mL)$), enabling inference on sequences far exceeding the training length ($L = 64$), whereas the Full variant scales quadratically. **(c) Long-Range Dependencies:** Robustness to increasing delayed-effects m. SEQ2CAUSE maintains F1 > 0.8 even as dependencies span one third of the sequence ($m = 20$), confirming the method's ability to capture distant causal mechanisms.

H.4.2 SAMPLE-LEVEL: THRESHOLD SENSITIVITY TO LAG-DECAYING CAUSAL STRENGTH

The adaptive threshold of Cor. 2, $\tau^*_{uv} = c_{\text{type}} + \sqrt{\epsilon_T/2} \cdot \eta_{uv}$, already ties the acceptance boundary for an edge $u \to v$ to a margin term rather than to a fixed constant: η_{uv} grows with the amount of matched evidence for that pair, so τ^*_{uv} self-adjusts to the causal strength actually observed. Fig. 10 shows the practical consequence at the population level: the uncorrected threshold collapses (Precision $\to 0$ as

N grows) while the corrected τ^*_{uv} keeps Precision above 0.80 across three orders of magnitude in N, confirming that a margin-aware threshold, not a fixed cutoff, is what makes the method robust to the amount of accumulated evidence.

This matters directly for lagged effects, since our synthetic SCM (Eq. 40) injects an exponentially decaying causal weight $e^{-(k-1)}$ at lag k: the true causal strength between $E_{t-\ell}$ and its effect decreases with ℓ by construction. A single, lag-independent threshold would therefore under-call deep lags long before it over-calls shallow ones. Fig. 9(c) shows this is not what happens in practice: SEQ2CAUSE maintains F1 > 0.80 up to $m = 20$ because the per-edge margin term η_{uv} (accumulated over all matched positions at that lag, Appendix D.5) shrinks together with the decaying signal, keeping the acceptance boundary proportionate to the remaining causal strength instead of penalizing deep lags with the same fixed cutoff used for shallow ones.

In practice, choosing a threshold without access to ϵ_T or to Cor. 2 (e.g. when auditing a population summary graph post hoc) reduces to picking a cutoff on the empirical CMI score distribution. Two choices matter more than the exact statistic used. First, scores should be pooled across all available sequences before fitting a single cutoff: a threshold fit per sequence is unstable at the modest N typically available per source system, since η_{uv} is itself accumulated across sequences. Second, the right unsupervised cutoff family depends on the expected sparsity of the true graph: rank/percentile-based cutoffs are more robust when true edges are a small fraction of candidate pairs (the common case for causal graphs), while variance-based cutoffs (e.g. Otsu, Gaussian-mixture) are competitive only when the two score populations are comparably sized. When even a small labeled subset is available, it should be used to validate the chosen cutoff rather than trusting any single unsupervised rule in isolation.

Table 6: **Double degeneracy of classical MB algorithms on the Sample-Level $\mathcal{X} \to \mathcal{Y}$.** Weighted F1 (%) on the vehicle diagnostic Markov-boundary recovery task, evaluated on restricted vocabularies ($|\mathcal{X}|$ = top-K most frequent event types) and reduced sample sizes n. At low (K, n), classical LSL algorithms run to completion but collapse to F1 $= 0$ because expected co-occurrence $\mathbb{E}[\#(u, Y_j)] \approx n/(K|\mathcal{Y}|)$ falls below the threshold required for any asymptotic CI test to reject independence. Scaling up either K or n to alleviate this sample-starvation regime immediately triggers the dual vocabulary-driven intractability regime (†: timeout >3 days on 4×A10G) GPUs. SEQ2CAUSE amortizes conditional density estimation across all (u, Y_j) pairs through a single AR forward pass and is the only method viable in both regimes.

$\lvert\mathcal{X}\rvert$	n	IAMB	CMB	MB-by-MB	PCDbyPCD	MI-MCF	SEQ2CAUSE
500	500	0.0	0.0	0.0	0.0	0.0	**58.4**
500	5,000	2.1	1.8	2.4	0.0	0.0	**59.7**
2,000	500	0.0	0.0	0.0	0.0	0.0	**52.2**
2,000	5,000	0.0	0.0	0.0	–†	0.0	**53.1**
2,000	50,000	–†	–†	–†	–†	–†	**55.3**
29,100	50,000	–†	–†	–†	–†	–†	**46.1**

H.4.3 POPULATION-LEVEL: NUMBER OF SEQUENCE n

H.4.4 SAMPLE-LEVEL: NUMBER OF PARTICLES N

We observe in Fig. 11 that the number of particles N is a crucial parameter to increase Recall and reduce SHD. After $N > 256$, however, we don't observe significant changes.

H.4.5 POPULATION-LEVEL: NUMBER OF COUNTERFACTUALS M FOR MEDIATORS

We ablate the number of counterfactuals for the mediators M in Fig. 12. We observe that we don't need exceeding number of counterfactuals and that generally 2 to 4 is sufficient.

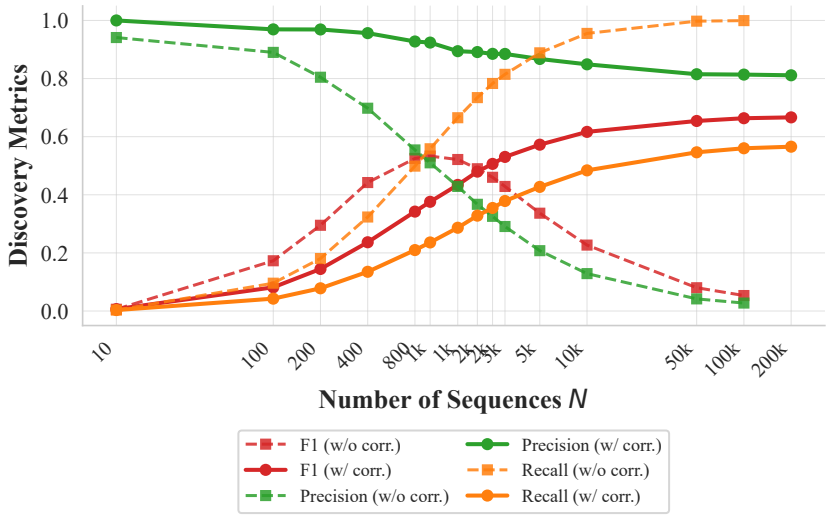


Figure 10: **Self-correcting MDL threshold τ^*_{uv} prevents precision collapse as evidence accumulates.** Population-level causal discovery metrics as a function of the number of sequences N, comparing the uncorrected MDL threshold (dashed, *w/o corr.*) against the generative-error-corrected threshold of Cor. 2 (solid, *w/ corr.*). *Without correction:* Precision degrades catastrophically beyond $N{\approx}1{,}000$ sequences — each additional sequence accumulates approximation error from the ϵ-oracle, progressively lowering the effective threshold until spurious edges flood the graph (Recall $\to$ 1, Precision $\to$ 0). *With correction:* The adaptive term $\sqrt{\epsilon_T/2} \cdot \eta_{uv}$ in τ^*_{uv} absorbs this drift, maintaining Precision above 0.80 across the full range $N \in [10, 200{,}000]$ while Recall grows steadily as genuine weak causal signals accumulate evidence. The crossover at $N{\approx}2{,}000$ marks the point where uncorrected evidence accumulation overtakes the model cost c_{type} (synthetic non-linear SCMs: $|\mathcal{X}|{=}1000, L = 64$).

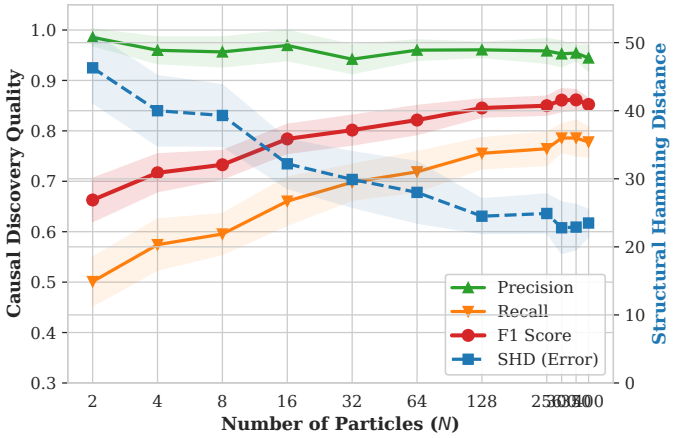


Figure 11: **Evolution of the Sample-Level $\mathcal{X} \to \mathcal{X}$ Causal Discovery in Function of the Number of Particles** N at $|\mathcal{X}| = 1000, m = 6, L = 64$ on non-linear SCMs.

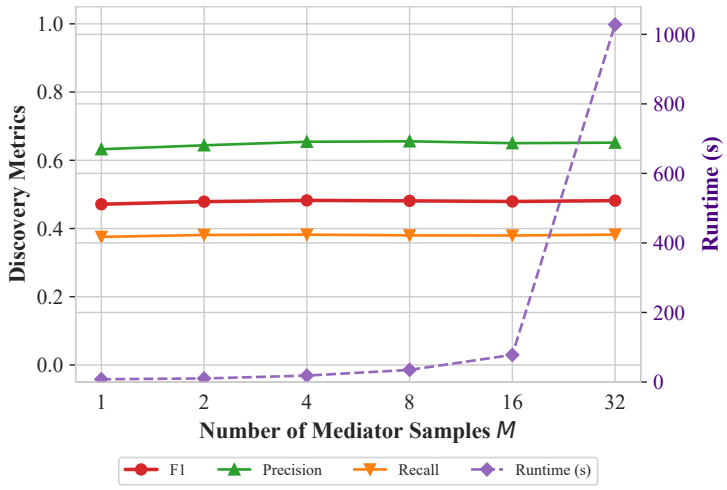


Figure 12: **Evolution of the Population Causal Discovery in Function of the Number of Counterfactuals sampled as Mediators** M at $|\mathcal{X}| = 1000, m = 6, L = 66, n = 2000$