LEARNING GRANGER STRUCTURE OVER TIME AND SYSTEM STATE FROM NON-STATIONARY EVENT SEQUENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Event logs from systems such as semiconductor fabrication lines exhibit non-stationarity in both background activity and causal dependence, including changes in event dependencies and their influence strength. We show that segmentation- and regime-based approaches can absorb within-segment background drift into influence estimates, leading to spurious edges. We introduce CAUSALFIELD, a point-process framework that represents causal influence as a continuous field $C(t, z)$ over time and system state. The framework parameterizes this field with a coordinate-based neural network and learns it directly from the point-process likelihood, allowing excitation to vary with system state while modeling background drift separately. Under the stated conditions, the population field identifies a direct state-conditional Granger graph on the visited (t, z) region, together with diagnostics for assessing whether recovery is supported by the data. On synthetic data, CAUSALFIELD recovers the influence field and reproduces the predicted failure mode of discretized alternatives. On a real alarm log, it remains competitive for thresholded static-graph recovery, while state conditioning improves held-out fit beyond static and time-only variants.

1 INTRODUCTION

Event logs arise in operational systems such as semiconductor fabrication lines, cellular networks, and microservice deployments, as well as in records of daily activities (Bae et al., 2023; Cüppers et al., 2024). Each event is represented by a discrete type and a continuous timestamp (Ordóñez et al., 2013). The need to understand dependencies in these logs has motivated methods for causal discovery among event types (Hu et al., 2026).

However, standard causal discovery methods are built around a different data representation. They assume an $n \times d$ design matrix, with every variable observed at every sampling instant, and operate on those columns (Zheng et al., 2018; Pamfil et al., 2020). An event log provides no such matrix: event types occur asynchronously, inter-event intervals are irregular, and there is no aligned vector of variable values on which to condition. Binning time to obtain such a matrix sacrifices the fine-grained timing information on which the task relies: propagation faster than the bin width appears simultaneous, whereas slower effects are spread across multiple bins.

Under classical Hawkes-based causal discovery, pairwise influence is typically represented as time-invariant, often through an excitation coefficient α_{ij} whose support determines the causal graph (Xu et al., 2016; Cai et al., 2021). This assumption can be violated easily in operational systems as equipment ages, workloads shift, and configurations change. An aging chamber, for example, may raise the background rates of two alarms without changing the dependency between them, yet a stationary model may attribute their coincident increase to excitation and infer a spurious edge (Mossel & Sridhar, 2026) as illustrated in Figure 1. More generally, non-stationarity means that the process generating observed events changes over time. Such change need not imply a change in causal structure: it may arise from variation in background event activity, changes in the strength of existing causal influences, or the appearance or disappearance of causal relations (Saggioro et al., 2020; Rahmani & Frossard, 2025). These distinct sources of non-stationarity can produce similar changes in the observed event sequence while implying fundamentally different root causes.

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

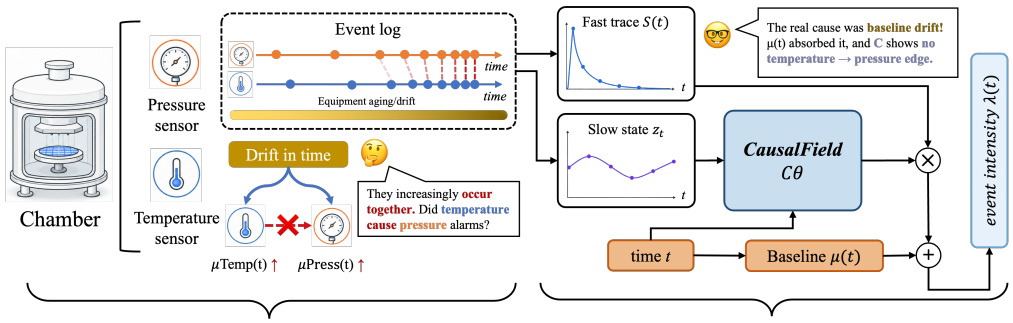

Figure 1: Overview of CAUSALFIELD. (a) Baseline drift makes temperature and pressure alarms co-occur; co-occurrence alone cannot establish causation. (b) The influence field $C_\theta(t, z_t)$, driven by slow state z_t, gates fast excitation $S(t)$ and is added to the baseline $\mu(t)$ to yield the intensity $\lambda(t)$.

Existing methods handle non-stationarity via temporal segmentation or discrete states (Liu et al., 2025), leaving three questions open: (**RQ1**) can background variation be separated from changes in cross-type influence; (**RQ2**) what does a continuous representation over time and history-derived state gain over discretized ones; and (**RQ3**) under what conditions is such structure identifiable.

We represent pairwise causal influence as a continuous field $C(t, z)$ within the point-process intensity,

$$\lambda_j(t) = \varphi\left(\mu_j(t) + \sum_i C_{ij}(t, z_t) S_i(t)\right).$$

S_i captures excitation from recent type-i events, the history-derived state z_t modulates its strength through $C_{ij}(t, z_t)$, and $\mu_j(t)$ models time-varying background activity. Instead of a separate latent-state process, we compute z_t from the observed event history. We parameterize $C(t, z)$ with a coordinate-based neural network, adapting neural-field representations to pairwise event influence over time and state (Mildenhall et al., 2021; Tancik et al., 2020). Our contributions are as follows:

- **Continuous influence field (RQ2).** We represent direct state-conditional Granger influence as a continuous field over time and state, with a separate time-varying background and no block width or regime count.
- **Identification conditions (RQ1, RQ3).** We show that background and influence are pointwise separated under non-degenerate conditional excitation variation, and characterize when this separation and time–state separation fail.
- **Empirical evidence.** CAUSALFIELD leads graph recovery on synthetic data, remains competitive on a real alarm log and across alternative generators, and reduces coefficient error relative to time-only and blocked variants.

2 RELATED WORK

Event sequence causal discovery. Classical causal discovery infers causal relations among a fixed set of variables from i.i.d. observations $\mathbf{X} \in \mathbb{R}^{n \times D}$, typically by recovering a directed acyclic graph over the D variables (Spirtes & Zhang, 2016; Zheng et al., 2018). In the structural causal model view, each variable is generated from its parents and independent noise (Huber, 2024); time-series extensions retain this aligned-sample format by treating each series as a variable and adding lagged dependencies (Assaad et al., 2022; Pamfil et al., 2020; Runge et al., 2019).

Event sequences fit neither setting. Operational systems such as manufacturing and online platforms produce irregular records $\{(e_i, t_i)\}_{i=1}^N$ of discrete event types $e_i \in \{1, \ldots, K\}$ and continuous timestamps $t_i \in \mathbb{R}^+$ (Bacry et al., 2015). Temporal point processes model such streams through conditional intensities, from mutually exciting Hawkes processes (Hawkes, 1971) to neural history-dependent models (Du et al., 2016; Mei & Eisner, 2017; Zuo et al., 2020; Shchur et al., 2021). Most

causal approaches adopt Granger causality (Granger, 1969): type i is non-causal for type j when the intensity of j is conditionally independent of the history of i (Didelez, 2008). For multivariate Hawkes processes, this graph is the support of the excitation kernels (Eichler et al., 2017; Xu et al., 2016). Existing estimators recover this support using penalized likelihood or moment matching (Zhou et al., 2013; Xu et al., 2016; Achab et al., 2017), cardinality regularization and discrete-time extensions (Idé et al., 2021; Qiao et al., 2023), or neural intensity and event-history representations (Zhang et al., 2020; Wu et al., 2024). Other work uses likelihood- or description-length-based graph selection and models flexible delays between individual cause and effect events (Cai et al., 2021; Jalaldoust et al., 2022; Cüppers et al., 2024; Cornanguer et al., 2026). Despite these differences, the resulting causal structure is typically a single graph summarizing the observation window.

Non-stationary event sequences. Prior work models changing event dynamics through a finite collection of locally stationary models, either by partitioning a sequence into stationary segments (Chen et al., 2022) or by conditioning excitation on latent or observed discrete states (Liu et al., 2025; Morariu-Patrichi & Pakkanen, 2021). Such representations impose a discrete resolution on temporal variation: changes can be localized only at the level of segment or state assignments, while within-cell background variation may be conflated with excitation. Recent studies investigate graph recovery for Hawkes processes with time-varying baselines (Mossel & Sridhar, 2026), and account for confounded excitation through latent sub-processes under partial observation (Jin & Huang, 2026). Parallel work on non-stationary time-series causal discovery considers regime-dependent graphs, smoothly varying modules, and MDL-based partitions (Saggioro et al., 2020; Huang et al., 2020; Rahmani & Frossard, 2025; Mameche et al., 2025), but they are not tailored for event-driven causal discovery. In contrast, we model influence as a continuous function of time and a history-derived system state, without choosing a block width, regime count, or latent-state posterior.

3 Causal Field

3.1 Problem and estimand

We observe M sequences of a K-type marked point process on $[0, T]$, $D^{(m)} = \{(t_k^{(m)}, e_k^{(m)})\}$, and recover which event types influence others as these relationships vary over time and system state. Let z_t denote a predictable summary of system condition constructed from the observed event history.

Definition 3.1 (Direct state-conditional Granger structure)**.** *For a visited point (t, z), we define*

$$G^\star(t, z) = \operatorname{supp} C^\star(t, z).$$

Its edge $i \to j$ represents direct excitation-mediated dependence of type j on type i conditional on the history-derived state $z_t = z$.

Under admissible nondegenerate variation of S_i conditional on (t, z, S_{-i}), $C^\star_{ij}(t, z) = 0$ is equivalent to invariance of the conditional intensity to changes in S_i with the remaining coordinates fixed. Because z_t is history-derived, $G^\star(t, z)$ captures excitation-mediated dependence conditional on state, excluding additional predictive effects mediated through changes in z_t. This is a localized direct-path version of the local-independence interpretation for marked point processes (Didelez, 2008). Segmentation-based approaches approximate changing dynamics using piecewise-constant parameters. When the baseline is held constant within a segment, unresolved background drift can be misattributed to excitation, potentially producing spurious positive influence estimates.

3.2 Model

We designed our framework as illustrated in Figure 1. We model the conditional intensity of type j as

$$\lambda_j(t \mid \mathcal{F}_{t^-}) = \varphi\Big(\mu_j(t) + \textstyle\sum_{i=1}^{K} C_{ij}(t, z_t)\, S_i(t)\Big), \qquad S_i(t) = \textstyle\sum_{t_m < t,\, e_m = i} e^{-\beta(t - t_m)}, \quad (1)$$

with $\varphi = \mathrm{softplus}$ and excitation kernel $\kappa(u) = e^{-\beta u}$, where the decay $\beta > 0$ is learned. We use (μ, C) for the generic model and $(\mu^\star, C^\star)$ for the population quantities. We adapt the nonlinear Hawkes construction in which a background component and history-dependent excitation are mapped through a positive link function (Xu et al., 2016). The normalization $\kappa(0) = 1$ fixes the kernel amplitude at zero lag. The model separates the estimand into two channels. The baseline $\mu_j(t)$

represents event activity not attributed to cross-type excitation, whereas $C : (t, z) \mapsto \mathbb{R}_{\geq 0}^{K \times K}$ is the *influence field*. Its entry $C_{ij}(t, z)$ specifies the strength of direct excitation from type i to type j at time t and system state z. We adapt the coordinate-based neural-field view, in which a neural network represents a signal as a function of continuous coordinates (Sitzmann et al., 2020; Mildenhall et al., 2021), to represent pairwise event influence over the coordinates (t, z). For example, the same recent type-i event history can produce different excitation contributions to type j under different system states because $C_{ij}(t, z_t)$ modulates $S_i(t)$. Meanwhile, $\mu_j(t)$ provides a separate time-varying background component, allowing the model to represent changes in background activity without requiring changes in the influence coefficients. The history-derived state is computed from the observed event stream as an exponentially weighted average of event embeddings, initialized at z_0 (Roberts, 1959; Ma et al., 2022). Without pre-specifying a regime count, each recovered edge is indexed by a history-derived system condition. At the population level, exact zeros of $C^\star(t, z)$ determine the absence of direct state-conditional influence (Xu et al., 2016). The fitted neural field is parameterized with a softplus output head, so $\widehat{C}_{\theta,ij}(t, z) > 0$.

History-derived system state. We construct the state from the observed history as

$$z_t = \frac{\sum_{t_k < t} w_k \, \mathrm{emb}(e_k)}{\sum_{t_k < t} w_k}, \qquad w_k = e^{-(t - t_k)/\tau_z}, \tag{2}$$

with a fixed initial state z_0 before the first event. The normalization makes z_t primarily sensitive to the composition of recent history rather than to the unnormalized accumulated event count. The parameter τ_z controls the state-memory horizon, whereas β^{-1} controls the excitation horizon. The baseline $\mu_j(t)$ captures excitation-independent variation, whereas $C_{ij}(t, z_t)$ controls how previous type-i events affect type j. Under the conditions of Proposition 3.1, the local baseline and influence vector are separately identified when the conditional excitation design is non-degenerate. The relevant distinction is not smoothness of the sample path $t \mapsto z_t$: the state in Eq. 2 can jump when an event arrives. Instead, what matters is the sensitivity of the state to perturbations of recent event timing. For the normalized encoder in Eq. 2, bounded event embeddings imply $\left\| \frac{\partial z_t}{\partial t_k} \right\| = O(\tau_z^{-1})$ for a local perturbation of an event time that preserves event ordering. Consequently, an $O(\beta^{-1})$ perturbation within the excitation horizon changes z_t by at most $O\big((\tau_z \beta)^{-1}\big)$, while the corresponding change in the excitation statistic can remain $O(1)$. For the normalized encoder, Lemma A.1 bounds the state change induced by an admissible event-time perturbation of size c/β by $2B_z c/(\tau_z \beta)$. This motivates choosing a longer state-memory timescale, but timescale separation is not required for Proposition 3.1 and does not guarantee the conditional excitation variation used there. Exponentially weighted summaries are standard tools for tracking changing process conditions (Song & Zhao, 2024), and related long-memory representations appear in temporal context models and slowness-based representation learning (Klindt et al., 2021). This encoder provides history-dependent context without giving the influence field direct access to the fast excitation statistic.

Parameterization. Each event type has a learned embedding used exclusively by the state encoder in Eq. 2, whereas event types enter the excitation pathway directly through $S \in \mathbb{R}^K$. We incorporate time into both the baseline and field networks via Fourier features (Tancik et al., 2020) to represent multi-frequency temporal variation without an explicit change-point parameterization. The fitted influence field is $\widehat{C}_\theta(t, z) = \psi\big(\mathrm{mat}\big(f_{\theta_C}(\mathrm{Fourier}(t), z)\big)\big)$, where f_{θ_C} is a small MLP with K^2 outputs and $\psi(u) = \mathrm{softplus}(u)$ ensures strictly positive fitted influence coefficients. Both ψ and φ are softplus functions: ψ enforces positivity of the fitted influence coefficients, while φ maps the combined background and excitation predictor to a positive conditional intensity. While the baseline network receives only $\mathrm{Fourier}(t)$ as input, the field network takes $(\mathrm{Fourier}(t), z_t)$ to model interaction coefficients as a coordinate-based neural field over (t, z) (Sitzmann et al., 2020).

The field network receives only $(\mathrm{Fourier}(t), z_t)$, so the pre-link predictor remains affine in $S(t)$ conditional on (t, z_t). Since the state and excitation are both history-derived, their conditional variation must still satisfy Assumption 2. $\widehat{C}_{\theta,ij}(t, z) > 0$ for every finite network output: exact zeros characterize the population field $C^\star$, whereas fitted graphs are obtained by thresholding $\widehat{C}_\theta$. The baseline network μ_{θ_μ} has an unconstrained output, since the outer link φ ensures a positive conditional intensity. The excitation decay $\beta > 0$ is learned as a shared scalar, whereas τ_z is fixed.

Estimation. Let $\theta = (\mathrm{emb}, \theta_C, \theta_\mu, \beta)$ collect the trainable parameters. We minimize the penalized point-process negative log-likelihood

$$\mathcal{L}(\theta) = \sum_{m=1}^{M} \left[- \sum_{k=1}^{|D^{(m)}|} \log \lambda^{(m)}_{e^{(m)}_k}(t^{(m)}_k) + \int_0^T \sum_{j=1}^{K} \lambda^{(m)}_j(t)\, dt \right] + \eta_1 \left\| \bar{C}_\theta \right\|_1 + \eta_{\mathrm{TV}} \sum_{p=1}^{P-1} \sum_{i,j} h_\delta \left(\left[\widehat{C}_\theta(g_{p+1}) - \widehat{C}_\theta(g_p) \right]_{ij} \right), \tag{3}$$

where $\bar{C}_\theta = \frac{1}{P} \sum_{p=1}^{P} \widehat{C}_\theta(g_p)$, $\{g_p\}_{p=1}^{P}$ denotes the evaluation points for regularization and graph read-out, and h_δ is the Huber penalty (quadratic for small differences, linear for large ones). For each sequence, the compensator is approximated by Monte Carlo integration (Mei & Eisner, 2017),

$$\int_0^T \sum_{j=1}^{K} \lambda^{(m)}_j(t)\, dt \approx \frac{T}{n_{\mathrm{mc}}} \sum_{q=1}^{n_{\mathrm{mc}}} \sum_{j=1}^{K} \lambda^{(m)}_j(u^{(m)}_q), \qquad u^{(m)}_q \sim \mathrm{Unif}[0, T]. \tag{4}$$

The ℓ_1 term shrinks globally weak influence coefficients toward zero. Because $\widehat{C}_\theta$ has a softplus output head, it does not induce literal zeros. The Huber-TV (Total Variation) term discourages unnecessary local variation while remaining less punitive for large field differences than a purely quadratic penalty (Hanhela et al., 2020; Hauzenberger et al., 2022). We restrict both regularizers to finite-sample regularization, excluding them entirely from the population identification argument.

Training. We use complete event sequences as training units. The excitation and state statistics are maintained recursively, avoiding repeated summation over the full event history, as in Algorithm 1. For fixed hidden dimension, the dominant dependence on the number of event types arises from the $K \times K$ output of the field network.

3.3 What the Model Recovers

We characterize pointwise separation of the population influence field over the visited time–state region under a well-defined population process, a fixed predictable state map, a correctly specified shared excitation kernel, and nondegenerate conditional variation in S. With β fixed, this variation suffices to separate the background from the influence coefficients. Lemma A.2 gives a separate sufficient condition for uniqueness of β.

Proposition 3.1 (Pointwise separation of background and influence)**.** *Fix the known link, population state map, and shared decay in Eq. 1. Under Assumption 2, any two parameter pairs (μ, C) and $(\widetilde{\mu}, \widetilde{C})$ with these same components that induce the same observed-process law satisfy*

$$\mu_j(t) = \widetilde{\mu}_j(t), \qquad C_{\cdot j}(t, z) = \widetilde{C}_{\cdot j}(t, z)$$

for every j and ν-almost every $(t, z) \in \mathcal{R}$, where ν is the occupation measure defined in Appendix A.

Thus, a non-degenerate conditional excitation design separates background activity from cross-type influence at the population level. Proposition 3.1 conditions on a fixed predictable state map $z_t = \Psi^\star(\mathcal{F}_{t^-})$: it identifies the population field relative to that representation, not the learned state coordinates themselves, and it does not imply consistency or exact support recovery of the jointly trained estimator $\widehat{C}_\theta$. This result addresses RQ1 within the population model. If $z_t = f(t)$ almost surely, only the composite $t \mapsto C^\star(t, f(t))$ is constrained along that trajectory. These conditions, together with the effects of degenerate conditional design, address RQ3 and are discussed in Appendix A.

Holding the history-derived state fixed, $\mathrm{supp}\, C^\star(t, z)$ defines the *direct* state-conditional Granger graph. Because z_t is history-derived, type i can also affect λ_j through changes in z_t. This state-mediated pathway lies outside our estimand (Didelez, 2008; Eichler et al., 2017).

Diagnostics and graph read-out. We report complementary diagnostics with distinct roles. The timescale ratio $r_{\mathrm{time}} = \tau_z \widehat{\beta}$ describes the encoder's memory relative to the excitation timescale. The

event-time covariance ratio ϱ_S measures global excitation conditioning, whereas local neighborhood summaries provide empirical proxies for the conditional rank condition in Assumption 2. The state-spread statistic ϱ_z describes variation across sequences at common times. None of these finite-sample diagnostics establishes the population condition. Empirical proxies for conditional excitation variation are reported in Remark C.1. For a specified threshold $\gamma \geq 0$, the estimated graph contains $i \to j$ when $\widehat{C}_{\theta,ij}(t,z) > \gamma$. The aggregate graph applies the same rule to $\bar{C}_\theta$.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Synthetic data. We use a self-consistent generator whose intensity matches the estimation model, $\lambda = \mathrm{softplus}(\mu^\star(t) + \langle C^\star(t,z), S\rangle)$, and whose latent state is produced by the same exponentially weighted encoder. The generator controls the amplitude of the time-varying baseline (a_μ) and the amplitude of the state-dependent gate (a_C) independently with a horizon of $T = 100$ and 20 random seeds. The average numbers of change points are 2.6, 2.1, and 2.3 for $K = 20$, 50, and 100.

Real-world data. We evaluate on the publicly released sample of the alarm dataset [1]: 30,691 events over 59 event types, grouped into 45 device sequences. The ground-truth graph has 147 directed edges and no self-loops; after applying the provided causal prior, 3,306 off-diagonal relations remain to be decided. This setting differs from our synthetic experiments in an important respect. The available ground truth is a single static summary graph and therefore cannot directly evaluate time- or state-conditional causal structure. Nevertheless, an event-based causal discovery method intended for operational logs should also recover meaningful structure from real event data, making the alarm dataset a complementary test of aggregate graph recovery. We report summary-graph recovery against the static reference, alongside the two fit-time diagnostics of Section 3.3, and treat the result as external validity for the method.

Hyperparameters. All models are trained with Adam at a learning rate of 5×10^{-3} for 150 epochs. Gradient norms are clipped at 5.0 and threshold is 0.3. The compensator of the point-process likelihood is estimated by Monte Carlo with 64 samples during training and 256 samples at evaluation time. The kernel decay β is the only kernel parameter and is learned jointly with the rest of the model. The sparsity coefficient is fixed at $\lambda_1 = 10^{-3}$ for the experiments. The temporal total-variation coefficient λ_{TV} is selected from $\{3 \times 10^{-3}, 10^{-2}, 3 \times 10^{-2}\}$ by validation log-likelihood on the model, and the selected value is reused for every ablation row of the same configuration. The Huber parameter of the TV penalty is $\delta = 0.05$. For the real alarm data we use a longer schedule (300 epochs, learning rate 10^{-2}) and a stronger sparsity range ($\lambda_1 \in \{3 \times 10^{-2}, 10^{-1}\}$), reflecting the larger number of event types. Other hyperparameters are deferred to Appendix B.

Evaluation metrics. We evaluate aggregate graph recovery using precision, TPR, F1, and Matthews correlation coefficient (MCC) (Chicco & Jurman, 2020). We additionally report ROC AUC (AUC) to assess edge ranking without selecting a decision threshold. Metric definitions and coefficient-level evaluation details are provided in Appendix B.2.

Baselines. Among event sequence causal discovery methods we include L0HAWKES (Idé et al., 2021), NPHC (Achab et al., 2017), CAUSE (Zhang et al., 2020), and CASCADE (Cüppers et al., 2024). Among neural methods we include THP (Cai et al., 2021) and NGC (Tank et al., 2022). Stationary baselines were provided with ground-truth change points, conferring an advantage over our model. Blocked-t is an internal control that isolates the contribution of the continuous field: it replaces the Fourier time features of our model with a one-hot block index and prevents C from reading the latent state, while keeping the softplus link, the kernel, the optimizer, and the penalties identical. To avoid advantaging our method through tuning, each baseline is tuned over a method-specific grid (Appendix B), and the best configuration is selected for each method on the data. We reimplemented StateHPs (Liu et al., 2025) (N$\in \{2, 3, 5\}$ states $\times$ 3 inits, init by held-out marginal likelihood).

Table 1: Results on synthetic data across different numbers of event types. Values are mean ± standard deviation over 20 seeds. Dark and bold entries denote the best mean, whereas light and underlined entries denote the second-best mean. * indicates that no oracle information is used.

K	Method	F_1	Precision	TPR	AUC	MCC
20	L0HAWKES	0.11 ± 0.04	0.07 ± 0.02	0.25 ± 0.09	0.48 ± 0.07	−0.03 ± 0.05
	NPHC	0.10 ± 0.06	0.09 ± 0.05	0.13 ± 0.08	0.53 ± 0.05	0.01 ± 0.07
	CAUSE	0.16 ± 0.02	0.09 ± 0.01	**0.96 ± 0.06**	0.66 ± 0.06	0.03 ± 0.04
	CASCADE	0.16 ± 0.03	0.09 ± 0.02	0.89 ± 0.10	0.52 ± 0.05	0.03 ± 0.04
	THP	0.15 ± 0.02	0.09 ± 0.02	0.51 ± 0.07	0.50 ± 0.05	0.00 ± 0.04
	NGC	0.30 ± 0.12	0.28 ± 0.12	0.36 ± 0.13	0.75 ± 0.05	0.24 ± 0.12
	StateHPs*	0.36 ± 0.08	0.34 ± 0.10	0.44 ± 0.12	0.79 ± 0.06	0.30 ± 0.08
	CAUSALFIELD (ours)*	**0.68 ± 0.09**	**0.62 ± 0.16**	0.80 ± 0.11	**0.96 ± 0.02**	**0.66 ± 0.09**
50	L0HAWKES	0.06 ± 0.01	0.03 ± 0.01	0.30 ± 0.08	0.53 ± 0.04	0.01 ± 0.03
	NPHC	0.04 ± 0.02	0.03 ± 0.02	0.06 ± 0.04	0.51 ± 0.03	0.01 ± 0.02
	CAUSE	0.06 ± 0.01	0.03 ± 0.00	**0.98 ± 0.03**	0.61 ± 0.05	0.01 ± 0.02
	CASCADE	0.06 ± 0.01	0.03 ± 0.00	0.95 ± 0.04	0.53 ± 0.04	0.01 ± 0.02
	THP	0.06 ± 0.01	0.03 ± 0.01	0.50 ± 0.06	0.50 ± 0.03	0.00 ± 0.02
	NGC	0.21 ± 0.05	0.15 ± 0.05	0.39 ± 0.06	0.77 ± 0.02	0.20 ± 0.04
	StateHPs*	0.16 ± 0.05	0.26 ± 0.14	0.13 ± 0.06	0.59 ± 0.05	0.15 ± 0.07
	CAUSALFIELD (ours)*	**0.53 ± 0.07**	**0.43 ± 0.10**	0.75 ± 0.08	**0.95 ± 0.02**	**0.54 ± 0.06**
100	L0HAWKES	0.04 ± 0.00	0.02 ± 0.00	0.28 ± 0.05	0.56 ± 0.03	0.03 ± 0.01
	NPHC	0.02 ± 0.01	0.01 ± 0.01	0.02 ± 0.02	0.48 ± 0.03	0.00 ± 0.01
	CAUSE	0.03 ± 0.00	0.01 ± 0.00	**1.00 ± 0.00**	0.59 ± 0.02	0.01 ± 0.01
	CASCADE	0.03 ± 0.00	0.02 ± 0.00	0.98 ± 0.02	0.51 ± 0.03	0.01 ± 0.01
	THP	0.03 ± 0.00	0.01 ± 0.00	0.50 ± 0.05	0.50 ± 0.03	0.00 ± 0.01
	NGC	0.20 ± 0.05	0.14 ± 0.05	0.40 ± 0.07	0.80 ± 0.02	0.21 ± 0.04
	StateHPs*	0.12 ± 0.04	0.34 ± 0.16	0.08 ± 0.03	0.58 ± 0.07	0.15 ± 0.05
	CAUSALFIELD (ours)*	**0.47 ± 0.06**	**0.37 ± 0.10**	0.71 ± 0.06	**0.96 ± 0.01**	**0.50 ± 0.05**

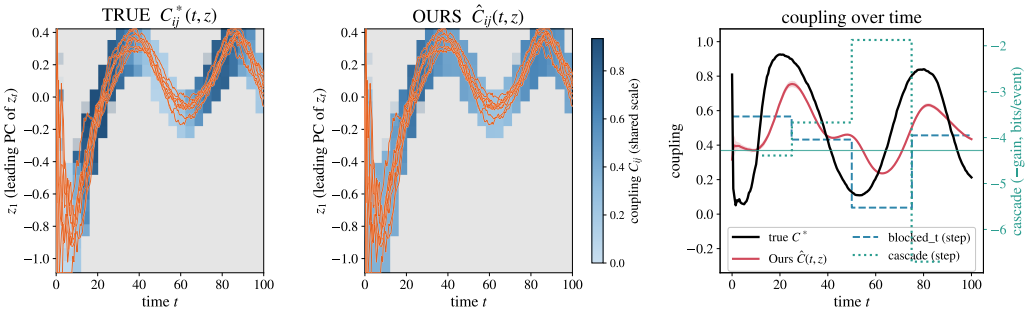

Figure 2: The fitted field versus a segmented model. Left: true $C^{\star}_{ij}(t,z)$; middle: recovered $\hat{C}_{ij}(t,z)$ on the same color scale, gray marking (t,z) visited fewer than $n_{\min}$ times. Right: CAUSALFIELD (red) tracks both oscillations of the true coupling (black), with a spurious inflection near t=35, while *blocked_t* and CASCADE are constrained to one value per block and average the oscillation away.

4.2 RESULTS

Synthetic data results. Table 1 reports aggregate graph-recovery performance for $K \in \{20, 50, 100\}$. CAUSALFIELD achieves the highest mean F1, precision, AUC, and MCC at every scale. For F1 and AUC, the reported mean ± standard deviation ranges do not overlap those of any baseline. Its F1 is approximately 1.9–2.5 times that of the strongest baseline at each scale: StateHPs at $K = 20$ and NGC at $K = 50$ and 100. Baselines lead only in TPR at low precision, largely by predicting nearly all candidate edges. CAUSALFIELD retains the strongest overall performance, with the same ordering under threshold-free evaluation and even under the baselines' best configuration, oracle threshold (Table 13 in Appendix C). As K increases, mean AUC remains stable at 0.95–0.96, whereas F1 decreases from 0.68 to 0.47, indicating stable ranking performance despite

[1] https://github.com/huawei-noah/trustworthyAI

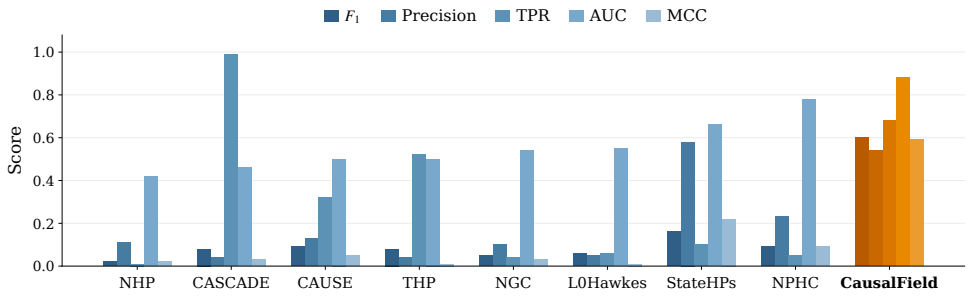

Figure 3: Graph recovery evaluation (F1, Precision, TPR, AUC, MCC) on the Alarm dataset.

Table 2: Coefficient errors on held-out synthetic histories in a separate re-fit ($K = 20$). Values are mean $\pm$ standard deviation over 20 seeds, each averaged over five held-out sequences. All columns report root mean squared error (RMSE).

Variant	Background $\downarrow$	True edges $\downarrow$	Non-edges $\downarrow$
Full $C(t, z)$	**0.626 $\pm$ 0.055**	**0.169 $\pm$ 0.017**	**0.059 $\pm$ 0.011**
Time-only $C(t)$	0.714 $\pm$ 0.060	0.183 $\pm$ 0.020	0.077 $\pm$ 0.014
Quantized, $K_z = 4$	0.694 $\pm$ 0.062	0.178 $\pm$ 0.019	0.074 $\pm$ 0.011
Blocked-time, $\Delta = 25$	1.481 $\pm$ 0.090	0.257 $\pm$ 0.019	0.111 $\pm$ 0.009
Static C	2.282 $\pm$ 0.553	0.184 $\pm$ 0.027	0.182 $\pm$ 0.004

lower thresholded graph-recovery performance. Figure 2 complements these aggregate results with an example of within-block coupling variation. CAUSALFIELD captures changes that a single coefficient per block cannot represent. Section 5.1 evaluates coefficient accuracy quantitatively on held-out synthetic histories.

Real-world data results. On the alarm dataset, we evaluate aggregate graph recovery against the static reference graph without using the equipment topology (Figure 3). CAUSALFIELD achieves the highest mean F1 (0.60), AUC (0.88), and MCC (0.59). CASCADE achieves a higher TPR only at a near-all-positive operating point. Among the methods with non-trivial precision, CAUSALFIELD achieves the highest recall as well as the highest AUC. These results support aggregate graph recovery, although the static reference cannot directly assess time- or state-dependent influence coefficients. Appendix C separately reports results using the 45×45 equipment topology. We complement this aggregate evaluation with background and influence-field RMSE on held-out synthetic histories, where the true coefficients are known.

5 DISCUSSION

5.1 CONTINUOUS FIELDS UNDER AGGREGATE EVALUATION

The alarm benchmark provides only a static reference graph and cannot directly evaluate time- or state-dependent influence. CAUSALFIELD remains effective under aggregate evaluation. On alarm, the full model yields small held-out NLL gains over time-only variants, including after capacity matching, whereas restricting the background to a constant incurs a much larger loss (Table 18 in Appendix C). Synthetic ablations also yield similar aggregate graph-recovery metrics across time-only and quantized-state variants (Table 9). Aggregate metrics average the field over the window and are therefore insensitive by construction to variation within a block, so the contribution of the field appears in the quantities that resolve such variation, namely coefficient error, background error, and held-out likelihood. We examine coefficient accuracy directly in a separate re-fit (Table 2), with evaluation details in Appendix C.2.

We compare the learned and generative composed fields, $\widehat{C}_\theta(t, \widehat{z}_t)$ and $C^\star(t, z_t^\star)$, on the same held-out histories without aligning state coordinates (Remark A.3). Relative to the time-only variant, the

Table 3: Robustness across data-generating processes (DGPs). Results are reported as F1 / AUC. Dark and bold entries denote the highest mean scores; light shading marks cells containing a second-highest score, which is underlined. Table 10 in Appendix C reports the full set of metrics.

Method	exp-Hawkes	delayed	power-law	logistic
L0HAWKES	0.14 ± 0.04 / 0.48 ± 0.07	0.15 ± 0.04 / 0.50 ± 0.08	0.15 ± 0.04 / 0.49 ± 0.09	0.16 ± 0.04 / 0.52 ± 0.10
NPHC	0.17 ± 0.04 / 0.54 ± 0.06	0.17 ± 0.04 / 0.54 ± 0.07	0.16 ± 0.02 / 0.53 ± 0.05	0.23 ± 0.04 / $\underline{0.64 \pm 0.06}$
CAUSE	0.51 ± 0.13 / 0.83 ± 0.10	0.29 ± 0.09 / 0.61 ± 0.09	0.39 ± 0.09 / 0.72 ± 0.07	0.24 ± 0.05 / 0.58 ± 0.08
CASCADE	0.11 ± 0.02 / 0.41 ± 0.05	0.12 ± 0.02 / 0.43 ± 0.05	0.13 ± 0.02 / 0.44 ± 0.05	0.12 ± 0.02 / 0.39 ± 0.06
THP	0.15 ± 0.04 / 0.49 ± 0.05	0.15 ± 0.04 / 0.49 ± 0.05	0.15 ± 0.04 / 0.49 ± 0.05	0.15 ± 0.04 / 0.49 ± 0.05
SHP	0.71 ± 0.08 / 0.84 ± 0.05	0.35 ± 0.08 / 0.64 ± 0.05	0.36 ± 0.08 / 0.64 ± 0.05	0.21 ± 0.07 / 0.55 ± 0.03
StateHPs	$\underline{0.85 \pm 0.16}$ / $\underline{0.98 \pm 0.02}$	$\underline{0.79 \pm 0.12}$ / $\underline{0.97 \pm 0.02}$	$\underline{0.80 \pm 0.08}$ / $\underline{0.97 \pm 0.02}$	**0.70 ± 0.12 / 0.92 ± 0.05**
CAUSALFIELD (ours)	**0.97 ± 0.01 / 1.00 ± 0.00**	**0.89 ± 0.04 / 0.99 ± 0.01**	**0.92 ± 0.04 / 0.99 ± 0.01**	$\underline{0.68 \pm 0.12}$ / **0.92 ± 0.05**

full model reduces pre-link background root mean squared error (RMSE) by 12%, with true-edge and non-edge RMSE lower by 8% and 23%. The reductions relative to the blocked-time variant are larger, at 58% for background RMSE and 34% for true-edge RMSE. These joint improvements provide empirical evidence relevant to background–influence separation in RQ1. The mechanism is confirmed directly when $C^\star$ is held constant and only the baselines drift (Figure 5 and Table 25 in Appendix C). Blocked-t absorbs the unexplained background as a positive coefficient offset of $+0.069$ (95% CI $[0.038, 0.099]$) that tracks its residual baseline error across seeds ($r = 0.90$), and a per-block Hawkes fit reaches an FDR of 0.20, whereas the full field shows neither. Gains over state quantization are modest, and $K_z = 16$ matches the full model on field RMSE (Table 9). The continuous state therefore removes the need to choose a regime count rather than outperforming every discretization, and strong edge ranking should not be read as accurate recovery of coefficient magnitudes.

5.2 BEYOND THE DATA-GENERATING PROCESS

To test robustness across diverse generative mechanisms and distributions, we evaluated CAUSALFIELD on four alternative data-generating processes under the same evaluation protocol (Table 3). CAUSALFIELD ranks first or second across all four DGPs. In particular, it leads on exp-Hawkes, delayed, and power-law processes, while remaining competitive on the logistic process. These results indicate that the observed advantage does not rely on an exact match between the model and the data-generating process. Instead, performance appears to depend more on whether the underlying coupling variation can be recovered than on the specific form of the triggering kernel.

The logistic generator is the only tested process on which CAUSALFIELD does not achieve the highest mean F1. Its F1 is nevertheless close to that of StateHPs (0.68 ± 0.12 versus 0.70 ± 0.12), with the same mean AUC. In this setting, 28.8% of the time steps fire with probability above 0.9, suggesting that saturation limits the information available to distinguish excitation effects. Reducing firing probabilities raises F1 to 0.90 despite roughly one third as many events. CAUSALFIELD then outperforms StateHPs by 0.107, with a paired 95% interval of $[0.058, 0.165]$ over 20 seeds. The excitation-design diagnostic ϱ_S remains comparable to that of the continuous-time generators. Together, these results support saturation as a likely explanation for the degradation (Tables 11 and 12 in Appendix C).

6 CONCLUSION

We introduced CAUSALFIELD, which replaces fixed or regime-wise causal graphs with a continuous influence field, parameterized as a coordinate-based neural field and learned within the point-process likelihood. Experiments demonstrated strong aggregate graph recovery on synthetic and real-world event data, while synthetic ablations showed improved background and influence-coefficient accuracy over the tested alternatives. Performance remained competitive across alternative data-generating processes, supporting robustness beyond the matched generative setting. These findings support a continuous-field perspective on non-stationary causal discovery that captures both the connections between event types and how their influence varies with time and system state.

ETHICS STATEMENT

Our work involves no human subjects, no personal or user-generated data, and no application domain in which fairness or bias concerns arise: the task is recovering dependency structure from machine-generated operational event logs. Our synthetic datasets are produced by our own generator and contain nothing beyond event times and event types, so they carry no privacy-relevant information by construction. Also the real dataset is the publicly released sample of the alarm dataset which likewise records only device-level alarm events and contains no personally identifiable information. We foresee no negative societal impacts specific to this work.

REPRODUCIBILITY STATEMENT

To support the reproducibility of this work, we will publicly release the complete source code used to conduct the experiments upon publication. The released materials will include the implementation of the proposed method, scripts for data preprocessing and training, evaluation code, configuration files, and instructions for reproducing the main experimental results reported in this paper. We also provide details of the experimental environment, including software dependencies, library versions, hardware specifications, and random seeds in Appendix B and B.4. These materials are intended to enable other researchers to reproduce our experimental setup and independently verify the reported results.

AI USE STATEMENT

AI tools were used to improve the writing, organization, and readability of this manuscript. They were also used to identify potential logical issues in the proposed ideas and proofs. During the experimental process, AI tools were also used for tasks such as assisting with code implementation, organizing code structure, and refining comments.

REFERENCES

Massil Achab, Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. Uncovering causality from multivariate hawkes integrated cumulants. *J. Mach. Learn. Res.*, 18(1): 6998–7025, January 2017. ISSN 1532-4435.

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.

Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005, 2015.

Young-Mok Bae, Young-Gwan Kim, Jeong-Woo Seo, Hyun-A Kim, Chang-Ho Shin, Jeong-Hwan Son, Gyu-Ho Lee, and Kwang-Jae Kim. Detecting abnormal behavior of automatic test equipment using autoencoder with event log data. *Comput. Ind. Eng.*, 183(C), September 2023. ISSN 0360-8352. doi: 10.1016/j.cie.2023.109547. URL https://doi.org/10.1016/j.cie.2023.109547.

Anna Bonnet, Miguel Martinez Herrera, and Maxime Sangnier. Inference of multivariate exponential hawkes processes with inhibition and application to neuronal activity. *Statistics and Computing*, 33 (4), June 2023. ISSN 1573-1375. doi: 10.1007/s11222-023-10264-w. URL http://dx.doi.org/10.1007/s11222-023-10264-w.

Pierre Brémaud and Laurent Massoulié. Stability of nonlinear hawkes processes. *The Annals of Probability*, 24(3):1563–1588, 1996. ISSN 00911798, 2168894X. URL http://www.jstor.org/stable/2244985.

Ruichu Cai, Siyu Wu, Jie Qiao, Zhifeng Hao, Keli Zhang, and Xi Zhang. Thps: Topological hawkes processes for learning causal structure on event sequences. *IEEE Transactions on Neural Networks and Learning Systems*, 35:479–493, 2021. URL https://api.semanticscholar.org/CorpusID:248811254.

Wei Chen, Jibin Chen, Ruichu Cai, Yuequn Liu, and Zhifeng Hao. Learning granger causality for non-stationary hawkes processes. *Neurocomput.*, 468(C):22–32, January 2022. ISSN 0925-2312. doi: 10.1016/j.neucom.2021.10.030. URL https://doi.org/10.1016/j.neucom.2021.10.030.

Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6, 2020.

Lénaïg Cornanguer, Joscha Cüppers, and Jilles Vreeken. Causal discovery from interval-based event sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20642–20649, 2026.

Joscha Cüppers, Sascha Xu, Ahmed Musa, and Jilles Vreeken. Causal discovery from event sequences by local cause-effect attribution. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 24216–24241. Curran Associates, Inc., 2024. doi: 10.52202/079017-0763. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2b22bacd7ad8677f4837b28a11fe496f-Paper-Conference.pdf.

Vanessa Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):245–264, 02 2008. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2007.00634.x. URL https://doi.org/10.1111/j.1467-9868.2007.00634.x.

Peter J. Diggle. Conditional intensity: A powerful tool for modelling and analysing point process data. *Australian & New Zealand Journal of Statistics*, 63(1):83–92, 2021. doi: https://doi.org/10.1111/anzs.12331. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/anzs.12331.

Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1555–1564, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939875. URL https://doi.org/10.1145/2939672.2939875.

Michael Eichler, Rainer Dahlhaus, and Johannes Dueck. Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242, 2017. doi: https://doi.org/10.1111/jtsa.12213. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jtsa.12213.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

Matti Hanhela, Mikko I. Kettunen, Olli H. J. Gröhn, Marko Vauhkonen, and Ville Kolehmainen. Temporal huber regularization for dce-mri. *Journal of Mathematical Imaging and Vision*, 62:1334 – 1346, 2020. URL https://api.semanticscholar.org/CorpusID:213176527.

Niko Hauzenberger, Florian Huber, Gary Koop, and Luca Onorante. Fast and flexible bayesian inference in time-varying parameter regression models. *Journal of Business & Economic Statistics*, 40(4):1904–1918, 2022. doi: 10.1080/07350015.2021.1990772. URL https://doi.org/10.1080/07350015.2021.1990772.

Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58 (1):83–90, 1971.

Junjie Hu, Yuting Yang, Zongkai Tian, Youchen Wang, Bin Lu, Wei Song, and Jiaxin Guo. A survey on event causality identification. In *Proceedings of the 2025 2nd International Conference on Artificial Intelligence, Digital Media Technology and Interaction Design*, ICADI '25, pp. 535–542, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400720178. doi: 10.1145/3795926.3796011. URL https://doi.org/10.1145/3795926.3796011.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *J. Mach. Learn. Res.*, 21(1), January 2020. ISSN 1532-4435.

Martin Huber. An introduction to causal discovery. *Swiss Journal of Economics and Statistics*, 160 (1):14, 2024.

Tsuyoshi Idé, Georgios Kollias, Dzung Phan, and Naoki Abe. Cardinality-regularized hawkes-granger model. *Advances in Neural Information Processing Systems*, 34:2682–2694, 2021.

Amirkasra Jalaldoust, Kateřina Hlaváčková-Schindler, and Claudia Plant. Causal discovery in hawkes processes by minimum description length. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6978–6987, 2022.

Songyao Jin and Biwei Huang. Causal structure learning in hawkes processes with complex latent confounder networks. In *International Conference on Learning Representations*, volume 2026, pp. 58550–58582, 2026.

David A. Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=EbIDjBynYJ8.

Yuequn Liu, Guangdong Sun, Ruichu Cai, Zijian Li, Keli Zhang, Lujia Pan, and Zhifeng Hao. Statehps: State hawkes processes for granger causal discovery from non-stationary event sequences. *Information Sciences*, 718:122422, 2025. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2025.122422. URL https://www.sciencedirect.com/science/article/pii/S0020025525005547.

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. Mega: Moving average equipped gated attention. *ArXiv*, abs/2209.10655, 2022. URL https://api.semanticscholar.org/CorpusID:252439127.

Sarah Mameche, Lénaïg Cornanguer, Urmi Ninad, and Jilles Vreeken. Spacetime: Causal discovery from non-stationary time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 19405–19413, 2025.

Hongyuan Mei and Jason Eisner. The neural hawkes process: a neurally self-modulating multivariate point process. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6757–6767, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, December 2021. ISSN 0001-0782. doi: 10.1145/3503250. URL https://doi.org/10.1145/3503250.

Maxime Morariu-Patrichi and Mikko S. Pakkanen. State-dependent hawkes processes and their application to limit order book modelling, 2021. URL https://arxiv.org/abs/1809.08060.

Elchanan Mossel and Anirudh Sridhar. Detecting mutual excitations in non-stationary hawkes processes. *arXiv preprint arXiv:2601.11717*, 2026.

Fco. Javier Ordóñez, Paula De Toledo, and Araceli Sanchis. Activity recognition using hybrid generative/discriminative models on home environments using binary sensors. *Sensors*, 13(5): 5460–5477, 2013. ISSN 1424-8220. doi: 10.3390/s130505460. URL https://www.mdpi.com/1424-8220/13/5/5460.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Paul Beaumont, Konstantinos Georgatzis, and Bryon Aragam. Dynotears: Structure learning from time-series data, 2020. URL https://arxiv.org/abs/2002.00498.

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.

Abdellah Rahmani and Pascal Frossard. Flow based approach for dynamic temporal causal models with non-gaussian or heteroscedastic noises, 2025. URL https://arxiv.org/abs/2506.17065.

S. W. Roberts. Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250, 1959. doi: 10.1080/00401706.1959.10489860. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1959.10489860.

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.

Elena Saggioro, Jana de Wiljes, Marlene Kretschmer, and Jakob Runge. Reconstructing regime-dependent causal relationships from observational time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(11), November 2020. ISSN 1089-7682. doi: 10.1063/5.0020538. URL http://dx.doi.org/10.1063/5.0020538.

Oleksandr Shchur, Ali Caner Türkmen, Tim Januschowski, and Stephan Günnemann. Neural temporal point processes: A review. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 4585–4593, Montreal, Canada, 8 2021. International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2021/623. URL https://doi.org/10.24963/ijcai.2021/623. Survey Track.

Zijian Shi and John Cartlidge. State dependent parallel neural hawkes process for limit order book event stream prediction and simulation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, pp. 1607–1615, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539462. URL https://doi.org/10.1145/3534678.3539462.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Pengyu Song and Chunhui Zhao. Slow down to go better: A survey on slow feature analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3):3416–3436, 2024. doi: 10.1109/TNNLS.2022.3201621.

Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pp. 3, Berlin, Germany, 2016. Springer, Springer.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B. Fox. Neural granger causality. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(8):4267–4279, August 2022. ISSN 0162-8828. doi: 10.1109/TPAMI.2021.3065601. URL https://doi.org/10.1109/TPAMI.2021.3065601.

Dongxia Wu, Tsuyoshi Idé, Georgios Kollias, Jiri Navratil, Aurelie Lozano, Naoki Abe, Yian Ma, and Rose Yu. Learning granger causality from instance-wise self-attentive hawkes processes. In *International Conference on Artificial Intelligence and Statistics*, pp. 415–423. PMLR, 2024.

Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. Learning granger causality for hawkes processes. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1717–1726, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/xuc16.html.

Wei Zhang, Thomas Kobber Panum, Somesh Jha, Prasad Chalasani, and David Page. Cause: Learning granger causality from event sequences using attribution methods, 2020. URL https://arxiv.org/abs/2002.07906.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with no tears: continuous optimization for structure learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 9492–9503, Red Hook, NY, USA, 2018. Curran Associates Inc.

Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial intelligence and statistics*, pp. 641–649. PMLR, 2013.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer hawkes process. In *International conference on machine learning*, pp. 11692–11702. PMLR, 2020.

A Assumptions and Supporting Statements

A.1 Notation

Table 4 summarizes the key notations and mathematical symbols used throughout this paper.

A.2 Assumptions

We state the conditions under which the population influence field $C^\star$ admits the interpretation used in the main text. Throughout this section, the history-derived state is treated as a fixed predictable map of the observed filtration.

Because the normalized state in Eq. 2 is a convex combination of event embeddings, its centered variation need not span the ambient embedding space. Let

$$\mathcal{Z} = \operatorname{span}\{\operatorname{emb}(e) - \operatorname{emb}(e_0) : e = 1, \dots, K\}$$

denote the realized state subspace, for an arbitrary reference type e_0, and let $Q_{\mathcal{Z}}$ be an orthonormal basis of $\mathcal{Z}$.

Assumption 1 (Regularity and stability)**.** *The pre-link baseline $\mu^\star$ and influence field $C^\star$ are bounded, and the state z_t is $\mathcal{F}_{t^-}$-predictable. Event embeddings are uniformly bounded,*

$$\sup_e \|\operatorname{emb}(e)\| \le B_z < \infty,$$

and there exists an entrywise nonnegative matrix $\overline{C}$ such that

$$0 \le C^\star(t, z) \le \overline{C}.$$

We assume the usual subcriticality condition

$$\rho\left(\overline{C} \int_0^\infty \kappa(u)\, du\right) < 1.$$

Assumption 2 (Visited support and conditional excitation variation)**.** *The recovery region $\mathcal{R}$ lies within the locally visited support of (t, z_t). Wherever pointwise identification is claimed,*

$$\operatorname{Var}(S \mid t, z) \succ 0.$$

Remark A.1 (Separating time and state dependence)**.** *Assumption 2 is sufficient for pointwise identification of the local background and influence vector. To distinguish dependence on z from dependence on t as separate coordinates, additional state variation is required. If $z_t^{(m)} = f(t)$ across replicated sequences, the population law identifies only the composite $t \mapsto C^\star(t, f(t))$ along the observed trajectory.*

Table 4: Notation used throughout the paper.

Data	
M	Number of replicated event sequences; the independent sampling unit.
K	Number of event types.
$[0, T]$	Observation window, with horizon T.
$D^{(m)}$	m-th sequence $\{(t_k^{(m)}, e_k^{(m)})\}$.
$\mathcal{F}_{t-}$	Event history immediately before time t.
Population model and estimand	
$\lambda_j(t \mid \mathcal{F}_{t-})$	Conditional intensity of event type j.
φ	Known link function on the conditional intensity; $\varphi = \mathrm{softplus}$.
ψ	Output link of the field network; $\psi = \mathrm{softplus}$, applied entrywise.
$\mu_j^\star(t)$	Population time-varying background component of type j.
$C^\star(t, z)$	Population influence field in $\mathbb{R}_{\geq 0}^{K \times K}$.
$C_{ij}^\star(t, z)$	Direct excitation coefficient from type i to type j at (t, z).
$G^\star(t, z)$	Population direct state-conditional Granger graph, $G^\star(t, z) = \operatorname{supp} C^\star(t, z)$.
$S_i(t)$	Fast excitation statistic $\sum_{t_k < t,\, e_k = i} e^{-\beta(t - t_k)}$.
$S(t)$	Vector $(S_1(t), \ldots, S_K(t))$.
β	Shared decay of the excitation kernel.
$\kappa(u)$	Exponential kernel $e^{-\beta u}$, with $\kappa(0) = 1$.
z_t	Predictable history-derived system state in eq. (2).
$\mathrm{emb}(e)$	Learned event-type embedding used only by the state encoder.
d_z	Dimension of the event embeddings and z_t.
τ_z	Fixed memory constant of the state encoder.
$\mathcal{R}$	Region of (t, z) visited by the population process.
Neural parameterization and estimation	
$\mathrm{Fourier}(t)$	Sinusoidal time features input to μ_{θ_μ} and C_θ.
$f_{\theta_C}(\mathrm{Fourier}(t), z)$	Unconstrained pre-softplus output of the field MLP (K^2 outputs).
$\mathrm{mat}(\cdot)$	Reshape $\mathbb{R}^{K^2} \to \mathbb{R}^{K \times K}$ applied to the field-MLP output.
$\widehat{C}_\theta(t, z)$	Fitted field $\psi\big(\mathrm{mat}(f_{\theta_C}(\mathrm{Fourier}(t), z))\big) > 0$.
$\widehat{G}_\gamma(t, z)$	Empirical graph obtained by thresholding $\widehat{C}_\theta(t, z)$ at threshold γ.
θ	Trainable parameters $(\mathrm{emb}, \theta_C, \theta_\mu, \beta)$.
$\mathcal{L}(\theta)$	Penalized point-process negative log-likelihood.
$\eta_1, \eta_{\mathrm{TV}}$	Weights of the ℓ_1 and Huber-TV regularizers.
$\bar{C}_\theta$	Fitted field averaged over the evaluation points $\{g_p\}$.
$\{g_p\}$	Fixed evaluation points used for field regularization and graph read-out.
h_δ	Huber function, quadratic for small and linear for large differences.
n_{mc}, u_q	Monte Carlo sample count and uniform integration points.

Empirical diagnostics for the conditional excitation variation in Assumption 2 are reported in Appendix C.

Assumption 1 is a standard sufficient stability condition for nonlinear mutually exciting point processes (Brémaud & Massoulié, 1996). This support-based interpretation is consistent with local-independence graphs for marked point processes and Granger graphical models for Hawkes processes (Didelez, 2008; Xu et al., 2016; Eichler et al., 2017). Assumption 2 is the local rank condition used for pointwise separation of the background and influence coefficients.

Conditioning on the state. Define the occupation measure of the time–state process by

$$\nu(A) = \frac{1}{T}\mathbb{E}\left[\int_0^T \mathbb{I}[(t, z_t) \in A]\, dt\right].$$

Pointwise statements below are understood with respect to a version of the regular conditional law given $(t, z_t) = (t, z)$, for ν-almost-every (t, z) in the visited region $\mathcal{R}$.

Lemma A.1 (Local timing sensitivity of the state encoder)**.** *Assume $\|\mathrm{emb}(e)\| \leq B_z$ for all event types and fix an evaluation time t with at least one preceding event. Let $W_t = \sum_{t_\ell < t} w_\ell$. For a perturbation of one event time t_ℓ that preserves event ordering and keeps $t_\ell < t$,*

$$\frac{\partial z_t}{\partial t_\ell} = \frac{w_\ell}{\tau_z W_t}\left(\mathrm{emb}(e_\ell) - z_t\right). \tag{5}$$

Consequently, $\left\|\frac{\partial z_t}{\partial t_\ell}\right\| \le \frac{2B_z}{\tau_z}$. *If* $|t'_\ell - t_\ell| \le \frac{c}{\beta}$ *and the perturbation preserves event membership and ordering, then*

$$\|z'_t - z_t\| \le \frac{2B_z c}{\tau_z \beta} = O\big((\tau_z\beta)^{-1}\big)\,.$$

Proof. Write $z_t = \frac{N_t}{W_t}$, $N_t = \sum_{\ell:t_\ell<t} w_\ell \,\mathrm{emb}(e_\ell)$. Because $\frac{\partial w_\ell}{\partial t_\ell} = \frac{w_\ell}{\tau_z}$, the quotient rule gives

$$\begin{aligned}\frac{\partial z_t}{\partial t_\ell} &= \frac{(w_\ell/\tau_z)\mathrm{emb}(e_\ell)W_t - N_t(w_\ell/\tau_z)}{W_t^2} \\ &= \frac{w_\ell}{\tau_z W_t}\left(\mathrm{emb}(e_\ell) - z_t\right).\end{aligned}$$

Since z_t is a convex combination of event embeddings, $\|z_t\| \le B_z$, and hence $\left\|\frac{\partial z_t}{\partial t_\ell}\right\| \le \frac{2B_z}{\tau_z}$. Integrating this local bound over a perturbation of magnitude at most c/β gives the final inequality. □

The result concerns local event-time sensitivity rather than smoothness of the sample path $t \mapsto z_t$, which can jump at event arrivals.

A.3 Shared Excitation Decay

Related identifiability results have been established for parametric multivariate exponential Hawkes models (Bonnet et al., 2023).

Lemma A.2 (Uniqueness of the shared decay on a conditional slice)**.** *Fix the population state map. Suppose a conditional family of histories at fixed* (t, z) *supports a source-event lag* u *on an open interval* I*, and its conditional intensity function is determined by the population law. Suppose every competing parameterization with decay* $b > 0$ *has inverse-linked intensity*

$$f_b(u) = a_b + c_b e^{-bu}, \qquad u \in I,$$

where a_b *and* c_b *do not depend on* u*. If the true coefficient* $c_\beta > 0$*, observationally equivalent parameterizations have* $b = \beta$*.*

Proof. Compare two observationally equivalent parameterizations with the same state map and link, but with decay parameters b and β. On the conditional slice in the lemma, equality of their conditional intensities gives $f_b = f_\beta$ on I. Their predictable intensities agree $dt \otimes P$-almost everywhere. The true function is nonconstant, so the competing coefficient is also nonzero. Differentiation yields

$$-\frac{f_b''(u)}{f_b'(u)} = b, \qquad -\frac{f_\beta''(u)}{f_\beta'(u)} = \beta.$$

Consequently, $b = \beta$. □

Similar results on recovering the decay of parametric multivariate exponential Hawkes models are given by Bonnet et al. (2023). We stress that this condition is a separate sufficient requirement that does not follow from the conditional rank condition or the state-sensitivity bound.

A.4 Pointwise Separation of Background and Influence

For a simple point process, the conditional intensity characterizes the process law (Diggle, 2021). Hence equality of the relevant conditional laws implies equality of their conditional intensities.

Proof of Proposition 3.1. Fix the shared decay β and a visited point $(t, z) \in \mathcal{R}$. Since φ is known and strictly monotone,

$$\varphi^{-1}(\lambda_j) = \mu_j^\star(t) + \left\langle C_{\cdot j}^\star(t, z), S\right\rangle.$$

Suppose two population pairs $(\mu_j^\star, C_{\cdot j}^\star)$ and $(\widetilde{\mu}_j, \widetilde{C}_{\cdot j})$ give the same conditional law. Then

$$\Delta\mu_j + \langle \Delta C_{\cdot j}, S\rangle = 0 \qquad \text{a.s. given } (t, z),$$

where

$$\Delta\mu_j = \widetilde{\mu}_j - \mu_j^\star, \qquad \Delta C_{\cdot j} = \widetilde{C}_{\cdot j} - C_{\cdot j}^\star.$$

Taking conditional covariance with S gives

$$\mathrm{Var}(S \mid t, z)\Delta C_{\cdot j} = 0.$$

By Assumption 2,

$$\Delta C_{\cdot j} = 0,$$

and substitution into the affine equality yields

$$\Delta\mu_j = 0.$$

Thus the local background and influence vector are jointly unique. □

Remark A.2 (Degenerate conditional design)**.** *At fixed (t, z), let $v \neq 0$ lie in the null space of $\mathrm{Var}(S \mid t, z)$ and set $a_v = \mathbb{E}[v^\top S \mid t, z]$. Then $v^\top S = a_v$ almost surely. Whenever $C_{\cdot j}^\star + \epsilon v \geq 0$ for some $\epsilon \neq 0$,*

$$(\mu_j^\star, C_{\cdot j}^\star) \longmapsto (\mu_j^\star - \epsilon a_v,\, C_{\cdot j}^\star + \epsilon v)$$

leaves the conditional predictor unchanged on this slice. This is local nonuniqueness with a free slice intercept; extending it to the full model requires compatibility with the baseline's restriction to t alone and the remaining model constraints.

Remark A.3 (Population target and fitted neural field)**.** *The population results condition on a fixed predictable state map $z_t = \Psi^\star(\mathcal{F}_{t^-})$ and therefore do not identify the coordinates of the learned state representation itself. In the empirical model, the event embeddings and field are learned jointly. Moreover,*

$$\widehat{C}_\theta(t, z) = \mathrm{softplus}(f_{\theta_C}(\mathrm{Fourier}(t), z)) > 0$$

for every finite pre-link output. Exact sparsity is therefore a property of the population target. Empirical graphs are obtained by thresholding $\widehat{C}_\theta$, as described in Section 3.3. The results above identify the population target; they do not constitute a consistency or exact-support-recovery claim for the global neural estimator. Although the coordinates of the learned state are not themselves identified, invertible linear reparameterizations of the embedding coordinates do not change the composed field values: if $z_t' = Az_t$ and $C'(t, z') = C^\star(t, A^{-1}z')$, then $C'(t, z_t') = C^\star(t, z_t)$. Hence the graph read-out is invariant within this coordinate-equivalence class.

B Additional Experimental Settings

B.1 Datasets and generation settings

Synthetic dataset. The state variable (z_t) is not exogenous. Past events drive the state, the state modulates the coupling, and the resulting coupling in turn governs the intensity of future events. This self-consistent feedback distinguishes our generator from designs in which (z_t) is supplied externally and reflects the endogenous state dependence observed in real event systems (Morariu-Patrichi & Pakkanen, 2021; Shi & Cartlidge, 2022). Table 5 summarizes the synthetic dataset settings. Events per pair is the total number of events divided by the number of ordered pairs, aggregated over all sequences, and measures the data density relative to the number of coupling functions to be estimated. Temporal variation is the standard deviation of $(C(t, z_t))$ over time, averaged across edges and expressed relative to the mean coupling magnitude of each edge. Figure 4 illustrates the ground-truth graphs across regimes for $K = 20$.

Real-world dataset. The alarm dataset consists of 30,691 events over 59 event types, grouped into 45 device sequences. The ground-truth graph has 147 directed edges and no self-loops; after applying the provided causal prior, 3,306 off-diagonal relations remain to be decided. Event times are measured in seconds. The median inter-event time is 5 seconds and the mean is 20 seconds, indicating a right-skewed inter-event-time distribution. Because multiple events can share the same one-second timestamp, we describe tie handling. The dataset also provides a topology over 45 devices. Its setting differs from ours. The ground truth is a single static summary graph, so structure metrics on it cannot resolve the time- or state-conditional claims that the synthetic studies test, but a method intended for operational logs must be shown to work on real ones.

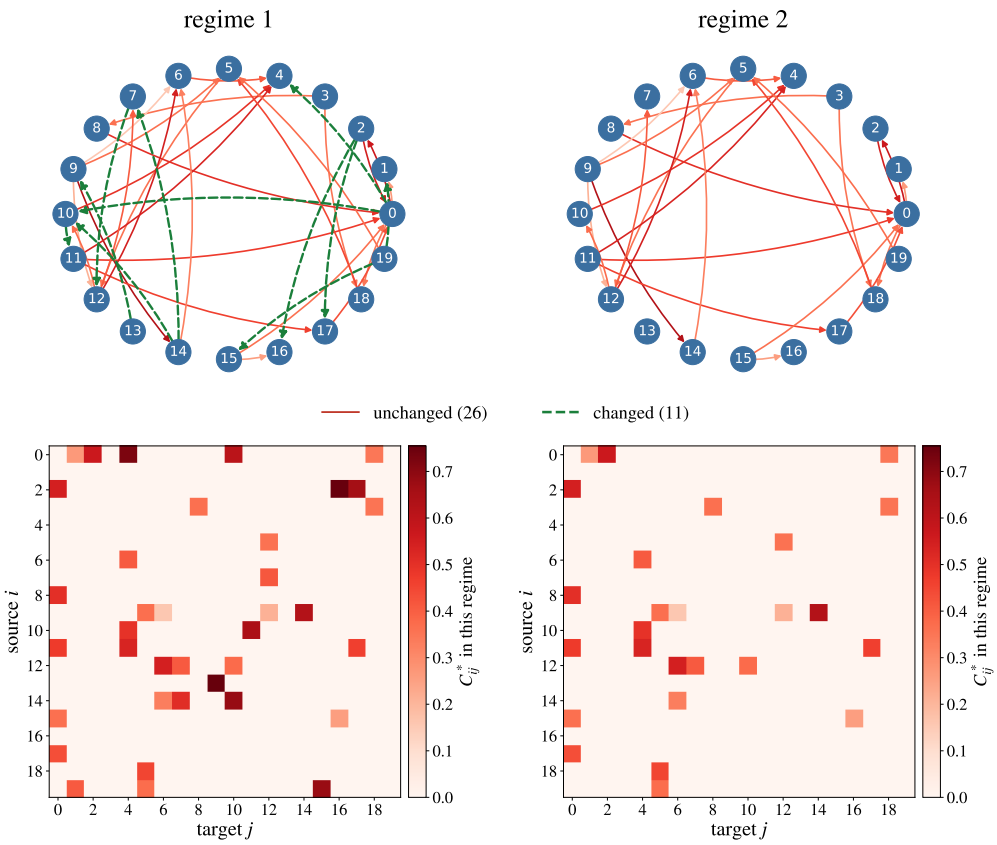

Figure 4: Ground-truth graphs across regimes for $K = 20$ event types.

Table 5: Synthetic dataset structure. As K increases, the data available per candidate pair decrease and the temporal variation becomes weaker, making recovery of the time-varying structure progressively more challenging.

	$K = 20$	$K = 50$	$K = 100$
Ordered pairs	380	2,450	9,900
Edges	32	77	137
Events per pair	36	13	6.3
Events per sequence	1,100	2,700	5,200
Temporal variation	17%	15%	9%

DGP experiments. To make the DGP comparison as controlled as possible, we use the same experimental configuration across all alternative generators. Each dataset contains (K=20) event types over a horizon of (T=80), with the same graph-generation rule and approximately the same edge density. We generate the sequences for each DGP and evaluate all methods over the same 20 random seeds. The scoring protocol is also fixed across DGPs, using best-F1 over 60 thresholds without an orientation oracle. CAUSALFIELD is run with the same configuration throughout, using the slow encoder with ($\tau_z = 20$), 150 training epochs, a learning rate of (5×10^{-3}), ($\lambda_1 = 3 \times 10^{-2}$), and ($\lambda_{\mathrm{TV}} = 10^{-2}$). The alternative DGPs provide approximately 30–120 events per candidate pair, avoiding a systematic disadvantage due to data scarcity.

We further control the structural properties of the generated data. All DGPs have nearly identical numbers of true edges and graph densities, while the delayed and power-law generators are rescaled when necessary to ensure stability. For the four time-gated generators, the magnitude of temporal variation in the coupling is also closely matched, with relative variation ranging from 0.170 to

0.173. The exponential-Hawkes generator is left unchanged because its coupling varies through its endogenous state, serving as the state-modulated reference case. These controls keep graph size, sparsity, information level, and the degree of non-stationarity broadly comparable while varying the underlying event-generation mechanism.

B.2 Evaluation Metrics

Graph recovery. Structure metrics are computed per seed and then averaged. Precision, recall (TPR), and F1 are defined as

$$\mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \qquad \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \qquad \mathrm{F1} = \frac{2\,\mathrm{Prec}\cdot\mathrm{TPR}}{\mathrm{Prec}+\mathrm{TPR}}.$$

The Matthews correlation coefficient (MCC) (Chicco & Jurman, 2020) is

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN}-\mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}.$$

ROC AUC and AUPRC evaluate the ranking induced by the estimated edge scores without selecting a decision threshold. We compute AUPRC using average precision. Results are reported as mean $\pm$ standard deviation over 20 seeds; deterministic methods on alarm are evaluated once.

Coefficient accuracy. On each held-out sequence, we compare $\widehat{C}(t,\widehat{z}_t)$ with $C^\star(t,z_t^\star)$ over a 200-point time grid using the same observed history, without aligning state coordinates. Let $\langle\cdot\rangle_{\mathcal{A}}$ denote averaging over grid points and edge pairs in a set $\mathcal{A}$. For the true-edge set $\mathcal{E}$, we compute

$$\mathrm{RMSE}_{\mathcal{A}} = \sqrt{\left\langle (\widehat{C}-C^\star)^2 \right\rangle_{\mathcal{A}}},$$
$$\mathrm{rel} = \frac{\mathrm{RMSE}_{\mathcal{E}}}{\sqrt{\langle (C^\star)^2\rangle_{\mathcal{E}}}},$$
$$\mathrm{bias} = \left\langle \widehat{C}-C^\star \right\rangle_{\mathcal{E}}.$$

RMSE is reported separately for true edges and non-edges. Background RMSE is computed between $\widehat{\mu}$ and $\mu^\star$ on the pre-link scale, averaging squared errors over grid points and event types before taking the square root. Temporal correlation is the Pearson correlation between estimated and true coefficient trajectories, averaged over edges with varying $C^\star$.

Sequence-level metrics are averaged over the five held-out sequences of each seed and reported as mean $\pm$ standard deviation over seeds. Comparisons with the full model use seed-paired bootstrap confidence intervals.

Predictive fit. Held-out negative log-likelihood (NLL) per sequence is

$$\mathrm{NLL} = -\sum_{k=1}^{N}\log\lambda_{e_k}(t_k) + \int_0^T \sum_{j=1}^{K}\lambda_j(t)\,dt.$$

Per-event NLL is obtained by dividing by the event count. Table 22 compares Monte Carlo estimates of the compensator with a 16,384-point midpoint rule.

B.3 Hyperparameters and Baselines

Table 6 lists the setting used in the main experiments. The field network C_θ is `Linear(`$8+d_z$`, 64) - SiLU - Linear(64, 64) - SiLU - Linear(64,` K^2`)` followed by a softplus output, with the final layer initialized to zero weight and bias -1.0 so that the initial field is small but not saturated. The baseline network is `Linear(8, 32) - SiLU - Linear(32,` K`)` with a zero-initialized final layer, and the scalar baseline intercept is initialized to -2.5. Gradient norms are clipped at 5.0. The latent state has dimension $d_z = 16$ and is computed as a normalized exponential moving average of past event embeddings with time constant $\tau_z = 20$; τ_z is held fixed rather than learned, so that the state time scale is decoupled from the excitation time scale. In the static mode, C is a free matrix initialized as $0.1\cdot\mathcal{N}(0,I) - 1.0$ before the softplus.

We extensively tuned all baselines using method-specific grids rather than relying on default configurations. For example, we swept discretization, delay window, and distribution for CASCADE (up to 36 configurations), binning and regularization for NGC (40 configurations), and state number, decay, and block size for StateHPs (24 configurations), with analogous sweeps for the remaining methods. These searches substantially improved several baselines while confirming that some methods, such as THP, remained insensitive to tuning. We also test SHP (Qiao et al., 2023) and NHP (Mei & Eisner, 2017). These baseline models are under the same configurations in Table 1.

Blocked-t. The full model's Fourier time representation is replaced with a one-hot block index, and the field is restricted from using the latent state. All other components, including the softplus link, kernel decay, optimizer, and regularization terms, are kept unchanged. Following the same procedure as for the full model, λ_{TV} is selected by validation likelihood for each configuration. In all cases, the largest candidate value, 3×10^{-2}, is selected. The resulting models contain between $10{,}020$ parameters at $T/2$ and $11{,}364$ parameters at $T/16$, compared with $33{,}000$ parameters in the full model.

Runtime constraints. For TTPM and THP we report results at hop 1; hop 2 requires approximately 3.3 hours per run and was not included in the sweep. L0HAWKES scales quadratically in the number of events, so we impose an event cap and a wall-clock timeout, both applied uniformly across configurations.

Hyperparameter tuning for baselines. To ensure a fair comparison, we performed extensive method-specific hyperparameter tuning for all baselines. We tuned the parameters most relevant to each method, including decay rates, regularization strengths, temporal binning, optimization settings, and method-specific structural parameters. StateHPs was evaluated over 24 configurations spanning the number of states, decay rate, and number of blocks, NGC over 40 configurations of temporal binning and regularization, and CASCADE over 36 configurations spanning discretization, delay window, distribution, alignment, and MDL variants. CAUSE, SHP, NPHC, L0HAWKES, and THP were likewise tuned over their relevant hyperparameters.

Tuning materially improved several baselines. On the synthetic setting, NGC improved from AUC 0.566 to (0.75$\pm$0.05), CAUSE from 0.571 to (0.66$\pm$0.06), and NPHC from 0.502 to (0.58$\pm$0.05). StateHPs reached (0.79$\pm$0.06) after tuning. Notably, its best configuration used ($\beta = 3$), showing that fixing the true generative decay ($\beta = 1$) does not necessarily yield the strongest baseline under model misspecification.

We followed the same procedure on the real alarm dataset, using separate method-specific search spaces and reporting the best configuration found for each baseline. NPHC, for example, reached F1 0.252 and AUC 0.783 after tuning, while methods such as TTPM, L0HAWKES, and NGC showed only limited gains across their tested configurations. These experiments therefore compare CAUSALFIELD against tuned baseline configurations. When a baseline remains weak, its reported result reflects the best performance obtained over the corresponding search space.

B.4 REPRODUCIBILITY AND COMPUTE

All experiments were run in a single conda environment with Python 3.11.5 and CUDA 12.8, on a SLURM cluster using one Titan GPU, 24 GB of memory, and a wall-clock limit of 3.5 hours per job; each random seed corresponds to one job.

The core stack is NumPy 2.2.6 for the simulators, metrics, and aggregation, PyTorch 2.10.0 for the model and all training, scikit-learn 1.8.0 for AUC computation, k-means quantization of the latent state, and the canonical correlation diagnostics, SciPy 1.16.3 for the perturbation tests and the Bessel normalization of the mean-preserving baseline, and pandas 2.3.3 and Matplotlib 3.10.7 for data loading and figures.

For causal discovery specifically we use gCastle, which provides both the TTPM/THP baselines and the `MetricsDAG` scorer used for all real-data results, together with networkx 3.5 for the topological graph structures required by TTPM. Auxiliary dependencies are Numba 0.61.2 for accelerating the CASCADE baseline, joblib 1.3.2 for parallel execution, psutil 5.9.0 for resource benchmarking and timeout handling, and TensorFlow 2.20.0, which is required only by the NPHC baseline.

Table 6: Complete hyperparameter settings. Values in the right column apply to the real alarm data where they differ from the default.

	Setting	Default	Alarm
Model	Latent state dimension d_z	16	—
	State time constant τ_z	20 (fixed)	5
	Fourier order	4 (8 dims)	—
	Kernel decay β	learned	—
Optimization	Optimizer	Adam	—
	Learning rate	5×10^{-3}	10^{-2}
	Epochs	150	300
	Batching	full-batch	—
	Gradient clipping (norm)	5.0	—
	Ensemble	3×120 ep.	—
Objective	Sparsity λ_1	10^{-3}	$3 \times 10^{-2}, 10^{-1}$
	TV λ_{TV}	val. NLL over $\{3\times10^{-3}, 10^{-2}, 3\times10^{-2}\}$	10^{-2}
	Huber δ	0.05	—
	Regularization grid	64 points	—
	MC samples (train / eval)	64 / 256	—
Read-out	Graph read-out	mean-C	—

Table 7: Alarm-data results without topology information at the reference-free threshold $\gamma = 0.3{\cdot}\max$. Results are averaged over 20 seeds where applicable, while deterministic methods are reported from a single run. AUC and AUPRC are threshold-free and computed on the 3,306 candidate pairs (145 true edges), so a random ranker scores AUPRC $= 0.044$; TTPM outputs a binary graph and has no ranking metric. Dark and bold entries denote the best value, whereas light and underlined entries denote the second-best value.

Method	F_1	Precision	TPR	AUC	AUPRC	MCC
CASCADE	0.08	0.04	**0.99**	0.46	0.18	0.03
NPHC	0.09	0.23	0.05	0.78	0.16	0.09
CAUSE	0.09 ± 0.02	0.13 ± 0.08	0.32 ± 0.33	0.50 ± 0.07	0.06 ± 0.01	0.05 ± 0.05
StateHPs	0.16 ± 0.00	0.58 ± 0.00	0.10 ± 0.00	0.66 ± 0.00	0.21 ± 0.00	0.22 ± 0.00
NGC	0.05 ± 0.01	0.10 ± 0.01	0.04 ± 0.01	0.54 ± 0.01	0.05 ± 0.00	0.03 ± 0.01
L0HAWKES	0.06	0.05	0.06	0.55	0.05	0.01
THP	0.08 ± 0.00	0.04 ± 0.00	0.52 ± 0.03	0.50 ± 0.02	0.05 ± 0.00	0.01 ± 0.01
NHP	0.02 ± 0.00	0.11 ± 0.00	0.01 ± 0.00	0.42 ± 0.00	0.04 ± 0.00	0.02 ± 0.00
TTPM	0.12	**1.00**	0.06	–	–	–
CAUSALFIELD (ours)	$\mathbf{0.60 \pm 0.01}$	0.54 ± 0.02	0.68 ± 0.02	$\mathbf{0.88 \pm 0.00}$	$\mathbf{0.66 \pm 0.01}$	$\mathbf{0.59 \pm 0.01}$

Reproducing the model and training results requires only NumPy, PyTorch, scikit-learn, and SciPy; reproducing the real-data scores additionally requires gCastle. Thresholded graph metrics are computed with the gCastle evaluation utilities, while ROC AUC and AUPRC are computed from continuous edge scores using scikit-learn.

C ADDITIONAL EXPERIMENTS

C.1 ADDITIONAL RESULTS

Alarm experiment with topology information. The alarm dataset includes a 45×45 equipment-connectivity matrix that provides auxiliary topology information for causal inference. This information is available to all methods and substantially improves several baselines, most notably CASCADE (best-F1 $0.29 \rightarrow 0.60$) and StateHPs ($0.35 \rightarrow 0.74$), as shown in Tables 14 and 15. CAUSALFIELD achieves the highest best-F1 both with and withouttopology information, indicating that its advantage does not depend on prior connectivity information.

THP in Table 3. THP yields nearly identical graph estimates across all four DGPs. A shuffled-time control reproduces essentially the same adjacency (correlation $= 1.0000$, maximum absolute

Table 8: Alarm-data results with topology information at the reference-free threshold $\gamma = 0.3 \cdot \max$. Results are averaged over 20 seeds where applicable, while deterministic methods are reported from a single run. AUC and AUPRC are threshold-free and computed on the 3,306 candidate pairs (145 true edges), so a random ranker scores AUPRC $= 0.044$; TTPM outputs a binary graph and has no ranking metric. $|\widehat{G}|$ is the mean number of predicted edges. Dark and bold entries denote the best value, whereas light and underlined entries denote the second-best value.

Method	F_1	Precision	TPR	AUC	AUPRC	MCC	$\lvert\widehat{G}\rvert$
CASCADE	0.23	**1.00**	0.13	0.96	0.66	0.35	19
NPHC	0.21	0.34	0.15	0.90	0.28	0.20	64
CAUSE	0.13 ± 0.12	0.19 ± 0.15	0.55 ± 0.44	0.87 ± 0.04	0.30 ± 0.04	0.12 ± 0.12	1,508
StateHPs	0.62 ± 0.00	0.92 ± 0.00	0.46 ± 0.00	0.93 ± 0.00	0.75 ± 0.00	0.64 ± 0.00	74
NGC	0.07 ± 0.01	0.09 ± 0.01	0.06 ± 0.01	0.48 ± 0.00	0.05 ± 0.00	0.04 ± 0.01	110
L0HAWKES	0.05	0.04	0.06	0.54	0.05	0.00	215
THP	0.08 ± 0.00	0.04 ± 0.00	0.51 ± 0.03	0.50 ± 0.02	0.05 ± 0.00	0.01 ± 0.01	1,676
NHP	0.08 ± 0.00	0.04 ± 0.00	0.62 ± 0.10	0.48 ± 0.03	0.05 ± 0.00	-0.01 ± 0.02	2,198
TTPM	0.14	**1.00**	0.07	–	–	0.27	–
CAUSALFIELD (ours)	**0.93 ± 0.01**	0.90 ± 0.02	**0.95 ± 0.01**	**1.00 ± 0.00**	**0.98 ± 0.00**	**0.92 ± 0.01**	155

Table 9: Ablation results for different influence-field variants (20 seeds).

Variant	Agg. AUC	F1 @ $0.3 \cdot \max$	Snapshot AUC	Held-out NLL	Params	Paired ΔNLL vs. full [95% CI]	Field rel. RMSE
Full $C(t, z)$	0.987 ± 0.013	0.859 ± 0.040	0.974 ± 0.015	1196.4	33,049	—	0.406 ± 0.043
Time-only $C(t)$	0.984 ± 0.014	0.838 ± 0.062	0.960 ± 0.018	1200.8	32,025	$+4.47$ $[+3.76, +5.20]^*$	0.441 ± 0.054
Quantized $K_z = 2$	0.986 ± 0.012	0.845 ± 0.058	0.965 ± 0.017	1199.6	31,833	$+3.22$ $[+2.61, +3.90]^*$	0.417 ± 0.052
Quantized $K_z = 4$	0.985 ± 0.014	0.836 ± 0.056	0.963 ± 0.017	1200.0	31,961	$+3.64$ $[+2.92, +4.36]^*$	0.427 ± 0.049
Quantized $K_z = 8$	0.986 ± 0.013	0.849 ± 0.054	0.967 ± 0.017	1198.8	32,217	$+2.44$ $[+1.87, +3.07]^*$	0.419 ± 0.053
Quantized $K_z = 16$	0.987 ± 0.012	0.863 ± 0.052	0.971 ± 0.017	1198.0	32,729	$+1.60$ $[+0.91, +2.29]^*$	0.404 ± 0.054

difference $= 1.97 \times 10^{-4}$), indicating a degenerate solution that is insensitive to event-timing information.

Ablation study on z-quantization: continuous vs. finite-regime representations. To evaluate the impact of latent discretization, we conduct an ablation study on the quantized z-space, contrasting a continuous-state representation with a finite-regime formulation as shown in Table 9.

Held-out NLL is the mean per held-out sequence with the compensator integrated by a deterministic 16 384-point rule. The paired ΔNLL is over the 100 held-out sequences with a 95% bootstrap CI, * = CI excludes 0. NLL is ≈1,200 nats per sequence (≈1,475 events), so a gain of 3–4 nats is approximately 0.3% of the reported NLL, or 0.002–0.003 nats per event: statistically clear, small in magnitude. Field error is the relative RMSE of the composed $\hat{C}(t, \hat{z})$ against $C^\star(t, z^\star)$ on the true edges of the held-out sequences. Parameter counts differ by less than 6%.

Full DGP results and analysis of the logistic generator. Table 10 shows the score of all evaluated metrics in the different DGP experiments. While performance remains broadly consistent across other settings, we observe a noticeable variation on the logistic generator. We examine the factors underlying this specific outcome as in Tables 11 and 12.

All runs use $K = 20$, $T = 80$, $M = 48$ sequences, 20 seeds, and the slow encoder with $\tau_z = 20$. The logistic generator is the only process on which CAUSALFIELD does not lead. Its performance is nevertheless within one standard deviation of StateHPs, with F1 0.68 ± 0.12 versus 0.70 ± 0.12 and identical AUC.

The degradation is caused by saturation rather than the discrete time grid. At the default bias the median per-step firing probability is 0.375, but 28.8% of the steps fire with probability above 0.9. Events on those steps are nearly certain and therefore carry little information about which edge triggered them. Table 12 shows that the excitation design remains at least as well conditioned as in the continuous-time processes because events are jittered within each step. This makes grid degeneracy unlikely to explain the performance drop. Table 11 provides a direct test of saturation by lowering the bias while keeping the grid fixed. As the median firing probability falls from 0.375 to 0.049, CAUSALFIELD recovers to F1 0.903 with roughly one third as many events (281 versus 798 per

Table 10: Performance across synthetic DGPs, reported as mean ± standard deviation over seeds. Rows correspond to algorithms and columns to DGPs. Bold values indicate the best performance within each DGP and metric.

Metric	Method	expHawkes	Delayed	Power-law	Logistic
F1	SHP	0.71 ± 0.08	0.35 ± 0.08	0.36 ± 0.08	0.21 ± 0.07
	CAUSE	0.51 ± 0.13	0.29 ± 0.09	0.39 ± 0.09	0.24 ± 0.05
	L0HAWKES	0.14 ± 0.04	0.15 ± 0.04	0.15 ± 0.04	0.16 ± 0.04
	NPHC	0.17 ± 0.04	0.17 ± 0.04	0.16 ± 0.02	0.23 ± 0.04
	CASCADE	0.11 ± 0.02	0.12 ± 0.02	0.13 ± 0.02	0.12 ± 0.02
	THP	0.15 ± 0.04	0.15 ± 0.04	0.15 ± 0.04	0.15 ± 0.04
	StateHPs	0.85 ± 0.16	0.79 ± 0.12	0.80 ± 0.08	**0.70 ± 0.12**
	CAUSALFIELD	**0.97 ± 0.01**	**0.89 ± 0.04**	**0.92 ± 0.04**	0.68 ± 0.12
Precision	SHP	0.80 ± 0.11	0.41 ± 0.15	0.42 ± 0.14	0.33 ± 0.11
	CAUSE	0.61 ± 0.18	0.31 ± 0.11	0.47 ± 0.21	0.25 ± 0.17
	L0HAWKES	0.10 ± 0.04	0.10 ± 0.03	0.09 ± 0.03	0.11 ± 0.03
	NPHC	0.12 ± 0.05	0.11 ± 0.04	0.10 ± 0.03	0.16 ± 0.05
	CASCADE	0.07 ± 0.02	0.08 ± 0.02	0.08 ± 0.02	0.08 ± 0.02
	THP	0.10 ± 0.04	0.10 ± 0.04	0.10 ± 0.04	0.10 ± 0.04
	StateHPs	0.80 ± 0.22	0.75 ± 0.16	0.79 ± 0.12	0.70 ± 0.18
	CAUSALFIELD	**0.97 ± 0.03**	**0.92 ± 0.05**	**0.94 ± 0.04**	**0.73 ± 0.15**
TPR	SHP	0.64 ± 0.08	0.33 ± 0.07	0.33 ± 0.08	0.16 ± 0.06
	CAUSE	0.50 ± 0.15	0.32 ± 0.13	0.40 ± 0.12	0.35 ± 0.14
	L0HAWKES	0.34 ± 0.12	0.43 ± 0.12	0.48 ± 0.13	0.40 ± 0.14
	NPHC	0.44 ± 0.16	0.43 ± 0.14	0.45 ± 0.11	0.49 ± 0.20
	CASCADE	0.32 ± 0.07	0.36 ± 0.11	0.37 ± 0.11	0.30 ± 0.09
	THP	0.40 ± 0.14	0.40 ± 0.14	0.40 ± 0.14	0.40 ± 0.14
	StateHPs	0.95 ± 0.07	0.85 ± 0.09	0.83 ± 0.08	**0.74 ± 0.12**
	CAUSALFIELD	**0.98 ± 0.02**	**0.86 ± 0.07**	**0.91 ± 0.06**	0.66 ± 0.13
MCC	SHP	0.69 ± 0.08	0.31 ± 0.10	0.32 ± 0.10	0.18 ± 0.08
	CAUSE	0.50 ± 0.12	0.25 ± 0.09	0.36 ± 0.09	0.18 ± 0.08
	L0HAWKES	0.03 ± 0.06	0.04 ± 0.06	0.04 ± 0.06	0.07 ± 0.05
	NPHC	0.08 ± 0.06	0.07 ± 0.07	0.05 ± 0.03	0.16 ± 0.06
	CASCADE	-0.02 ± 0.04	-0.00 ± 0.04	0.00 ± 0.05	-0.01 ± 0.05
	THP	0.04 ± 0.06	0.04 ± 0.06	0.04 ± 0.06	0.04 ± 0.06
	StateHPs	0.85 ± 0.16	0.78 ± 0.12	0.79 ± 0.09	**0.68 ± 0.13**
	CAUSALFIELD	**0.97 ± 0.01**	**0.88 ± 0.04**	**0.92 ± 0.04**	0.66 ± 0.14
AUC	SHP	0.84 ± 0.05	0.64 ± 0.05	0.64 ± 0.05	0.55 ± 0.03
	CAUSE	0.83 ± 0.10	0.61 ± 0.09	0.72 ± 0.07	0.58 ± 0.08
	L0HAWKES	0.48 ± 0.07	0.50 ± 0.08	0.49 ± 0.09	0.52 ± 0.10
	NPHC	0.54 ± 0.06	0.54 ± 0.07	0.53 ± 0.05	0.64 ± 0.06
	CASCADE	0.41 ± 0.05	0.43 ± 0.05	0.44 ± 0.05	0.39 ± 0.06
	THP	0.49 ± 0.05	0.49 ± 0.05	0.49 ± 0.05	0.49 ± 0.05
	StateHPs	0.98 ± 0.02	0.97 ± 0.02	0.97 ± 0.02	**0.92 ± 0.05**
	CAUSALFIELD	**1.00 ± 0.00**	**0.99 ± 0.01**	**0.99 ± 0.01**	**0.92 ± 0.05**

sequence) and outperforms StateHPs by $+0.107$, with a paired 95% interval of $[+0.058, +0.165]$ over 20 seeds. The paired 95% confidence interval excludes zero at $b_0 = -2.5, -3.0, -3.3$, and -3.6, but includes zero at $b_0 = -2.7$. Recovery improves as saturation decreases despite fewer events, supporting saturation as a contributor to the performance gap. The shrinking λ/p column indicates that the link misreading vanishes together with saturation, so we treat it as a consequence of saturation rather than a competing cause.

We report $\varrho_S = \lambda_{\min}/\lambda_{\max}$ of $\mathrm{Cov}(S)$ at event times and the fraction of distinct realized S values at $\beta = 1$ in Table 12. The logistic generator is at least as well conditioned as the continuous-time generators because within-step jitter produces continuous realized lags.

Threshold and read-out sensitivity. Table 7 gives the full results summarized in Figure 3, and Table 8 reports the same evaluation with the device topology. CAUSALFIELD attains the highest F1, AUC, AUPRC, and MCC in both settings; baselines exceed it only in precision or TPR at extreme operating points.

Table 11: Logistic generator: saturation and de-saturation. p is the per-step firing probability measured from the generator alone; λ/p is the misreading of a continuous-time intensity at the median p. F1 is best-F1 over 20 seeds. The default row is the logistic column of Table 3. The last column is the paired Ours−StateHPs difference with 95% bootstrap CI; bold marks rows in which the interval excludes zero.

b_0	median p	p>0.9	λ/p	events/seq	Ours	StateHPs	paired ΔF1 [95% CI]
-2.0 (default)	0.375	28.8%	1.25×	798	0.683	0.701	-0.018 $[-0.073, +0.039]$
-2.3	0.216	24.7%	1.13×	697	0.735	0.760	-0.026 $[-0.076, +0.021]$
-2.5	0.123	19.6%	1.07×	589	**0.801**	0.749	$+0.052$ $[+0.002, +0.098]$
-2.7	0.078	14.9%	1.04×	485	0.839	0.804	$+0.035$ $[-0.008, +0.078]$
-3.0	0.049	5.6%	1.03×	281	**0.903**	0.796	$+0.107$ $[+0.058, +0.165]$
-3.3	0.036	2.4%	1.02×	165	**0.913**	0.791	$+0.122$ $[+0.059, +0.186]$
-3.6	0.027	0.5%	1.01×	81	**0.869**	0.831	$+0.038$ $[+0.006, +0.075]$

Table 12: Conditioning of the excitation design using generated events only (20 seeds).

	ϱ_S			distinct-S
DGP	β=0.3	β=1	β=3	fraction
exp-Hawkes	1.6×10^{-2}	3.4×10^{-2}	3.8×10^{-2}	0.747
delayed	1.1×10^{-2}	3.8×10^{-2}	6.8×10^{-2}	0.937
power-law	1.7×10^{-2}	5.1×10^{-2}	8.1×10^{-2}	0.930
logistic	1.1×10^{-2}	5.9×10^{-2}	8.9×10^{-2}	0.959

Tables 13, 14 and 15 further show that CAUSALFIELD retains the strongest performance even under the most favorable seed and threshold selection for each method. This comparison is intended as a robustness check against threshold choice: even when each method is evaluated at its best operating point, CAUSALFIELD remains the top performer.

Table 17 shows that the F1 advantage over CASCADE holds on every seed under every model-based read-out, while the read-out choice trades ranking (AUC) against thresholded accuracy.

Held-out likelihood provides a complementary evaluation of predictive fit. On alarm, the full $C(t, z)$ model achieves lower NLL than static and time-only variants, including the capacity-matched time-only model (Table 18). These gains are small, whereas replacing the time-varying background $\mu_j(t)$ with a constant μ_j causes a much larger increase in NLL. Synthetic ablations likewise show lower held-out NLL than time-only and coarsely quantized-state variants, despite similar aggregate graph-recovery metrics across the time-only model and k-means quantizations with 2 to 16 clusters (Table 9 in Appendix C). Figure 2 separately illustrates how the continuous field captures changes in coupling within a block.

Support of the drift in RQ1. A piecewise-constant background cannot capture drift within a segment. Because the intensity is given by $\varphi(\mu + \langle C, S\rangle)$, residual background variation can instead be absorbed into C. Figure 5 tests this effect in a setting where $C^\star_{ij}$ is constant and the baselines drift with a common phase, so systematic variation in $\hat{C}_{ij}$ reflects background leakage.

The four-step background in $blocked_t$ leaves 69% of the variance in $\mu^\star_j(t)$ unexplained, with a residual s.d. of 1.14 versus 0.50 for the full field (paired difference $+0.65$, $[0.44, 0.85]$). This residual is reflected in $\hat{C}$ as a positive offset (paired $+0.069$, $[0.038, 0.099]$), and the two errors are strongly correlated across seeds ($r = 0.90$). In contrast, the full field tracks the baseline drift through its μ component and shows neither a comparable offset nor substantial co-movement with $\mu^\star_j(t)$ ($\bar{\rho} = +0.11$). This provides a coefficient-level explanation for the false edges produced by segmented estimators in Table 25. Under independent drift (bottom), the same leakage mechanism remains, but without shared temporal variation its effect does not appear as systematic correlation. The remaining variation is consistent with finite-sample error.

Ablation studies of the encoders. We further conducted an encoder ablation study by replacing the default slow encoder with an attention encoder (`attn`) to assess whether the results depend on the choice of state encoder as in Table 19. Under `attn`, AUC and F1 are numerically identical

Table 13: Best F1 results on synthetic data across different numbers of event types, evaluated under the best hyperparameter setting over 20 independent seeds based on the highest F1 score. Values are mean ± standard deviation. Dark and bold entries denote the best mean, whereas light and underlined entries denote the second-best mean. * indicates that no oracle information is used.

K	Method	F_1	Precision	TPR	AUC	MCC
20	L0HAWKES	0.20 ± 0.02	0.12 ± 0.02	0.52 ± 0.12	0.57 ± 0.04	0.10 ± 0.04
	NPHC	0.21 ± 0.04	0.15 ± 0.04	0.44 ± 0.14	0.58 ± 0.05	0.12 ± 0.05
	CAUSE	0.28 ± 0.06	0.22 ± 0.07	0.46 ± 0.14	0.66 ± 0.06	0.21 ± 0.07
	CASCADE	0.17 ± 0.04	0.11 ± 0.03	0.46 ± 0.11	0.52 ± 0.05	0.05 ± 0.06
	THP	0.17 ± 0.03	0.11 ± 0.02	0.42 ± 0.10	0.50 ± 0.05	0.06 ± 0.05
	NGC	0.37 ± 0.10	0.32 ± 0.16	0.54 ± 0.14	0.75 ± 0.05	0.32 ± 0.11
	StateHPs	0.41 ± 0.08	0.39 ± 0.09	0.47 ± 0.13	0.79 ± 0.06	0.36 ± 0.07
	CAUSALFIELD (ours)	$\mathbf{0.75 \pm 0.06}$	$\mathbf{0.78 \pm 0.08}$	$\mathbf{0.74 \pm 0.12}$	$\mathbf{0.96 \pm 0.02}$	$\mathbf{0.73 \pm 0.06}$
50	L0HAWKES	0.08 ± 0.02	0.05 ± 0.02	0.32 ± 0.14	0.55 ± 0.02	0.05 ± 0.02
	NPHC	0.10 ± 0.02	0.06 ± 0.02	0.29 ± 0.12	0.57 ± 0.03	0.07 ± 0.02
	CAUSE	0.11 ± 0.03	0.09 ± 0.07	0.25 ± 0.14	0.61 ± 0.05	0.09 ± 0.04
	CASCADE	0.07 ± 0.01	0.04 ± 0.01	0.43 ± 0.15	0.53 ± 0.04	0.03 ± 0.02
	THP	0.06 ± 0.01	0.04 ± 0.01	0.34 ± 0.15	0.50 ± 0.03	0.02 ± 0.02
	NGC	0.24 ± 0.05	0.21 ± 0.06	0.32 ± 0.10	0.77 ± 0.02	0.22 ± 0.04
	StateHPs	0.18 ± 0.05	0.23 ± 0.11	0.18 ± 0.09	0.59 ± 0.05	0.16 ± 0.06
	CAUSALFIELD (ours)	$\mathbf{0.58 \pm 0.05}$	$\mathbf{0.56 \pm 0.08}$	$\mathbf{0.61 \pm 0.08}$	$\mathbf{0.95 \pm 0.02}$	$\mathbf{0.57 \pm 0.06}$
100	L0HAWKES	0.05 ± 0.01	0.03 ± 0.01	0.17 ± 0.08	0.58 ± 0.02	0.04 ± 0.01
	NPHC	0.04 ± 0.01	0.02 ± 0.01	0.22 ± 0.14	0.53 ± 0.02	0.02 ± 0.01
	CAUSE	0.06 ± 0.01	0.04 ± 0.02	0.18 ± 0.09	0.60 ± 0.03	0.05 ± 0.02
	CASCADE	0.03 ± 0.00	0.02 ± 0.00	0.37 ± 0.13	0.51 ± 0.03	0.01 ± 0.01
	THP	0.03 ± 0.00	0.02 ± 0.00	0.22 ± 0.13	0.50 ± 0.03	0.01 ± 0.01
	NGC	0.26 ± 0.04	0.27 ± 0.08	0.26 ± 0.06	0.80 ± 0.02	0.25 ± 0.04
	StateHPs	0.12 ± 0.04	0.18 ± 0.13	0.11 ± 0.05	0.58 ± 0.07	0.12 ± 0.05
	CAUSALFIELD (ours)	$\mathbf{0.54 \pm 0.04}$	$\mathbf{0.55 \pm 0.10}$	$\mathbf{0.56 \pm 0.08}$	$\mathbf{0.96 \pm 0.01}$	$\mathbf{0.54 \pm 0.04}$

Table 14: Best-F1 results on the alarm dataset without device topology. Results are averaged over 20 seeds where applicable, while deterministic methods are reported from a single run. Dark and bold entries denote the best value, whereas light and underlined entries denote the second-best value.

Method	F_1	Precision	TPR	AUC	AUPRC	MCC
CASCADE	0.29	0.66	0.18	0.46	0.18	0.33
NPHC	0.25	0.18	0.44	0.78	0.16	0.23
CAUSE	0.12 ± 0.02	0.10 ± 0.04	0.23 ± 0.11	0.50 ± 0.07	0.06 ± 0.01	0.08 ± 0.02
StateHPs	0.35 ± 0.00	0.46 ± 0.00	0.28 ± 0.00	0.66 ± 0.00	0.21 ± 0.00	0.33 ± 0.00
NGC	0.10 ± 0.00	0.06 ± 0.00	0.40 ± 0.11	0.54 ± 0.01	0.05 ± 0.00	0.04 ± 0.01
L0HAWKES	0.10	0.06	0.29	0.55	0.05	0.05
THP	0.09 ± 0.01	0.05 ± 0.01	0.37 ± 0.11	0.50 ± 0.02	0.05 ± 0.00	0.03 ± 0.02
NHP	0.06 ± 0.00	0.03 ± 0.00	0.37 ± 0.05	0.42 ± 0.00	0.04 ± 0.00	-0.04 ± 0.01
TTPM	0.12	**1.00**	0.06	–	–	–
CAUSALFIELD (ours)	$\mathbf{0.62 \pm 0.01}$	0.64 ± 0.06	$\mathbf{0.62 \pm 0.06}$	$\mathbf{0.88 \pm 0.00}$	$\mathbf{0.66 \pm 0.01}$	$\mathbf{0.61 \pm 0.01}$

at $\tau_z = 20$ and $\tau_z = 5$, while the diagnostic moves from 19.80 to 4.95. This occurs because the attention encoder does not use τ_z, making the timescale-ratio diagnostic vacuous outside the `slow` encoder.

Remark C.1 (Conditional excitation variation)**.** *Although z_t and $S(t)$ are both functions of the event history, conditioning on z_t need not determine $S(t)$. Since z_t is a low-dimensional summary of the history, distinct histories can reach the same time–state neighborhood with different excitation profiles. After conditioning on t, we compare the conditional covariance of S within (t, z) neighborhoods against a null that permutes state-bin labels within each time bin. At M=48, the observed-to-null ratio of the median $\lambda_{\min}$ ranges from 0.34 to 0.61 across generators, rising to 0.55 at M=96 for the reference cell, while the trace ratio ranges from 0.78 to 0.84 as in Table 20. The between-replicate covariance is positive definite wherever it is estimable. Because the neighborhoods condition only on the two leading principal coordinates of z_t, these statistics are a coarsened proxy and may overstate the variation remaining after conditioning on the full state.*

Table 15: Best-F1 results on the alarm data with topology information, evaluated under the best hyperparameter setting over 20 independent seeds based on the highest F1 score. Results are averaged over 20 seeds where applicable, while deterministic methods are reported from a single run. Dark and bold entries denote the best value, whereas light and underlined entries denote the second-best value.

Method	F_1	Precision	TPR	AUC	AUPRC	MCC
CASCADE	0.60	0.79	0.48	0.96	0.66	0.60
NPHC	0.38	0.30	0.52	0.90	0.28	0.36
CAUSE	0.47 ± 0.05	0.40 ± 0.05	0.58 ± 0.08	0.87 ± 0.04	0.30 ± 0.04	0.45 ± 0.06
StateHPs	0.74 ± 0.00	0.74 ± 0.00	0.73 ± 0.00	0.93 ± 0.00	0.75 ± 0.00	0.73 ± 0.00
NGC	0.09 ± 0.00	0.08 ± 0.01	0.10 ± 0.02	0.48 ± 0.00	0.05 ± 0.00	0.04 ± 0.01
L0HAWKES	0.10	0.06	0.43	0.54	0.05	0.05
THP	0.09 ± 0.01	0.05 ± 0.01	0.37 ± 0.11	0.50 ± 0.02	0.05 ± 0.00	0.03 ± 0.02
NHP	0.09 ± 0.01	0.06 ± 0.01	0.23 ± 0.11	0.48 ± 0.03	0.05 ± 0.00	0.03 ± 0.02
TTPM	0.14	**1.00**	0.07	–	–	0.27
CAUSALFIELD (ours)	$\mathbf{0.93 \pm 0.01}$	0.93 ± 0.01	$\mathbf{0.93 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.98 \pm 0.00}$	$\mathbf{0.93 \pm 0.01}$

Table 16: Synthetic results across scale at $K = 150, 200$.

K	Method	AUC	F1	Precision	TPR
150	**CAUSALFIELD**	$\mathbf{0.953 \pm 0.007}$	$\mathbf{0.395 \pm 0.040}$	$\mathbf{0.382 \pm 0.066}$	$\mathbf{0.424 \pm 0.070}$
	CASCADE	0.456 ± 0.019	0.019 ± 0.003	0.010 ± 0.002	0.263 ± 0.105
200	**CAUSALFIELD**	$\mathbf{0.952 \pm 0.008}$	$\mathbf{0.350 \pm 0.045}$	$\mathbf{0.313 \pm 0.044}$	$\mathbf{0.398 \pm 0.047}$
	CASCADE	0.462 ± 0.019	0.015 ± 0.002	0.008 ± 0.001	0.272 ± 0.123

Two qualifications apply. First, the condition cannot be assessed when fewer sequences than active coordinates visit a neighborhood, as at $M{=}12$. Second, the minimum-over-neighborhoods form of ϱ_S in Section 3.3 is identically zero at these sample sizes because some type is dormant in at least one neighborhood. We report the median over neighborhoods on active coordinates.

C.2 FIELD-ERROR AND LIKELIHOOD-STABILITY ANALYSIS

Evaluation protocol. We re-fit the ablation of Table 9 on the synthetic setting with $K = 20$ over 20 seeds. We reuse all other settings from the earlier ablation, except that the blocked-time variant selects its weight by held-out NLL and the static variant uses no TV penalty.

Field accuracy. For each held-out history, we compare $\widehat{C}_\theta(t, \widehat{z}_t)$ with $C^\star(t, z_t^\star)$ on the same 200-point time grid. Each state is computed by its own encoder from the same observed events, so no alignment of state coordinates is performed. We report field RMSE separately for true edges and non-edges, pre-link background RMSE, and temporal correlation on the 30% of true edges with varying coefficients.

The seed-paired increases in true-edge RMSE relative to the full model are 0.015 for the time-only variant (95% bootstrap CI: $[0.010, 0.020]$) and 0.009 for the quantized-state variant ($[0.006, 0.013]$). These intervals are computed over the 20 seeds. The corresponding temporal-correlation intervals include zero, so the clearest improvement is in coefficient accuracy. Recovery remains imperfect: the full model has a relative true-edge RMSE of 0.406 and a mean signed offset of -0.060.

Likelihood stability. We re-evaluate the same fitted models using independent 256-point Monte Carlo integration, shared Monte Carlo points across variants, and a deterministic 16,384-point midpoint rule as shown in Table 22. Shared-point evaluations use 256, 1024, and 4096 points, with eight, four, and two repetitions, respectively, and the independent 256-point evaluation also uses eight repetitions. We report mean paired differences $\Delta\mathrm{NLL} = \mathrm{NLL}_{\mathrm{variant}} - \mathrm{NLL}_{\mathrm{full}}$ over the 100 held-out sequences.

The time-only minus full difference ranges from 4.22 to 4.68 nats per sequence across the evaluated schemes, compared with 4.47 under deterministic quadrature. The ordering of the variants is unchanged, supporting the numerical stability of the synthetic likelihood comparison.

Table 17: Alarm read-out study without the device topology, 20 seeds, all read-outs computed from one fit per seed (CASCADE is deterministic). Denied the topology, CASCADE does not rank the candidate pairs at all (AUC 0.457), so every read-out dominates it on every seed. The read-out choice trades ranking against thresholded accuracy, and mean-C – the read-out the main table reports – is best on neither axis.

Read-out	AUC	F_1 (mean, min)	Precision	SHD	$F_1 >$ CASCADE
rank-sum(C, delay)	0.861	**0.663 ± 0.010 (0.648)**	0.648	98	**20/20**
$C\times$ delay	0.853	0.661 ± 0.011 (0.642)	0.654	98	**20/20**
mean-C (model only)	0.877	0.624 ± 0.014 (0.594)	0.643	105	**20/20**
per-row-normalized mean-C	**0.895**	0.580 ± 0.011 (0.553)	0.560	119	**20/20**
delay-only (data statistic)	0.848	0.634	**0.712**	**95**	—
perturbation $\Delta\lambda$ (do-drop)	0.757	0.303	0.249	218	—
CASCADE, no topology (fixed)	0.457	0.287	0.658	126	—

Table 18: Held-out NLL per event on alarm (4-fold device CV). Δ is the paired difference $\mathrm{NLL}_{\text{variant}} - \mathrm{NLL}_{\text{full}}$ with a 95% bootstrap CI over folds, so positive values favor the full model. All CIs exclude zero, and the gain over time-only survives capacity matching.

Variant	NLL/event	Paired Δ vs. full	95% CI
full $C(t, z)$	**4.527**	—	—
time-only $C(t)$	4.581	$+0.054$	$[+0.046, +0.066]$
time-only, capacity-matched	4.602	$+0.076$	$[+0.061, +0.091]$
static C	4.668	$+0.141$	$[+0.128, +0.155]$
constant background μ_j	5.966	$+1.439$	$[+1.187, +1.752]$

C.3 Time and Memory Complexity

Tables 23 and 24 summarize computational scalability on the synthetic dataset. CausalField completes all configurations up to $K = 200$, with runtime increasing from 24 seconds at $K = 20$ to 491 seconds at $K = 200$. In contrast, several baselines become computationally impractical at larger K: SHP repeatedly times out from $K = 100$, TTPM times out at larger dimensions, and CAUSE reaches 85 GB of host memory at $K = 150$ before running out of memory at $K = 200$. We emphasize that MaxRSS measures host RAM rather than GPU VRAM; GPU memory consumption is evaluated separately under a controlled benchmark.

D Limitations

The non-negativity constraint $C \geq 0$ fixes the sign of every interaction, not its magnitude. On an inhibitory generator this does not destroy the signal — the true suppressive edges are driven to the $C{=}0$ boundary while null pairs take small positive values, so they are recovered almost perfectly under a sign-flipped read-out (AUC 0.999, F_1 0.990) and not at all under the standard one (0.001, 0.000; 20 seeds). What the model cannot do is tell an inhibitory edge from no edge without knowing the sign convention in advance, and a single flip cannot rescue a graph with both signs.

Recovery after a structural change is fast but not instantaneous. The accumulated read-out continues to reflect the pre-change support for ≈ 3.5 time units, so the graph estimate is transiently degraded near a regime boundary as illustrated in Figure 6. The continuous field shortens this window by $\approx 5\times$ relative to fixed-Δ segmentation (floor $\Delta/2 = 25$) but does not remove it, which is a real operational cost in low-data or rapidly switching regimes.

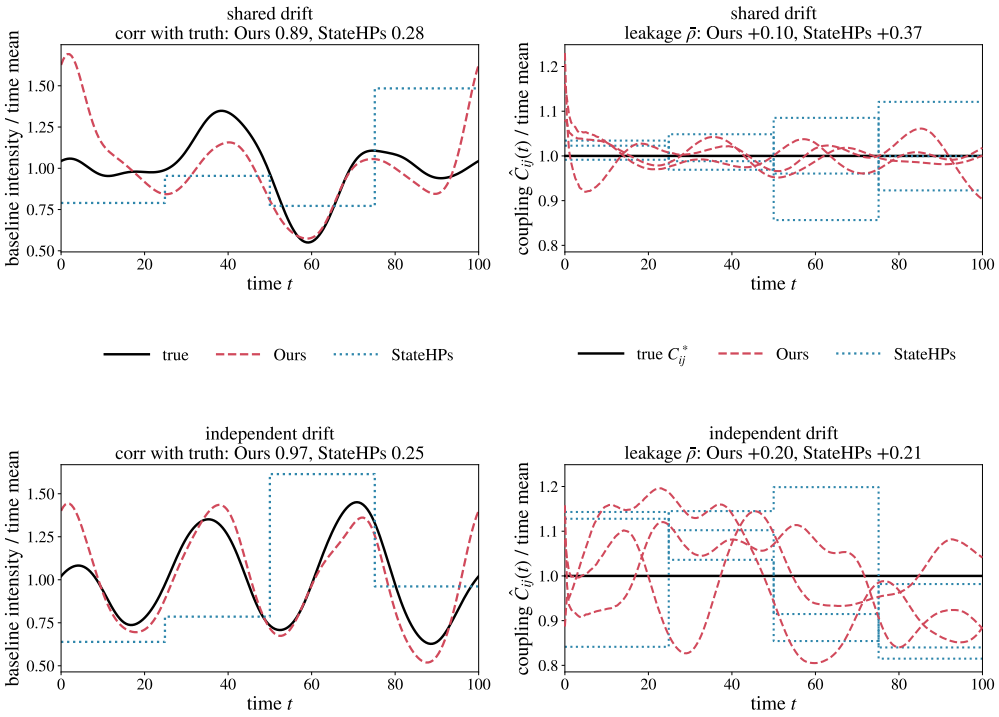

Figure 5: Constant true coupling under shared (top) and independent (bottom) baseline drift across 20 seeds. Curves are normalized by their own time means: baseline intensity (left) and coupling on the three strongest edges (right; truth $= 1$). Titles report baseline correlation and leakage $\bar{\rho}$. CAUSALFIELD tracks drift more accurately, with lower leakage under shared drift and comparable leakage under independent drift.

Table 19: Encoder decision experiment. The slow encoder yields comparable AUC and higher alarm F1 than the attention encoder in these experiments.

Data	Encoder	τ_z	AUC	Held-out F1	$\tau_z\hat{\beta}$
Synthetic	**`slow`**	20	**0.957 ± 0.015**	—	17.22
Synthetic	`attn`	20	0.950 ± 0.016	—	17.31
Alarm	**`slow`**	**20**	**0.877 ± 0.002**	**0.625 ± 0.014**	**18.99 (pass)**
Alarm	`slow`	5	0.873 ± 0.001	0.610 ± 0.013	4.59 (fires)
Alarm	`attn`	20	0.878 ± 0.004	0.588 ± 0.021	19.80
Alarm	`attn`	5	0.878 ± 0.004	0.588 ± 0.021	4.95 (fires)

Table 20: Residual excitation variation within time–state neighborhoods. Between-replicate minimum eigenvalues are reported only when the number of sequences exceeds the number of active coordinates in the synthetic setting of Table 1.

M	obs/null $\lambda_{\min}$	obs/null trace	full-rank neighborhoods	$\lambda_{\min}$ between replicates
12	0.24 ± 0.17	0.78	0	—
48	0.45 ± 0.15	0.84	27	5.0×10^{-3}
96	0.55 ± 0.11	0.82	37	8.2×10^{-3}

Table 21: Field and background errors in the re-fitted synthetic ablation. Values are mean ± standard deviation over 20 seeds, with each seed averaged over five held-out sequences. Temporal correlation is computed on the gated edges and is undefined for the static field.

Variant	Field RMSE ↓		Background RMSE ↓	Temporal corr. ↑
	True edges	Non-edges		
Full $C(t, z)$	0.169 ± 0.017	0.059 ± 0.011	0.626 ± 0.055	0.584 ± 0.158
Time-only $C(t)$	0.183 ± 0.020	0.077 ± 0.014	0.714 ± 0.060	0.534 ± 0.121
Quantized, $K_z = 4$	0.178 ± 0.019	0.074 ± 0.011	0.694 ± 0.062	0.599 ± 0.111
Blocked-time, $\Delta = 25$	0.257 ± 0.019	0.111 ± 0.009	1.481 ± 0.090	0.331 ± 0.167
Static C	0.184 ± 0.027	0.182 ± 0.004	2.282 ± 0.553	—

Table 22: Sensitivity of the synthetic NLL comparison to compensator integration. Entries are mean paired NLL differences in nats per sequence. Monte Carlo entries are averaged over repeated evaluations.

Variant	Independent	Shared Monte Carlo			Midpoint
	256	256	1,024	4,096	16,384
Time-only	4.47	4.68	4.22	4.45	4.47
Quantized, $K_z = 4$	3.51	3.69	3.56	3.70	3.64
Blocked-time	82.10	83.14	81.83	82.22	82.29
Static	88.60	86.96	89.33	88.54	89.06

Table 23: Runtime scalability on the self-consistent synthetic setting. We report the fitting time per seed. TO and OOM denote observed timeout and out-of-memory failures, respectively.

K	**Ours**	NGC	NPHC	L0HAWKES	THP	SHP	CAUSE	TTPM (δ=0.1)	TTPM (δ=1.0+cap)
20	24 s	–	29 s	117 s	841 s	–	–	∼300 s	32 s
50	84 s	32 s	29 s	117 s	522 s	983 s	11 s	∼6.6 h	–
100	164 s	52 s	29 s	117 s	572 s	TO (×19)	5,834 s	TO	7,564 s
150	323 s	148 s	29 s	117 s	639 s	TO (×20)	11,967 s	TO	TO
200	491 s	148 s	29 s	117 s	736 s	TO (×19)	OOM	TO	TO

Table 24: Host-memory usage on the self-consistent synthetic setting. We report the median MaxRSS recorded by `sacct`.

K	**Ours**	THP	CAUSE	TTPM (δ=0.1)	TTPM (δ=1.0+cap)
20	–	1.5 GB	–	0.4 GB	0.4 GB
50	–	1.6 GB	–	0.5 GB	–
100	0.7 GB	1.7 GB	–	1.5 GB	1.2 GB
150	1.4 GB	1.9 GB	85 GB	4.9 GB	4.3 GB
200	1.7 GB	2.3 GB	OOM	13.5 GB	12.5 GB

Table 25: Background drift without structural change. F1 performance of the continuous field and two segmented estimators as the baseline variation amplitude a_μ increases while the true coupling remains fixed. The segmented models convert background drift into spurious edges, whereas the continuous field remains robust.

Baseline drift	CAUSALFIELD	seg-ADM4	blocked_t
none (stationary)	**1.000**	0.933	0.933
moderate ($a_\mu = 1$)	0.968	0.968	0.882
strong ($a_\mu = 2$)	**1.000**	0.788 (FDR 0.20)	0.812

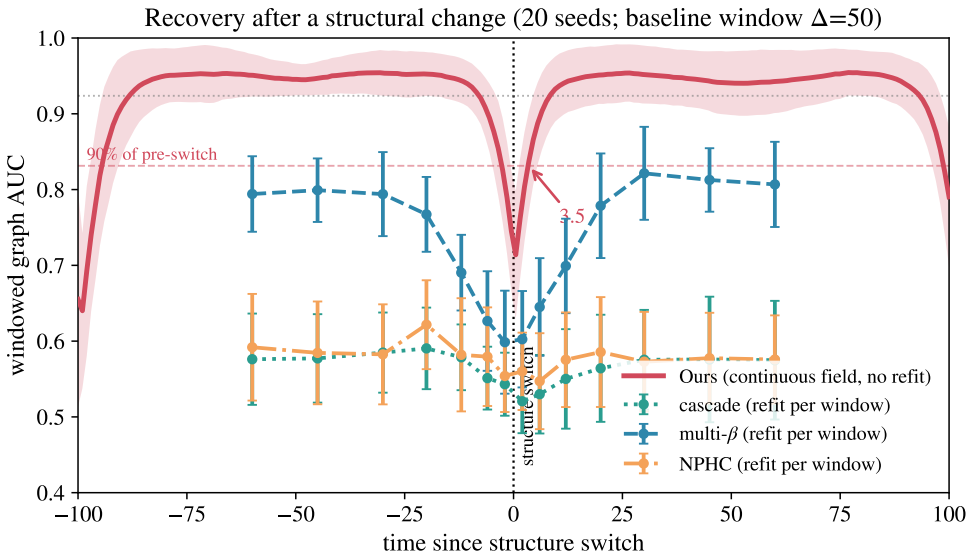

Figure 6: Performance dynamics around change points. While performance temporarily drops immediately following a change point, it recovers rapidly as a few samples accumulate.

E PSEUDOCODE AND COMPLEXITY

Per iteration, cost is linear in the number of events: the excitation and state statistics are maintained recursively, avoiding repeated summation over the full event history, as summarized in Algorithm 1.

Algorithm 1 CAUSALFIELD: Estimation

Require: sequences $\{D^{(m)}\}_{m=1}^{M}$ on $[0,T]$; fixed τ_z; evaluation grid $\{g_p\}_{p=1}^{P}$; $\lambda_1, \lambda_{\mathrm{TV}}$; MC sample count n_{mc}

Ensure: fitted parameters $\widehat{\theta}$ and influence field $\widehat{C}_\theta$

initialize $\theta = (\mathrm{emb}, \theta_C, \theta_\mu, \beta)$

repeat

  Select a batch $\mathcal{B}$ of complete sequences ▷ sequence is the sampling unit

  for all $m \in \mathcal{B}$ **do**

    initialize $N_0^{(m)} \leftarrow \mathbf{0}$, $W_0^{(m)} \leftarrow 0$, and $S^{(m)}(0^+) \leftarrow \mathbf{0}$

    $\mathrm{NLL}_m \leftarrow 0$

    for $k = 1, \ldots, |D^{(m)}|$ **do**

      $\Delta_k \leftarrow t_k^{(m)} - t_{k-1}^{(m)}$ ▷ $t_0^{(m)} = 0$

      decay the history statistics to t_k^-:

$$N_k^- \leftarrow e^{-\Delta_k/\tau_z} N_{k-1}^+, \quad W_k^- \leftarrow e^{-\Delta_k/\tau_z} W_{k-1}^+, \quad S^{(m)}(t_k^-) \leftarrow e^{-\beta\Delta_k} S^{(m)}(t_{k-1}^+)$$

      compute

$$z_k^- \leftarrow \begin{cases} N_k^-/W_k^-, & W_k^- > 0, \\ z_0, & \text{otherwise}, \end{cases} \qquad \widehat{C}_k \leftarrow \widehat{C}_\theta(t_k^{(m)}, z_k^-)$$

      compute $\lambda_j^{(m)}(t_k^{(m)})$ from Eq. 1 using $(z_k^-, \widehat{C}_k, S^{(m)}(t_k^-))$

      $\mathrm{NLL}_m \leftarrow \mathrm{NLL}_m - \log \lambda_{e_k^{(m)}}^{(m)}(t_k^{(m)})$

      update after observing $e_k^{(m)}$:

$$S_i^{(m)}(t_k^+) \leftarrow S_i^{(m)}(t_k^-) + \mathbb{I}[e_k^{(m)} = i],$$

$$N_k^+ \leftarrow N_k^- + \mathrm{emb}(e_k^{(m)}), \qquad W_k^+ \leftarrow W_k^- + 1$$

    end for

    draw $u_q^{(m)} \sim \mathrm{Unif}[0,T]$, $q = 1, \ldots, n_{\mathrm{mc}}$, and estimate the compensator using the same history statistics:

$$\mathrm{NLL}_m \leftarrow \mathrm{NLL}_m + \frac{T}{n_{\mathrm{mc}}} \sum_{q=1}^{n_{\mathrm{mc}}} \sum_{j=1}^{K} \lambda_j^{(m)}(u_q^{(m)})$$

  end for

  compute

$$\bar{C}_\theta \leftarrow \frac{1}{P} \sum_{p=1}^{P} \widehat{C}_\theta(g_p)$$

  compute

$$\mathcal{P} \leftarrow \lambda_1 \|\bar{C}_\theta\|_1 + \lambda_{\mathrm{TV}} \sum_{p=1}^{P-1} h_\delta\left(\widehat{C}_\theta(g_{p+1}) - \widehat{C}_\theta(g_p)\right)$$

  update θ using gradients of

$$\sum_{m \in \mathcal{B}} \mathrm{NLL}_m + \mathcal{P}$$

until the prescribed training budget is reached ▷ no ground-truth structure metric is consulted

return $\widehat{\theta}$ and $\widehat{C}_\theta$