From Jumps to Signatures: A Generative Method for Temporal Point Processes

Niels Cariou-Kotlarek[♠,★] and **Vasileios Lampos**[♠]

[♠] Centre for Artificial Intelligence, Computer Science, UCL
[★] Department of Mathematics, UCL
{niels.kotlarek.23, v.lampos}@ucl.ac.uk

Abstract

Rough path signatures are a universal feature map for continuous paths and, via the expected signature, characterise path distributions. These guarantees do not directly extend to càdlàg paths of Temporal Point Processes (TPPs), limiting the use of signature methods for event sequences. Furthermore, neural TPP models, including recent generative approaches, optimise per-event objectives with no global sequence-level loss, while evaluation of variable-length event sequences lacks distributional discrepancy measures. This paper proposes a common pathwise framework for addressing these limitations. We introduce the interarrival embedding, a stable, injective lift from jump paths to continuous paths of bounded variation, extending signature methods to discrete event sequences. Our theoretical contributions give rise to SIGTPP, the first signature-based generative model for TPPs, trained using a path-level loss on complete trajectories. We further analyse the space of counting paths and derive 3 distributional discrepancies, providing mathematically justified tools for evaluating generative TPP models. Across 12 tasks, SIGTPP achieves the best mean rank based on 8 complementary metrics, outperforms or is within one standard error of the strongest baseline (which varies based on performance) in 57% of the dataset-metric pairs, and according to a relative score, improves against each baseline by at least 24% on average.

Source code and datasets: github.com/Code-Cornelius/sigtpp

1 Introduction

Temporal Point Processes (TPPs) model discrete events occurring in continuous time [12, 24]. They underpin applications across finance [3], seismology [14], information diffusion [25], and health [20, 53]. Neural TPPs (NTPPs) parametrise the conditional or interarrival density with expressive neural networks [16, 39, 44, 58], including continuous-time state-space architectures [5], and are often trained via log-likelihood maximisation. Yet, likelihood-based training introduces fundamental constraints [51]. It either restricts models to tractable likelihood families [50] or relies on expensive numerical approximations [39, 44], limiting both expressiveness and scalability. This has led to a growing interest in alternative generative paradigms, such as diffusion-based methods [35, 57], normalising flows [50], adversarial training [55], and reinforcement learning [30], with frameworks [32] unifying several of these perspectives. At the same time, the signature transform from rough path theory [19, 38] has become central to time series modelling [6, 11, 28, 45]. Acting as a universal feature map, it encodes pathwise information into a tensor series up to tree-like equivalence [23, 38], while the expected signature becomes a moment generating function on path space [7, 8], supporting generative models such as the Signature-Wasserstein GAN [31, 41].

Despite this activity, the current landscape has significant structural gaps. The prevailing training objectives are local as they decompose into a sum of per-event losses, each comparing one event to the

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

ground truth conditional on preceding events. This is also true for recent non-autoregressive models as they fit reverse denoising transitions [34], transport fields [27], or edit rates at sampled intermediate states [36], rather than directly minimising a single global discrepancy between generated and empirical trajectories. Consequently, global structural properties, such as autocorrelation, long-range inter-event dependencies, and joint timing distributions, receive no direct loss signal.

Furthermore, the evaluation methods for generative TPP models are underdeveloped [40, 51]. Negative log-likelihood (NLL) measures local conditional fit and is insensitive to global sequence structure. Low NLL does not preclude mode collapse, misspecified tails, or distorted inter-event dependencies [32, 33, 40]. One-step-ahead pointwise metrics (e.g. mean absolute error) share this limitation and additionally favour underdispersed predictions [42, 43, 59–61]. A practical shift towards sequence-level distributional losses, such as the maximum mean discrepancy (MMD) with a Gaussian kernel [35, 50] and the Wasserstein-1 distance [34, 36], has emerged, yet whether these losses define meaningful distances on the space of counting paths has not been theoretically established. Signatures would be a suitable candidate to address the aforementioned shortcomings, as they can naturally handle variable-length sequences. However, the determinacy and universal nonlinearity that underpin signature-based generative modelling [7, 8, 23] are formulated for continuous paths of bounded variation. TPP sample paths are càdlàg step functions, on which these guarantees fail.

This paper attempts to address these limitations with the following contributions:

1. We introduce the interarrival embedding, a stable, injective lift to continuous paths of bounded variation, enabling signature methods on jump processes (Section 4.1).
2. Building on this embedding approach, we propose SigTPP, a signature-based generative model for TPPs that replaces per-event objectives with a global signature loss on counting paths (Section 4.2).
3. We provide 3 theoretically grounded discrepancies on counting-path distributions, namely the energy ($\mathcal{E}$), Wasserstein-1 (W_1), and signature-Wasserstein-1 distances ($\mathrm{Sig\text{-}W}_1$), for evaluating generative TPP models (Section 4.4). We also showcase the utility of other diagnostic metrics for NTPP evaluation, including moment and histogram-based summaries.
4. We demonstrate empirically that SigTPP is a strong generative model for TPPs. It achieves the best relative score on 5 out of the 8 metrics. SigTPP achieves the best performance or is within one standard error of the strongest baseline in 57% of all metric-dataset pairs (Section 5).

2 Related Work

Likelihood-based NTPPs [5, 16, 39, 58] compute the optimisation target as the product of the conditional densities and the survival probability of the final interval [12]. Exact evaluation requires tractable densities [44, 49, 50]. Otherwise, models integrate the intensity numerically [39]. Beyond likelihood, many autoregressive models train on the next interarrival time [30, 32, 55]. Our proposed method (SigTPP) is also autoregressive but compares the expected signatures of complete sequences. This provides a sample-based training objective without likelihood evaluation or a learned critic. On the other hand, non-autoregressive models generate whole sequences through intermediate states [34–36]. SigTPP retains autoregressive sampling and trains directly on complete generated paths, without fitting an auxiliary denoising, editing, or transport process. In parallel, signature objectives have been used to train generators of continuous paths. Signature-Wasserstein models match the expected signatures [31, 41], and signature-kernel scores train neural stochastic differential equations [26]. Applying signature objectives to event sequences requires a suitable representation. The Marcus signature extends universality to continuous functions of càdlàg paths by interpolating jumps of the counting path [11], while our proposed interarrival embedding records waiting times as the values of a continuous path. With time augmentation, it also combines event times and interarrival durations, the temporal quantities used in neural TPPs [16, 32].

3 Preliminaries

Temporal Point Processes [12]**.** Fix a finite horizon $T_{\max} \in \mathbb{R}_{>0}$. A realisation of a TPP on $[0, T_{\max})$ is a finite sequence of strictly increasing event times, $0 < t_1 < \cdots < t_m < T_{\max}$. To define a metric on event sequences and study its topological properties, we work with the equivalent counting-path representation. Let $\mathcal{N}$ denote the set of nondecreasing, càdlàg functions $\eta\colon [0, T_{\max}] \to \mathbb{N}$ with $\eta_0 = 0$, $\eta_{T_{\max}} < \infty$, unit jumps only, and no jump at $T_{\max}$. η and $(t_k)_{k=1}^m$ are related by

$\eta_t = \#\{k : t_k \le t\}$. A TPP is an $\mathcal{N}$-valued random variable N on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The conditional intensity $\lambda(t \mid \mathcal{H}_t) := \lim_{\Delta t \downarrow 0} \Delta t^{-1} \mathbb{E}\big[N_{t+\Delta t} - N_t \mid \mathcal{H}_t\big]$, where $\mathcal{H}_t := \sigma(N_s : s < t)$, characterises the instantaneous rate of events given history.

We adopt the following boundary convention throughout, since the embedding and its stability analysis use the same augmented event grid.

Convention 1 (Boundary Convention)**.** *For $\eta \in \mathcal{N}$ with event times $0 < t_1 < \cdots < t_m < T_{\max}$, we augment the sequence by setting $t_0 := 0$ and $t_{m+1} := T_{\max}$, and define the corresponding interarrival times $\tau_0 := 0$, $\tau_k := t_k - t_{k-1}$ where $k = 1, \ldots, m$, and $\tau_{m+1} := T_{\max} - t_m$.*

Rough Path Signatures [19, 38]**.** The signature transform provides an injective representation, up to tree-like equivalence, of continuous paths of finite variation into the tensor algebra, $T((\mathbb{R}^d))$. We restrict attention to continuous paths of bounded variation (as opposed to finite p-variation).

Definition 2 (Signature)**.** *Let $X : [0, T_{\max}] \to \mathbb{R}^d$ be a continuous path of bounded variation. The signature of X, denoted by $S(X)$, is defined as the infinite sequence of iterated integrals*

$$S(X) = (1, S(X)^1, S(X)^2, \ldots), \quad \text{where} \quad S(X)^k = \int_{0 < u_1 < \cdots < u_k < T_{\max}} \mathrm{d}X_{u_1} \otimes \cdots \otimes \mathrm{d}X_{u_k}. \quad (1)$$

$S^M(X)$ denotes the truncated signature of X of degree M, i.e. $S^M(X) = (1, S(X)^1, \ldots, S(X)^M)$.

Theorem 3 (Determinacy [7])**.** *Let X and Y be continuous random paths of bounded variation. If $\mathbb{E}[S(X)] = \mathbb{E}[S(Y)]$ and $\mathbb{E}[S(X)]$ has infinite radius of convergence, then $S(X)$ and $S(Y)$ have the same distribution.*

4 Signature-based Generative TPPs (SigTPP)

The signature transform is defined on continuous paths of bounded variation, whereas counting paths in $\mathcal{N}$ are càdlàg step functions. Our approach to addressing this rests on two components. First, we introduce a deterministic embedding Φ into the space $\mathcal{C} := C([0, T_{\max}], \mathbb{R})$ of continuous functions on $[0, T_{\max}]$. Φ maps a discrete event sequence to a continuous, piecewise-linear path on which the signature transform applies (Section 4.1). Then, we train the proposed SigTPP on the Euclidean distance between the truncated expected signatures of the embedded model-generated and observed counting paths (Section 4.2). We also provide theoretical guarantees ensuring the signature loss reflects genuine discrepancies between counting-path distributions (Section 4.3), and we propose principled evaluation metrics for generative TPPs (Section 4.4).

4.1 The Interarrival Embedding Φ

Definition 4 (Interarrival Embedding)**.** *For $\eta \in \mathcal{N}$ with m events, let $0 =: t_0 < t_1 < \cdots < t_m < t_{m+1} := T_{\max}$ be the augmented grid and $\tau_k := t_k - t_{k-1}$ the interarrival times. The interarrival embedding is the continuous, bounded-variation, piecewise-linear function $x := \Phi(\eta) \in \mathcal{C}$ satisfying*

$$\begin{aligned} &x(t_k) = \tau_k \quad \text{for } k = 0, \ldots, m+1, \text{ and} \\ &\Phi(\eta)_t = \tau_k + \frac{t - t_k}{t_{k+1} - t_k}\big(\tau_{k+1} - \tau_k\big), \quad \text{where } t \in [t_k, t_{k+1}],\ k = 0, \ldots, m. \end{aligned} \quad (2)$$

Equivalently, $\Phi(\eta)$ is the unique continuous function that is affine on every interval and satisfies $\Phi(\eta)_{t_k} = \tau_k$, $k \le m+1$. Figure 1 illustrates this on a Poisson sample: the counting process (blue, càdlàg) records cumulative event counts, the embedding points (orange) record the duration of the most recent interarrival interval, and $\Phi(\eta)$ (green dashed line) is the piecewise-linear interpolant of this staircase on the augmented event grid. The embedded path is continuous and of bounded variation, satisfying the assumptions required by the signature transform (Definition 2). This piecewise-linear lift is the standard parameter-free construction for signatures [6]. It produces a continuous path of bounded variation, as required by the Hambly-Lyons uniqueness theorem and the Chevyrev-Lyons expected-signature determinacy result [8, 23]. We also

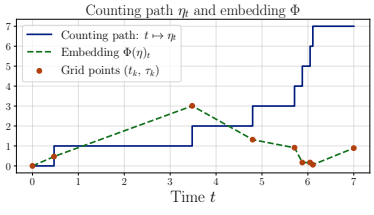


Figure 1: Illustration of the interarrival embedding Φ (Definition 4) for a sampled Poisson process ($\lambda = 1$, $T_{\max} = 7$). Orange dots show the duration of the most recent interarrival interval.

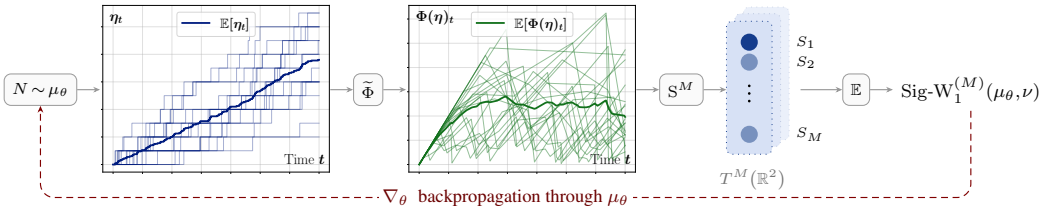


Figure 2: SIGTPP's optimisation. A counting path sampled from μ_θ is embedded with $\widetilde{\Phi}$ to a continuous path, mapped to its truncated signature in $T^M(\mathbb{R}^2)$, and summarised by its expectation under the model. The loss (Sig-W_1) compares this against the expected signature of the embedded targets' paths, opposed to the usual per-event loss. Loss gradient flows back through μ_θ.

prove that its affine structure yields the global Lipschitz bound (Theorem 6). In practice, we apply the signature transform to the time-augmented path $t \mapsto \widetilde{\Phi}(\eta)_t := \big(t, \Phi(\eta)_t\big)$ in $\mathbb{R}^2$. This augmentation is necessary because, for a scalar path X, every signature level satisfies $S(X)^k = (X_{T_{\max}} - X_0)^k / k!$. The signature of $\Phi(\eta)$ alone therefore depends only on its total increment and cannot distinguish event sequences. By contrast, the strictly increasing time coordinate makes $\widetilde{\Phi}(\eta)$ unique in its class and hence identifiable from its signature [23]. The resulting two-channel path combines absolute event times and interarrival durations, the standard temporal inputs to neural TPPs [16, 32]. These quantities already determine the counting path, and injectivity of Φ ensures that no event information is lost. The stability of the map Φ and a constructive inverse $\Psi := \Phi^{-1}$ are established in Theorem 6.

4.2 Model Optimisation

Our goal is to train a parametric generator μ_θ whose output distribution faithfully reproduces the unknown data distribution ν, without specifying a parametric intensity model. A natural loss between distributions μ_θ and ν is the Wasserstein-1 distance used in Wasserstein generative adversarial network (GAN) models [55], expressible via Kantorovich-Rubinstein duality as

$$\mathrm{W}_1(\mu_\theta, \nu) = \sup_{\|f\|_{\mathrm{Lip}} \leq 1} \mathbb{E}_{x \sim \mu_\theta}\big[f(x)\big] - \mathbb{E}_{y \sim \nu}\big[f(y)\big]. \tag{3}$$

Approximating the supremum over 1-Lipschitz functions often requires training a neural network discriminator with gradient penalties or weight clipping. This adversarial procedure can lead to training instability and convergence issues [2], especially when modelling TPPs [55]. Recent work on generative modelling of continuous paths proposes replacing the adversarial discriminator with a signature-based metric, denoted by Sig-W_1 [31, 41]. By leveraging the universal nonlinearity signature property, any 1-Lipschitz test function can be approximated by a linear functional on the signature. Consequently, the optimisation over the infinite-dimensional function space reduces to a linear operation on the (truncated for numerical purposes) tensor algebra $T^M(\mathbb{R}^d)$. Given that TPP sample paths are càdlàg step functions, we apply the signature loss to their time-augmented embeddings $\widetilde{\Phi}(\eta)$. Hence, the loss for optimising the proposed model, SIGTPP, is given by:

$$\text{Sig-}\mathrm{W}_1^{(M)}(\mu_\theta, \nu) := \Big\| \mathbb{E}_{N_\theta \sim \mu_\theta}\big[S^M(\widetilde{\Phi}(N_\theta))\big] - \mathbb{E}_{N \sim \nu}\big[S^M(\widetilde{\Phi}(N))\big] \Big\|_2. \tag{4}$$

Here $\|\cdot\|_2$ denotes the ℓ^2 norm on the truncated tensor algebra $T^M(\mathbb{R}^2)$. The training procedure is summarised in Figure 2. The expected signature under the data measure ν is precomputed once on the training set. Each gradient step requires only computing signatures under μ_θ.

SIGTPP is a 1-step autoregressive generator that follows the standard NTPP design [32]. A single-layer LSTM encodes the history. Each log-interarrival time $\log \tau_i$ is first mapped to a learnable sinusoidal positional encoding, and processed recurrently to produce the history state. The decoder is a shallow perceptron that predicts $\log \tau_{i+1}$ from the recurrent state h_i, the current cumulative time, and an exogenous noise variable $\varepsilon \sim \log(\mathrm{Exp}(1))$. The core temporal training procedure is given in Algorithm 2 of Appendix C; it repeatedly draws free-running trajectories from the generator via the sampling routine of Algorithm 1. Appendix G gives further implementation and training details.

4.3 Theoretical Guarantees

To analyse the stability of the embedding Φ and to quantify discrepancies between counting-path distributions, we propose to work with the following metric.

Definition 5 (Counting Path Metric)**.** *We define the pathwise TPP metric between two counting processes $\eta,\xi\in\mathcal{N}$ on $[0,T_{\max})$ by $d_{\mathcal{N}}(\eta,\xi):=\int_0^{T_{\max}}|\eta_t-\xi_t|\,\mathrm{d}t$.*

Given that counting paths are càdlàg, distinct event sets $\eta\neq\xi$ yield paths that differ on an interval of positive Lebesgue measure. Hence, $d_{\mathcal{N}}(\eta,\xi)=0$ implies $\eta=\xi$, and $d_{\mathcal{N}}$ is a genuine metric on $\mathcal{N}$ (see also Proposition S3). Appendix A.3 identifies $d_{\mathcal{N}}$ with Wasserstein-1 between event-time point measures padded at $T_{\max}$. This result puts the optimal transport perspective of Xiao et al. [55] on a rigorous footing, yielding a genuine metric on $\mathcal{N}$. The space $\mathcal{N}$ is not a linear subspace of $L^1([0,T_{\max}])$. The difference $\eta_t-\xi_t$ is signed, scalar multiples of counting paths are not counting paths, and the sum or difference of two counting paths need not have unit jumps. The metric structure is the natural level at which to study stability. Consequently, standard tools from linear and functional analysis are not directly available. Most notably, there is no differentiable structure on which to apply an inverse function theorem, no open mapping theorem to promote a continuous bijection to a bi-Lipschitz one, and no compact embedding to fall back on since event counts are unbounded. We therefore work throughout with L^1-type metrics on both $\mathcal{N}$ and $\mathcal{C}$, which is also natural for signature-based losses that are computed through integrals.

Hence, the embedding chain from counting paths to the tensor algebra is given by:

$$(\mathcal{N},d_{\mathcal{N}})\xrightarrow{\Phi}(\mathcal{C},d_1)\xrightarrow{\mathrm{t}\mapsto(\mathrm{t},\Phi(\eta)_{\mathrm{t}})}C([0,T_{\max}],\mathbb{R}^2)\xrightarrow{\text{Signature}}T((\mathbb{R}^2)), \tag{5}$$

where $\mathcal{N}$ is equipped with a metric $d_{\mathcal{N}}$ and $(\mathcal{C},d_1)$ denotes the space of continuous paths $[0,T_{\max}]\to\mathbb{R}$ equipped with the L^1 metric $d_1(x,y):=\int_0^{T_{\max}}|x(t)-y(t)|\,\mathrm{d}t$. Theorem 6 establishes that Φ is Lipschitz and injective, with a controlled inverse under mild assumptions. The Lipschitz bound ensures that the signature loss (Eq. 4) inherits regularity from the data distribution. Closeness in counting-path space implies closeness of expected signatures. Injectivity guarantees that no event information is lost in the embedding, which is what allows Sig-$\mathrm{W}_1^{(M)}$ to separate distinct laws on $\mathcal{N}$. Finally, the Hölder bound gives the converse direction under a mild separation assumption that is satisfied by any temporal data recorded at finite precision.

Theorem 6 (Stability of the Embedding, full proofs in Appendix B)**.** *The embedding $\Phi\colon(\mathcal{N},d_{\mathcal{N}})\to(\mathcal{C},d_1)$ satisfies the following properties:*

1. ***Lipschitz continuity.*** *For all $\eta,\xi\in\mathcal{N}$, $d_1\big(\Phi(\eta),\Phi(\xi)\big)\le 5\cdot T_{\max}\cdot d_{\mathcal{N}}(\eta,\xi)$. Hence, it is continuous and measurable* (Appendix B.1, Theorem S11)*. The constant 5 is not tight but sufficient.*
2. ***Bijection.*** *Φ is injective on $\mathcal{N}$, hence a bijection $\Phi\colon\mathcal{N}\xrightarrow{\sim}\Phi(\mathcal{N})$. The inverse $\Psi:=\Phi^{-1}$ is recovered constructively by a level-set recursion* (Appendix B.2, Theorem S16)*.*

Moreover, on counting paths whose interarrival times $\tau_1,\dots,\tau_{m+1}$ (Convention 1) are bounded below by any fixed $\delta>0$, the inverse Ψ is Hölder continuous of exponent $1/2$ (see also Appendix B.3, Theorem S29)*.*

4.4 Evaluation Metrics for Counting-Path Distributions

Evaluating generative TPPs requires discrepancy measures between distributions on $\mathcal{N}$. Current practice borrows tools from continuous data settings, e.g. MMD with Gaussian kernels [35, 50] or W_1 on event times [34, 36], without verifying that these are valid distances on counting-path space. We show that the metric $d_{\mathcal{N}}$ gives rise to mathematically valid distributional discrepancies on $\mathcal{N}$.

Throughout this section, uppercase $P,Q\in\mathcal{P}_1(\mathcal{N})$ denote probability distributions over counting paths $\mathcal{N}$ with finite first $d_{\mathcal{N}}$-moment. The energy ($\mathcal{E}$) and Wasserstein-1 (W_1) distances use $d_{\mathcal{N}}$ as ground cost. The signature discrepancy Sig-$\mathrm{W}_1^{(M)}$ compares expected features of the embedded paths.

Energy distance ($\mathcal{E}$). The energy distance ($\mathcal{E}$) is a standard metric induced by an underlying distance function [21, 52] (Appendix D). Nonnegativity of $\mathcal{E}$ requires that $(\mathcal{N},d_{\mathcal{N}})$ be of negative type [52]. The map $\eta\mapsto(t\mapsto\eta_t)$ is an isometry from $(\mathcal{N},d_{\mathcal{N}})$ into $L^1([0,T_{\max}])$, which is of negative type, and the property is inherited by subspaces (Appendix A.1). Hence $D:=\sqrt{\mathcal{E}}$ is a well-defined discrepancy. We further show in Appendix A.4 that $d_{\mathcal{N}}$ is not of strong negative type, so D is a pseudometric rather than a metric. For the purpose of ranking models, this is sufficient.

Wasserstein-1 distance (W_1). Since $(\mathcal{N}, d_{\mathcal{N}})$ is a metric space, W_1 [47, 54] is a well-defined metric on $\mathcal{P}_1(\mathcal{N})$. We recall its definition in Appendix D.

Signature discrepancy (Sig-W_1). The loss $\mathrm{Sig}\text{-}\mathrm{W}_1^{(M)}$ (Eq. 4) can also be used as an evaluation metric in the continuous setting and we adapt it to TPPs with our embedding. Since Φ is injective and measurable (Theorem 6), its time augmentation $\widetilde{\Phi}$ is also injective and measurable. Thus, distinct laws $P \neq Q$ on $\mathcal{N}$ induce distinct pushforwards $\widetilde{\Phi}_\sharp P \neq \widetilde{\Phi}_\sharp Q$. If these pushforward laws satisfy the moment condition of Theorem 3, then the expected signature determines the law, and hence $\mathbb{E}_{\eta\sim P}[S(\widetilde{\Phi}(\eta))] \neq \mathbb{E}_{\xi\sim Q}[S(\widetilde{\Phi}(\xi))]$. Consequently, for any such pair $P \neq Q$, there exists a finite signature degree M such that $\mathrm{Sig}\text{-}\mathrm{W}_1^{(M)}(P,Q) > 0$. For each fixed M with finite expected signature coordinates, $\mathrm{Sig}\text{-}\mathrm{W}_1^{(M)}$, being an ℓ^2 distance between expected truncated signatures, is symmetric and satisfies the triangle inequality, and is therefore a pseudometric.

Pointwise Losses. As mentioned in the introduction, pointwise losses fail in a quantifiable, easily reproducible way on the very benchmarks the community uses to evaluate generative TPPs. For an interarrival distribution P with finite first moment, mean absolute error (MAE) is minimised by $\hat{\tau} = \mathrm{median}(P)$. With a finite second moment, mean squared error (MSE) is minimised by $\hat{\tau} = \mathbb{E}[P]$ [21]. Both optima are Dirac measures. Neither can, in principle, distinguish a calibrated generative model from a deterministic regressor. See Tables S2 and S3 for an empirical confirmation, where DETER has the lowest sample-based MAE on all 12 datasets despite poor pathwise scores.

5 Experimental Results

We evaluate SIGTPP on controlled synthetic regimes and real-world tasks. We assess performance across multiple axes, from path-level distributional fidelity to moment structure, and examine the reliability of pointwise metrics as evaluation criteria. We further study the dependence of SIGTPP on the signature truncation degree.

5.1 Datasets, Baselines, and Evaluation Metrics

We use 12 datasets in total, 6 synthetic and 6 real-world. Table S11 and Figures S13-S24 summarise their basic characteristics, and Appendix F describes each dataset in more detail. The synthetic corpus covers the canonical regimes of TPP modelling: Poisson process (PS), inhomogeneous Poisson (IP), univariate (H1) and 3-dimensional (H3) Hawkes processes, self-correcting (SC), and stationary renewal (SR) processes. The real-world corpus encompasses datasets from prior work, namely Earthquake (EQ), Stack Overflow (SO), Taobao (TB), and Taxi (TX) from EasyTPP [56], and Yelp Mississauga (YLP) and Yelp Airport (YLA) from the benchmark proposed by Lüdke et al. [34].

We compare SIGTPP against 5 baselines spanning the main paradigms for generative TPP modelling. We focus on autoregressive baselines that have a neural architecture similar to SIGTPP (similar number of parameters) to make comparisons fair. These include the denoising diffusion probabilistic model (DDPM) and variational autoencoder (VAE) [32], the Wasserstein GAN (WGAN) [55], and a maximum-likelihood Gamma fit [9, 10] as a parametric reference (GAMMA). To isolate the contribution of distributional modelling, we also include DETER [32], a deterministic regressor whose predictive law is a Dirac measure and which by construction cannot capture distributional structure.

We evaluate the goodness of fit along the following axes: pathwise discrepancies via $\mathcal{E}$ (energy distance) and W_1 under our proposed metric $d_{\mathcal{N}}$, signature-based discrepancy via $\mathrm{Sig}\text{-}\mathrm{W}_1$, pointwise probabilistic accuracy via the continuous ranked probability score (CRPS), marginal and temporal structure via histograms of interarrival times ($L_{\log(\tau)}$) and intensity (L_λ), pairwise correlation discrepancy (PCD), and autocorrelation discrepancy (ACD) across sequence components. Lower is better for all reported metrics. Metric definitions and a more expansive justification for their appropriateness are deferred to Appendix D. The main benchmark tables report mean test performance, with standard errors in parentheses, computed as the sample standard deviation over 100 bootstrap replicate estimates.

For each dataset, hyperparameters are selected on the validation split using a rank-based aggregate criterion over the validation diagnostics. The criterion sums within-dataset ranks across the ranked validation metrics. During training, we select checkpoints based on the validation $L_{\log(\tau)}$. The aggregate rank criterion is used only for model selection, and not as a performance measure.

Table 1: Performance of SIGTPP and baselines on synthetic point processes. Results for SC and SR are in Table S1. Values report standard error in parentheses over 100 bootstrap replicates. Lower is better (↓). Best in **bold**, second best underlined. Metric definitions in Appendix D.

	Model	PS	IP	H1	H3		PS	IP	H1	H3
$\mathcal{E}$ ($\times 10^{-3}$)	SIGTPP	**0.53** (0.93)	**0.42** (0.32)	**1.16** (0.74)	**1.32** (1.60)	W_1 ($\times 10^{-2}$)	**5.26** (0.20)	**6.55** (0.14)	**3.75** (0.31)	**10.1** (0.42)
	VAE	8.64 (3.47)	2.24 (0.64)	16.8 (3.85)	20.0 (6.11)		7.12 (0.64)	6.76 (0.18)	6.84 (0.57)	13.8 (1.11)
	DDPM	2.72 (1.11)	2.56 (0.63)	10.7 (2.73)	1.61 (1.35)		6.15 (0.24)	7.27 (0.15)	5.93 (0.49)	10.5 (0.50)
	WGAN	4.99 (1.83)	3.20 (1.03)	16.9 (3.93)	11.4 (4.00)		6.43 (0.40)	7.68 (0.29)	6.97 (0.61)	12.8 (0.83)
	DETER	330 (12.8)	108 (4.85)	510 (15.7)	640 (25.3)		28.3 (0.65)	18.2 (0.34)	38.1 (0.62)	54.6 (1.10)
	GAMMA	4.38 (2.58)	20.3 (3.22)	17.3 (3.87)	25.5 (5.92)		6.33 (0.54)	9.50 (0.49)	7.05 (0.59)	15.1 (1.07)
Sig-W_1 ($\times 10^{-5}$)	SIGTPP	**1.42** (1.52)	1.20 (1.05)	**44.0** (29.2)	1.02 (0.75)	CRPS ($\times 10^{-1}$)	4.64 (0.08)	3.04 (0.03)	10.2 (0.27)	2.40 (0.03)
	VAE	13.4 (6.42)	**0.83** (0.89)	155 (28.1)	**0.25** (0.26)		**4.56** (0.08)	3.07 (0.03)	**9.53** (0.26)	2.33 (0.03)
	DDPM	129 (24.3)	64.0 (9.18)	302 (65.9)	10.9 (2.70)		4.81 (0.09)	3.27 (0.04)	10.1 (0.30)	**2.23** (0.03)
	WGAN	17.4 (6.92)	1.01 (0.70)	141 (30.9)	0.80 (0.70)		4.58 (0.08)	**3.03** (0.03)	9.59 (0.28)	2.36 (0.03)
	DETER	62.7 (12.7)	8.01 (1.76)	862 (112)	1.09 (0.39)		6.29 (0.11)	4.34 (0.05)	12.7 (0.37)	3.23 (0.04)
	GAMMA	5.76 (4.25)	23.8 (4.84)	284 (51.3)	0.63 (0.29)		4.60 (0.08)	3.21 (0.03)	9.95 (0.26)	2.39 (0.03)
L_λ ($\times 10^{-2}$)	SIGTPP	3.35 (0.46)	**3.09** (0.35)	**6.09** (1.04)	2.68 (0.40)	$L_{\log(\tau)}$ ($\times 10^{-2}$)	2.95 (0.60)	**1.97** (0.37)	**3.84** (0.84)	7.64 (0.56)
	VAE	4.46 (0.57)	4.90 (0.43)	13.1 (1.22)	**2.25** (0.40)		4.16 (0.58)	5.47 (0.46)	4.07 (0.67)	4.01 (0.48)
	DDPM	7.93 (0.64)	7.18 (0.41)	15.2 (1.05)	5.10 (0.50)		6.08 (0.72)	4.10 (0.45)	5.45 (0.89)	2.32 (0.47)
	WGAN	7.31 (0.67)	4.47 (0.37)	13.3 (1.31)	3.37 (0.49)		3.74 (0.55)	4.70 (0.46)	5.33 (0.74)	4.70 (0.56)
	DETER	12.6 (0.71)	12.8 (0.43)	20.2 (1.25)	4.17 (0.52)		60.7 (5.26)	48.4 (1.46)	53.5 (2.86)	66.6 (6.25)
	GAMMA	**3.17** (0.45)	11.0 (0.47)	8.11 (1.08)	3.41 (0.43)		**2.86** (0.64)	3.21 (0.48)	6.74 (1.00)	**1.71** (0.59)
ACD ($\times 10^{-2}$)	SIGTPP	1.35 (0.46)	**0.70** (0.20)	1.70 (0.45)	**1.54** (0.39)	PCD ($\times 10^{-2}$)	**7.36** (0.53)	**6.01** (0.42)	9.07 (0.85)	8.56 (0.40)
	VAE	1.53 (0.42)	0.71 (0.20)	2.73 (0.62)	2.44 (0.45)		7.43 (0.53)	6.40 (0.43)	8.82 (0.86)	8.39 (0.40)
	DDPM	1.86 (0.52)	1.64 (0.25)	2.54 (0.47)	1.76 (0.44)		7.54 (0.51)	6.27 (0.43)	**8.25** (0.97)	**8.33** (0.42)
	WGAN	2.39 (0.49)	1.94 (0.28)	**1.60** (0.39)	5.03 (0.45)		7.49 (0.53)	6.15 (0.43)	8.87 (0.84)	8.82 (0.41)
	DETER	49.3 (0.53)	35.3 (0.28)	52.5 (0.71)	63.4 (0.45)		96.5 (0.43)	82.7 (0.28)	59.7 (0.83)	97.9 (0.22)
	GAMMA	**1.32** (0.44)	1.28 (0.26)	4.08 (0.71)	6.12 (0.45)		7.39 (0.53)	6.39 (0.42)	9.03 (0.85)	8.74 (0.41)

Since evaluation metrics can differ substantially in scale across datasets, we report the relative score computed as a 2-step process (see also [1, 13, 22]). First, we compute, for each model, a relative score defined as the ratio between the model's score and that of a baseline model, here DETER. We then aggregate these relative scores across tasks using the geometric mean [17], which is the appropriate summary statistic for multiplicative relative comparisons, and makes the model ordering independent of the choice of the baseline (see Appendix D.4).

5.2 Recovering Known Synthetic Processes

We first use 6 synthetic point processes with known data-generating mechanisms to evaluate SIGTPP. Table 1 presents the discrepancy metrics $\mathcal{E}$ and W_1 under $d_{\mathcal{N}}$, Sig-W_1, CRPS, $L_{\log(\tau)}$, L_λ, ACD, and PCD. The results for SC and SR appear in Table S1. Our first observation is that SIGTPP improves distributional fidelity (Section 4.4) relative to the strongest baselines. SIGTPP obtains the lowest $\mathcal{E}$ and W_1 on 5 out of the 6 datasets, and on 3 out of 6 w.r.t. Sig-W_1. The relative score (Appendix D.4) of our model is the lowest under $\mathcal{E}$, W_1, and Sig-W_1. Compared to the second-best model for each metric (DDPM for $\mathcal{E}$ and W_1, and VAE for Sig-W_1), SIGTPP improves the performance by 68%, 13%, and 34%, respectively (Figure S2 and Table S6). These gains are consistent with SIGTPP's training objective, which targets the joint temporal law rather than local event-level behaviour.

However, it is also striking that while SIGTPP improves the distributional fit, it occasionally underperforms according to CRPS (4% and 1% error increase compared to VAE and WGAN, but 7% decrease compared to DDPM). This suggests that optimising a path-level distributional objective favours global distributional alignment, but does not necessarily improve conditional calibration relative to methods trained directly with conditional losses. Nevertheless, a number of summary-statistic discrepancies indicate that SIGTPP remains competitive. $L_{\log(\tau)}, L_\lambda$, ACD, and PCD exhibit overlapping standard errors for many methods. SIGTPP is either the best method or within one standard error of the best baseline in 20 out of 24 metric-dataset pairs: in $2(+2)/6$ datasets for $L_{\log(\tau)}$, $4(+1)/6$ for L_λ, $3(+2)/6$ for ACD, and $3(+3)/6$ for PCD, where the number in parentheses denotes outcomes within one standard error of the best. When ranking methods separately across each dataset and metric, SIGTPP has the best average rank of 2.00, followed by VAE with an average rank of 2.88 (Table S5).

Figures 3(a) and 3(b) illustrate two of these statistics. On SR, whose lognormal interarrival times are heavy-tailed, the QQ plot shows that SIGTPP and DDPM stay closest to the diagonal in the upper tail, whereas VAE, WGAN, and GAMMA fall below it and generate too few long gaps. DDPM has

Table 2: Performance of SIGTPP and baselines on real-world tasks. Results for YLA are in Table S1. Values report standard error in parentheses over 100 bootstrap replicates. Lower is better (↓). Best in **bold**, second best underlined. Metric definitions in Appendix D.

	Model	EQ	SO	TB	TX	YLP		EQ	SO	TB	TX	YLP
$\mathcal{E}$ ($\times 10^{-3}$)	SIGTPP	32.4 (0.65)	12.0 (1.71)	512 (25.2)	**76.3** (2.95)	38.4 (14.9)	W_1 ($\times 10^{-2}$)	**2.91** (0.04)	7.44 (0.42)	58.2 (2.56)	8.02 (0.22)	**15.0** (1.86)
	VAE	34.5 (0.64)	**9.64** (2.55)	545 (24.7)	77.7 (2.73)	**33.6** (17.6)		2.91 (0.05)	**7.08** (0.76)	59.8 (2.40)	8.13 (0.23)	16.5 (2.90)
	DDPM	**26.0** (0.66)	24.5 (6.37)	**455** (21.5)	76.3 (2.77)	57.5 (20.5)		3.02 (0.05)	10.9 (1.22)	**52.8** (2.31)	**7.74** (0.18)	20.0 (2.06)
	WGAN	32.6 (0.71)	26.5 (3.34)	502 (22.4)	81.9 (2.80)	49.2 (24.3)		2.99 (0.04)	10.3 (0.64)	56.1 (2.07)	8.38 (0.27)	18.8 (2.78)
	DETER	47.2 (0.92)	478 (22.6)	958 (32.1)	161 (5.46)	2402 (90)		3.74 (0.07)	36.2 (0.80)	86.0 (2.48)	14.1 (0.40)	143 (3.55)
	GAMMA	35.8 (0.66)	31.4 (3.78)	539 (23.5)	85.6 (2.99)	250 (59.0)		3.11 (0.05)	11.0 (0.66)	60.3 (2.15)	8.79 (0.28)	35.2 (3.68)
Sig-W_1 ($\times 10^{-5}$)	SIGTPP	409 (36.3)	**0.09** (0.10)	2.27 (0.94)	4.74 (1.22)	0.77 (0.86)	CRPS ($\times 10^{-1}$)	5.61 (0.11)	5.43 (0.10)	0.37 (0.01)	1.13 (0.02)	3.22 (0.13)
	VAE	397 (38.1)	3.70 (0.52)	**0.45** (0.30)	9.63 (1.61)	1.08 (1.10)		5.13 (0.12)	4.64 (0.09)	0.36 (0.01)	**1.06** (0.02)	**2.74** (0.12)
	DDPM	1144 (56)	2.58 (0.44)	9.78 (1.99)	5.52 (1.21)	1.70 (1.08)		**4.57** (0.11)	**4.63** (0.10)	**0.34** (0.01)	1.20 (0.02)	2.76 (0.11)
	WGAN	603 (45.9)	3.95 (0.64)	7.81 (1.44)	6.80 (1.51)	**0.42** (0.47)		5.85 (0.10)	4.93 (0.09)	0.37 (0.01)	1.15 (0.02)	2.91 (0.11)
	DETER	**146** (14.5)	0.19 (0.16)	49.7 (4.45)	24.5 (2.82)	7.78 (0.61)		7.05 (0.16)	6.51 (0.12)	0.42 (0.01)	1.49 (0.02)	3.66 (0.15)
	GAMMA	572 (38.1)	6.66 (0.81)	12.3 (3.47)	**3.07** (1.16)	10.6 (1.40)		6.16 (0.09)	4.91 (0.08)	0.37 (0.01)	1.15 (0.02)	3.19 (0.11)
L_λ ($\times 10^{-2}$)	SIGTPP	19.3 (0.99)	9.00 (0.68)	**3.77** (0.90)	6.31 (1.05)	13.3 (1.14)	$L_{\log(\tau)}$ ($\times 10^{-2}$)	12.9 (0.84)	10.5 (0.41)	7.33 (0.39)	4.35 (0.44)	4.73 (0.99)
	VAE	**17.9** (1.12)	8.94 (0.69)	9.57 (1.06)	6.99 (1.04)	**10.8** (1.14)		14.8 (0.97)	10.6 (0.56)	**6.29** (0.53)	4.36 (0.40)	**4.02** (0.68)
	DDPM	19.0 (0.62)	9.95 (0.63)	7.01 (1.08)	**4.57** (1.04)	16.9 (1.11)		**9.33** (0.82)	5.22 (0.63)	10.6 (0.72)	**2.02** (0.40)	6.47 (0.80)
	WGAN	24.1 (1.13)	**6.25** (0.77)	14.0 (1.24)	5.06 (0.74)	13.3 (1.03)		9.58 (0.58)	**2.72** (0.62)	7.86 (0.70)	6.72 (0.55)	5.69 (0.93)
	DETER	20.3 (2.84)	7.38 (0.37)	41.6 (3.32)	18.0 (1.33)	25.3 (0.93)		48.5 (3.15)	54.9 (7.16)	81.0 (2.02)	56.9 (6.67)	45.5 (1.62)
	GAMMA	27.0 (1.12)	11.5 (0.60)	16.2 (1.08)	7.05 (0.73)	31.7 (1.20)		9.93 (0.83)	5.70 (0.75)	15.1 (1.26)	4.84 (0.41)	13.2 (0.85)
ACD ($\times 10^{-2}$)	SIGTPP	**3.95** (0.68)	8.93 (0.51)	**0.94** (0.20)	2.97 (0.43)	**2.11** (0.40)	PCD ($\times 10^{-2}$)	**7.51** (0.53)	12.4 (0.33)	5.94 (0.26)	6.11 (0.23)	12.9 (0.69)
	VAE	7.17 (0.76)	**2.79** (0.51)	2.48 (0.40)	2.26 (0.42)	2.61 (0.53)		9.92 (0.74)	**10.2** (0.28)	**5.83** (0.27)	5.85 (0.22)	**12.4** (0.71)
	DDPM	6.95 (0.76)	5.17 (0.51)	1.41 (0.35)	**1.87** (0.45)	3.35 (0.60)		8.44 (0.69)	11.1 (0.32)	5.84 (0.25)	**5.61** (0.22)	12.8 (0.68)
	WGAN	6.95 (0.76)	6.57 (0.51)	3.20 (0.41)	3.48 (0.47)	5.51 (0.46)		7.55 (0.54)	11.0 (0.31)	5.92 (0.28)	6.25 (0.22)	13.1 (0.69)
	DETER	61.9 (0.76)	72.7 (0.51)	74.2 (0.41)	60.7 (0.50)	64.4 (0.62)		73.4 (0.81)	79.3 (0.27)	94.8 (0.34)	43.4 (0.31)	72.4 (0.80)
	GAMMA	16.8 (0.76)	8.03 (0.51)	2.92 (0.41)	3.51 (0.47)	4.61 (0.61)		18.8 (0.81)	12.0 (0.35)	5.86 (0.27)	6.24 (0.22)	13.3 (0.69)

the lowest $L_{\log(\tau)}$ on SR, with SIGTPP within one standard error (Table S1). On H1, the intensity of SIGTPP stays close to the reference, whereas the intensities of VAE and DDPM rise towards the end of the window. SIGTPP has the lowest L_λ on H1 (Table 1). Figures S5 and S6 give the full qualitative comparison on SC and SR, including the correlation and autocorrelation structures.

Overall, across all synthetic metric-dataset pairs, SIGTPP is best in $25/48$ cases and within one standard error of the best in another $11/48$ cases, making it competitive in about 75% of them. This is further supported by the relative score, with SIGTPP achieving the best performance on 6 out of 8 metrics. In pairwise comparisons against each baseline, averaged over datasets and metrics, SIGTPP outperforms WGAN by 50%, VAE by 39%, and DDPM by 46%.

5.3 Real-World Tasks

We now focus our empirical analysis on 6 real-world tasks under the same discrepancy metrics. The results are enumerated in Tables 2 and S1 and, again, indicate strong performance for SIGTPP (see Table S6). It achieves the lowest W_1 or lies within the best model's standard error on $4/6$ real-world

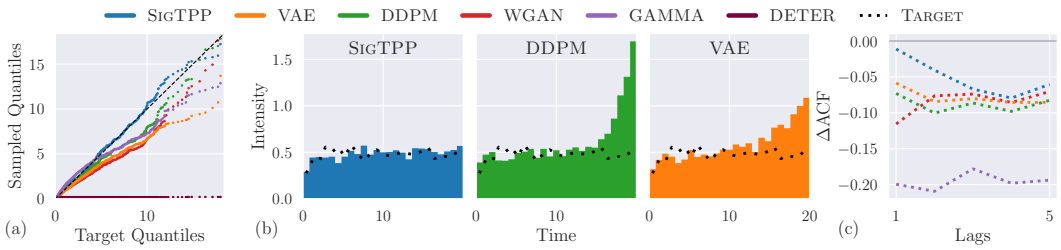


Figure 3: Qualitative fit on 3 datasets, one per summary statistic: (a) $L_{\log(\tau)}$, (b) L_λ, and (c) ACD. Colours mark models as in the legend; black marks the reference data. (a) Quantile-quantile (QQ) plot of interarrival times for the 6 models on SR. (b) Empirical intensity functions $\hat{\lambda}(t)$ on H1 for SIGTPP and the two baselines with the best overall mean rank, DDPM, and VAE. The dotted line is the reference intensity. (c) Signed autocorrelation error of interarrival times on EQ for lags 1 to 5. DETER is omitted because its much larger error compresses the scale.

datasets. This is confirmed by the relative scores, for which it obtains the lowest $\mathcal{E}$, W_1, and Sig-W_1 among all models. The second-best model on real-world tasks is VAE, which is outperformed on geometric average by 7% across all metrics and datasets. The improvements in $\mathcal{E}$ and W_1 support pathwise distributional agreement under criteria distinct from the signature-based training objective. The contrast with the synthetic setting is informative. Synthetic data is generated by a known mechanism, so distributional recovery is assessed under correct specification. The real-world datasets instead combine heterogeneous trajectories, variable observation windows, and latent contextual effects, so the empirical path distribution is only an approximation to a single data-generating law. This makes path-level comparison intrinsically harder and likely explains why the gains of SIGTPP are smaller and less uniform than in the synthetic experiments.

The remaining summary-statistic discrepancies provide a more nuanced comparison. As on the synthetic datasets, $L_{\log(\tau)}$, L_λ, ACD, and PCD exhibit larger sampling variability, and several methods have overlapping standard errors. Across these metrics, SIGTPP performs best in $5/24$ metric-dataset pairs and lies within one standard error of the best method in an additional $4/24$ cases. Thus, although CRPS favours the conditional baselines on aggregate, with SIGTPP trailing VAE, WGAN, and DDPM by 9%, 2%, and 10%, respectively, it remains competitive on real-world data. In a separate ablation (Appendix E.3), an auxiliary next-event regression loss is added to SIGTPP and narrows the CRPS gap to VAE, the strongest baseline on CRPS (from trailing by 20% to 3%). This regression loss variant improves SIGTPP's $\mathcal{E}$, W_1, and Sig-W_1 by 26%, 12%, and 30%, respectively. Despite its weaker CRPS results and mixed performance on summary statistics, SIGTPP achieves the lowest geometric relative score overall, and specifically on 5 out of 8 metrics (Figure 4). In pairwise comparisons, SIGTPP improves over WGAN by 20%, VAE by 7%, and DDPM by 17%.

Figure 3(c) shows the short range dependence that ACD measures on EQ. All plotted models underestimate the autocorrelation of interarrival times, but SIGTPP stays closest to the reference at every lag, by a wide margin at lags 1 and 2. ACD averages the absolute error over these lags, and SIGTPP has the lowest ACD on EQ (Table 2). Appendix Figures S7-S11 provide dataset-level qualitative comparisons of the 6 models and further illustrate these trends.

Average ranks provide additional insights (Table S5). SIGTPP achieves the second-best average rank of 2.54, behind VAE (2.38). Relative scoring nevertheless favours SIGTPP (see Figure S3). Ranks record only the ordering of models, whereas geometric score ratios also reflect the size of performance differences. Overall, the real-world geometric scores mirror the synthetic results, but with smaller margins and greater dataset-level variability. The persistence of improvements over pairwise comparisons, and average ranks in $\mathcal{E}$, W_1, and Sig-W_1 suggests that the signature-based objective continues to provide useful distributional alignment under misspecification.

More generally, across all 12 datasets and 8 metrics, SIGTPP has the lowest overall mean rank among the 6 models (2.27), followed by VAE (2.62). It has the lowest geometric relative score on 5 of the 8 metrics, the 2^{nd}-lowest on $L_{\log(\tau)}$ and PCD (behind DDPM), and the 4^{th}-lowest on CRPS (Figure 4). We also compare SIGTPP with 5 methods from EasyTPP [56] on a subset of 6 datasets (Appendix E). Consistent with the main benchmark, SIGTPP reduces scores by at least 28% against every EasyTPP baseline on geometric average across the available dataset-metric pairs and achieves the best overall average rank of 2.40, followed by NHP with rank 2.94 (Table S8).

Finally, Tables S2 and S3 illustrate why pointwise errors are not suitable primary metrics for generative TPP evaluation. The deterministic baseline DETER outputs a single prediction for each event time and is therefore directly optimised for pointwise accuracy rather than distributional fidelity. Although it performs poorly on most path-level and distributional metrics, in some cases by several orders of magnitude, it achieves rank 1.00 on MAE, outperforming every distributional model on every dataset. For $\mathrm{MSE}_{\text{mean}}$, which evaluates the predictive mean, its average ranks are 4.33 on synthetic tasks and 5.17 on real-world tasks. The median and mean summarise only part of the predictive law, so accurate point predictions do not establish a good fit. Hence, we do not use them as metrics for generative TPP evaluation, and instead recommend path-level distributional distances. For forecasting-focused tasks, longer-horizon predictive metrics may also be appropriate.

5.4 Ablation of the Signature Degree

In Table 3, we isolate the effect of the signature degree M on H3, TX, and SO, 3 datasets with different statistical properties and sizes (see Table S11). Although the number of signature coordinates

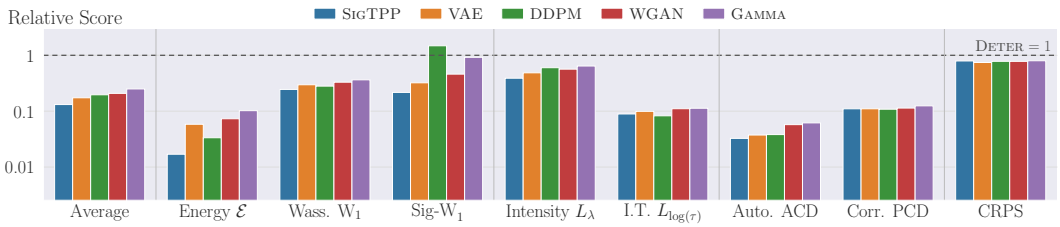


Figure 4: Geometric means of per-dataset score ratios to DETER over all 12 datasets. The vertical axis is logarithmic. The dashed line marks DETER $=1$. Lower is better (↓). The exact values are enumerated in Table S6. Metric definitions are provided in Appendix D.

grows as $\mathcal{O}(d^M)$, gains depend on the dataset and metric. On H3, increasing from $M=2$ to $M=3$ improves all 4 metrics shown. $M=4$ gives the lowest $\mathcal{E}$ and W_1, while $M=5$ gives the lowest Sig-W_1 and CRPS. The identical scores at $M=6$ and $M=8$ are consistent with the diagnostic's first inactive degree of 6 on H3 (Table S4). On TX, most of the improvement occurs from $M=2$ to $M=3$; larger values produce only small changes in W_1 and CRPS. On SO, $\mathcal{E}$ and W_1 are lowest at $M=3$ and worsen at $M=4$ and $M=5$, while CRPS is highest at $M=4$. Scores at larger M are non-monotone.

No tested degree is best across all 4 metrics on all 3 datasets. To understand that better, note that increasing M changes the set of coordinates used for training rather than simply refining the same discrepancy. Low-order coordinates often dominate the signature scale, while higher-order iterated integrals decay factorially for bounded variation paths, despite the exponential growth in their number [38]. Consequently, a large M can introduce weakly varying or poorly conditioned features, increasing estimation noise without necessarily improving the learned path distribution. This is consistent with the dead-degree statistics in Table S4, showing the first inactive degree at 6 on H3. Overall, the ablation supports $M=3$ for the evaluation and suggests that using a degree higher than 8 does not yield further improvement.

Table 3: Ablation of the training signature degree M for SIGTPP on H3, TX, and SO. Values are mean test scores over 100 bootstrap replicates. Lower is better (↓). Best in **bold**, second best underlined. Metric definitions in Appendix D.

	$\mathcal{E}$ ($\times 10^{-3}$) ↓						W_1 ($\times 10^{-2}$) ↓						Sig-W_1 ($\times 10^{-5}$) ↓						CRPS ($\times 10^{-1}$) ↓					
M	2	3	4	5	6	8	2	3	4	5	6	8	2	3	4	5	6	8	2	3	4	5	6	8
H3	192.1	5.21	**1.88**	5.48	2.58	2.58	31.09	11.644	**10.30**	11.76	10.79	10.79	25.50	0.685	1.344	**0.656**	1.845	1.845	3.369	2.437	2.384	**2.379**	2.411	2.411
TX	86.95	**79.31**	79.46	79.91	80.39	79.68	9.491	8.343	8.270	8.533	**8.163**	8.172	1.363	0.796	0.664	**0.341**	0.456	1.627	1.257	1.250	1.168	1.166	1.152	**1.147**
SO	19.93	**8.73**	10.68	24.81	19.54	11.78	8.619	**6.321**	8.225	10.02	9.432	7.050	0.514	0.980	0.456	0.144	0.189	**0.136**	5.446	**5.259**	10.55	5.409	5.375	6.165

6 Conclusion

This paper introduces a common pathwise framework for generative modelling and evaluation of TPPs. The interarrival embedding Φ provides a stable, injective lift from càdlàg counting paths to continuous bounded-variation paths (Theorem 6), making signature-based tools applicable to discrete event sequences. Building on Φ, SIGTPP optimises a global signature Wasserstein loss on whole paths, rather than a sum of local conditional losses. The same counting-path metric $d_{\mathcal{N}}$ supports the energy distance ($\mathcal{E}$), W_1, as well as Sig-W_1 through the embedding. It also clarifies a practical failure mode of pointwise evaluation: MAE can reward deterministic regressors even when they fail to capture distributional variation. Across 12 datasets, SIGTPP achieves the best average rank on 6 of 8 evaluation criteria, best average relative score on 5 of them, and outperforms every other baseline by at least 24% on average across all metrics. The gains are largest on pathwise distributional criteria, while diffusion and variational baselines remain competitive on marginal calibration. The limitations of our work are discussed in Appendix H.

Acknowledgements

NCK is supported by EPSRC grants EP/T517793/1 and EP/W524335/1. VL would like to thank all levels of support from grant EP/X031276/1 (EPSRC) and the SOFAIR Lab (UKRI).

References

[1] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the Language of Time Series. *Transactions on Machine Learning Research*, 2024.

[2] Martin Arjovsky and Léon Bottou. Towards Principled Methods for Training Generative Adversarial Networks. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[3] Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes Processes in Finance. *Market Microstructure and Liquidity*, 1(1), 2015.

[4] Jean Bretagnolle, Didier Dacunha-Castelle, and Jean-Louis Krivine. Lois stables et espaces L^p. In *Symposium on Probability Methods in Analysis*, pages 48–54, 1967.

[5] Yuxin Chang, Alex James Boyd, Cao Xiao, Taha Kass-Hout, Parminder Bhatia, Padhraic Smyth, and Andrew Warrington. Deep Continuous-Time State-Space Models for Marked Event Sequences. In *Advances in Neural Information Processing Systems*, 2025.

[6] Ilya Chevyrev and Andrey Kormilitzin. *A Primer on the Signature Method in Machine Learning*, pages 3–64. Springer, 2026.

[7] Ilya Chevyrev and Terry Lyons. Characteristic Functions of Measures on Geometric Rough Paths. *The Annals of Probability*, 44(6):4049–4082, 2016.

[8] Ilya Chevyrev and Harald Oberhauser. Signature Moments to Characterize Laws of Stochastic Processes. *Journal of Machine Learning Research*, 23(176):7928–7969, 2022.

[9] Andrea Coletta, Aymeric Moulin, Svitlana Vyetrenko, and Tucker Balch. Learning to simulate realistic limit order book markets from data as a World Agent. In *Proceedings of the Third ACM International Conference on AI in Finance*, pages 428–436, 2022.

[10] Rama Cont, Mihai Cucuringu, Jonathan Kochems, and Felix Prenzel. Limit Order Book Simulation with Generative Adversarial Networks. *SSRN working paper 4512356*, 2023.

[11] Christa Cuchiero, Francesca Primavera, and Sara Svaluto-Ferro. Universal approximation theorems for continuous functions of càdlàg paths and Lévy-type signature models. *Finance and Stochastics*, 29(2): 289–342, 2025.

[12] Daryl J. Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer, 2003.

[13] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A Decoder-Only Foundation Model for Time-Series Forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, pages 10148–10167, 2024.

[14] Kelian Dascher-Cousineau, Oleksandr Shchur, Emily E. Brodsky, and Stephan Günnemann. Using Deep Learning for Flexible and Scalable Earthquake Forecasting. *Geophysical Research Letters*, 50(17): e2023GL103909, 2023.

[15] Michel Marie Deza and Monique Laurent. *Geometry of Cuts and Metrics*. Springer, 1997.

[16] 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*, pages 1555–1564, 2016.

[17] Philip J. Fleming and John J. Wallace. How Not to Lie with Statistics: The Correct Way to Summarize Benchmark Results. *Communications of the ACM*, 29(3):218–221, 1986.

[18] Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley, 1999.

[19] Peter K. Friz and Nicolas B. Victoir. *Multidimensional Stochastic Processes as Rough Paths: Theory and Applications*. Cambridge University Press, 2010.

[20] Álvaro Gajardo and Hans-Georg Müller. Point process models for COVID-19 cases and deaths. *Journal of Applied Statistics*, 50(11-12):2294–2309, 2023.

[21] Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[22] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large Language Models Are Zero-Shot Time Series Forecasters. In *Advances in Neural Information Processing Systems*, 2023.

[23] Ben Hambly and Terry Lyons. Uniqueness for the Signature of a Path of Bounded Variation and the Reduced Path Group. *Annals of Mathematics*, 171(1):109–167, 2010.

[24] Alan G. Hawkes. Spectra of Some Self-Exciting and Mutually Exciting Point Processes. *Biometrika*, 58 (1):83–90, 1971.

[25] William Hua, Hongyuan Mei, Sarah Zohar, Magali Giral, and Yanxun Xu. Personalized Dynamic Treatment Regimes in Continuous Time: A Bayesian Approach for Optimizing Clinical Decisions with Timing. *Bayesian Analysis*, 17(3):849–878, 2022.

[26] Zeyang Issa, Blanka Horvath, Maud Lemercier, and Cristopher Salvi. Non-adversarial training of Neural SDEs with signature kernel scores. In *Advances in Neural Information Processing Systems*, pages 11102–11126, 2023.

[27] Gavin Kerrigan, Kai Nelson, and Padhraic Smyth. EventFlow: Forecasting Temporal Point Processes with Flow Matching. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, 2026.

[28] Patrick Kidger, Patric Bonnier, Imanol Perez Arribas, Cristopher Salvi, and Terry Lyons. Deep Signature Transforms. In *Advances in Neural Information Processing Systems*, pages 3082–3092, 2019.

[29] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[30] Shuang Li, Shuai Xiao, Shixiang Zhu, Nan Du, Yao Xie, and Le Song. Learning Temporal Point Processes via Reinforcement Learning. In *Advances in Neural Information Processing Systems*, pages 10804–10814, 2018.

[31] Shujian Liao, Hao Ni, Marc Sabate-Vidales, Lukasz Szpruch, Magnus Wiese, and Baoren Xiao. Sig-Wasserstein GANs for Conditional Time Series Generation. *Mathematical Finance*, 34(2):622–670, 2024.

[32] Haitao Lin, Lirong Wu, Guojiang Zhao, Pai Liu, and Stan Z. Li. Exploring Generative Neural Temporal Point Process. *Transactions on Machine Learning Research*, 2022.

[33] Haitao Lin, Cheng Tan, Lirong Wu, Zhangyang Gao, Zicheng Liu, and Stan Z. Li. An Extensive Survey With Empirical Studies on Deep Temporal Point Processes. *IEEE Transactions on Knowledge and Data Engineering*, 37(4):1599–1619, 2025.

[34] David Lüdke, Marin Biloš, Oleksandr Shchur, Marten Lienen, and Stephan Günnemann. Add and Thin: Diffusion for Temporal Point Processes. In *Advances in Neural Information Processing Systems*, pages 56784–56801, 2023.

[35] David Lüdke, Enric Rabasseda Raventós, Marcel Kollovieh, and Stephan Günnemann. Unlocking Point Processes through Point Set Diffusion. In *Proceedings of the 13th International Conference on Learning Representations*, 2025.

[36] David Lüdke, Marten Lienen, Marcel Kollovieh, and Stephan Günnemann. Edit-Based Flow Matching for Temporal Point Processes. In *Proceedings of the 14th International Conference on Learning Representations*, 2026.

[37] Russell Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284–3305, 2013.

[38] Terry J. Lyons, Michael Caruana, and Thierry Lévy. *Differential Equations Driven by Rough Paths: École d'Été de Probabilités de Saint-Flour XXXIV - 2004*. Springer, 2007.

[39] Hongyuan Mei and Jason M. Eisner. The Neural Hawkes Process: A Neurally Self-Modulating Multivariate Point Process. In *Advances in Neural Information Processing Systems*, pages 6754–6764, 2017.

[40] Sumantrak Mukherjee, Mouad Elhamdi, George Mohler, David A. Selby, Yao Xie, Sebastian Vollmer, and Gerrit Grossmann. Neural Spatiotemporal Point Processes: Trends and Challenges. *Transactions on Machine Learning Research*, 2025.

[41] Hao Ni, Lukasz Szpruch, Marc Sabate-Vidales, Baoren Xiao, Magnus Wiese, and Shujian Liao. Sig-Wasserstein GANs for Time Series Generation. In *Proceedings of the Second ACM International Conference on AI in Finance*, 2021.

[42] Maya Okawa, Tomoharu Iwata, Takeshi Kurashima, Yusuke Tanaka, Hiroyuki Toda, and Naonori Ueda. Deep Mixture Point Processes: Spatio-temporal Event Prediction with Rich Contextual Information. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 373–383, 2019.

[43] Maya Okawa, Tomoharu Iwata, Yusuke Tanaka, Takeshi Kurashima, Hiroyuki Toda, and Hisashi Kashima. Context-aware spatio-temporal event prediction via convolutional Hawkes processes. *Machine Learning*, 111(8):2929–2950, 2022.

[44] Takahiro Omi, Naonori Ueda, and Kazuyuki Aihara. Fully Neural Network Based Model for General Temporal Point Processes. In *Advances in Neural Information Processing Systems*, pages 2120–2129, 2019.

[45] Imanol Perez Arribas, Guy M. Goodwin, John R. Geddes, Terry Lyons, and Kate E. A. Saunders. A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. *Translational Psychiatry*, 8(274), 2018.

[46] Romain Pic, Clément Dombry, Philippe Naveau, and Maxime Taillardat. Proper scoring rules for multivariate probabilistic forecasts based on aggregation and transformation. *Advances in Statistical Climatology, Meteorology and Oceanography*, 11(1):23–58, 2025.

[47] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests. *Entropy*, 19(2):47, 2017.

[48] Isaac J. Schoenberg. Metric Spaces and Positive Definite Functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

[49] Oleksandr Shchur, Marin Biloš, and Stephan Günnemann. Intensity-Free Learning of Temporal Point Processes. In *International Conference on Learning Representations*, 2020.

[50] Oleksandr Shchur, Nicholas Gao, Marin Biloš, and Stephan Günnemann. Fast and Flexible Temporal Point Processes with Triangular Maps. In *Advances in Neural Information Processing Systems*, pages 73–84, 2020.

[51] Oleksandr Shchur, Ali Caner Türkmen, Tim Januschowski, and Stephan Günnemann. Neural Temporal Point Processes: A Review. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 4585–4593, 2021.

[52] Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

[53] Bjørnar Vassøy, Massimiliano Ruocco, Eliezer de Souza da Silva, and Erlend Aune. Time is of the Essence: A Joint Hierarchical RNN and Point Process Model for Time and Item Predictions. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 591–599, 2019.

[54] Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

[55] Shuai Xiao, Mehrdad Farajtabar, Xiaojing Ye, Junchi Yan, Xiaokang Yang, Le Song, and Hongyuan Zha. Wasserstein Learning of Deep Generative Point Process Models. In *Advances in Neural Information Processing Systems*, pages 3247–3257, 2017.

[56] Siqiao Xue, Xiaoming Shi, Zhixuan Chu, Yan Wang, Hongyan Hao, Fan Zhou, Caigao Jiang, Chen Pan, James Y. Zhang, Qingsong Wen, Jun Zhou, and Hongyuan Mei. EasyTPP: Towards Open Benchmarking Temporal Point Processes. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.

[57] Yuan Yuan, Jingtao Ding, Chenyang Shao, Depeng Jin, and Yong Li. Spatio-Temporal Diffusion Point Processes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3173–3184, 2023.

[58] Qiang Zhang, Aldo Lipani, Omer Kirnap, and Emine Yilmaz. Self-Attentive Hawkes Process. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11183–11193, 2020.

[59] Zihao Zhou and Rose Yu. Automatic Integration for Spatiotemporal Neural Point Processes. In *Advances in Neural Information Processing Systems*, pages 50237–50253, 2023.

[60] Zihao Zhou, Xingyi Yang, Ryan Rossi, Handong Zhao, and Rose Yu. Neural Point Process for Learning Spatiotemporal Event Dynamics. In *Proceedings of the 4th Annual Learning for Dynamics and Control Conference*, pages 777–789, 2022.

[61] Shixiang Zhu, Shuang Li, Zhigang Peng, and Yao Xie. Imitation Learning of Neural Spatio-Temporal Point Processes. *IEEE Transactions on Knowledge and Data Engineering*, 34(11):5391–5402, 2022.

[62] Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer Hawkes Process. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 11692–11702, 2020.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction are supported by the theoretical work from Section 4 and the empirical experiments from Section 5 .

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work are discussed in a dedicated paragraph in Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All the theoretical results are presented clearly in Section 4 with all the assumptions summarised in the theorems. The complete proofs can be found in the extensive Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All the main practical information used to run the experiments are clearly stated in the main text. Further details are provided in Appendix G.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The source code for reproducing our experimental results can be obtained from a publicly accessible repository under an open access license: https://github.com/Code-Cornelius/sigtpp.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All the details about the architecture and the parameters we used to perform the experiments are given in the main text, and more details can be found in Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All tables report bootstrap standard errors, which quantify uncertainty in the reported results. Section 5 describes the experimental settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All the information about the resources used to run the experiments can be found in Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research presented in this paper is compliant with the NeurIPS code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Due to the theoretical nature of the paper, there is no direct broader impact to declare.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not pose such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the openly public datasets from EasyTPP [56], EdiFlow [36] and Add and Thin [34], that were used to carry parts of the numerical experiments, as well as the origin of the models used as baselines.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, documentation is provided in the linked repository: `https://github.com/Code-Cornelius/sigtpp`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The present research does not involve LLMs as an important component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Appendix

A A Pathwise Metric on Counting Paths

A.1 Distances induced by the Pathwise Metric on Counting Paths

Definition S1 (Metric of Negative Type)**.** *A metric space (X,d) is of negative type if, for every $n\in\mathbb{N}$, all points $x_1,\dots,x_n\in X$, and all real coefficients $a_1,\dots,a_n\in\mathbb{R}$ satisfying $\sum_{i=1}^n a_i=0$, the following inequality holds:*

$$\sum_{i,j=1}^{n} a_i a_j\, d(x_i,x_j)\le 0\,. \tag{S1}$$

Geometrically, a metric d is of negative type if and only if the metric space $(X,\sqrt{d})$ can be embedded isometrically into a Hilbert space [48], i.e. if there exists a map $\iota: X\to\mathcal{H}$ into a Hilbert space $\mathcal{H}$ such that

$$d(x,y)=\|\iota(x)-\iota(y)\|_{\mathcal{H}}^2\,. \tag{S2}$$

Proposition S2 (L^1 is of Negative Type [4, 15, 37])**.** *Let $(\Omega,\mathcal{F},\mu)$ be a measure space. Then the metric space $(L^1(\mu),\|\cdot\|_{L^1})$ is of negative type. Equivalently, for all $f_1,\dots,f_n\in L^1(\mu)$ and all $a_1,\dots,a_n\in\mathbb{R}$ such that $\sum_i a_i=0$, we have*

$$\sum_{i,j=1}^{n} a_i a_j\, \|f_i-f_j\|_{L^1}\le 0\,. \tag{S3}$$

Proposition S3 (Subspace Inheritance [4, 15])**.** *Let (X,d) be a metric space of negative type, and let $Y\subset X$. Then the restricted metric space $(Y,d|_{Y\times Y})$ is also of negative type.*

Theorem S4 (Negative Type Property)**.** *The metric space $(\mathcal{N},d_{\mathcal{N}})$ is of negative type.*

Proof. Identifying each $\eta\in\mathcal{N}$ with its counting path $t\mapsto\eta_t$ gives an isometric embedding of $(\mathcal{N},d_{\mathcal{N}})$ into $L^1([0,T_{\max}])$, since $d_{\mathcal{N}}(\eta,\xi)=\|\eta-\xi\|_{L^1}$. Proposition S2 states that L^1 is of negative type, and Proposition S3 shows that negative type is inherited by subspaces. Therefore, $(\mathcal{N},d_{\mathcal{N}})$ is of negative type. □

A.2 Direct Integration

We begin with the following exact discrete formulation of $d_{\mathcal{N}}$ for efficient computation.

Proposition S5 (Discrete Computation)**.** *Let $\eta,\xi\in\mathcal{N}$ with event times $(t_k)_{k=1}^m$ and $(s_k)_{k=1}^n$. Assuming without loss of generality that $m\le n$, the distance is given by*

$$d_{\mathcal{N}}(\eta,\xi)=\sum_{k=1}^{m}|t_k-s_k|+(n-m)T_{\max}-\sum_{k=m+1}^{n}s_k\,. \tag{S4}$$

Proof. We prove the discrete expression from the integral representation, as written in Eq. S4. Assuming $m\le n$, we prove that

$$d_{\mathcal{N}}(\eta,\xi)=\sum_{i=1}^{m}|t_i-s_i|+(n-m)T_{\max}-\sum_{i=m+1}^{n}s_i\,. \tag{S5}$$

Let

$$f(t):=\eta_t-\xi_t,\quad I(T_{\max}):=\int_0^{T_{\max}}|\eta_t-\xi_t|\,\mathrm{d}t=\int_0^{T_{\max}}|f(t)|\,\mathrm{d}t\,. \tag{S6}$$

f records the imbalance between the number of events of η and ξ that have occurred up to time t: it is constant between event times and changes by $+1$ whenever some t_i is crossed and by -1 whenever some s_j is crossed. Pairing events by rank and setting $t_i:=+\infty$ for $i=m+1,\dots,n$ (the process η has no further events), we may write

$$f(t)=\eta_t-\xi_t=\sum_{i=1}^{n}\left(\mathbb{1}_{\{t\ge t_i\}}-\mathbb{1}_{\{t\ge s_i\}}\right). \tag{S7}$$

The i-th summand equals $+1$ for $t \in [t_i, s_i)$ and -1 for $t \in [s_i, t_i)$. In either case its magnitude is the indicator of $[t_i \wedge s_i, t_i \vee s_i)$. Since both (t_i) and (s_i) are nondecreasing, two summands of opposite sign cannot be nonzero at the same t: if term k contributes $+1$ and term l contributes -1 with $k < l$, then $s_l \le t < s_k$ contradicts $s_k \le s_l$ (and symmetrically for $k > l$). Hence at every t the nonzero summands share a common sign, so the absolute value passes through the sum,

$$|f(t)| = \sum_{i=1}^{n} \left| \mathbb{1}_{\{t \ge t_i\}} - \mathbb{1}_{\{t \ge s_i\}} \right| = \sum_{i=1}^{n} \mathbb{1}_{[t_i \wedge s_i,\, t_i \vee s_i)}(t). \tag{S8}$$

Splitting off the unpaired indices $i > m$, for which $t_i = +\infty$ and hence $[t_i \wedge s_i, t_i \vee s_i) \cap [0, T_{\max}] = [s_i, T_{\max}]$, gives, for all $t \in [0, T_{\max}]$,

$$|f(t)| = \sum_{i=1}^{m} \mathbb{1}_{[t_i \wedge s_i,\, t_i \vee s_i)}(t) + \sum_{j=m+1}^{n} \mathbb{1}_{[s_j,\, T_{\max}]}(t). \tag{S9}$$

The first sum counts rank-paired times (t_i, s_i), for which exactly one of the two events has occurred by time t, while the second sum accounts for the unpaired events s_j with $j > m$. Integrating both sides of Eq. S9, we obtain

$$\begin{aligned} I(T_{\max}) &= \sum_{i=1}^{m} (t_i \vee s_i - t_i \wedge s_i) + \sum_{j=m+1}^{n} (T_{\max} - s_j) \\ &= \sum_{i=1}^{m} |t_i - s_i| + (n-m) T_{\max} - \sum_{j=m+1}^{n} s_j = d_{\mathcal{N}}(\eta, \xi)\,, \end{aligned} \tag{S10}$$

which establishes Eq. S4 for $m \le n$. The case $m > n$ follows by symmetry. □

A.3 Optimal Transport Interpretation

Let $\eta, \xi \in \mathcal{N}$ with event times $(t_k)_{k=1}^{m}$ and $(s_k)_{k=1}^{n}$ respectively. For the optimal transport argument, associate to each the point measure $\mu := \sum_{k=1}^{m} \delta_{t_k}$ and $\nu := \sum_{k=1}^{n} \delta_{s_k}$. We write $\delta_{T_{\max}}$ for the Dirac measure at $T_{\max}$, and use $(\cdot)^+ = \max(\cdot, 0)$.

We establish that the Wasserstein distance [47] for the deterministic measures with anchor points at $T_{\max}$ is indeed the L^1 distance between the respective counting processes:

$$\mathrm{W}_1(\tilde{\mu}, \tilde{\nu}) = \int_0^{T_{\max}} |\eta_t - \xi_t| \, \mathrm{d}t, \tag{S11}$$

where, assuming that $\eta_{T_{\max}} = m, \xi_{T_{\max}} = n$,

$$\tilde{\mu} := \mu + (n-m)^+ \delta_{T_{\max}}\,, \qquad \tilde{\nu} := \nu + (m-n)^+ \delta_{T_{\max}}\,. \tag{S12}$$

This establishes the correspondence between our definition and the optimal transport interpretation.

Proof. By construction, both padded measures have equal total mass. Since their supports are contained in $[0, T_{\max}]$, both $\tilde{\mu}$ and $\tilde{\nu}$ have finite first moments. The one-dimensional transport identity extends to finite positive measures of equal mass and finite first moment [54]:

$$\mathrm{W}_1(\tilde{\mu}, \tilde{\nu}) \triangleq \inf_{\gamma \in \Gamma(\tilde{\mu}, \tilde{\nu})} \int_{\mathbb{R} \times \mathbb{R}} |x - y| \, \gamma(\mathrm{d}x, \mathrm{d}y) = \int_{\mathbb{R}} |F_{\tilde{\mu}}(t) - F_{\tilde{\nu}}(t)| \, \mathrm{d}t. \tag{S13}$$

For $t < T_{\max}$, no padded atoms have appeared. Hence, $F_{\tilde{\mu}}(t) = \tilde{\mu}((-\infty, t]) = \mu((-\infty, t]) = \eta_t$, and analogously, $F_{\tilde{\nu}}(t) = \xi_t$. Thus,

$$|F_{\tilde{\mu}}(t) - F_{\tilde{\nu}}(t)| = \begin{cases} |\eta_t - \xi_t|, & t \in [0, T_{\max})\,, \\ 0, & t \notin [0, T_{\max})\,. \end{cases} \tag{S14}$$

Therefore, we get the equality

$$\mathrm{W}_1(\tilde{\mu}, \tilde{\nu}) = \int_{\mathbb{R}} |F_{\tilde{\mu}}(t) - F_{\tilde{\nu}}(t)| \, \mathrm{d}t = \int_0^{T_{\max}} |\eta_t - \xi_t| \, \mathrm{d}t. \tag{S15}$$

□

A.4 Counterexample for Strong Negative Type

The TPP distance is of negative type but not of strong negative type. Consider four realised TPPs with two events each:

$$\xi_1 = (0,2), \qquad \xi_2 = (1,3), \qquad \xi_3 = (0,3), \qquad \text{and} \qquad \xi_4 = (1,2). \tag{S16}$$

Then, with coefficients $a_1 = a_2 = 1$ and $a_3 = a_4 = -1$ (such that $\sum_i a_i = 0$) one has

$$\begin{aligned}\sum_{i,j=1}^{4} a_i a_j \, d_{\mathcal{N}}(\xi_i,\xi_j) &= 2\,d_{\mathcal{N}}(\xi_1,\xi_2) + 2\,d_{\mathcal{N}}(\xi_3,\xi_4) \\ &\quad - 2\cdot\big(d_{\mathcal{N}}(\xi_1,\xi_3) + d_{\mathcal{N}}(\xi_1,\xi_4) + d_{\mathcal{N}}(\xi_2,\xi_3) + d_{\mathcal{N}}(\xi_2,\xi_4)\big) \\ &= 2\cdot 2 + 2\cdot 2 - 2(1+1+1+1) \\ &= 0,\end{aligned} \tag{S17}$$

with the ξ_i not all equal, which violates strong negative type. In general, L^1 spaces are of negative type but not of strong negative type [37].

B Proofs of Theorem 6

Let $\mathcal{N}_\delta$ denote the counting paths whose interarrival times, including the terminal interval, are all at least $\delta > 0$. Following Theorem 6, we prove the following statements.

1. The map Φ is Lipschitz (Theorem S11).
2. The map Φ is injective on $\mathcal{N}$, hence a bijection between $\mathcal{N}$ and its image $\Phi(\mathcal{N})$ (Theorem S16).
3. The inverse map $\Psi = \Phi|_{\mathcal{N}}^{-1}$ is continuous on $\Phi(\mathcal{N}_\delta)$ for each $\delta > 0$. More precisely, it satisfies a $\frac{1}{2}$-Hölder bound on the separated subset $\Phi(\mathcal{N}_\delta)$ defined in Appendix B.3 (Theorem S29). Continuity fails on all of $\Phi(\mathcal{N})$. See Remark S17.

B.1 Φ is Lipschitz

Lemma S6 (Affine L^1 Bound)**.** *Let $a < b$ and let $h\colon [a,b] \to \mathbb{R}$ be an affine function. Then*

$$\int_a^b |h(t)|\,\mathrm{d}t \le \frac{b-a}{2}\big(|h(a)| + |h(b)|\big). \tag{S18}$$

Proof. The composition of a convex function with an affine function is convex. Since $t \mapsto |h(t)|$ is the composition of the convex function $|\cdot|$ with the affine function h, it is convex on $[a,b]$. Every convex function on an interval is bounded above by the chord through its endpoint values:

$$|h(t)| \le \frac{b-t}{b-a}\,|h(a)| + \frac{t-a}{b-a}\,|h(b)|, \qquad t \in [a,b]. \tag{S19}$$

Integrating over $[a,b]$ gives the result. □

We recall that a right-continuous function of bounded variation induces a Lebesgue-Stieltjes (total variation) measure $|Dy|$. See e.g. Folland [18], Sections 1.5 and 3.5.

Lemma S7 (Bounded Variation Time-Warp Inequality)**.** *Let $y\colon [0,T_{\max}] \to \mathbb{R}$ be (continuous[1] and) of bounded variation, with total variation $\mathrm{TV}(y) := \mathrm{TV}(y;[0,T_{\max}])$. Let $\lambda\colon [0,T_{\max}] \to [0,T_{\max}]$ be measurable and set $\|\lambda - \mathrm{Id}\|_\infty := \sup_{t\in[0,T_{\max}]} |\lambda(t) - t|$. Then*

$$\|y(\lambda(t)) - y(t)\|_{L^1} \le 2\,\|\lambda - \mathrm{Id}\|_\infty \,\mathrm{TV}(y). \tag{S20}$$

Proof. Let $|Dy|$ denote the total variation measure of y, i.e. the Lebesgue-Stieltjes measure associated with the total variation function $\mathrm{TV}(y;[0,t])$. For any $a,b \in [0,T_{\max}]$,

$$|y(b) - y(a)| \le |Dy|\big([a\wedge b,\, a\vee b]\big). \tag{S21}$$

[1] Continuity is not required for the proof but makes the argument cleaner, and we do not require a more general statement for our proofs.

Set $\delta := \|\lambda - \mathrm{Id}\|_\infty$ (smaller than or equal to $T_{\max}$). For each $t \in [0, T_{\max}]$ we have $\lambda(t) \in [t-\delta, t+\delta] \cap [0, T_{\max}]$, so that $[t \wedge \lambda(t), t \vee \lambda(t)] \subseteq [t-\delta, t+\delta] \cap [0, T_{\max}]$. By monotonicity of measures,

$$|y(\lambda(t)) - y(t)| \le |Dy|\left([t \wedge \lambda(t), t \vee \lambda(t)]\right) \le |Dy|\left([t-\delta, t+\delta] \cap [0, T_{\max}]\right). \tag{S22}$$

Integrating over $[0, T_{\max}]$ and writing the right-hand side as a double integral,

$$\int_0^{T_{\max}} |y(\lambda(t)) - y(t)|\,\mathrm{d}t \le \int_0^{T_{\max}} \int_{[0,T_{\max}]} \mathbb{1}_{\{|u-t| \le \delta\}}\, |Dy|(\mathrm{d}u)\,\mathrm{d}t. \tag{S23}$$

The integrand is nonnegative and jointly measurable (the set $\{(t,u) : |u-t| \le \delta\}$ is closed in $\mathbb{R}^2$), so Tonelli's theorem gives the desired inequality

$$\begin{aligned}\int_0^{T_{\max}} |y(\lambda(t)) - y(t)|\,\mathrm{d}t &\le \int_{[0,T_{\max}]} \left(\int_0^{T_{\max}} \mathbb{1}_{\{|u-t| \le \delta\}}\,\mathrm{d}t \right) |Dy|(\mathrm{d}u) \\ &\le \int_{[0,T_{\max}]} 2\delta\, |Dy|(\mathrm{d}u) = 2\delta\, \mathrm{TV}(y). \end{aligned} \tag{S24}$$

□

Lemma S8 (Interarrival Telescoping Bound). *Let $0 = t_0 < t_1 < \cdots < t_m < t_{m+1} = T_{\max}$ and $0 = s_0 < s_1 < \cdots < s_n < s_{n+1} = T_{\max}$ be two augmented grids (in the sense of Convention 1) with $m \le n$. Define the interarrival times*

$$\tau_k := t_k - t_{k-1}, \quad k = 1, \ldots, m+1, \qquad \varsigma_k := s_k - s_{k-1}, \quad k = 1, \ldots, n+1, \tag{S25}$$

and set $\tau_0 := \varsigma_0 := 0$. Extend the shorter grid by padding: $t_i := T_{\max}$ for $i = m+1, \ldots, n$, so that both sequences $(t_i)_{i=1}^n$ and $(s_i)_{i=1}^n$ have the same length.

Then the first $m+2$ interarrival differences are controlled by the event-time differences:

$$\sum_{k=0}^{m+1} |\tau_k - \varsigma_k| \le 2 \sum_{i=1}^{n} |t_i - s_i|. \tag{S26}$$

Proof. For $k = 1, \ldots, m+1$, the telescoping identity is

$$\tau_k - \varsigma_k = (t_k - t_{k-1}) - (s_k - s_{k-1}) = (t_k - s_k) - (t_{k-1} - s_{k-1}). \tag{S27}$$

The triangle inequality gives

$$|\tau_k - \varsigma_k| \le |t_k - s_k| + |t_{k-1} - s_{k-1}|. \tag{S28}$$

Summing over $k = 1, \ldots, m+1$ yields

$$\sum_{k=1}^{m+1} |\tau_k - \varsigma_k| \le \sum_{k=1}^{m+1} |t_k - s_k| + \sum_{k=0}^{m} |t_k - s_k|. \tag{S29}$$

Since $\tau_0 = \varsigma_0 = 0$ and $t_0 = s_0 = 0$, the boundary terms vanish. Noting that all summands are nonnegative, we obtain

$$\sum_{k=0}^{m+1} |\tau_k - \varsigma_k| \le 2 \sum_{k=1}^{m+1} |t_k - s_k| \le 2 \sum_{k=1}^{n} |t_k - s_k|. \tag{S30}$$

When $m = n$, the sum includes the extra term $|t_{n+1} - s_{n+1}| = T_{\max} - T_{\max} = 0$. □

Lemma S9 (Grid Mismatch Controlled by a Time-Change). *Let $\xi \in \mathcal{N}$ with n events and augmented grid $0 = s_0 < s_1 < \cdots < s_{n+1} = T_{\max}$, and let $y = \Phi(\xi) \in \mathcal{C}$ be the piecewise-linear interpolant through $(s_k, \varsigma_k)_{k=0}^{n+1}$, where $\varsigma_0 := 0$ and $\varsigma_k := s_k - s_{k-1}$ for $k = 1, \ldots, n+1$ are the interarrival times (Convention 1).*

Let $0 = t_0 < t_1 < \cdots < t_{m+1} = T_{\max}$ be a second augmented grid with $m \le n$. Define:

- *(a) the piecewise-linear map $\lambda \colon [0, T_{\max}] \to [0, s_{m+1}]$ by $\lambda(t_k) = s_k$ for $k = 0, \ldots, m+1$, affine on each $[t_k, t_{k+1}]$.*
- *(b) the re-gridded interpolant $\tilde{y}$ on the t-grid: $\tilde{y}(t_k) := \varsigma_k$ for $k = 0, \ldots, m+1$, affine between nodes.*

Extend the shorter grid by padding $t_i := T_{\max}$ for $i = m+1, \ldots, n$. Then:

(i) $\tilde{y}(t) = y(\lambda(t))$ *for all* $t \in [0, T_{\max}]$.

(ii) $\|\tilde{y} - y\|_{L^1} \le 4T_{\max} \|\lambda - \mathrm{Id}\|_\infty$.

(iii) $\|\lambda - \mathrm{Id}\|_\infty \le \sum_{k=1}^{n} |t_k - s_k|$.

Consequently,

$$\|\tilde{y} - y\|_{L^1} \le 4T_{\max} \sum_{k=1}^{n} |t_k - s_k| \,. \tag{S31}$$

Proof. Part (i): Identity $\tilde{y} = y \circ \lambda$. Fix $k \in \{0, \dots, m\}$ and $t \in [t_k, t_{k+1}]$. Since the grids are strictly increasing, $t_{k+1} - t_k > 0, k \le m$ and $s_{k+1} - s_k > 0, k \le n$. The fact that the grid is strictly increasing also ensures that on the interval $[t_k, t_{k+1}]$, the map λ is strictly increasing and affine with $\lambda(t_k) = s_k$ and $\lambda(t_{k+1}) = s_{k+1}$, hence

$$\lambda\big([t_k, t_{k+1}]\big) = [s_k, s_{k+1}] \,, \tag{S32}$$

so in particular $\lambda(t) \in [s_k, s_{k+1}]$ and $y(\lambda(t))$ is given by the k-th linear piece of y.

By definition of $\tilde{y}$,

$$\tilde{y}(t) = \varsigma_k + \frac{\varsigma_{k+1} - \varsigma_k}{t_{k+1} - t_k}(t - t_k) \,. \tag{S33}$$

Since λ is linear on $[t_k, t_{k+1}]$,

$$\lambda(t) = s_k + \frac{s_{k+1} - s_k}{t_{k+1} - t_k}(t - t_k) \,, \qquad \text{equivalently} \qquad \frac{\lambda(t) - s_k}{s_{k+1} - s_k} = \frac{t - t_k}{t_{k+1} - t_k} \,. \tag{S34}$$

Substituting this into the expression for y on $[s_k, s_{k+1}]$ yields

$$y(\lambda(t)) = \varsigma_k + \frac{\varsigma_{k+1} - \varsigma_k}{s_{k+1} - s_k}\big(\lambda(t) - s_k\big) = \varsigma_k + \frac{\varsigma_{k+1} - \varsigma_k}{t_{k+1} - t_k}(t - t_k) = \tilde{y}(t) \,, \tag{S35}$$

which proves (i).

Part (ii): L^1-bound via Lemma S7. The map $\lambda : [0, T_{\max}] \to [0, T_{\max}]$ is piecewise linear, hence measurable and y is piecewise linear hence continuous and of bounded variation, so all hypotheses of Lemma S7 are fulfilled. Eq. S20 together with Part (i) gives

$$\|\tilde{y} - y\|_{L^1} = \|y \circ \lambda - y\|_{L^1} \le 2\,\mathrm{TV}(y)\, \|\lambda - \mathrm{Id}\|_\infty \,. \tag{S36}$$

Since y is piecewise linear with nodes $\varsigma_0 = 0$, $\varsigma_k \ge 0$, and $\sum_{k=1}^{n+1} \varsigma_k = T_{\max}$,

$$\mathrm{TV}(y) = \sum_{k=0}^{n} |\varsigma_{k+1} - \varsigma_k| \le \sum_{k=0}^{n} (\varsigma_{k+1} + \varsigma_k) \le 2 \sum_{k=1}^{n+1} \varsigma_k = 2T_{\max} \,. \tag{S37}$$

This yields (ii):

$$\|\tilde{y} - y\|_{L^1} \le 4T_{\max} \|\lambda - \mathrm{Id}\|_\infty \,. \tag{S38}$$

Part (iii): Control of the time change. Since $t \mapsto \lambda(t) - t$ is affine on each interval $[t_k, t_{k+1}]$, the function $|\lambda(t) - t|$ is convex on each such interval. Therefore, its maximum over $[0, T_{\max}]$ is attained at a node. Because all terms are nonnegative, we may bound the maximum by the sum

$$\|\lambda - \mathrm{Id}\|_\infty = \max_{0 \le k \le m+1} |s_k - t_k| \le \sum_{k=0}^{m+1} |s_k - t_k| \,. \tag{S39}$$

The term at $k = 0$ vanishes since $s_0 = t_0 = 0$. Finally, using the padding convention $t_i := T_{\max}$ for $i = m+1, \dots, n$, the sum $\sum_{i=1}^{n} |t_i - s_i|$ contains all nonzero terms $|s_k - t_k|$ for $k = 1, \dots, m+1$. When $m = n$, the omitted endpoint term is $|t_{n+1} - s_{n+1}| = 0$. It also contains additional nonnegative terms for $i = m+2, \dots, n$. Hence,

$$\|\lambda - \mathrm{Id}\|_\infty \le \sum_{k=1}^{m+1} |s_k - t_k| \le \sum_{i=1}^{n} |t_i - s_i| \,, \tag{S40}$$

which proves (iii).

Conclusion. Combining (ii) and (iii) gives

$$\|\tilde{y}-y\|_{L^1} \le 4T_{\max}\sum_{i=1}^{n}|t_i-s_i|\,. \tag{S41}$$

□

Remark S10 (On the Map λ)**.** *When $m<n$, the map λ satisfies $\lambda(T_{\max})=s_{m+1}<T_{\max}$, so λ is not surjective onto $[0,T_{\max}]$. It aligns the grids only up to the $(m+1)$-th node and compresses the remaining time $[s_{m+1},T_{\max}]$ to the single point $\lambda(T_{\max})$. When $m=n$, λ is a bijection of $[0,T_{\max}]$.*

Theorem S11 (Global Lipschitz continuity of Φ)**.** *Equip $\mathcal{N}$ with $d_{\mathcal{N}}$ and $\mathcal{C}$ with $d_1(x,y):=\int_0^{T_{\max}}|x(t)-y(t)|\,\mathrm{d}t$. Then $\Phi\colon(\mathcal{N},d_{\mathcal{N}})\to(\mathcal{C},d_1)$ is globally Lipschitz:*

$$d_1\big(\Phi(\eta),\Phi(\xi)\big)\le 5T_{\max}\,d_{\mathcal{N}}(\eta,\xi)\qquad \textit{for all } \eta,\xi\in\mathcal{N}\,. \tag{S42}$$

Proof. Let $\eta\in\mathcal{N}$ have m events at times $(t_k)_{k=1}^m$ and $\xi\in\mathcal{N}$ have n events at times $(s_k)_{k=1}^n$, with $m\le n$ (without loss of generality, by symmetry of $d_{\mathcal{N}}$). Write $x:=\Phi(\eta)$, $y:=\Phi(\xi)$.

Construct the re-gridded interpolant $\tilde{y}$ on the t-grid with nodal values $\tilde{y}(t_k):=\varsigma_k$ for $k=0,\dots,m+1$, as in Lemma S9. By the triangle inequality,

$$\|x-y\|_{L^1}\le \underbrace{\|x-\tilde{y}\|_{L^1}}_{\text{(I): vertical}}+\underbrace{\|\tilde{y}-y\|_{L^1}}_{\text{(II): horizontal}}\,. \tag{S43}$$

Bound (I): Vertical mismatch (same grid, different nodal values). On each $[t_k,t_{k+1}]$, $x-\tilde{y}$ is affine. By Lemma S6,

$$\int_{t_k}^{t_{k+1}}|x(t)-\tilde{y}(t)|\,\mathrm{d}t\le\frac{\Delta_k}{2}\big(|\tau_k-\varsigma_k|+|\tau_{k+1}-\varsigma_{k+1}|\big)\,, \tag{S44}$$

where $\Delta_k:=t_{k+1}-t_k$. Summing over $k=0,\dots,m$, each $|\tau_j-\varsigma_j|$ appears with coefficient at most $T_{\max}/2$ (since $\Delta_{j-1}+\Delta_j\le T_{\max}$). Hence

$$\|x-\tilde{y}\|_{L^1}\le\frac{T_{\max}}{2}\sum_{k=0}^{m+1}|\tau_k-\varsigma_k|\le T_{\max}\sum_{k=1}^{n}|t_k-s_k|=T_{\max}\,d_{\mathcal{N}}(\eta,\xi)\,, \tag{S45}$$

where the second inequality applies Lemma S8.

Bound (II): Horizontal mismatch (different grids). Recall that $y=\Phi(\xi)$ is the piecewise-linear interpolant on the grid $0=s_0<s_1<\cdots<s_{n+1}=T_{\max}$ with nodal values $y(s_k)=\varsigma_k$, and that $\tilde{y}$ is the piecewise-linear interpolant on $0=t_0<t_1<\cdots<t_{m+1}=T_{\max}$ with the same nodal values $\tilde{y}(t_k)=\varsigma_k$ for $k=0,\dots,m+1$.

Define the piecewise-linear increasing map $\lambda\colon[0,T_{\max}]\to[0,T_{\max}]$ by $\lambda(t_k)=s_k$ for $k=0,\dots,m+1$, linear on each $[t_k,t_{k+1}]$. λ is the unique increasing piecewise-linear map sending until index m each knot t_k to s_k. Then Lemma S9 gives $\tilde{y}(t)=y(\lambda(t))$ for all t and

$$\|\tilde{y}-y\|_{L^1}\le 4T_{\max}\sum_{k=1}^{n}|t_k-s_k|=4T_{\max}\,d_{\mathcal{N}}(\eta,\xi)\,. \tag{S46}$$

Conclusion. Combining Eq. S43 with Bounds (I) and (II):

$$d_1\big(\Phi(\eta),\Phi(\xi)\big)\le(1+4)T_{\max}\,d_{\mathcal{N}}(\eta,\xi)=5T_{\max}\,d_{\mathcal{N}}(\eta,\xi)\,. \tag{S47}$$

□

Remark S12 (Sharpness of the Lipschitz Constant)**.** *The constant $5T_{\max}$ in Theorem S11 is not sharp. Several intermediate bounds (notably $\max\le\sum$ in Lemma S9(iii)) are deliberately coarse. The structurally relevant feature is the linear scaling $\mathcal{O}(T_{\max})$, which is natural: the L^1 metrics on both spaces integrate over $[0,T_{\max}]$.*

B.2 Φ is a Bijection

Definition S13 (Backshift Transform)**.** *For $x\in\mathcal{C}$, the backshift transform of x is*

$$f_x(t):=t-x(t)\,,\qquad t\in[0,T_{\max}]\,. \tag{S48}$$

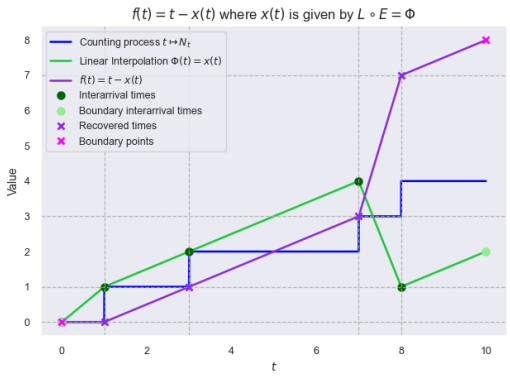


Figure S1: Recovery of event times with $f_x(t) = t - x(t)$ for $x = \Phi(\eta)$ by iterative level-hitting. Here, the TPP η is constructed artificially with times $t_k = [1,3,7,8]$, interarrival times $\tau_k = [1,2,4,1]$, and $T_{\max} = 10$. Starting from $t_0 = 0$, each event time is obtained as $t_k = \sup\{t \in (t_{k-1}, T_{\max}) : f_x(t) = t_{k-1}\}$. Horizontal dashed lines denote thresholds t_{k-1}, and vertical dashed lines mark the corresponding hit times t_k. The figure illustrates the constructive recovery underlying the bijectivity of Φ.

When $x = \Phi(\eta)$ for some $\eta \in \mathcal{N}$ with event times $0 < t_1 < \cdots < t_m < T_{\max}$ and interarrival times $\tau_k = t_k - t_{k-1}$ (Convention 1), the key observation is that $f_x(t_k) = t_k - \tau_k = t_{k-1}$. That is, f_x maps each event time to its predecessor. This is the basis for the level-set recovery of Lemma S14.

Figure S1 illustrates the level-set recursion for a single path. This construction detail supports the inverse proof and is included in the appendix rather than the main overview.

Lemma S14 (Level-Set Recovery)**.** *Let $\eta \in \mathcal{N}$ have event times $0 < t_1 < \cdots < t_m < T_{\max}$, augmented by $t_0 := 0$ and $t_{m+1} := T_{\max}$ (Convention 1), interarrival times $\tau_k := t_k - t_{k-1}$, and set $x := \Phi(\eta)$. If $m = 0$ then $f_x \equiv 0$ on $[0, T_{\max}]$. Assume henceforth $m \geq 1$. Then the backshift transform $f_x(t) = t - x(t)$ satisfies the following properties.*

(i) *f_x is continuous and piecewise-linear on $[0, T_{\max}]$.*

(ii) *$f_x(0) = 0$, $f_x(t_k) = t_{k-1}$ for $k = 1, \ldots, m$, and $f_x(T_{\max}) = t_m$.*

(iii) *$f_x \equiv 0$ on $[0, t_1]$.*

(iv) *f_x is strictly increasing on $[t_1, T_{\max}]$.*

(v) *For $k = 1, \ldots, m$, $t_k = \sup\{t \in (t_{k-1}, T_{\max}) : f_x(t) = t_{k-1}\}$.*

Proof. *(i)* Since $x \in \mathcal{C}$ is continuous and piecewise-linear on the grid $(t_k)_{k=0}^{m+1}$ and the map $t \mapsto t$ is linear, their difference f_x inherits the same properties.

(ii) At each event time t_k,

$$x(t_k) = \tau_k = t_k - t_{k-1}, \tag{S49}$$

hence

$$f_x(t_k) = t_k - \tau_k = t_{k-1}. \tag{S50}$$

At the origin, $x(0) = \tau_0 = 0$, so $f_x(0) = 0$.

(iii) On $[0, t_1]$, x has slope equal to 1 by definition and is linear with $x(0) = 0$. So $x(t) = t$ and $f_x \equiv 0$.

(iv) On each $[t_k, t_{k+1}]$ for $k = 1, \ldots, m$, the function f_x is affine with

$$f_x(t_{k+1}) - f_x(t_k) = t_k - t_{k-1} = \tau_k > 0, \tag{S51}$$

where strict positivity follows from simplicity of η (i.e. $\tau_k > 0$ for $k = 1, \ldots, m$). Since f_x is affine and strictly increasing on each subinterval and continuous across nodes, it is strictly increasing on all of $[t_1, T_{\max}]$.

(v) For $k=1$: by (iii), $f_x \equiv 0$ on $[0,t_1]$, and by (iv), $f_x > 0$ on $(t_1, T_{\max}]$. Hence $\{t \in (0,T_{\max}) : f_x(t) = 0\} = (0,t_1]$ and $\sup = t_1$. For $k \geq 2$: by (iv), f_x is strictly increasing on $[t_{k-1}, T_{\max}] \subset [t_1, T_{\max}]$. Since $f_x(t_k) = t_{k-1}$ and f_x is strictly increasing on this interval, t_k is the unique solution of $f_x(t) = t_{k-1}$ in $[t_{k-1}, T_{\max}]$, hence the supremum. □

Lemma S15 (Inverse Recovery)**.** *For $x \in \Phi(\mathcal{N})$, define the sequence $(t_k)_{k\geq 0}$ by*

$$t_0 := 0, \qquad t_k := \sup\{t \in (t_{k-1}, T_{\max}) : f_x(t) = t_{k-1}\}, \quad k \geq 1, \tag{S52}$$

and let $m \geq 0$ be the first index for which the level set $\{t \in (t_m, T_{\max}) : f_x(t) = t_m\}$ is empty or has supremum equal to $T_{\max}$. Then:

- *(i) The recovered sequence satisfies $0 < t_1 < t_2 < \cdots < t_m < T_{\max}$ and $m < \infty$. Moreover, $m = 0$ if and only if $x(t) = t$ for all t, the embedding of the eventless path.*
- *(ii) The map $\Psi : \Phi(\mathcal{N}) \to \mathcal{N}$ defined by*

$$\Psi(x)_t := \#\{k \in \{1,\ldots,m\} : t_k \leq t\}, \tag{S53}$$

  is a well-defined element of $\mathcal{N}$.
- *(iii) If $x = \Phi(\eta)$ for some $\eta \in \mathcal{N}$ with event times $(s_k)_{k=1}^n$, then $m = n$ and $t_k = s_k$ for $k = 1,\ldots,n$. Equivalently, $\Psi(\Phi(\eta)) = \eta$.*

Proof. Fix $\eta \in \mathcal{N}$ with event times $0 < s_1 < \cdots < s_n < T_{\max}$, augmented by $s_0 := 0$ and $s_{n+1} := T_{\max}$ (Convention 1), interarrival times $\varsigma_k := s_k - s_{k-1}$, and set $x := \Phi(\eta)$. If $n = 0$, then $x(t) = t$ and $f_x \equiv 0$, so the index-0 level set is $(0, T_{\max})$, whose supremum is $T_{\max}$: the recursion stops immediately with $m = 0$ and $\Psi(x) \equiv 0 = \eta$. Assume henceforth $n \geq 1$.

We show by induction that the recursion in Eq. S52 recovers $t_k = s_k$ for every k. Well-definedness of Ψ is then immediate.

Part (i): Strict monotonicity and finiteness. Base case ($k = 1$). On $[0, s_1]$, the path x interpolates linearly from $(0,0)$ to $(s_1, \varsigma_1) = (s_1, s_1)$, hence $x(t) = t$ and $f_x(t) = 0$ for all $t \in [0, s_1]$. By Lemma S14(iv), f_x is strictly increasing on $[s_1, T_{\max}]$, so $f_x(t) > 0$ for all $t > s_1$. Therefore

$$\{t \in (0, T_{\max}) : f_x(t) = 0\} = (0, s_1], \tag{S54}$$

and the supremum yields $t_1 = s_1$.

Inductive step. Assume $t_j = s_j$ for all $j \leq k-1$. By Lemma S14, the equation $f_x(t) = s_{k-1}$ has a unique solution in $(s_{k-1}, T_{\max})$, namely s_k. Hence $t_k = \sup\{s_k\} = s_k$, and in particular $t_k > t_{k-1}$. For each $k = 1,\ldots,n$ the level set has supremum $s_k < T_{\max}$, so the stopping rule does not trigger before index n.

Termination. It remains to show that no time is recorded after $t_n = s_n$. By Lemma S14(ii) to (iv), $f_x \leq f_x(s_n) = s_{n-1} < s_n$ on $[0, s_n]$, and f_x is strictly increasing on $[s_n, T_{\max}]$ with $f_x(T_{\max}) = s_n$. Hence $f_x(t) < s_n$ for every $t \in (s_n, T_{\max})$. The level set $\{t \in (s_n, T_{\max}) : f_x(t) = s_n\}$ is therefore empty, and the recursion terminates with $m = n < \infty$.

Part (ii): Well-definedness of $\Psi(x)$. By (i), the sequence $(t_k)_{k=1}^m$ is finite and strictly increasing. The process defined in Eq. S53 is therefore nondecreasing, càdlàg, has unit jumps only, satisfies $\Psi(x)_0 = 0$, and has $\Psi(x)_{T_{\max}} = m < \infty$. Hence $\Psi(x) \in \mathcal{N}$. When $m = 0$, $\Psi(x) \equiv 0$ is the eventless path, which lies in $\mathcal{N}$ trivially. □

Theorem S16 (Bijective Embedding)**.** *The following identities hold.*

- *(i) $\Psi \circ \Phi = \mathrm{Id}_{\mathcal{N}}$.*
- *(ii) $\Phi \circ \Psi = \mathrm{Id}_{\Phi(\mathcal{N})}$.*

In particular, $\Phi : \mathcal{N} \to \Phi(\mathcal{N})$ is a bijection with inverse Ψ.

Proof. We continue with the notation of the previous lemmas. First we prove $\Psi \circ \Phi = \mathrm{Id}_{\mathcal{N}}$. Let $\eta \in \mathcal{N}$ have event times $(t_k)_{k=1}^m$. Lemma S15 shows that applying the recursion to $x = \Phi(\eta)$ recovers all event times exactly, with no extra times. Therefore,

$$\Psi(\Phi(\eta))_t = \#\{k \in \{1,\ldots,m\} : t_k \leq t\} = \eta_t, \qquad t \in [0, T_{\max}]. \tag{S55}$$

Next we prove $\Phi \circ \Psi = \mathrm{Id}_{\Phi(\mathcal{N})}$. Let $x \in \Phi(\mathcal{N})$, so $x = \Phi(\eta)$ for some $\eta \in \mathcal{N}$ with event times $(t_k)_{k=1}^m$. By the previous statement, $\Psi(x)$ recovers exactly these event times. Reapplying Φ yields the same piecewise-linear interpolation through (t_k, τ_k), hence $\Phi(\Psi(x)) = x$.

The maps Φ and Ψ are therefore mutual inverses, and $\Phi : \mathcal{N} \to \Phi(\mathcal{N})$ is a bijection. □

B.3 Ψ is Continuous on Each Sieve $\Phi(\mathcal{N}_\delta)$

Hölder stability of the inverse. Fix $\delta > 0$ and let

$$\mathcal{N}_\delta := \{\eta \in \mathcal{N} : \tau_k \geq \delta \text{ for all } k = 1, \ldots, m+1\}, \tag{S56}$$

where $m = \eta_{T_{\max}}$ and $\tau_1, \ldots, \tau_{m+1}$ are the interarrival times (Convention 1).

The separation parameter δ excludes near-simultaneous events, which cause the embedding's slopes to become indistinguishable. In practice, temporal data recorded at finite precision satisfies $\eta \in \mathcal{N}_\delta$ for some resolution-dependent δ. The Lipschitz bound and bijectivity hold without any separation assumption.

Remark S17 (Separation is necessary)**.** *The inverse Ψ is not continuous on all of $\Phi(\mathcal{N})$, so the sieve $\mathcal{N}_\delta$ cannot be dispensed with. Let $\eta \in \mathcal{N}$ have event times $0 < t_1 < \cdots < t_m < T_{\max}$ (possibly $m = 0$, with the convention $t_1 := T_{\max}$), and for $0 < h < t_1$ let $\eta^{(h)} \in \mathcal{N}$ have event times $\{h, t_1, \ldots, t_m\}$. Then $d_{\mathcal{N}}(\eta^{(h)}, \eta) = T_{\max} - h$, while $\|\Phi(\eta^{(h)}) - \Phi(\eta)\|_\infty \leq h$, so $d_1(\Phi(\eta^{(h)}), \Phi(\eta)) \leq T_{\max} h \to 0$ as $h \downarrow 0$. Hence $\Phi(\eta^{(h)}) \to \Phi(\eta)$ in $(\mathcal{C}, d_1)$ while $d_{\mathcal{N}}(\eta^{(h)}, \eta) \to T_{\max}$, and Ψ admits no modulus of continuity at any point of $\Phi(\mathcal{N})$. On $\mathcal{N}_\delta$ this construction is unavailable: it forces the first interarrival time $\tau_1 = h < \delta$.*

Lemma S18 (Quantitative Bounds for f_x on the Sieve)**.** *Let $\delta > 0$, $M_\delta := T_{\max}/\delta$. Let $\eta \in \mathcal{N}_\delta$ have $m \geq 1$ events, and write $(t_k)_{k=0}^{m+1}$ for its augmented event-time grid and $(\tau_k)_{k=0}^{m+1}$ for the corresponding interarrival times (Convention 1), and set $x := \Phi(\eta)$, $f_x(t) := t - x(t)$. Then on each interval $[t_k, t_{k+1}]$, $k = 1, \ldots, m$, the function f_x is affine with slope*

$$r_k := \frac{\tau_k}{\tau_{k+1}} \in \left[M_\delta^{-1}, M_\delta\right]. \tag{S57}$$

In particular, f_x is M_δ-Lipschitz on $[0, T_{\max}]$.

Proof. Fix $k \in \{1, \ldots, m\}$. By construction, $x = \Phi(\eta)$ is affine on $[t_k, t_{k+1}]$ with values $x(t_k) = \tau_k$ and $x(t_{k+1}) = \tau_{k+1}$. Its slope is $(\tau_{k+1} - \tau_k)/\tau_{k+1}$, so $f_x = \mathrm{id} - x$ is affine on $[t_k, t_{k+1}]$ with slope

$$1 - \frac{\tau_{k+1} - \tau_k}{\tau_{k+1}} = \frac{\tau_k}{\tau_{k+1}}. \tag{S58}$$

Since $\eta \in \mathcal{N}_\delta$, every interarrival time lies in $[\delta, T_{\max}]$, giving $r_k = \tau_k/\tau_{k+1} \in [\delta/T_{\max}, T_{\max}/\delta] = [1/M_\delta, M_\delta]$.

Finally, f_x is continuous and piecewise affine. On $[0, t_1]$ the slope is 0. On each subsequent piece it is $r_k \leq M_\delta$. Hence f_x is M_δ-Lipschitz. □

Proposition S19 (Restricted Invertibility of the Backshift Transform on the Sieve)**.** *Let $\delta > 0$, $M_\delta := T_{\max}/\delta$. Let $\eta \in \mathcal{N}_\delta$ have $m \geq 1$ events, and write $(t_k)_{k=0}^{m+1}$ for its augmented event-time grid (Convention 1), and set $x := \Phi(\eta)$, $f_x(t) := t - x(t)$. Then:*

(i) The restriction $f_x\big|_{[t_1, T_{\max}]} : [t_1, T_{\max}] \to [0, t_m]$ is a continuous, strictly increasing bijection.

(ii) Its inverse $F_x := \left(f_x\big|_{[t_1, T_{\max}]}\right)^{-1}$ satisfies, for all $u, v \in [0, t_m]$,

$$|F_x(u) - F_x(v)| \leq M_\delta \, |u - v|. \tag{S59}$$

Proof. By Lemma S18, f_x is affine on each $[t_k, t_{k+1}]$ ($k = 1, \ldots, m$) with strictly positive slope $r_k \geq M_\delta^{-1}$. In particular, f_x is continuous and strictly increasing on $[t_1, T_{\max}]$, hence a bijection onto $[f_x(t_1), f_x(T_{\max})]$.

It remains to identify the endpoints. By Convention 1, $\tau_1 = t_1 - t_0 = t_1$ and $\tau_{m+1} = t_{m+1} - t_m = T_{\max} - t_m$. Since $x = \Phi(\eta)$ interpolates the nodes (t_k, τ_k), we have $x(t_1) = \tau_1 = t_1$ and $x(T_{\max}) = x(t_{m+1}) = \tau_{m+1} = T_{\max} - t_m$. Therefore,

$$f_x(t_1) = t_1 - x(t_1) = 0, \qquad f_x(T_{\max}) = T_{\max} - x(T_{\max}) = t_m, \tag{S60}$$

so the range is exactly $[0, t_m]$.

On each image interval $f_x([t_k, t_{k+1}])$, the inverse F_x is affine with slope $1/r_k \le M_\delta$. Since F_x is continuous and piecewise-affine on a compact interval with all slopes at most M_δ, it is globally M_δ-Lipschitz. □

Lemma S20 (L^1-L^∞ Interpolation for Lipschitz functions)**.** *Let $h : [0, T_{\max}] \to \mathbb{R}$ be L-Lipschitz with $h(0) = 0$ and $L > 0$. Then*

$$\|h\|_\infty^2 \le 4L \|h\|_1 , \tag{S61}$$

where $\|h\|_1 := \int_0^{T_{\max}} |h(t)| \, \mathrm{d}t$ and $\|h\|_\infty := \max_{t \in [0, T_{\max}]} |h(t)|$.

Proof. Let $M := \|h\|_\infty$ and choose $t^* \in [0, T_{\max}]$ with $|h(t^*)| = M$. If $M = 0$ the claim is trivial, so assume $M > 0$. Since $h(0) = 0$ and h is L-Lipschitz, $M = |h(t^*) - h(0)| \le L t^*$, so $t^* \ge M/L$. Consequently, the interval $[t^* - M/(2L), t^*]$ has length $M/(2L)$ and is contained in $[0, T_{\max}]$. For t in this interval, the Lipschitz bound gives

$$|h(t)| \ge |h(t^*)| - L|t - t^*| \ge M - L \cdot \frac{M}{2L} = \frac{M}{2} . \tag{S62}$$

Therefore

$$\|h\|_1 \ge \int_{t^* - M/(2L)}^{t^*} |h(t)| \, \mathrm{d}t \ge \frac{M}{2} \cdot \frac{M}{2L} = \frac{M^2}{4L} . \tag{S63}$$

□

Proposition S21 (Sup-norm Control)**.** *Let $\eta, \xi \in \mathcal{N}_\delta$ and set $x := \Phi(\eta)$, $y := \Phi(\xi)$, $g := x - y$. Assume $M_\delta := T_{\max}/\delta \ge 2$. When $M_\delta < 2$, the sieve $\mathcal{N}_\delta$ contains at most the eventless path, so the only available pair is $x = y$ and all pairwise bounds are trivial.*

(i) Each of x and y is $(M_\delta - 1)$-Lipschitz on $[0, T_{\max}]$, and g satisfies $g(0) = 0$ and is $2(M_\delta - 1)$-Lipschitz.

(ii) $\|g\|_\infty \le 2\sqrt{2(M_\delta - 1) \, d_1(x, y)} \le 2\sqrt{2 M_\delta \, d_1(x, y)}$.

Proof. We treat three cases according to the event counts of η and ξ.

Case 1: Both η and ξ have zero events. Then $x(t) = y(t) = t$ for all $t \in [0, T_{\max}]$, so $g \equiv 0$ and both claims hold trivially.

Case 2: Both η and ξ have at least one event. Let $m \ge 1$ denote the event count of η, with augmented grid $0 = t_0 < t_1 < \cdots < t_m < t_{m+1} = T_{\max}$ and interarrival times $\tau_k := t_k - t_{k-1}$.

Part (i): Lipschitz bound. By Convention 1, $x(0) = \tau_0 = 0$ and $y(0) = \varsigma_0 = 0$, so $g(0) = 0$.

We show $x = \Phi(\eta)$ is $(M_\delta - 1)$-Lipschitz. For $k \ge 1$, the slope of x on $[t_k, t_{k+1}]$ is

$$\alpha_k := \frac{\tau_{k+1} - \tau_k}{\tau_{k+1}} = 1 - \frac{\tau_k}{\tau_{k+1}} . \tag{S64}$$

Since $\eta \in \mathcal{N}_\delta$, every interarrival time lies in $[\delta, T_{\max}]$, so $\tau_k / \tau_{k+1} \in [M_\delta^{-1}, M_\delta]$ and hence

$$|\alpha_k| \le \max\big(|1 - M_\delta^{-1}|, \, |1 - M_\delta|\big) = M_\delta - 1 . \tag{S65}$$

For $k = 0$, the slope is $\alpha_0 = \tau_1 / \tau_1 = 1$. Since $m \ge 1$, we have $T_{\max} \ge (m+1)\delta \ge 2\delta$, so $M_\delta \ge 2$ and $|\alpha_0| = 1 \le M_\delta - 1$.

Thus x is $(M_\delta - 1)$-Lipschitz on $[0, T_{\max}]$. The same argument applies to y (which also has ≥ 1 event), so y is $(M_\delta - 1)$-Lipschitz. Their difference g is therefore $2(M_\delta - 1)$-Lipschitz.

Case 3: Exactly one of η, ξ has zero events. Without loss of generality, suppose η has $m = 0$ events and ξ has $n \ge 1$ events. Then $x(t) = \Phi(\eta)(t) = t$ for all $t \in [0, T_{\max}]$, so x has constant slope 1. Since ξ has $n \ge 1$ events, $T_{\max} \ge (n+1)\delta \ge 2\delta$, so $M_\delta \ge 2$. Therefore $|1| = 1 \le M_\delta - 1$, and x is $(M_\delta - 1)$-Lipschitz.

The function $y = \Phi(\xi)$ has $n \ge 1$ events in $\mathcal{N}_\delta$, so by the slope analysis in Case 2 (applied to ξ), y is also $(M_\delta - 1)$-Lipschitz.

Since $g(0)=x(0)-y(0)=0-0=0$ and g is the difference of two $(M_\delta-1)$-Lipschitz functions, g is $2(M_\delta-1)$-Lipschitz. This establishes Part (i).

Part (ii): Sup-norm bound (all cases). In all cases, Part (i) gives that g is L-Lipschitz with $L=2(M_\delta-1)>0$ (recall $M_\delta\geq 2$) and $g(0)=0$. By Lemma S20,

$$\|g\|_\infty^2 \leq 4L\,\|g\|_1 = 8(M_\delta-1)\,d_1(x,y). \tag{S66}$$

Taking square roots and using $M_\delta-1\leq M_\delta$ gives both bounds. □

Lemma S22 (Separation of Distinct-Count Images)**.** *Assume $M_\delta\geq 2$. Since the $m+1$ interarrival times of any $\eta\in\mathcal{N}_\delta$ with m events are each at least δ and sum to $T_{\max}$, we have $(m+1)\delta\leq T_{\max}$, i.e. $m\leq M_\delta-1$. As m is an integer, $m\leq\lfloor M_\delta\rfloor-1$. For $n\in\{0,\dots,\lfloor M_\delta\rfloor-1\}$, let $\mathcal{N}_\delta^{(n)}:=\{\eta\in\mathcal{N}_\delta:\eta_{T_{\max}}=n\}$ denote the subset of $\mathcal{N}_\delta$ with exactly n events. Each such class is nonempty (the equispaced grid $t_k=k\delta$, $k=1,\dots,n$, is admissible because $(n+1)\delta\leq\lfloor M_\delta\rfloor\delta\leq T_{\max}$), and these classes partition $\mathcal{N}_\delta$. For each pair $m\neq n$ in $\{0,\dots,\lfloor M_\delta\rfloor-1\}$,*

$$d_1\Big(\Phi(\mathcal{N}_\delta^{(m)}),\Phi(\mathcal{N}_\delta^{(n)})\Big)>0. \tag{S67}$$

Consequently,

$$d_{\min}(\delta):=\min_{\substack{m,n\in\{0,\dots,\lfloor M_\delta\rfloor-1\}\\ m\neq n}} d_1\Big(\Phi(\mathcal{N}_\delta^{(m)}),\Phi(\mathcal{N}_\delta^{(n)})\Big)>0. \tag{S68}$$

Proof. Fix $n\in\{0,\dots,\lfloor M_\delta\rfloor-1\}$. Identify $\mathcal{N}_\delta^{(n)}$ with the subset of $\mathbb{R}^n$

$$\Big\{(t_1,\dots,t_n)\in[0,T_{\max}]^n : t_k-t_{k-1}\geq\delta \text{ for } k=1,\dots,n,\, T_{\max}-t_n\geq\delta\Big\},\quad t_0=0, \tag{S69}$$

and equip $\mathcal{N}_\delta^{(n)}$ with the subspace topology inherited from $\mathbb{R}^n$. The defining constraints are finite intersections of closed half-spaces, hence the set is closed in the compact cube $[0,T_{\max}]^n$. Therefore $\mathcal{N}_\delta^{(n)}$ is compact (the case $n=0$ being trivial). Furthermore, on $\mathcal{N}_\delta^{(n)}$, $d_{\mathcal{N}}$ coincides with the ℓ^1 distance between vectors, and so do their topologies. Φ being Lipschitz with respect to $d_{\mathcal{N}}$ (Theorem S11), the restriction on $\mathcal{N}_\delta^{(n)}$ is continuous. Thus, $\Phi(\mathcal{N}_\delta^{(n)})$ is compact in $(\mathcal{C},d_1)$.

Injectivity of Φ implies that $\Phi(\mathcal{N}_\delta^{(m)})\cap\Phi(\mathcal{N}_\delta^{(n)})=\varnothing$ for $m\neq n$.

The continuous function $(x,y)\mapsto d_1(x,y)$ attains its minimum on the compact product of the two images $\Phi(\mathcal{N}_\delta^{(m)})\times\Phi(\mathcal{N}_\delta^{(n)})$. Since the sets are disjoint, this minimum is strictly positive. Since the index set $\{0,\dots,\lfloor M_\delta\rfloor-1\}$ is finite and, by $M_\delta\geq 2$, contains at least the two indices 0 and 1, the minimum in the definition of $d_{\min}(\delta)$ is taken over a finite, nonempty collection of strictly positive quantities. This yields $d_{\min}(\delta)>0$. □

Remark S23. *The lemma shows that under the minimal spacing condition, the event count becomes a topologically stable feature of the embedding, as distinct counts are separated by a strictly positive d_1-distance.*

Corollary S24 (Count stability)**.** *Assume $M_\delta\geq 2$ and let $\eta,\xi\in\mathcal{N}_\delta$ satisfy $d_1\big(\Phi(\eta),\Phi(\xi)\big)<d_{\min}(\delta)$, the separation constant of Lemma S22. Then η and ξ have the same number of events.*

Proof. This is immediate from Lemma S22: if the event counts differed, then $\Phi(\eta)$ and $\Phi(\xi)$ would lie in distinct image sets whose d_1-distance is at least $d_{\min}(\delta)$. □

Lemma S25 (Base-Case Perturbation)**.** *Let $\eta,\xi\in\mathcal{N}_\delta$ share event count $m\geq 1$, with first event times t_1 and s_1, respectively. Set $x:=\Phi(\eta)$, $y:=\Phi(\xi)$, $\epsilon:=\|x-y\|_\infty$, and $M_\delta:=T_{\max}/\delta$. Then*

$$|s_1-t_1|\leq M_\delta\epsilon. \tag{S70}$$

Proof. Assume $s_1\geq t_1$ (the reverse case follows by symmetry). By level-set recovery, $f_x(t_1)=0$ and $f_y(s_1)=0$. Since f_x has slope $\geq 1/M_\delta$ on $[t_1,T_{\max}]$ (Lemma S18),

$$f_x(s_1)\geq f_x(t_1)+\tfrac{1}{M_\delta}(s_1-t_1)=\tfrac{1}{M_\delta}|s_1-t_1|. \tag{S71}$$

Furthermore, $|f_x(s_1)-f_y(s_1)|=|f_x(s_1)-0|=f_x(s_1)\leq\|f_x-f_y\|_\infty=\|x-y\|_\infty=\epsilon$. □

Lemma S26 (Recursive Error Propagation)**.** *Let $\eta,\xi\in\mathcal{N}_\delta$ share event count $m\geq 1$, with augmented event-time grids $(t_k)_{k=0}^{m+1}$ and $(s_k)_{k=0}^{m+1}$, respectively. Set $x:=\Phi(\eta)$, $y:=\Phi(\xi)$, and $\epsilon:=\|x-y\|_\infty$. Suppose $\epsilon<\delta$, and $M_\delta:=T_{\max}/\delta$. Then for every $k=2,\dots,m$,*

$$|s_k-t_k|\leq M_\delta|s_{k-1}-t_{k-1}|+M_\delta\epsilon. \tag{S72}$$

Proof. By Proposition S19, the restricted inverses $F_x:=(f_x|_{[t_1,T_{\max}]})^{-1}$ and $F_y:=(f_y|_{[s_1,T_{\max}]})^{-1}$ are M_δ-Lipschitz, and the event times satisfy $t_k=F_x(t_{k-1})$, $s_k=F_y(s_{k-1})$ for $k\geq 2$.

Step 1: $t_k\geq s_1$. Suppose for contradiction that $t_k<s_1$. Then $f_y(t_k)=0$ (since $f_y\equiv 0$ on $[0,s_1]$), while by Lemma S14 $f_x(t_k)=t_{k-1}\geq t_1\geq\delta$. But $\|f_x-f_y\|_\infty=\epsilon<\delta$, a contradiction.

Step 2: $t_{k-1}\in[0,s_m]$. Since $f_x(T_{\max})=t_m$ and $f_y(T_{\max})=s_m$, we have $|s_m-t_m|=|(f_x-f_y)(T_{\max})|\leq\epsilon$. The δ-separation of η gives $t_{k-1}\leq t_m-\delta$. Combining:

$$t_{k-1}\leq t_m-\delta<t_m-\epsilon\leq s_m\,, \tag{S73}$$

where the second inequality uses $\epsilon<\delta$ and the third uses $s_m\geq t_m-\epsilon$. In other words $t_{k-1}\in[0,s_m]$ as required to evaluate $F_y(t_{k-1})$.

Step 3: Bounding $|F_y(t_{k-1})-F_x(t_{k-1})|$. By Step 1, $F_x(t_{k-1})=t_k\geq s_1$. Since $t_{k-1}\in[0,s_m]$ (Step 2), $F_y(t_{k-1})$ is well-defined and lies in $[s_1,T_{\max}]$ (the range of F_y). Thus both $F_y(t_{k-1})$ and $F_x(t_{k-1})$ lie in $[s_1,T_{\max}]$, where f_y has slope at least M_δ^{-1}, so

$$M_\delta^{-1}\,|F_y(t_{k-1})-F_x(t_{k-1})|\leq|f_y(F_y(t_{k-1}))-f_y(F_x(t_{k-1}))|\,. \tag{S74}$$

Since $f_y(F_y(t_{k-1}))=t_{k-1}=f_x(F_x(t_{k-1}))$ and $|f_x(F_x(t_{k-1}))-f_y(F_x(t_{k-1}))|\leq\|x-y\|_\infty=\epsilon$, we obtain $|F_y(t_{k-1})-F_x(t_{k-1})|\leq M_\delta\epsilon$.

Step 4: Triangle inequality.

$$\begin{aligned}|s_k-t_k|&=|F_y(s_{k-1})-F_x(t_{k-1})|\\&\leq\underbrace{|F_y(s_{k-1})-F_y(t_{k-1})|}_{\leq M_\delta|s_{k-1}-t_{k-1}|}+\underbrace{|F_y(t_{k-1})-F_x(t_{k-1})|}_{\leq M_\delta\epsilon}\\&\leq M_\delta|s_{k-1}-t_{k-1}|+M_\delta\epsilon.\end{aligned} \tag{S75}$$

□

Lemma S27 (Large-Perturbation Regime)**.** *Let $\eta,\xi\in\mathcal{N}_\delta$ share event count $m\geq 1$, and set $x:=\Phi(\eta)$, $y:=\Phi(\xi)$. Set $\epsilon:=\|x-y\|_\infty$ and suppose $\epsilon\geq\delta$. Then*

$$d_{\mathcal{N}}(\eta,\xi)\leq m\sqrt{8}M_\delta^{3/2}d_1(x,y)^{1/2}\,. \tag{S76}$$

Proof. Since both measures have m events in $[0,T_{\max})$, the discrete representation gives the trivial bound $d_{\mathcal{N}}(\eta,\xi)\leq mT_{\max}$. By Proposition S21, $d_1(x,y)\geq\epsilon^2/(8M_\delta)\geq\delta^2/(8M_\delta)$. Hence

$$d_{\mathcal{N}}(\eta,\xi)\leq mT_{\max}\cdot\frac{\sqrt{d_1(x,y)}}{\sqrt{d_1(x,y)}}\leq mT_{\max}\cdot\frac{\sqrt{8M_\delta}\,d_1(x,y)^{1/2}}{\delta}=m\sqrt{8}M_\delta^{3/2}d_1(x,y)^{1/2}. \tag{S77}$$

□

Proposition S28 (Same-Count Hölder Bound)**.** *Let $\eta,\xi\in\mathcal{N}_\delta$ have the same event count $m\geq 0$. Set $x:=\Phi(\eta)$, $y:=\Phi(\xi)$, $\epsilon:=\|x-y\|_\infty$, and $M_\delta:=T_{\max}/\delta$. Then*

$$d_{\mathcal{N}}(\eta,\xi)\leq C_m(M_\delta)\epsilon,\quad C_m(M_\delta):=\frac{M_\delta^2(M_\delta^m-1)}{(M_\delta-1)^2}\,, \tag{S78}$$

and consequently

$$d_{\mathcal{N}}(\eta,\xi)\leq 2\sqrt{2}C_m(M_\delta)M_\delta^{1/2}d_1(x,y)^{1/2}\leq\frac{4M_\delta^{m+3}}{(M_\delta-1)^2}d_1(x,y)^{1/2}. \tag{S79}$$

Proof. The case $m=0$ is trivial. Assume $m\geq 1$, then $M_\delta\geq m+1\geq 2$. Write $(t_k)_{k=0}^{m+1}$ and $(s_k)_{k=0}^{m+1}$ for the augmented event-time grids of η and ξ (Convention 1).

Step 1: $\epsilon < \delta$. Set $a_k := |s_k - t_k|$. The boundary convention gives $a_0 = 0$, and Lemmas S25 and S26 yield

$$a_k \le M_\delta a_{k-1} + M_\delta \epsilon, \quad k = 1, \dots, m. \tag{S80}$$

Unrolling from $a_0 = 0$,

$$a_k \le M_\delta \epsilon \sum_{j=0}^{k-1} M_\delta^j = \frac{M_\delta (M_\delta^k - 1)}{M_\delta - 1} \epsilon, \quad k = 1, \dots, m. \tag{S81}$$

By Eq. S4, because both processes share the same event count, $d_{\mathcal{N}}(\eta, \xi) = \sum_{k=1}^m a_k$. Summing:

$$\sum_{k=1}^m a_k \le \frac{M_\delta \epsilon}{M_\delta - 1} \sum_{k=1}^m (M_\delta^k - 1) = \frac{M_\delta \epsilon}{M_\delta - 1} \left(\frac{M_\delta (M_\delta^m - 1)}{M_\delta - 1} - m \right) \le \frac{M_\delta^2 (M_\delta^m - 1)}{(M_\delta - 1)^2} \epsilon, \tag{S82}$$

where the last inequality drops the term $-m < 0$. Substituting $\epsilon \le 2\sqrt{2 M_\delta d_1(x, y)} \le 4 M_\delta \sqrt{d_1(x, y)}$ (Proposition S21) into Eq. S78 yields Eq. S79. The final constant relaxes $M_\delta^m - 1$ to M_δ^m for a cleaner closed form.

Step 2: $\epsilon \ge \delta$. We first verify that Eq. S78 continues to hold in this regime. Since both processes have m events, Eq. S4 gives the trivial bound $d_{\mathcal{N}}(\eta, \xi) \le m T_{\max}$, while

$$C_m(M_\delta) \epsilon \ge C_m(M_\delta) \delta = \frac{M_\delta (M_\delta^m - 1)}{(M_\delta - 1)^2} T_{\max} \ge m T_{\max}, \tag{S83}$$

where the last inequality uses $\frac{M_\delta^m - 1}{M_\delta - 1} = 1 + M_\delta + \cdots + M_\delta^{m-1} \ge m$ together with $\frac{M_\delta}{M_\delta - 1} \ge 1$. Hence Eq. S78 holds unconditionally.

We now establish Eq. S79 in this regime. For $m \ge 1$ and $\epsilon \ge \delta$, by Lemma S27, $d_{\mathcal{N}}(\eta, \xi) \le m\sqrt{8} M_\delta^{3/2} d_1(x, y)^{1/2}$. Since $(M_\delta - 1)^2 \le M_\delta^2$, it suffices to show $m\sqrt{8} \le 4 M_\delta^{m - 1/2}$, i.e. $m \le \sqrt{2} M_\delta^{m - 1/2}$. We verify this by induction on m:

- *Base case* $m = 1$: need $1 \le \sqrt{2} M_\delta^{1/2}$, which holds since $M_\delta \ge 2$.
- *Inductive step*: if $m \le \sqrt{2} M_\delta^{m - 1/2}$, then $(m+1) \le 2m \le 2\sqrt{2} M_\delta^{m - 1/2} \le \sqrt{2} M_\delta^{m + 1/2}$ for $M_\delta \ge 2$, where the last inequality uses $2 M_\delta^{-1} \le 1$.

Thus, the constant holds for all $m \ge 1$ and when $M_\delta \ge 2$. Hence Eqs. S78 and S79 hold in both cases, completing the proof. □

Theorem S29 (Hölder stability of Φ^{-1})**.** *Let $\delta > 0$ and $M_\delta := T_{\max}/\delta$, and assume $M_\delta \ge 2$. When $M_\delta < 2$, the sieve $\mathcal{N}_\delta$ contains at most the eventless path and the statement holds trivially for any $\kappa > 0$. There exists a constant $\kappa = \kappa(\delta, T_{\max}) > 0$ such that for all $\eta, \xi \in \mathcal{N}_\delta$,*

$$d_{\mathcal{N}}(\eta, \xi) \le \kappa d_1\big(\Phi(\eta), \Phi(\xi)\big)^{1/2}. \tag{S84}$$

Explicitly, one may take

$$\kappa = \max\left(\frac{4 M_\delta^{\lfloor M_\delta \rfloor + 2}}{(M_\delta - 1)^2}, \frac{\lfloor M_\delta \rfloor T_{\max}}{\sqrt{d_{\min}(\delta)}} \right), \tag{S85}$$

where $d_{\min}(\delta) > 0$ is the separation constant from Lemma S22. Since $d_{\min}(\delta)$ is obtained by a compactness argument and is not quantified, we make no claim on the growth rate of κ as $\delta \downarrow 0$.

Proof. Let m and n denote the event counts of η and ξ, and set $x := \Phi(\eta)$, $y := \Phi(\xi)$.

Case 1: $m = n$. Proposition S28 gives $d_{\mathcal{N}}(\eta, \xi) \le C_{\text{rec}} d_1(x, y)^{1/2}$ with $C_{\text{rec}} := 4 M_\delta^{\lfloor M_\delta \rfloor + 2} / (M_\delta - 1)^2$, using $m = n \le \lfloor M_\delta \rfloor - 1$ (Lemma S22) together with $M_\delta \ge 2$ to bound $M_\delta^{m+3} \le M_\delta^{\lfloor M_\delta \rfloor + 2}$. When $m = 0$, $d_{\mathcal{N}}(\eta, \xi) = 0$ and the bound is trivial.

Case 2: $m \ne n$. By Lemma S22, $d_1(x, y) \ge d_{\min}(\delta) > 0$. Moreover, both event counts are at most $\lfloor M_\delta \rfloor - 1$ (Lemma S22), so Proposition S5 gives $d_{\mathcal{N}}(\eta, \xi) \le \max(m, n) T_{\max} \le \lfloor M_\delta \rfloor T_{\max}$. Hence

$$d_{\mathcal{N}}(\eta, \xi) \le \lfloor M_\delta \rfloor T_{\max} = \frac{\lfloor M_\delta \rfloor T_{\max}}{\sqrt{d_{\min}(\delta)}} \sqrt{d_{\min}(\delta)} \le \frac{\lfloor M_\delta \rfloor T_{\max}}{\sqrt{d_{\min}(\delta)}} d_1(x, y)^{1/2}. \tag{S86}$$

Taking $\kappa := \max\big(C_{\mathrm{rec}}, \lfloor M_\delta \rfloor T_{\max}/\sqrt{d_{\min}(\delta)}\big)$ completes the proof. □

Remark S30 (Interpretation and sharpening of the inverse constant)**.** *The same-count term in Eq. S85 satisfies $C_{\mathrm{rec}} = \mathcal{O}(M_\delta^{\lfloor M_\delta \rfloor})$. This is a conservative proof constant and does not establish instability of the inverse on empirical paths. Its exponential dependence comes from unrolling the worst-case recursive error propagation $a_k \le M_\delta\, a_{k-1} + M_\delta \epsilon$ in Proposition S28. Finite recording precision and the strictly increasing event-time convention ensure a positive minimum interarrival $\delta_{\mathrm{data}} > 0$ on each finite dataset used in this work, so the separated-sieve guarantee applies to the observed paths. If one additionally requires that consecutive slopes of f_x differ by at least some fixed $\gamma > 0$ (a slope-gap condition), a local, non-recursive argument gives the bound*

$$\frac{\gamma}{4\lfloor M_\delta \rfloor} d_{\mathcal{N}}(\eta,\xi)^2 \le d_1\big(\Phi(\eta), \Phi(\xi)\big)\,. \tag{S87}$$

This yields $\kappa = \mathcal{O}(\sqrt{\lfloor M_\delta \rfloor/\gamma})$. The argument goes along the following lines. When two breakpoint grids are shifted by a_k, interlacing ensures that on a neighbourhood of each t_k, one embedding has a kink while the other is linear. The slope-gap condition γ guarantees this kink has opening $\ge \gamma$, producing an L^1 contribution greater or equal to $\gamma/4\, a_k^2$. Since the neighbourhoods are disjoint (spacing larger than δ, shifts smaller or equal to $\delta/4$), the contributions sum independently and Cauchy-Schwarz converts $\sum a_k^2 \ge d_{\mathcal{N}}(\eta,\xi)^2/\lfloor M_\delta \rfloor$. No error propagation occurs, hence the polynomial dependence.

C Sampling and Training Procedures for SigTPP

The following algorithms describe the free-running temporal sampling and training procedures used by SigTPP (Section 4.2). Algorithm 1 defines how to generate each next interarrival from the previously generated history and fresh noise, subject to the observation horizon and rollout cap. Algorithm 2 calls this routine at each optimisation step, embeds the generated counting paths, and computes their truncated signatures for backpropagation.

Algorithm 1 Free-running sampling procedure for SigTPP. Indices start at 1 to match the augmented time grid of Definition 4, where $t_0 := 0$ denotes the start of the horizon.

Require: Generator G_θ, horizon $T_{\max}$, batch size B, maximum rollout length $L_{\max}$
Ensure: A batch of generated event sequences on $[0, T_{\max})$, each containing at most $L_{\max}$ events

procedure SampleFreeRunning($G_\theta, T_{\max}, B, L_{\max}$)
  for $b = 1, \dots, B$ **do**
    Set $t_{b,0} \leftarrow 0$ and $i \leftarrow 1$
    while $t_{b,i-1} < T_{\max} \wedge i \le L_{\max}$ **do**
      Draw $z_{b,i} \sim \mathrm{Exp}(1)$ and set $\varepsilon_{b,i} \leftarrow \log z_{b,i}$
      $\ell_{b,i} \leftarrow G_\theta(\tau_{b,1:i-1}, t_{b,i-1}, \varepsilon_{b,i})$
      $\tau_{b,i} \leftarrow \exp(\ell_{b,i})$ and $t_{b,i} \leftarrow t_{b,i-1} + \tau_{b,i}$
      $i \leftarrow i + 1$
    end while
    $n_b \leftarrow \max(\{j \in \{1, \dots, i-1\} : t_{b,j} < T_{\max}\} \cup \{0\})$
  end for
  return $\{\tau_{b,1:n_b}\}_{b=1}^{B}$
end procedure

D Evaluation Metrics

We evaluate generative TPP models along several complementary axes. Each metric targets a distinct structural property of the process, and together they provide a comprehensive picture of model quality. For all metrics a lower value is better.

Algorithm 2 Core free-running temporal optimisation for SIGTPP.

Require: Training set $\mathcal{D}$, horizon $T_{\max}$, signature degree M, optimisation steps N_{steps}, batch size B, maximum rollout length $L_{\max}$
Ensure: Trained generator parameters θ

```
Set N ← |D| and index D as (η_i)_{i=1}^N
(s_i)_{i=1}^N ← ( S^M(Φ̃(η_i)) )_{i=1}^N
Let μ_D and σ_D be the coordinate-wise mean and standard deviation of (s_i)_{i=1}^N
Initialise the generator G_θ and Adam optimiser
for k = 1, ..., N_steps do
    {τ_{b,1:n_b}}_{b=1}^B ← SAMPLEFREERUNNING(G_θ, T_max, B, L_max)
    X̂_b ← Φ̃(η̂_b), where η̂_b is induced by τ_{b,1:n_b}
    μ̂_θ ← B^{-1} Σ_{b=1}^B S^M(X̂_b)
    L_{Sig-W_1} ← ‖ (μ̂_θ − μ_D) / σ_D ‖_2^2
    Update θ with Adam using L_{Sig-W_1}
end for
```

D.1 Theory-Motivated Metrics

We define and theoretically justify 3 theory-motivated metrics in Section 4.4. We recall them briefly here and describe their empirical estimators. Throughout this section, we assume access to N_P real samples $\eta^{(1)},\dots,\eta^{(N_P)} \sim P$ and N_Q generated samples $\widehat{\xi}^{(1)},\dots,\widehat{\xi}^{(N_Q)} \sim Q$.

Signature Wasserstein-1 (Sig-W_1). Sig-W_1 measures discrepancy between path distributions through their expected truncated signatures, capturing higher-order temporal-dependence structure. It serves simultaneously as our training objective for SIGTPP and as an evaluation metric. Replacing the population expectations in Eq. 4 by sample means yields

$$\widehat{\text{Sig-W}_1}^{(M)}(P,Q) = \left\| \frac{1}{N_P}\sum_{i=1}^{N_P} S^M\big(\widetilde{\Phi}(\eta^{(i)})\big) - \frac{1}{N_Q}\sum_{i=1}^{N_Q} S^M\big(\widetilde{\Phi}(\widehat{\xi}^{(i)})\big) \right\|_2 . \tag{S88}$$

We report results at signature degree 3 with unscaled signature components. During training, signature features are standardised (common setting in linear regressions).

Energy Distance ($\mathcal{E}$). The energy distance is defined as

$$\mathcal{E} := 2\cdot\mathbb{E}\left[d_{\mathcal{N}}\left(N,\widetilde{N}\right)\right] - \mathbb{E}\left[d_{\mathcal{N}}\left(N,N'\right)\right] - \mathbb{E}\left[d_{\mathcal{N}}\left(\widetilde{N},\widetilde{N}'\right)\right], \quad P,Q\in\mathcal{P}_1(\mathcal{N}), \tag{S89}$$

with $N,N'\sim P$ and $\widetilde{N},\widetilde{N}'\sim Q$ independent [21, 52]. The energy distance is a distribution-level discrepancy on $\mathcal{P}_1(\mathcal{N},d_{\mathcal{N}})$ that compares the average inter-sample distance between P and Q against the within-distribution distances. Its empirical estimator is

$$\begin{aligned}\widehat{\mathcal{E}}(P,Q) = {} & \frac{2}{N_P N_Q}\sum_{i,j} d_{\mathcal{N}}\left(\eta^{(i)},\widehat{\xi}^{(j)}\right) - \frac{1}{N_P(N_P-1)}\sum_{i\neq i'} d_{\mathcal{N}}\left(\eta^{(i)},\eta^{(i')}\right) \\ & - \frac{1}{N_Q(N_Q-1)}\sum_{j\neq j'} d_{\mathcal{N}}\left(\widehat{\xi}^{(j)},\widehat{\xi}^{(j')}\right).\end{aligned} \tag{S90}$$

Although $\mathcal{E}\geq 0$ at the population level (since $d_{\mathcal{N}}$ is of negative type), the empirical estimator $\widehat{\mathcal{E}}(P,Q)$ can take small negative values in finite samples, especially when P and Q are close. We compute it on batches of 2,000 sequences so that all pairwise distances fit in memory. Remark that $\mathcal{E}$ reduces to the CRPS [46] in the univariate case, providing a natural score for unconditional generation that handles variable-length sequences.

Wasserstein-1 (W_1). Since $(\mathcal{N},d_{\mathcal{N}})$ is a metric space, the Wasserstein-1 distance [47, 54]

$$\mathrm{W}_1(P,Q) := \inf_{\pi\in\Pi(P,Q)} \int_{\mathcal{N}\times\mathcal{N}} d_{\mathcal{N}}(N,N')\,\mathrm{d}\pi(N,N'), \tag{S91}$$

where $\Pi(P,Q)$ denotes the set of couplings of P and Q, is a well-defined metric on $\mathcal{P}_1(\mathcal{N})$. W_1 lifts the path-level metric $d_{\mathcal{N}}$ to a metric between distributions. It is the minimum cost of transporting generated mass onto real mass, with per-pair cost $d_{\mathcal{N}}$. With empirical measures $\mu = \frac{1}{N_P}\sum_{i=1}^{N_P}\delta_{\eta^{(i)}}$ and $\nu = \frac{1}{N_Q}\sum_{j=1}^{N_Q}\delta_{\widehat{\xi}^{(j)}}$, it is estimated by solving

$$\widehat{W}_1(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \sum_{i=1}^{N_P}\sum_{j=1}^{N_Q}\pi_{ij}\, d_{\mathcal{N}}\Big(\eta^{(i)},\widehat{\xi}^{(j)}\Big)\,, \tag{S92}$$

where $\Pi(\mu,\nu)$ denotes the set of couplings between μ and ν. As for the energy distance, we solve this linear program exactly on batches of $2{,}000$ sequences, chosen so the full cost matrix fits in memory.

Continuous Ranked Probability Score (CRPS). CRPS evaluates the calibration and sharpness of 1-step-ahead predictive distributions over interarrival times. For each true event at position l in sequence n, the model defines a predictive distribution $F_{n,l}$ conditional on the true past. We approximate it by the empirical cumulative distribution function $\widehat{F}_{n,l}$ of the S Monte Carlo samples $\widehat{\mathcal{T}}_{:,n,l}$. We report the per-event average

$$\overline{\mathrm{CRPS}} = \frac{1}{\sum_{n=1}^{N_{\mathrm{seq}}}\ell_n}\sum_{n=1}^{N_{\mathrm{seq}}}\sum_{l=1}^{\ell_n}\mathrm{CRPS}\big(\widehat{F}_{n,l},\mathcal{T}_{n,l}\big)\,, \tag{S93}$$

where the standard energy-form definition for a sample y

$$\mathrm{CRPS}(\widehat{F},y) = \mathbb{E}_{X\sim\widehat{F}}\big[|X-y|\big] - \tfrac{1}{2}\,\mathbb{E}_{X,X'\sim\widehat{F}}\big[|X-X'|\big]\,, \tag{S94}$$

where X, X' are i.i.d. draws from $\widehat{F}$.

Scaling. For comparability across datasets with different time horizons, we report $\widehat{\mathcal{E}}/T_{\max}^2$ (resp. $\widehat{W}_1/T_{\max}^2$) in all tables. The remaining metrics operate at the per-event level on interarrival times. Let N_{seq} denote the number of evaluation sequences and ℓ_n the (unpadded) length of sequence n, and write $\ell_{\max} = \max_n \ell_n$. We collect interarrivals in $\mathcal{T}\in\mathbb{R}^{N_{\mathrm{seq}}\times\ell_{\max}}$ and S Monte Carlo samples of generated interarrivals in $\widehat{\mathcal{T}}\in\mathbb{R}^{S\times N_{\mathrm{seq}}\times\ell_{\max}}$, where entries beyond each sequence's true length are set to `NaN`. All estimators below use pairwise deletion of `NaN` entries (per-lag for autocorrelations). We write $\mathcal{T}_{n,l}$ for the l-th interarrival of sequence n and $\widehat{\mathcal{T}}_{:,n,l}\in\mathbb{R}^S$ for the vector of S Monte Carlo predictions at position (n,l).

D.2 Heuristic Metrics

The following metrics target specific structural properties, second-order temporal dependence and marginal distributional shape, without formal guarantees. They are adapted from Ni et al. [41] to the TPP setting.

Pairwise Correlation Discrepancy (PCD). PCD assesses whether the model reproduces second-order dependence across event positions: it compares the Pearson correlation matrix of interarrival times at different positions in real vs. generated samples, catching models whose marginals are correct but whose cross-position structure is not. Writing $\mathcal{T}_{:,p}$ for the vector of position-p interarrivals across the N_{seq} sequences (and likewise, we slice 1 Monte Carlo sample $\widehat{\mathcal{T}}_{0,:,p}$), let

$$\rho_{p,q} = \mathrm{Corr}\big(\mathcal{T}_{:,p},\mathcal{T}_{:,q}\big)\,,\qquad \widehat{\rho}_{p,q} = \mathrm{Corr}\big(\widehat{\mathcal{T}}_{0,:,p},\widehat{\mathcal{T}}_{0,:,q}\big)\,, \tag{S95}$$

denote the Pearson correlations, computed by pairwise deletion of `NaN` entries. Then

$$\mathrm{PCD} = \frac{1}{|\mathcal{V}|}\sum_{(p,q)\in\mathcal{V}}\big|\rho_{p,q}-\widehat{\rho}_{p,q}\big|\,, \tag{S96}$$

where $\mathcal{V}$ collects position pairs (p,q) with $1\le p<q\le\ell^\star$ for which both correlations are well-defined. To stabilise estimation, we restrict $\ell^\star$ to the largest prefix length such that every position $p\le\ell^\star$ retains at least 50 valid (non-padded) samples in both $\mathcal{T}$ and $\widehat{\mathcal{T}}$.

Autocorrelation Discrepancy (ACD). ACD probes the temporal memory of the process, how strongly an interarrival depends on those a few events earlier. For lag $\tau\ge 1$, let $\mathrm{AC}_\tau(\mathcal{T})$ denote the lag-τ sample autocorrelation of interarrivals, computed by pooling all valid pairs $(\mathcal{T}_{n,l},\mathcal{T}_{n,l+\tau})$ across sequences (with `NaN`-padded positions excluded), and define $\mathrm{AC}_\tau(\widehat{\mathcal{T}})$ analogously by slicing one

Monte Carlo sample. We restrict to short lags, where temporal dependence is most informative and estimation is most reliable: we set $\tau_{\max} = 5$ in general, and $\tau_{\max} = 10$ on long-sequence datasets ($\ell_{\max} \geq 100$). We report

$$\mathrm{ACD} = \frac{1}{\tau_{\max}} \sum_{\tau \leq \tau_{\max}} \left| \mathrm{AC}_\tau(\mathcal{T}) - \mathrm{AC}_\tau(\widehat{\mathcal{T}}) \right|. \tag{S97}$$

Histogram Distances ($L_{\log(\tau)}$, L_λ). These metrics compare marginal distributional shape at two complementary levels: $L_{\log(\tau)}$ targets the interarrival times density (how long between events?) and L_λ the event-count density on $[0, T_{\max}]$ (when do events happen?). Both are L^1 distances between empirical histograms computed on B shared bins. B in our setting follows a cube-root heuristic $2n^{1/3}$. For $L_{\log(\tau)}$, let $\widehat{p}_{\mathrm{true}}$ and $\widehat{p}_{\mathrm{gen}}$ be histogram estimators of the marginal density of $u = \log\tau$, where τ is an interarrival time, on bins covering the empirical range of $\log\tau$ in $\mathcal{T}$:

$$L_{\log(\tau)} = \int |\widehat{p}_{\mathrm{true}}(u) - \widehat{p}_{\mathrm{gen}}(u)| \, \mathrm{d}u, \qquad u = \log\tau. \tag{S98}$$

For L_λ, let $\widehat{\lambda}^{\mathrm{emp}}_{\mathrm{true}}$ and $\widehat{\lambda}^{\mathrm{emp}}_{\mathrm{gen}}$ be histogram estimators of the empirical event rate per unit time on $[0, T_{\max}]$ (distinct from the conditional intensity used elsewhere in the paper):

$$L_\lambda = \int_0^{T_{\max}} \left| \widehat{\lambda}^{\mathrm{emp}}_{\mathrm{true}}(s) - \widehat{\lambda}^{\mathrm{emp}}_{\mathrm{gen}}(s) \right| \mathrm{d}s. \tag{S99}$$

D.3 1-Step-Ahead Point Estimates (MAE)

MAE and MSE evaluate the accuracy of 1-step-ahead point predictions of interarrival times. For each true event at position l in sequence n, the model defines a predictive distribution $F_{n,l}$ conditional on the true past, from which we draw the S Monte Carlo samples $\widehat{\mathcal{T}}_{:,n,l} \in \mathbb{R}^S$ already used for CRPS. The natural point predictors are the median of $F_{n,l}$ for MAE and its mean for MSE, which we estimate by their sample counterparts

$$\widehat{x}^{\mathrm{MAE}}_{n,l} = \mathrm{median}\big(\widehat{\mathcal{T}}_{:,n,l}\big), \qquad \widehat{x}^{\mathrm{MSE}}_{n,l} = \frac{1}{S} \sum_{s=1}^{S} \widehat{\mathcal{T}}_{s,n,l}. \tag{S100}$$

We report the per-event averages

$$\mathrm{MAE}_{\mathrm{med}} = \frac{1}{\sum_{n=1}^{N_{\mathrm{seq}}} \ell_n} \sum_{n=1}^{N_{\mathrm{seq}}} \sum_{l=1}^{\ell_n} \left| \widehat{x}^{\mathrm{MAE}}_{n,l} - \mathcal{T}_{n,l} \right|, \qquad \mathrm{MSE}_{\mathrm{mean}} = \frac{1}{\sum_{n=1}^{N_{\mathrm{seq}}} \ell_n} \sum_{n=1}^{N_{\mathrm{seq}}} \sum_{l=1}^{\ell_n} \big(\widehat{x}^{\mathrm{MSE}}_{n,l} - \mathcal{T}_{n,l} \big)^2. \tag{S101}$$

The sample median used in $\widehat{x}^{\mathrm{MAE}}_{n,l}$ is a consistent but finite-S biased plug-in for the population median of $F_{n,l}$. We take S large enough that this bias is negligible relative to the sequence-level Monte Carlo error.

We also report the 1-step-ahead forecasting error:

$$\widehat{\mathrm{MAE}} = \frac{1}{\sum_{n=1}^{N_{\mathrm{seq}}} \ell_n} \sum_{n=1}^{N_{\mathrm{seq}}} \sum_{l=1}^{\ell_n} \left(\frac{1}{S} \sum_{s=1}^{S} \left| \widehat{\mathcal{T}}_{s,n,l} - \mathcal{T}_{n,l} \right| \right). \tag{S102}$$

D.4 Relative Score

Let $s_A^{(i)}$ denote the score of model A on dataset $i \in \{1, \ldots, D\}$, and $s_{\mathrm{ref}}^{(i)}$ the score of a fixed reference model. The normalised score of model A is the geometric mean of per-dataset ratios:

$$\bar{s}_A = \left(\prod_{i=1}^{D} \frac{s_A^{(i)}}{s_{\mathrm{ref}}^{(i)}} \right)^{\frac{1}{D}} = \exp\left(\frac{1}{D} \sum_{i=1}^{D} \log \frac{s_A^{(i)}}{s_{\mathrm{ref}}^{(i)}} \right). \tag{S103}$$

A value $\bar{s}_A < 1$ indicates that A outperforms the reference on average. Crucially, comparing two models A and B via their normalised score ratio is independent of the choice of reference, i.e.

$$\frac{\bar{s}_A}{\bar{s}_B} = \frac{\left(\prod_i s_A^{(i)} / s_{\mathrm{ref}}^{(i)} \right)^{1/D}}{\left(\prod_i s_B^{(i)} / s_{\mathrm{ref}}^{(i)} \right)^{1/D}} = \left(\prod_{i=1}^{D} \frac{s_A^{(i)}}{s_B^{(i)}} \right)^{\frac{1}{D}}, \tag{S104}$$

where the $s_{\text{ref}}^{(i)}$ cancel [17].

E Additional Results

E.1 Performance Comparison

Table S1 completes the 12-dataset benchmark with 3 datasets: a self-correcting process (SC), a stationary renewal process (SR), and Yelp-Airport (YLA). The main text considers them when reporting results. We enumerate them here for space reasons. Appendix F describes them, Table S11 gives their statistics, and Figures S17, S18, and S23 show their diagnostics.

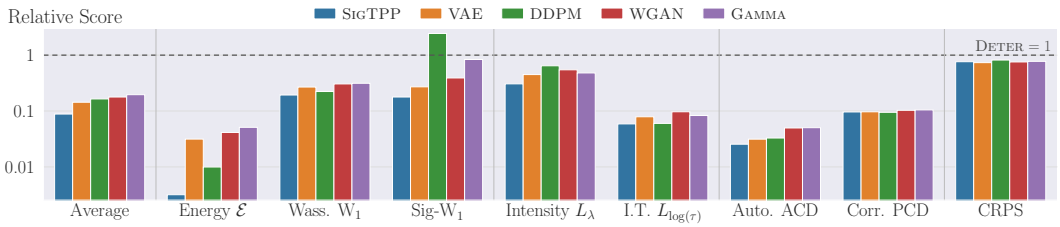


Figure S2: Geometric means of per-dataset score ratios to DETER over the 6 synthetic datasets. The vertical axis is logarithmic. The dashed line marks DETER $= 1$. Lower is better ($\downarrow$). The exact values are enumerated in Table S6. Metric definitions are provided in Appendix D.

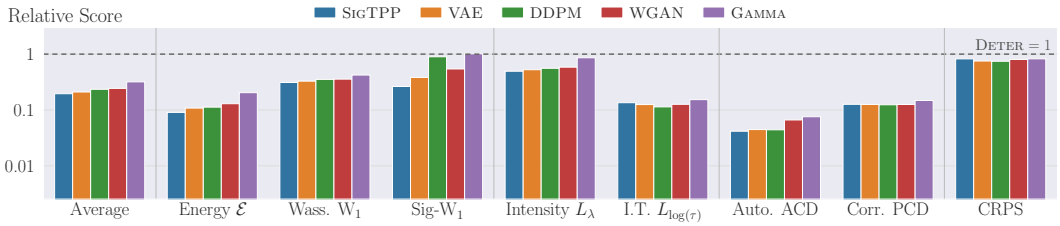


Figure S3: Geometric means of per-dataset score ratios to DETER over the 6 real-world datasets. The vertical axis is logarithmic. The dashed line marks DETER $= 1$. Lower is better ($\downarrow$). The exact values are enumerated in Table S6. Metric definitions are provided in Appendix D.

E.2 Comparison to EasyTPP

We also compare SIGTPP with 5 methods provided by the EasyTPP benchmark [56]: the recurrent Neural Hawkes Process (NHP) [39] and Recurrent Marked Temporal Point Process (RMTPP) [16], the attention-based Transformer Hawkes Process (THP) [62] and Self-Attentive Hawkes Process (SAHP) [58], and the continuous-time State-Space Point Process (S2P2) [5]. These models retain their likelihood training objectives, but we select the best model on the same criterion as the other method for fairness.

Computational constraints limit this comparison to half of the 12 benchmark datasets: homogeneous Poisson (PS), inhomogeneous Poisson (IP), univariate Hawkes (H1), Earthquake (EQ), Yelp-Airport (YLA), and Taobao (TB). Appendix F describes the datasets. We exclude the S2P2 run on TB because thinning exceeded the GPU time budget. Table S7 reports mean test scores with standard errors in parentheses. These results do not enter the main benchmark aggregates. Nonetheless, the findings are in line with the main results.

SIGTPP achieves the lowest geometric mean relative score on 6 of the 8 metrics (Figure S12), and the best overall average rank of 2.40, followed by NHP at 2.94, see Table S8. The EasyTPP baselines obtain lower aggregate CRPS, consistent with the conditional predictive advantage observed in the main comparison.

Table S1: Performance of SIGTPP and baselines across all 12 datasets (including the 3 extra datasets not shown in the main text). Values report standard error in parentheses over 100 bootstrap replicates. Lower is better (↓). Best in **bold**, second best underlined. Metric definitions in Appendix D.

Metric	Model	PS	IP	H1	H3	SC	SR	EQ	SO	TB	TX	YLP	YLA
$\mathcal{E}$ ($\times 10^{-3}$)	SIGTPP	**0.53** (0.93)	**0.42** (0.32)	**1.16** (0.74)	**1.32** (1.60)	1.01 (0.25)	**2.18** (2.09)	32.4 (0.65)	12.0 (1.71)	512 (25.2)	**76.3** (2.95)	38.4 (14.9)	**7.77** (4.47)
	VAE	8.64 (3.47)	2.24 (0.64)	16.8 (3.85)	20.0 (6.11)	1.51 (0.37)	73.19 (14.4)	34.5 (0.64)	**9.64** (2.55)	545 (24.7)	77.7 (2.73)	**33.6** (17.6)	27.89 (13.9)
	DDPM	2.72 (1.11)	2.56 (0.63)	10.7 (2.73)	1.61 (1.35)	**0.76** (0.16)	7.93 (4.93)	**26.0** (0.66)	24.5 (6.37)	**455** (21.5)	76.3 (2.77)	57.5 (20.5)	13.08 (6.52)
	WGAN	4.99 (1.83)	3.20 (1.03)	16.9 (3.93)	11.4 (4.00)	17.66 (1.58)	68.33 (14.9)	32.6 (0.71)	26.5 (3.34)	502 (22.4)	81.9 (2.80)	49.2 (24.3)	22.22 (12.4)
	DETER	330 (12.8)	108 (4.85)	510 (15.7)	640 (25.3)	22.62 (1.15)	2824 (44.0)	47.2 (0.92)	478 (22.6)	958 (32.1)	161 (5.46)	2402 (90)	984.2 (48.2)
	GAMMA	4.38 (2.58)	20.3 (3.22)	17.3 (3.87)	25.5 (5.92)	24.93 (2.13)	13.54 (3.40)	35.8 (0.66)	31.4 (3.78)	539 (23.5)	85.6 (2.99)	250 (59.0)	45.93 (15.7)
W_1 ($\times 10^{-2}$)	SIGTPP	**5.26** (0.20)	**6.55** (0.14)	**3.75** (0.31)	**10.1** (0.42)	2.55 (0.05)	**10.25** (0.60)	**2.91** (0.04)	7.44 (0.42)	58.2 (2.56)	8.02 (0.22)	**15.0** (1.86)	**8.48** (0.73)
	VAE	7.12 (0.64)	6.76 (0.18)	6.84 (0.57)	13.8 (1.11)	**2.42** (0.04)	22.31 (1.99)	2.91 (0.05)	**7.08** (0.76)	59.8 (2.40)	8.13 (0.23)	16.5 (2.90)	10.83 (1.45)
	DDPM	6.15 (0.24)	7.27 (0.15)	5.93 (0.49)	10.5 (0.50)	2.48 (0.03)	11.80 (0.95)	3.02 (0.05)	10.9 (1.22)	**52.8** (2.31)	**7.74** (0.18)	20.0 (2.06)	10.27 (0.76)
	WGAN	6.43 (0.40)	7.68 (0.29)	6.97 (0.61)	12.8 (0.83)	5.12 (0.18)	23.36 (2.44)	2.99 (0.04)	10.3 (0.64)	56.1 (2.07)	8.38 (0.27)	18.8 (2.78)	10.76 (1.48)
	DETER	28.3 (0.65)	18.2 (0.34)	38.1 (0.62)	54.6 (1.10)	3.66 (0.08)	166.21 (1.85)	3.74 (0.07)	36.2 (0.80)	86.0 (2.48)	14.1 (0.40)	143 (3.55)	61.20 (1.98)
	GAMMA	6.33 (0.54)	9.50 (0.49)	7.05 (0.59)	15.1 (1.07)	7.27 (0.23)	13.24 (0.90)	3.11 (0.05)	11.0 (0.66)	60.3 (2.15)	8.79 (0.28)	35.2 (3.68)	12.64 (1.56)
Sig-W_1 ($\times 10^{-5}$)	SIGTPP	**1.42** (1.52)	1.20 (1.05)	**44.0** (29.2)	1.02 (0.75)	2.81 (2.29)	**2.32** (1.91)	409 (36.3)	**0.09** (0.10)	2.27 (0.94)	4.74 (1.22)	0.77 (0.86)	4.64 (4.30)
	VAE	13.4 (6.42)	**0.83** (0.89)	155 (28.1)	**0.25** (0.26)	**0.57** (0.62)	25.37 (4.20)	397 (38.1)	3.70 (0.52)	**0.45** (0.30)	9.63 (1.61)	1.08 (1.10)	**2.01** (2.33)
	DDPM	129 (24.3)	64.0 (9.18)	302 (65.9)	10.9 (2.70)	19.35 (5.75)	6.37 (2.06)	1144 (56)	2.58 (0.44)	9.78 (1.99)	5.52 (1.21)	1.70 (1.08)	8.39 (5.68)
	WGAN	17.4 (6.92)	1.01 (0.70)	141 (30.9)	0.80 (0.70)	1.59 (1.19)	17.56 (4.25)	603 (45.9)	3.95 (0.64)	7.81 (1.44)	6.80 (1.51)	**0.42** (0.47)	2.11 (2.45)
	DETER	62.7 (12.7)	8.01 (1.76)	862 (112)	1.09 (0.39)	2.84 (2.07)	11.69 (3.14)	**146** (14.5)	0.19 (0.16)	49.7 (4.45)	24.5 (2.82)	7.78 (0.61)	16.56 (7.03)
	GAMMA	5.76 (4.25)	23.8 (4.84)	284 (51.3)	0.63 (0.29)	4.94 (1.64)	43.94 (9.84)	572 (38.1)	6.66 (0.81)	12.3 (3.47)	**3.07** (1.16)	10.6 (1.40)	3.11 (1.96)
CRPS ($\times 10^{-1}$)	SIGTPP	4.64 (0.08)	3.04 (0.03)	10.2 (0.27)	2.40 (0.03)	3.69 (0.03)	4.42 (0.10)	5.61 (0.11)	5.43 (0.10)	0.37 (0.01)	1.13 (0.02)	3.22 (0.13)	4.78 (0.14)
	VAE	**4.56** (0.08)	3.07 (0.03)	**9.53** (0.26)	2.33 (0.03)	**3.37** (0.02)	**4.30** (0.10)	5.13 (0.12)	4.64 (0.09)	0.36 (0.01)	**1.06** (0.02)	**2.74** (0.12)	4.67 (0.14)
	DDPM	4.81 (0.09)	3.27 (0.04)	10.1 (0.30)	**2.23** (0.03)	5.50 (0.05)	4.37 (0.10)	**4.57** (0.11)	**4.63** (0.10)	**0.34** (0.01)	1.20 (0.02)	2.76 (0.11)	**4.62** (0.14)
	WGAN	4.58 (0.08)	**3.03** (0.03)	9.59 (0.28)	2.36 (0.03)	3.94 (0.03)	4.32 (0.10)	5.85 (0.10)	4.93 (0.09)	0.37 (0.01)	1.15 (0.02)	2.91 (0.11)	4.81 (0.14)
	DETER	6.29 (0.11)	4.34 (0.05)	12.7 (0.37)	3.23 (0.04)	4.94 (0.03)	5.22 (0.12)	7.05 (0.16)	6.51 (0.12)	0.42 (0.01)	1.49 (0.02)	3.66 (0.15)	6.15 (0.19)
	GAMMA	4.60 (0.08)	3.21 (0.03)	9.95 (0.26)	2.39 (0.03)	3.96 (0.03)	4.45 (0.10)	6.16 (0.09)	4.91 (0.08)	0.37 (0.01)	1.15 (0.02)	3.19 (0.11)	4.83 (0.14)
L_Λ ($\times 10^{-2}$)	SIGTPP	3.35 (0.46)	**3.09** (0.35)	**6.09** (1.04)	2.68 (0.40)	**2.44** (0.39)	**4.01** (0.72)	19.3 (0.99)	9.00 (0.68)	**3.77** (0.90)	6.31 (1.05)	13.3 (1.14)	13.55 (1.20)
	VAE	4.46 (0.57)	4.90 (0.43)	13.1 (1.22)	**2.25** (0.40)	2.66 (0.36)	9.61 (0.76)	**17.9** (1.12)	8.94 (0.69)	9.57 (1.06)	6.99 (1.04)	**10.8** (1.14)	**9.34** (0.94)
	DDPM	7.93 (0.64)	7.18 (0.41)	15.2 (1.05)	5.10 (0.50)	3.77 (0.34)	8.71 (0.77)	19.0 (0.62)	9.95 (0.63)	7.01 (1.08)	**4.57** (1.04)	16.9 (1.11)	15.09 (0.95)
	WGAN	7.31 (0.67)	4.47 (0.37)	13.3 (1.31)	3.37 (0.49)	4.07 (0.36)	8.96 (0.81)	24.1 (1.13)	**6.25** (0.77)	14.0 (1.24)	5.06 (0.74)	13.3 (1.03)	14.34 (0.77)
	DETER	12.6 (0.71)	12.8 (0.43)	20.2 (1.25)	4.17 (0.52)	5.78 (0.41)	25.51 (1.09)	20.3 (2.84)	7.38 (0.37)	41.6 (3.32)	18.0 (1.33)	25.3 (0.93)	18.17 (1.25)
	GAMMA	**3.17** (0.45)	11.0 (0.47)	8.11 (1.08)	3.41 (0.43)	4.32 (0.32)	5.92 (0.67)	27.0 (1.12)	11.5 (0.60)	16.2 (1.08)	7.05 (0.73)	31.7 (1.20)	18.58 (0.80)
$L_{\log(\tau)}$ ($\times 10^{-2}$)	SIGTPP	2.95 (0.60)	**1.97** (0.37)	**3.84** (0.84)	7.64 (0.56)	3.75 (0.61)	2.31 (0.38)	12.9 (0.84)	10.5 (0.41)	7.33 (0.39)	4.35 (0.44)	4.73 (0.99)	6.34 (1.12)
	VAE	4.16 (0.58)	5.47 (0.46)	4.07 (0.67)	4.01 (0.48)	6.86 (0.64)	3.39 (0.44)	14.8 (0.97)	10.6 (0.56)	**6.29** (0.53)	4.36 (0.40)	**4.02** (0.68)	**4.70** (0.97)
	DDPM	6.08 (0.72)	4.10 (0.45)	5.45 (0.89)	2.32 (0.47)	**2.57** (0.57)	**2.07** (0.35)	**9.33** (0.82)	5.22 (0.63)	10.6 (0.72)	**2.02** (0.40)	6.47 (0.80)	6.82 (1.25)
	WGAN	3.74 (0.55)	4.70 (0.46)	5.33 (0.74)	4.70 (0.56)	15.78 (1.96)	4.28 (0.35)	9.58 (0.58)	**2.72** (0.62)	7.86 (0.70)	6.72 (0.55)	5.69 (0.93)	11.09 (1.42)
	DETER	60.7 (5.26)	48.4 (1.46)	53.5 (2.86)	66.6 (6.25)	42.19 (8.80)	81.21 (3.83)	48.5 (3.15)	54.9 (7.16)	81.0 (2.02)	56.9 (6.67)	45.5 (1.62)	38.25 (1.16)
	GAMMA	**2.86** (0.64)	3.21 (0.48)	6.74 (1.00)	**1.71** (0.59)	9.02 (0.52)	12.32 (0.47)	9.93 (0.83)	5.70 (0.75)	15.1 (1.26)	4.84 (0.41)	13.2 (0.85)	5.07 (0.90)
ACD ($\times 10^{-2}$)	SIGTPP	1.35 (0.46)	**0.70** (0.20)	1.70 (0.45)	**1.54** (0.39)	**1.44** (0.32)	0.98 (0.22)	**3.95** (0.68)	8.93 (0.51)	**0.94** (0.20)	2.97 (0.43)	**2.11** (0.40)	2.45 (0.56)
	VAE	1.53 (0.42)	0.71 (0.20)	2.73 (0.62)	2.44 (0.45)	2.37 (0.33)	0.72 (0.20)	7.17 (0.76)	**2.79** (0.51)	2.48 (0.40)	2.26 (0.42)	2.61 (0.53)	2.61 (0.58)
	DDPM	1.86 (0.52)	1.64 (0.25)	2.54 (0.47)	1.76 (0.44)	1.82 (0.32)	0.65 (0.20)	6.95 (0.76)	5.17 (0.51)	1.41 (0.35)	**1.87** (0.45)	3.35 (0.60)	**2.27** (0.56)
	WGAN	2.39 (0.49)	1.94 (0.28)	**1.60** (0.39)	5.03 (0.45)	6.83 (0.33)	0.76 (0.21)	6.95 (0.76)	6.57 (0.51)	3.20 (0.41)	3.48 (0.47)	5.51 (0.46)	2.93 (0.62)
	DETER	49.3 (0.53)	35.3 (0.28)	52.5 (0.71)	63.4 (0.45)	29.58 (0.35)	75.24 (0.25)	61.9 (0.76)	72.7 (0.51)	74.2 (0.41)	60.7 (0.50)	64.4 (0.62)	74.03 (0.62)
	GAMMA	**1.32** (0.44)	1.28 (0.26)	4.08 (0.71)	6.12 (0.45)	7.95 (0.23)	**0.62** (0.19)	16.8 (0.76)	8.03 (0.51)	2.92 (0.41)	3.51 (0.47)	4.61 (0.61)	2.86 (0.59)
PCD ($\times 10^{-2}$)	SIGTPP	**7.36** (0.53)	**6.01** (0.42)	9.07 (0.85)	8.56 (0.40)	5.47 (0.30)	**7.72** (0.44)	**7.51** (0.53)	12.4 (0.33)	5.94 (0.26)	6.11 (0.23)	12.9 (0.69)	12.35 (0.73)
	VAE	7.43 (0.53)	6.40 (0.43)	8.82 (0.86)	8.39 (0.40)	5.51 (0.30)	7.76 (0.43)	9.92 (0.74)	**10.2** (0.28)	**5.83** (0.27)	5.85 (0.22)	**12.4** (0.71)	**12.15** (0.77)
	DDPM	7.54 (0.51)	6.27 (0.43)	**8.25** (0.97)	**8.33** (0.42)	**5.28** (0.31)	7.73 (0.44)	8.44 (0.69)	11.1 (0.32)	5.84 (0.25)	**5.61** (0.22)	12.8 (0.68)	12.35 (0.76)
	WGAN	7.49 (0.53)	6.15 (0.43)	8.87 (0.84)	8.82 (0.41)	7.63 (0.28)	7.92 (0.43)	7.55 (0.54)	11.0 (0.31)	5.92 (0.28)	6.25 (0.22)	13.1 (0.69)	12.77 (0.76)
	DETER	96.5 (0.43)	82.7 (0.28)	59.7 (0.83)	97.9 (0.22)	55.48 (0.22)	70.57 (1.06)	73.4 (0.81)	79.3 (0.27)	94.8 (0.34)	43.4 (0.31)	72.4 (0.80)	76.31 (1.13)
	GAMMA	7.39 (0.53)	6.39 (0.42)	9.03 (0.85)	8.74 (0.41)	8.34 (0.27)	7.74 (0.43)	18.8 (0.81)	12.0 (0.35)	5.86 (0.27)	6.24 (0.22)	13.3 (0.69)	12.75 (0.76)

Table S2: Pointwise errors (MAE, MAE_{med}, and MSE_{mean}) for SIGTPP and baselines on PS, IP, H1, H3, EQ, SO, TB, TX, and YLP. Values report standard error in parentheses over 100 bootstrap replicates. Lower is better (↓). Best in **bold**, second best underlined. Metric definitions in Appendix D.

Model	PS	IP	H1	H3	EQ	SO	TB	TX	YLP
	MAE ($\times 10^{-1}$)				MAE ($\times 10^{-1}$)				
SIGTPP	9.51 (0.08)	6.09 (0.04)	22.39 (0.41)	4.79 (0.04)	10.75 (0.20)	9.88 (0.22)	0.69 (0.01)	2.21 (0.02)	6.12 (0.23)
VAE	8.57 (0.08)	6.06 (0.03)	18.16 (0.30)	4.42 (0.04)	10.33 (0.20)	8.67 (0.14)	0.67 (0.01)	2.05 (0.02)	4.91 (0.15)
DDPM	9.53 (0.09)	6.08 (0.04)	18.76 (0.32)	4.51 (0.04)	12.55 (0.26)	11.08 (0.17)	0.71 (0.01)	2.20 (0.02)	6.00 (0.16)
WGAN	8.68 (0.09)	5.98 (0.04)	18.18 (0.35)	4.66 (0.04)	13.98 (0.14)	10.41 (0.12)	0.73 (0.01)	2.21 (0.02)	5.49 (0.16)
DETER	**6.29** (0.11)	**4.34** (0.05)	**12.66** (0.37)	**3.23** (0.04)	**7.05** (0.16)	**6.51** (0.12)	**0.42** (0.01)	**1.49** (0.02)	**3.66** (0.15)
GAMMA	9.17 (0.08)	6.53 (0.03)	19.72 (0.26)	4.68 (0.03)	14.22 (0.09)	10.09 (0.08)	0.71 (0.01)	2.16 (0.02)	6.52 (0.11)
	MAE_{med} ($\times 10^{-1}$)				MAE_{med} ($\times 10^{-1}$)				
SIGTPP	6.41 (0.10)	4.23 (0.04)	13.36 (0.36)	3.26 (0.04)	7.71 (0.15)	7.25 (0.14)	0.43 (0.01)	1.46 (0.02)	4.30 (0.17)
VAE	6.31 (0.11)	4.28 (0.04)	12.94 (0.35)	3.18 (0.04)	6.95 (0.16)	6.37 (0.13)	0.42 (0.01)	**1.39** (0.02)	3.68 (0.14)
DDPM	6.56 (0.11)	4.43 (0.05)	12.81 (0.35)	**3.07** (0.04)	**6.02** (0.15)	**6.28** (0.13)	**0.40** (0.01)	1.55 (0.02)	**3.60** (0.15)
WGAN	6.35 (0.10)	**4.22** (0.04)	12.86 (0.37)	3.21 (0.04)	7.58 (0.16)	6.61 (0.13)	0.43 (0.01)	1.49 (0.02)	3.75 (0.14)
DETER	**6.29** (0.11)	4.34 (0.05)	**12.66** (0.37)	3.23 (0.04)	7.05 (0.16)	6.51 (0.12)	0.42 (0.01)	1.49 (0.02)	3.66 (0.15)
GAMMA	6.34 (0.11)	4.41 (0.04)	13.32 (0.34)	3.25 (0.04)	8.18 (0.13)	6.63 (0.11)	0.43 (0.01)	1.50 (0.02)	3.91 (0.16)
	MSE_{mean} ($\times 10^{-1}$)				MSE_{mean} ($\times 10^{-1}$)				
SIGTPP	8.42 (0.32)	3.87 (0.10)	51.20 (2.68)	2.51 (0.08)	16.39 (0.72)	14.52 (0.66)	0.16 (0.01)	0.90 (0.05)	11.63 (1.00)
VAE	**8.13** (0.35)	3.89 (0.10)	**43.84** (2.63)	2.43 (0.08)	14.71 (0.70)	**9.82** (0.37)	0.16 (0.01)	**0.82** (0.05)	**7.77** (0.73)
DDPM	9.03 (0.37)	4.49 (0.12)	49.96 (3.04)	**2.26** (0.08)	**13.06** (0.60)	10.20 (0.36)	**0.15** (0.01)	0.89 (0.05)	8.16 (0.76)
WGAN	8.17 (0.34)	**3.83** (0.10)	44.21 (2.71)	2.48 (0.08)	17.99 (0.73)	11.47 (0.37)	0.16 (0.01)	0.90 (0.06)	8.55 (0.72)
DETER	8.84 (0.40)	3.89 (0.10)	50.52 (3.26)	2.65 (0.09)	17.13 (0.84)	11.52 (0.47)	0.17 (0.01)	0.93 (0.06)	9.32 (0.81)
GAMMA	8.29 (0.35)	4.29 (0.11)	46.80 (2.68)	2.53 (0.09)	18.99 (0.66)	10.86 (0.34)	0.16 (0.01)	0.90 (0.06)	9.65 (0.77)

Table S3: Pointwise errors (MAE, MAE_{med}, and MSE_{mean}) for SIGTPP and baselines on SC, SR, and YLA, completing Table S2. Values report standard error in parentheses over 100 bootstrap replicates. Lower is better (↓). Best in **bold**, second best underlined. Metric definitions in Appendix D.

Model	SC	SR	YLA	SC	SR	YLA	SC	SR	YLA
	MAE ($\times 10^{-1}$)			MAE_{med} ($\times 10^{-1}$)			MSE_{mean} ($\times 10^{-1}$)		
SIGTPP	6.79 (0.04)	11.77 (0.08)	9.47 (0.15)	5.31 (0.05)	5.24 (0.12)	6.25 (0.18)	4.25 (0.08)	19.51 (0.94)	14.44 (0.93)
VAE	6.26 (0.03)	8.13 (0.10)	7.97 (0.16)	**4.86** (0.04)	5.22 (0.12)	6.17 (0.17)	**3.69** (0.06)	**17.22** (1.01)	13.82 (0.98)
DDPM	8.76 (0.05)	9.65 (0.09)	9.76 (0.16)	7.62 (0.06)	5.24 (0.12)	**6.07** (0.18)	8.75 (0.14)	18.11 (1.01)	**13.75** (0.98)
WGAN	7.45 (0.03)	8.59 (0.12)	9.06 (0.16)	5.65 (0.05)	5.23 (0.12)	6.22 (0.18)	4.81 (0.08)	17.37 (1.01)	14.50 (1.00)
DETER	**4.94** (0.03)	**5.22** (0.12)	**6.15** (0.19)	4.94 (0.03)	**5.22** (0.12)	6.15 (0.19)	3.98 (0.06)	18.76 (1.10)	15.87 (1.10)
GAMMA	7.74 (0.03)	9.90 (0.10)	9.00 (0.14)	5.69 (0.04)	5.27 (0.12)	6.21 (0.17)	5.00 (0.09)	17.71 (0.99)	14.75 (1.00)

E.3 Effect of Auxiliary Conditional Regression

Tables 1, 2, and S1 show a CRPS gap between SIGTPP and the conditional baselines. The signature objective matches complete-trajectory distributions but does not directly supervise the one-step conditional distributions assessed by CRPS. We test whether adding conditional supervision narrows this gap while retaining pathwise fidelity. Starting from SIGTPP, we keep the same generator and add a teacher-forced conditional regression term,

$$\mathcal{L}_{\text{hybrid}} = \mathcal{L}_{\text{Sig-W}_1} + \lambda \mathcal{L}_{\text{cond}}, \tag{S105}$$

where $\lambda \geq 0$ weights the auxiliary loss, $\mathcal{L}_{\text{Sig-W}_1}$ is the signature loss in Algorithm 2, and $\mathcal{L}_{\text{cond}}$ is one-step error on the next transformed log-interarrival given the observed history. We denote this hybrid model by SIGREG. Appendix G describes the loss and the search grid.

We train SIGREG on all datasets with $\lambda \in \{0, 0.1, 1\}$. Setting $\lambda = 0$ recovers SIGTPP. These runs use a different search grid from the main benchmark. Tables S9 and S10 report the test scores on the 12 datasets.

We summarise each metric by its geometric mean relative improvement over SIGTPP ($\lambda = 0$) across the 12 datasets in Tables S9 and S10. With $\lambda = 1$, SIGREG lowers CRPS on 10 datasets. The 2 increases, on PS and SR, are below 0.2%. The mean CRPS improvement is 14%. The largest gains are on SO (61%) and YLP (41%), where SIGTPP ($\lambda = 0$) trails the best conditional baseline the most. These gains therefore partly reflect weak SIGTPP runs on these datasets. The conditional

Table S4: First inactive signature degree M^{all} per dataset, defined as the first degree m for which all D^m coordinates have empirical standard deviation below 10^{-8} across the training set.

Dataset	PS	IP	H1	H3	SC	SR	EQ	SO	TB	TX	YLP	YLA
M^{all}	8	8	>8	6	7	7	>8	6	6	7	6	6

Table S5: Ranks for SIGTPP and baselines by metric, for all datasets. Values are means, with standard deviations in parentheses. Lower is better (↓). Best in **bold**, second best underlined. Metric definitions in Appendix D.

Metric	SIGTPP			VAE			DDPM		
	Synth.	Real	All	Synth.	Real	All	Synth.	Real	All
$\mathcal{E}$	**1.17** (0.41)	**1.83** (0.75)	**1.50** (0.67)	3.67 (1.21)	3.00 (1.67)	3.33 (1.44)	2.00 (0.63)	2.17 (1.17)	2.08 (0.90)
W_1	**1.33** (0.82)	**1.67** (0.82)	**1.50** (0.80)	3.17 (1.47)	2.67 (1.21)	2.92 (1.31)	2.17 (0.41)	2.67 (1.51)	2.42 (1.08)
Sig-W_1	**2.17** (1.33)	**2.33** (1.03)	**2.25** (1.14)	2.33 (1.63)	2.67 (1.63)	2.50 (1.57)	5.17 (1.60)	4.17 (1.17)	4.67 (1.44)
CRPS	3.67 (1.37)	3.67 (1.21)	3.67 (1.23)	**1.50** (0.84)	**1.67** (0.52)	**1.58** (0.67)	4.00 (1.79)	1.83 (1.60)	2.92 (1.98)
L_λ	**1.33** (0.52)	2.67 (1.03)	**2.00** (1.04)	2.83 (1.33)	**2.17** (1.33)	2.50 (1.31)	4.33 (1.21)	3.00 (1.55)	3.67 (1.50)
$L_{\log(\tau)}$	**2.17** (1.47)	2.83 (0.98)	**2.50** (1.24)	3.33 (1.03)	**2.67** (1.97)	3.00 (1.54)	2.67 (1.63)	**2.67** (1.51)	2.67 (1.50)
ACD	**2.00** (1.55)	2.17 (1.60)	**2.08** (1.51)	3.00 (0.63)	2.50 (1.05)	2.75 (0.87)	2.83 (0.98)	**2.00** (0.89)	2.42 (1.00)
PCD	**2.17** (1.60)	3.17 (1.60)	2.67 (1.61)	3.17 (1.17)	**1.67** (1.21)	2.42 (1.38)	**2.17** (1.60)	2.33 (0.82)	**2.25** (1.22)
Avg.	**2.00** (1.35)	2.54 (1.25)	**2.27** (1.33)	2.88 (1.28)	**2.38** (1.36)	2.62 (1.34)	3.17 (1.65)	2.60 (1.40)	2.89 (1.55)
Metric	WGAN			GAMMA			DETER		
	Synth.	Real	All	Synth.	Real	All	Synth.	Real	All
$\mathcal{E}$	3.83 (0.41)	3.17 (0.75)	3.50 (0.67)	4.50 (1.22)	4.83 (0.41)	4.67 (0.89)	5.83 (0.41)	6.00 (0.00)	5.92 (0.29)
W_1	4.17 (0.75)	3.00 (0.63)	3.58 (0.90)	4.50 (1.22)	5.00 (0.00)	4.75 (0.87)	5.67 (0.82)	6.00 (0.00)	5.83 (0.58)
Sig-W_1	2.83 (0.98)	3.33 (1.63)	3.08 (1.31)	4.00 (1.67)	4.17 (1.94)	4.08 (1.73)	4.50 (1.05)	4.33 (2.25)	4.42 (1.68)
CRPS	2.17 (0.75)	3.50 (0.55)	2.83 (0.94)	3.83 (0.75)	4.33 (0.82)	4.08 (0.79)	5.83 (0.41)	6.00 (0.00)	5.92 (0.29)
L_λ	3.50 (0.84)	2.83 (1.47)	3.17 (1.19)	3.17 (1.72)	5.67 (0.52)	4.42 (1.78)	5.83 (0.41)	4.67 (1.51)	5.25 (1.22)
$L_{\log(\tau)}$	3.83 (0.75)	3.17 (1.60)	3.50 (1.24)	3.00 (1.90)	3.67 (1.21)	3.33 (1.56)	6.00 (0.00)	6.00 (0.00)	6.00 (0.00)
ACD	3.83 (1.47)	4.00 (1.26)	3.92 (1.31)	3.33 (1.97)	4.33 (0.52)	3.83 (1.47)	6.00 (0.00)	6.00 (0.00)	6.00 (0.00)
PCD	3.83 (1.17)	3.67 (1.37)	3.75 (1.22)	3.67 (1.03)	4.17 (0.75)	3.92 (0.90)	6.00 (0.00)	6.00 (0.00)	6.00 (0.00)
Avg.	3.50 (1.07)	3.33 (1.19)	3.42 (1.13)	3.75 (1.48)	4.52 (1.05)	4.14 (1.33)	5.71 (0.68)	5.62 (1.10)	5.67 (0.91)

term also improves path-level scores on average. Across datasets, $\mathcal{E}$, W_1, and Sig-W_1 improve by 26%, 12%, and 30%, respectively. All 8 metrics improve, with a relative improvement of 18% across the 96 dataset-metric pairs. These aggregates do not imply uniform gains. In particular, the relative improvement in Sig-W_1 is -14% on synthetic tasks and 58% on real-world tasks. On SO, energy distance and Wasserstein distance improve, but L_λ and Sig-W_1 worsen. These results favour $\lambda = 1$ in this ablation, but do not establish a monotone effect of the regression weight.

With the best of VAE, DDPM, and WGAN on each dataset as reference, the CRPS gap falls from 23% at $\lambda = 0$ to 6% at $\lambda = 1$. SIGREG still trails that baseline on all 12 datasets. Therefore, this combined loss improves one-step predictive scores while retaining or improving path-level fidelity on average, with dataset-specific trade-offs. Further experiments are needed to understand the mechanism and to improve the signature loss.

Table S6: Geometric means of per-dataset score ratios to DETER, by metric, for all datasets. Lower is better (↓). Best in **bold**, second best underlined. Metric definitions in Appendix D.

Metric	SIGTPP			VAE			DDPM			WGAN			GAMMA		
	Synth.	Real	All	Synth.	Real	All	Synth.	Real	All	Synth.	Real	All	Synth.	Real	All
$\mathcal{E}$	**0.003**	**0.091**	**0.017**	0.031	0.108	0.058	0.010	0.112	0.033	0.041	0.130	0.073	0.051	0.204	0.102
W_1	**0.193**	**0.310**	**0.245**	0.269	0.328	0.297	0.223	0.353	0.281	0.305	0.357	0.330	0.314	0.422	0.364
Sig-W_1	**0.178**	**0.263**	**0.217**	0.271	0.383	0.322	2.442	0.896	1.479	0.390	0.542	0.460	0.835	1.014	0.920
CRPS	0.763	0.822	0.792	**0.734**	0.752	**0.743**	0.817	**0.745**	0.780	0.755	0.803	0.779	0.774	0.823	0.798
L_λ	**0.306**	**0.494**	**0.389**	0.449	0.526	0.486	0.645	0.557	0.599	0.546	0.584	0.565	0.480	0.860	0.643
$L_{\log(\tau)}$	**0.059**	0.135	0.089	0.079	0.125	0.099	0.060	**0.114**	**0.083**	0.097	0.126	0.111	0.083	0.153	0.113
ACD	**0.025**	**0.042**	**0.033**	0.031	0.045	0.037	0.033	0.044	0.038	0.050	0.066	0.057	0.050	0.076	0.062
PCD	0.096	0.126	0.110	0.097	0.126	0.110	**0.095**	**0.124**	**0.108**	0.103	0.125	0.114	0.105	0.148	0.125
Avg.	**0.088**	**0.196**	**0.131**	0.144	0.210	0.174	0.164	0.235	0.196	0.178	0.243	0.208	0.196	0.319	0.250

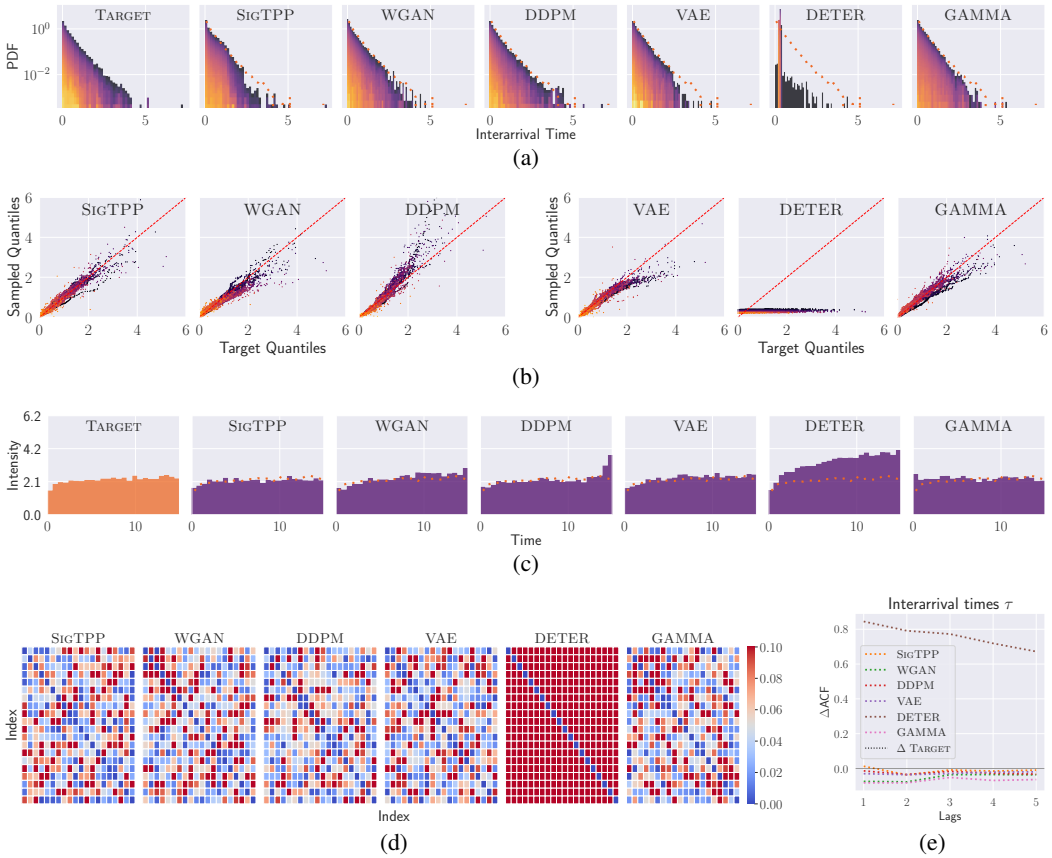


Figure S4: Qualitative evaluation on H3. Model panels are ordered SIGTPP, WGAN, DDPM, VAE, DETER, and GAMMA. (a) Histogram estimates of the interarrival times density $p(x)$ on a logarithmic scale. The leftmost panel shows the target distribution, the others the distributions of samples generated by each model. (b) Quantile-quantile (QQ) plots. Each subplot shows empirical quantiles of the reference data (x-axis) against generated sample quantiles (y-axis). The red dashed line is the identity $y = x$. (c) Empirical intensity functions $\hat{\lambda}(t)$ estimated from the same sequences. (d) Absolute Pearson correlation-matrix differences for up to the first 20 event positions. Each heatmap entry shows the absolute difference between the empirical correlation of two sequence components in the generated samples and in the reference data. (e) Signed autocorrelation error for 5 lags. Each curve shows the difference between the ACF of the indicated model and the empirical ACF of the reference process, so that the target corresponds to the horizontal line at zero.

Table S7: Performance of SIGTPP and 5 EasyTPP baselines on PS, IP, H1, EQ, YLA, and TB. Dashes denote the excluded S2P2 run on TB, which exceeded the GPU time budget. Values report standard error in parentheses over 100 bootstrap replicates. Lower is better (↓). Best in **bold**, second best underlined. Metric definitions in Appendix D.

	Model	PS	IP	H1	EQ	YLA	TB
$\mathcal{E}$ ($\times 10^{-3}$)	SIGTPP	**0.53** (0.93)	**0.42** (0.32)	**1.16** (0.74)	32.40 (0.65)	**7.77** (4.47)	511.8 (25.25)
	NHP	4.02 (2.39)	2.07 (0.62)	10.8 (3.61)	31.65 (0.62)	56.57 (19.67)	**455.9** (20.96)
	RMTPP	2.13 (1.28)	3.37 (0.87)	25.6 (5.03)	31.77 (0.62)	10.43 (7.81)	579.3 (33.24)
	SAHP	5.91 (2.47)	2.75 (1.13)	8.97 (3.09)	33.47 (0.58)	27.34 (17.65)	541.5 (24.80)
	THP	2.86 (1.28)	3.75 (1.45)	26.3 (5.29)	**29.74** (0.80)	12.99 (7.39)	585.2 (25.41)
	S2P2	21.31 (5.00)	2.89 (0.67)	10.1 (3.73)	32.29 (0.63)	50.03 (18.22)	–
W_1 ($\times 10^{-2}$)	SIGTPP	**5.26** (0.20)	**6.55** (0.14)	**3.75** (0.31)	2.91 (0.04)	**8.48** (0.73)	58.17 (2.56)
	NHP	6.26 (0.48)	7.12 (0.18)	5.95 (0.55)	2.95 (0.05)	14.28 (1.91)	**53.02** (2.04)
	RMTPP	5.99 (0.34)	7.47 (0.24)	8.11 (0.76)	2.94 (0.05)	9.06 (0.87)	66.74 (2.61)
	SAHP	6.76 (0.52)	7.38 (0.28)	5.74 (0.58)	2.96 (0.05)	11.48 (1.81)	59.83 (2.45)
	THP	5.89 (0.29)	7.75 (0.29)	8.34 (0.60)	**2.81** (0.05)	8.60 (0.92)	64.28 (2.22)
	S2P2	9.23 (0.93)	7.48 (0.23)	5.92 (0.64)	3.04 (0.05)	14.73 (2.32)	–
Sig-W_1 ($\times 10^{-5}$)	SIGTPP	**1.42** (1.52)	**1.20** (1.05)	44.0 (29.2)	409.2 (36.30)	4.64 (4.30)	2.27 (0.94)
	NHP	10.33 (6.19)	13.0 (4.04)	98.1 (25.0)	513.4 (41.31)	**1.52** (1.83)	18.28 (2.76)
	RMTPP	13.08 (7.49)	23.1 (5.39)	304 (66.1)	979.5 (59.51)	5.93 (4.85)	17.41 (2.62)
	SAHP	35.39 (12.05)	3.92 (2.18)	98.7 (26.9)	**406.4** (37.45)	3.55 (3.62)	**0.53** (0.36)
	THP	21.23 (8.72)	17.2 (4.72)	614 (103)	570.6 (46.18)	27.20 (11.58)	7.01 (1.35)
	S2P2	10.01 (5.94)	9.94 (3.50)	**40.7** (12.6)	626.1 (43.81)	2.38 (2.51)	–
CRPS ($\times 10^{-1}$)	SIGTPP	4.64 (0.08)	3.04 (0.03)	10.2 (0.27)	5.61 (0.11)	4.78 (0.14)	0.37 (0.01)
	NHP	4.50 (0.08)	2.99 (0.03)	**9.33** (0.27)	**5.22** (0.11)	4.74 (0.14)	**0.37** (0.01)
	RMTPP	4.51 (0.08)	2.98 (0.03)	9.50 (0.27)	5.33 (0.12)	4.58 (0.14)	0.37 (0.01)
	SAHP	4.50 (0.08)	2.99 (0.03)	9.42 (0.28)	5.23 (0.12)	4.65 (0.14)	0.37 (0.01)
	THP	**4.49** (0.08)	**2.98** (0.03)	9.58 (0.27)	5.29 (0.12)	**4.49** (0.14)	0.40 (0.01)
	S2P2	4.55 (0.09)	3.00 (0.03)	9.36 (0.27)	5.26 (0.12)	4.71 (0.15)	–
L_λ ($\times 10^{-2}$)	SIGTPP	**3.35** (0.46)	**3.09** (0.35)	**6.09** (1.04)	19.34 (0.99)	13.55 (1.20)	**3.77** (0.90)
	NHP	4.31 (0.54)	5.01 (0.41)	13.1 (1.13)	16.73 (0.92)	18.68 (0.82)	10.37 (1.21)
	RMTPP	4.34 (0.56)	5.99 (0.46)	16.2 (1.15)	21.30 (0.91)	11.37 (1.13)	21.93 (1.54)
	SAHP	5.64 (0.59)	4.09 (0.37)	13.6 (1.17)	18.22 (1.08)	10.51 (0.99)	8.37 (0.99)
	THP	5.37 (0.61)	4.32 (0.39)	18.2 (1.12)	**16.09** (1.12)	**10.27** (1.07)	16.73 (1.00)
	S2P2	3.36 (0.54)	5.11 (0.50)	6.32 (1.05)	22.35 (0.81)	18.61 (1.04)	–
$L_{\log(\tau)}$ ($\times 10^{-2}$)	SIGTPP	**2.95** (0.60)	1.97 (0.37)	**3.84** (0.84)	12.86 (0.84)	6.34 (1.12)	**7.33** (0.39)
	NHP	4.12 (0.53)	2.55 (0.39)	6.34 (0.85)	13.29 (1.01)	3.92 (0.91)	20.76 (0.72)
	RMTPP	3.66 (0.64)	3.60 (0.45)	7.64 (0.82)	21.02 (0.93)	**3.67** (0.90)	24.60 (0.91)
	SAHP	4.52 (0.59)	**1.72** (0.32)	5.47 (0.76)	14.39 (1.02)	3.67 (0.98)	20.15 (0.75)
	THP	3.44 (0.54)	4.96 (0.46)	7.64 (0.89)	17.62 (1.02)	4.09 (0.90)	36.72 (0.64)
	S2P2	4.36 (0.60)	2.92 (0.41)	4.48 (0.68)	**11.79** (0.77)	3.70 (0.92)	–
ACD ($\times 10^{-2}$)	SIGTPP	1.35 (0.46)	**0.70** (0.20)	**1.70** (0.45)	3.95 (0.68)	**2.45** (0.56)	0.94 (0.20)
	NHP	1.49 (0.42)	1.71 (0.26)	2.04 (0.42)	**1.78** (0.63)	2.46 (0.58)	**0.83** (0.25)
	RMTPP	**1.30** (0.41)	2.38 (0.27)	2.67 (0.41)	5.14 (0.76)	2.62 (0.58)	1.53 (0.36)
	SAHP	1.70 (0.43)	1.07 (0.23)	1.83 (0.44)	4.01 (0.76)	3.13 (0.60)	0.86 (0.24)
	THP	1.69 (0.47)	0.91 (0.22)	2.38 (0.40)	9.25 (0.76)	5.96 (0.62)	1.20 (0.28)
	S2P2	1.78 (0.49)	1.80 (0.28)	2.16 (0.57)	6.20 (0.76)	2.53 (0.59)	–
PCD ($\times 10^{-2}$)	SIGTPP	**7.36** (0.53)	**6.01** (0.42)	9.07 (0.85)	7.51 (0.53)	12.35 (0.73)	5.94 (0.26)
	NHP	7.39 (0.52)	6.04 (0.44)	**8.19** (0.92)	**5.06** (0.41)	12.78 (0.79)	**5.82** (0.26)
	RMTPP	7.42 (0.52)	6.06 (0.43)	8.25 (0.93)	6.39 (0.52)	**12.20** (0.74)	5.82 (0.26)
	SAHP	7.42 (0.51)	6.10 (0.43)	8.47 (0.88)	5.88 (0.52)	12.60 (0.76)	5.96 (0.24)
	THP	7.40 (0.52)	6.05 (0.43)	8.35 (0.92)	10.11 (0.76)	12.45 (0.73)	6.00 (0.24)
	S2P2	7.45 (0.51)	6.04 (0.43)	8.47 (0.87)	9.65 (0.71)	12.70 (0.79)	–

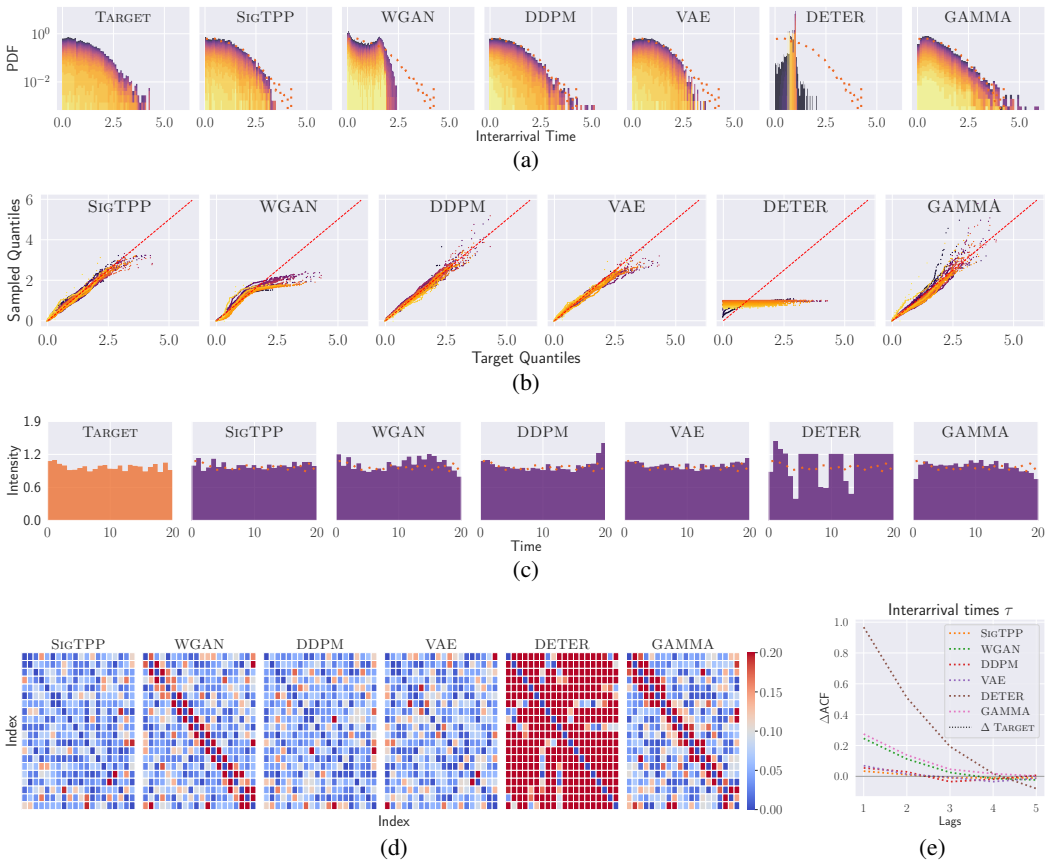


Figure S5: Qualitative evaluation on SC. Model panels are ordered SIGTPP, WGAN, DDPM, VAE, DETER, and GAMMA. (a) Histogram estimates of the interarrival times density $p(x)$ on a logarithmic scale. The leftmost panel shows the target distribution, the others the distributions of samples generated by each model. (b) Quantile-quantile (QQ) plots. Each subplot shows empirical quantiles of the reference data (x-axis) against generated sample quantiles (y-axis). The red dashed line is the identity $y = x$. (c) Empirical intensity functions $\hat{\lambda}(t)$ estimated from the same sequences. (d) Absolute Pearson correlation-matrix differences for up to the first 20 event positions. Each heatmap entry shows the absolute difference between the empirical correlation of two sequence components in the generated samples and in the reference data. (e) Signed autocorrelation error for 5 lags. Each curve shows the difference between the ACF of the indicated model and the empirical ACF of the reference process, so that the target corresponds to the horizontal line at zero.

Table S8: Ranks for SIGTPP and 5 EasyTPP baselines by metric, on PS, IP, H1, EQ, YLA, and TB. S2P2 uses 5 datasets (40 pairs), while the other models use 6 (48 pairs). Values are means, with standard deviations in parentheses. Lower is better (↓). Best in **bold**, second best underlined. Metric definitions in Appendix D.

Metric	SIGTPP	NHP	RMTPP	SAHP	THP	S2P2
$\mathcal{E}$	**1.83** (1.60)	3.17 (1.83)	3.50 (1.38)	3.83 (1.47)	4.00 (2.00)	4.40 (1.14)
W_1	**1.33** (0.52)	3.33 (1.51)	3.83 (0.98)	3.67 (1.21)	3.50 (2.17)	5.20 (1.30)
Sig-W_1	**2.00** (1.10)	3.17 (1.33)	5.00 (0.89)	2.83 (1.94)	4.83 (1.17)	2.60 (1.52)
CRPS	5.33 (1.63)	**2.17** (1.60)	3.50 (1.22)	3.00 (0.63)	2.83 (2.04)	3.80 (1.30)
L_λ	**2.00** (1.55)	3.50 (1.38)	4.67 (1.03)	3.17 (1.60)	3.33 (2.07)	4.00 (1.87)
$L_{\log(\tau)}$	**2.17** (1.94)	3.50 (0.55)	4.17 (1.94)	3.00 (1.79)	4.67 (1.37)	3.00 (1.58)
ACD	**1.67** (0.82)	2.33 (1.21)	4.33 (1.86)	3.33 (1.37)	4.50 (1.52)	4.60 (1.14)
PCD	2.83 (1.94)	**2.33** (1.97)	3.00 (1.67)	4.17 (1.33)	4.00 (1.26)	4.40 (1.52)
Avg.	**2.40** (1.81)	2.94 (1.46)	4.00 (1.46)	3.38 (1.42)	3.96 (1.74)	4.00 (1.54)

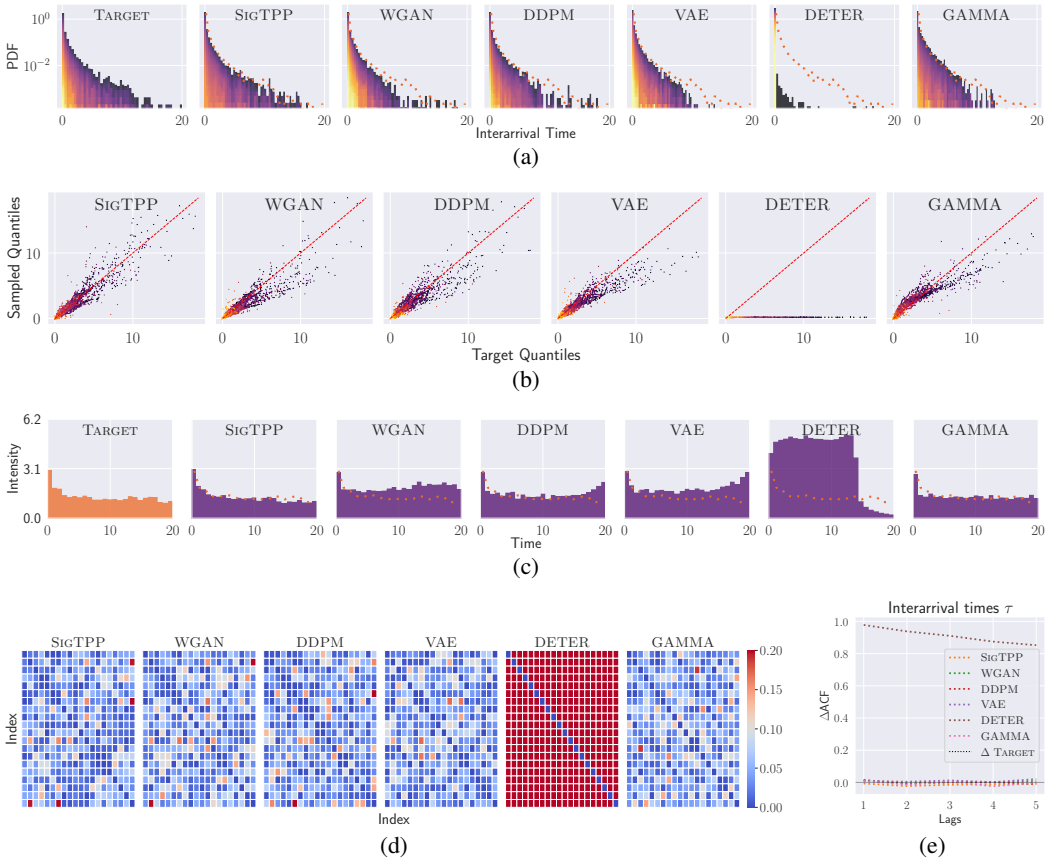


Figure S6: Qualitative evaluation on SR. Model panels are ordered SIGTPP, WGAN, DDPM, VAE, DETER, and GAMMA. (a) Histogram estimates of the interarrival times density $p(x)$ on a logarithmic scale. The leftmost panel shows the target distribution, the others the distributions of samples generated by each model. (b) Quantile-quantile (QQ) plots. Each subplot shows empirical quantiles of the reference data (x-axis) against generated sample quantiles (y-axis). The red dashed line is the identity $y = x$. (c) Empirical intensity functions $\hat{\lambda}(t)$ estimated from the same sequences. (d) Absolute Pearson correlation-matrix differences for up to the first 20 event positions. Each heatmap entry shows the absolute difference between the empirical correlation of two sequence components in the generated samples and in the reference data. (e) Signed autocorrelation error for 5 lags. Each curve shows the difference between the ACF of the indicated model and the empirical ACF of the reference process, so that the target corresponds to the horizontal line at zero.

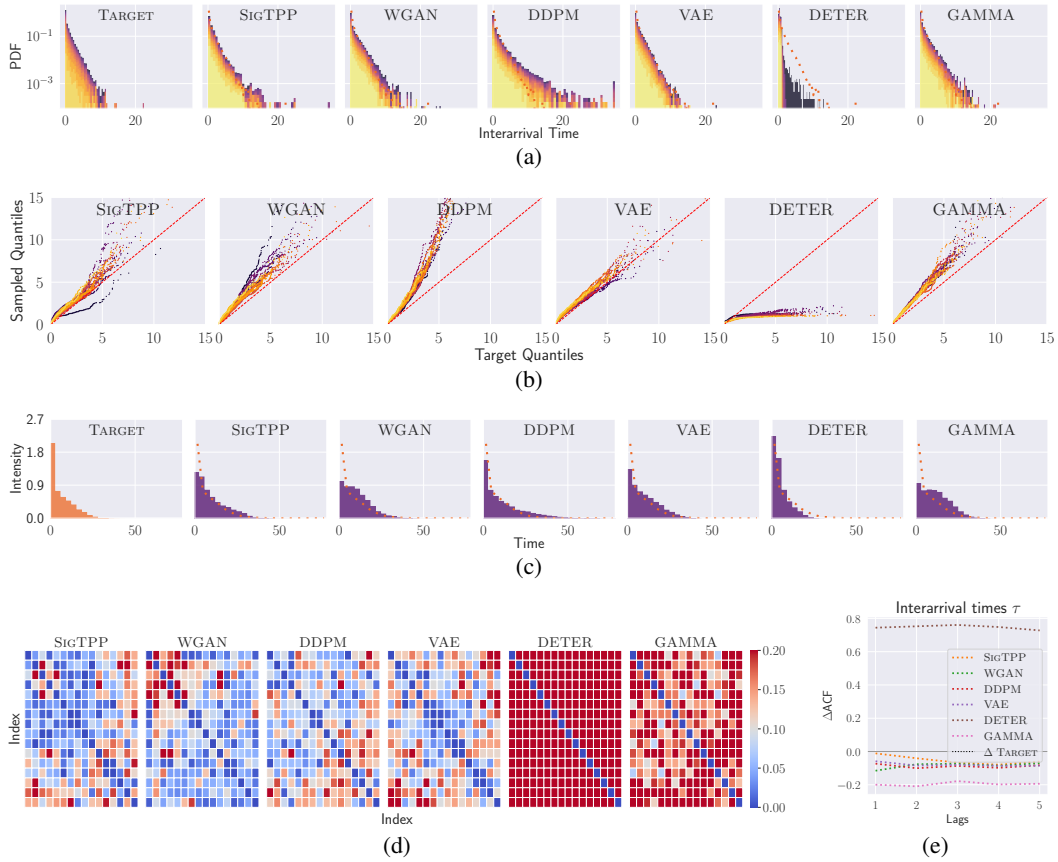


Figure S7: Qualitative evaluation on EQ. Model panels are ordered SIGTPP, WGAN, DDPM, VAE, DETER, and GAMMA. (a) Histogram estimates of the interarrival times density $p(x)$ on a logarithmic scale. The leftmost panel shows the target distribution, the others the distributions of samples generated by each model. (b) Quantile-quantile (QQ) plots. Each subplot shows empirical quantiles of the reference data (x-axis) against generated sample quantiles (y-axis). The red dashed line is the identity $y = x$. (c) Empirical intensity functions $\hat{\lambda}(t)$ estimated from the same sequences. (d) Absolute Pearson correlation-matrix differences for up to the first 20 event positions. Each heatmap entry shows the absolute difference between the empirical correlation of two sequence components in the generated samples and in the reference data. (e) Signed autocorrelation error for 5 lags. Each curve shows the difference between the ACF of the indicated model and the empirical ACF of the reference process, so that the target corresponds to the horizontal line at zero.

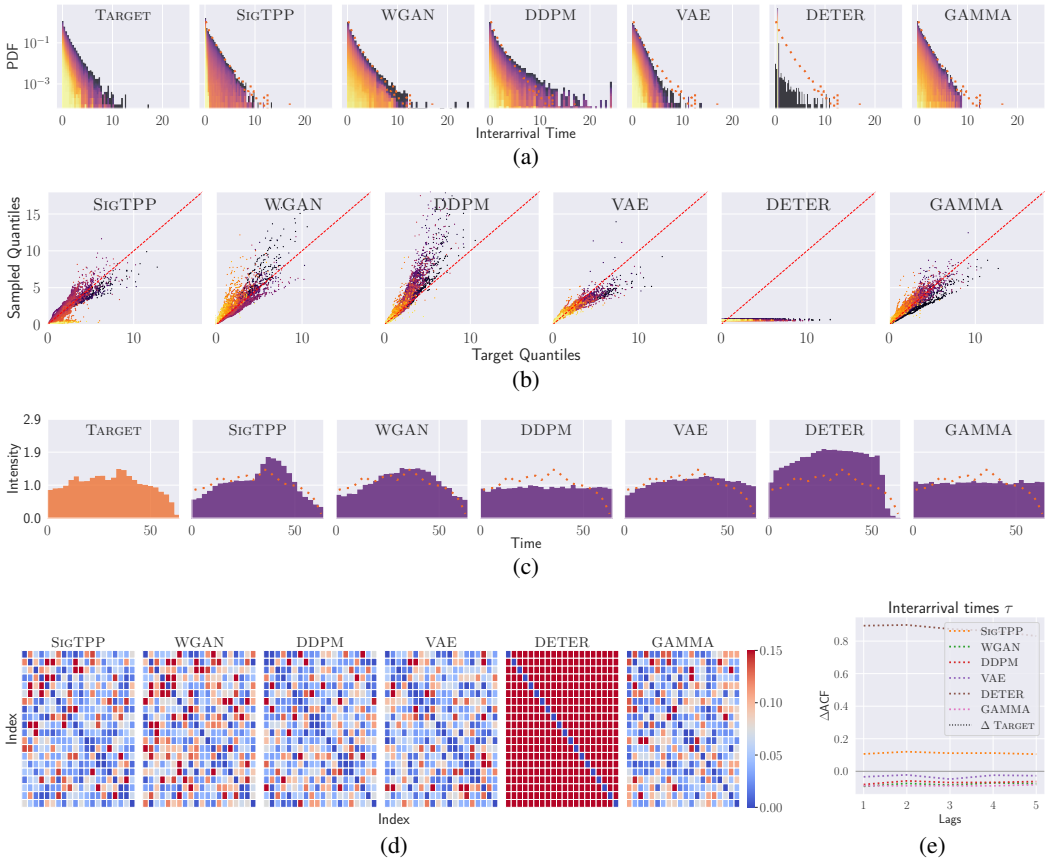


Figure S8: Qualitative evaluation on SO. Model panels are ordered SIGTPP, WGAN, DDPM, VAE, DETER, and GAMMA. (a) Histogram estimates of the interarrival times density $p(x)$ on a logarithmic scale. The leftmost panel shows the target distribution, the others the distributions of samples generated by each model. (b) Quantile-quantile (QQ) plots. Each subplot shows empirical quantiles of the reference data (x-axis) against generated sample quantiles (y-axis). The red dashed line is the identity $y = x$. (c) Empirical intensity functions $\hat{\lambda}(t)$ estimated from the same sequences. (d) Absolute Pearson correlation-matrix differences for up to the first 20 event positions. Each heatmap entry shows the absolute difference between the empirical correlation of two sequence components in the generated samples and in the reference data. (e) Signed autocorrelation error for 5 lags. Each curve shows the difference between the ACF of the indicated model and the empirical ACF of the reference process, so that the target corresponds to the horizontal line at zero.

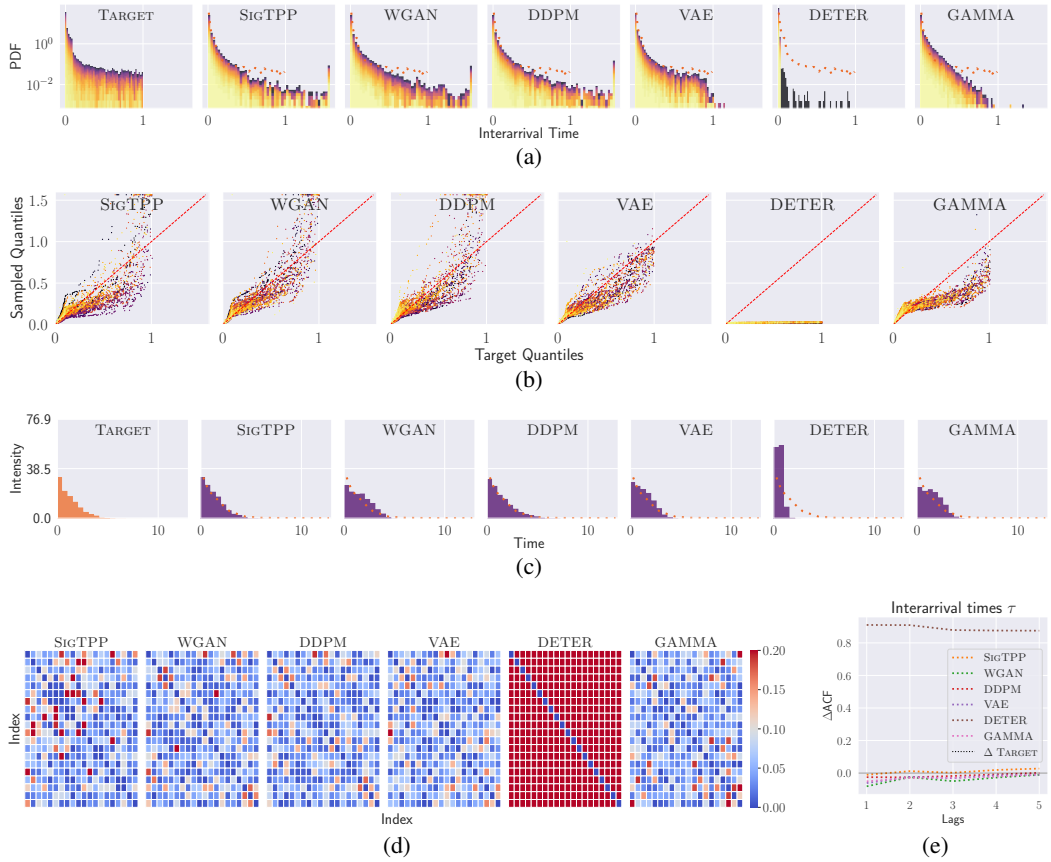


Figure S9: Qualitative evaluation on TB. Model panels are ordered SIGTPP, WGAN, DDPM, VAE, DETER, and GAMMA. (a) Histogram estimates of the interarrival times density $p(x)$ on a logarithmic scale. The leftmost panel shows the target distribution, the others the distributions of samples generated by each model. (b) Quantile-quantile (QQ) plots. Each subplot shows empirical quantiles of the reference data (x-axis) against generated sample quantiles (y-axis). The red dashed line is the identity $y = x$. (c) Empirical intensity functions $\hat{\lambda}(t)$ estimated from the same sequences. (d) Absolute Pearson correlation-matrix differences for up to the first 20 event positions. Each heatmap entry shows the absolute difference between the empirical correlation of two sequence components in the generated samples and in the reference data. (e) Signed autocorrelation error for 5 lags. Each curve shows the difference between the ACF of the indicated model and the empirical ACF of the reference process, so that the target corresponds to the horizontal line at zero.

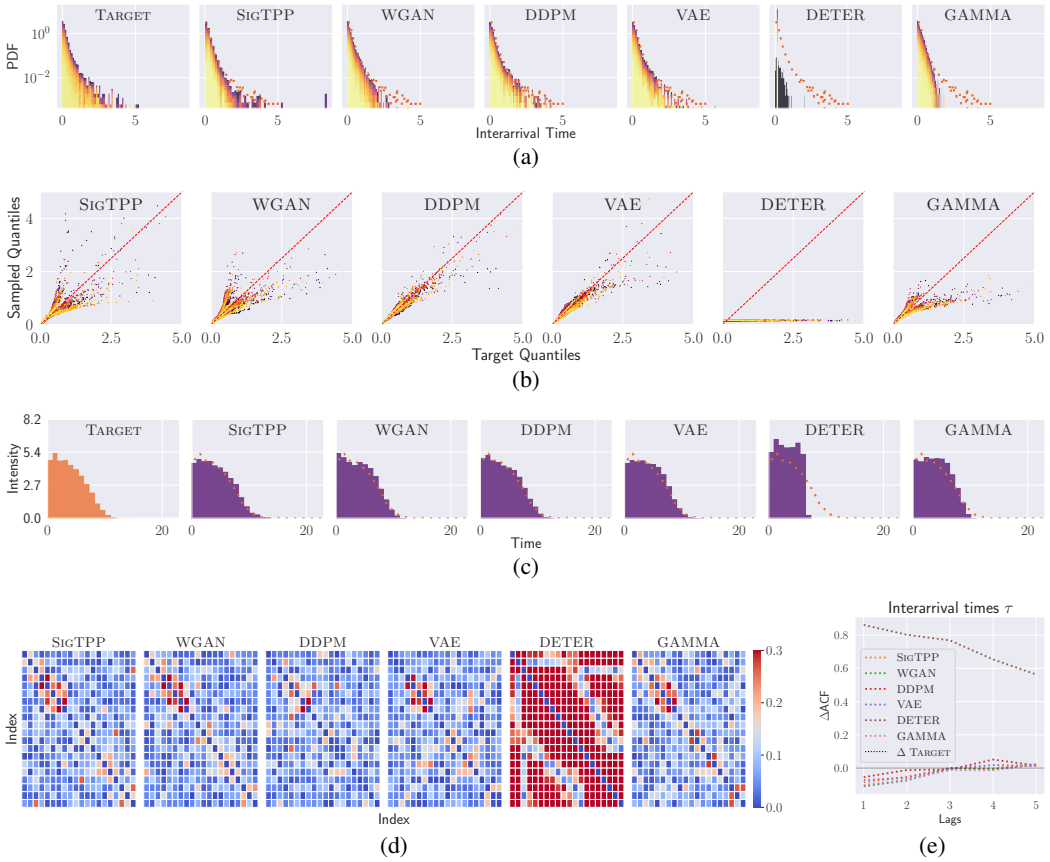


Figure S10: Qualitative evaluation on TX. Model panels are ordered SIGTPP, WGAN, DDPM, VAE, DETER, and GAMMA. (a) Histogram estimates of the interarrival times density $p(x)$ on a logarithmic scale. The leftmost panel shows the target distribution, the others the distributions of samples generated by each model. (b) Quantile-quantile (QQ) plots. Each subplot shows empirical quantiles of the reference data (x-axis) against generated sample quantiles (y-axis). The red dashed line is the identity $y = x$. (c) Empirical intensity functions $\hat{\lambda}(t)$ estimated from the same sequences. (d) Absolute Pearson correlation-matrix differences for up to the first 20 event positions. Each heatmap entry shows the absolute difference between the empirical correlation of two sequence components in the generated samples and in the reference data. (e) Signed autocorrelation error for 5 lags. Each curve shows the difference between the ACF of the indicated model and the empirical ACF of the reference process, so that the target corresponds to the horizontal line at zero.

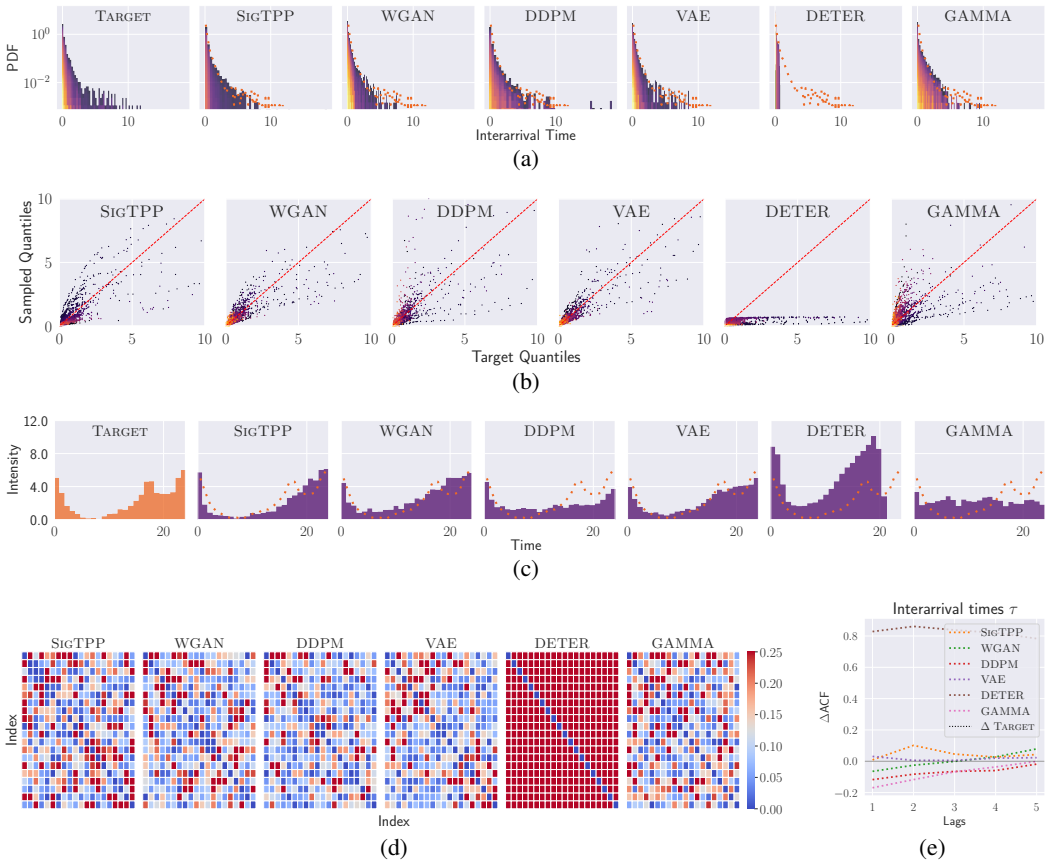


Figure S11: Qualitative evaluation on YLP. Model panels are ordered SIGTPP, WGAN, DDPM, VAE, DETER, and GAMMA. (a) Histogram estimates of the interarrival times density $p(x)$ on a logarithmic scale. The leftmost panel shows the target distribution, the others the distributions of samples generated by each model. (b) Quantile-quantile (QQ) plots. Each subplot shows empirical quantiles of the reference data (x-axis) against generated sample quantiles (y-axis). The red dashed line is the identity $y = x$. (c) Empirical intensity functions $\hat{\lambda}(t)$ estimated from the same sequences. (d) Absolute Pearson correlation-matrix differences for up to the first 20 event positions. Each heatmap entry shows the absolute difference between the empirical correlation of two sequence components in the generated samples and in the reference data. (e) Signed autocorrelation error for 5 lags. Each curve shows the difference between the ACF of the indicated model and the empirical ACF of the reference process, so that the target corresponds to the horizontal line at zero.

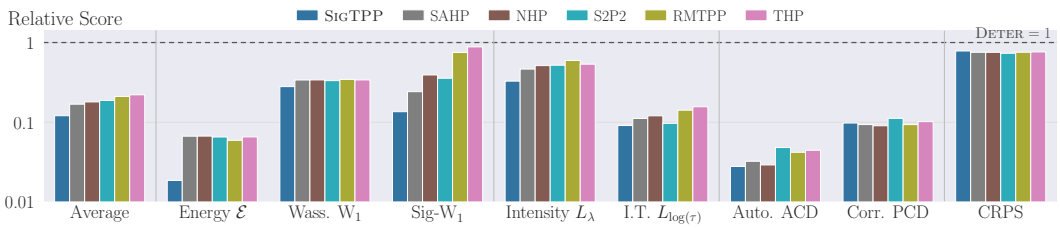


Figure S12: Geometric means of per-dataset score ratios to DETER over the 6 datasets PS, IP, H1, EQ, YLA, and TB. The vertical axis is logarithmic. The dashed line marks DETER $= 1$. Lower is better (↓). Metric definitions are provided in Appendix D.

Table S9: Performance of SIGREG for auxiliary-loss weights $\lambda \in \{0, 0.1, 1\}$ on PS, IP, H1, H3, SC, and SR. Setting $\lambda = 0$ recovers SIGTPP without regression. Values report standard error in parentheses over 100 bootstrap replicates. Lower is better (↓). Best in **bold**, second best underlined. Metric definitions in Appendix D.

Metric	λ	PS	IP	H1	H3	SC	SR
$\mathcal{E}$ ($\times 10^{-3}$)	0	1.52 (1.42)	1.37 (0.37)	8.23 (2.33)	3.64 (2.04)	1.10 (0.26)	1.88 (2.51)
	0.1	**0.81** (1.04)	1.69 (1.07)	5.63 (1.79)	**3.33** (1.95)	**0.92** (0.34)	**0.75** (1.50)
	1	1.08 (0.73)	**1.27** (0.40)	**1.28** (0.90)	3.85 (2.37)	0.99 (0.37)	3.67 (3.30)
W_1 ($\times 10^{-2}$)	0	5.59 (0.34)	6.92 (0.14)	8.44 (0.85)	11.0 (0.61)	2.53 (0.04)	11.1 (0.84)
	0.1	**5.37** (0.25)	**6.73** (0.23)	7.72 (0.92)	**10.9** (0.60)	**2.52** (0.04)	**9.79** (0.49)
	1	5.81 (0.27)	6.97 (0.13)	**3.96** (0.31)	11.3 (0.85)	2.54 (0.04)	10.2 (0.89)
Sig-W_1 ($\times 10^{-5}$)	0	4.59 (3.93)	2.75 (1.76)	22.4 (19.2)	1.01 (0.75)	0.51 (0.63)	3.40 (2.54)
	0.1	**1.39** (1.51)	**0.51** (0.56)	**17.1** (14.6)	1.77 (1.03)	**0.48** (0.68)	4.60 (2.95)
	1	4.60 (4.28)	5.68 (2.66)	39.7 (26.3)	**0.78** (0.69)	0.53 (0.78)	**2.51** (1.78)
CRPS ($\times 10^{-1}$)	0	**4.65** (0.08)	3.06 (0.03)	10.3 (0.24)	2.39 (0.03)	3.93 (0.05)	**4.42** (0.10)
	0.1	4.68 (0.08)	3.05 (0.03)	10.1 (0.24)	2.38 (0.03)	3.56 (0.03)	4.43 (0.10)
	1	4.65 (0.08)	**3.04** (0.03)	**10.0** (0.25)	**2.34** (0.03)	**3.50** (0.03)	4.43 (0.10)
L_λ ($\times 10^{-2}$)	0	3.30 (0.45)	3.27 (0.38)	5.44 (1.14)	3.04 (0.47)	4.09 (0.42)	3.89 (0.83)
	0.1	**3.13** (0.50)	**3.00** (0.35)	**5.08** (1.07)	3.21 (0.47)	3.13 (0.41)	4.22 (0.86)
	1	3.18 (0.48)	3.23 (0.38)	6.04 (1.06)	**2.71** (0.44)	**2.57** (0.40)	**3.73** (0.68)
$L_{\log(\tau)}$ ($\times 10^{-2}$)	0	2.90 (0.60)	2.79 (0.45)	16.4 (0.59)	7.99 (0.45)	5.50 (0.69)	2.15 (0.36)
	0.1	2.54 (0.51)	2.99 (0.46)	14.6 (0.60)	7.17 (0.44)	**3.40** (0.79)	**1.86** (0.41)
	1	**2.47** (0.64)	**2.19** (0.40)	**3.22** (0.74)	**2.11** (0.54)	4.28 (1.05)	2.00 (0.36)
ACD ($\times 10^{-2}$)	0	1.61 (0.47)	0.94 (0.22)	6.82 (0.71)	4.05 (0.45)	2.04 (0.38)	**0.66** (0.20)
	0.1	1.46 (0.45)	**0.73** (0.21)	6.42 (0.71)	**3.81** (0.45)	**1.55** (0.37)	0.66 (0.19)
	1	**1.35** (0.33)	1.05 (0.24)	**1.52** (0.39)	4.02 (0.45)	1.56 (0.37)	0.75 (0.20)
PCD ($\times 10^{-2}$)	0	7.34 (0.51)	6.08 (0.43)	10.3 (1.00)	8.58 (0.41)	5.87 (0.30)	7.85 (0.45)
	0.1	**7.33** (0.52)	**6.03** (0.43)	9.97 (0.96)	8.55 (0.41)	5.62 (0.30)	7.78 (0.44)
	1	7.38 (0.52)	6.09 (0.44)	**9.14** (0.86)	**8.47** (0.41)	**5.52** (0.30)	**7.76** (0.44)

Table S10: Performance of SIGREG for auxiliary-loss weights $\lambda \in \{0, 0.1, 1\}$ on EQ, SO, TB, TX, YLP, and YLA. Setting $\lambda = 0$ recovers SIGTPP without regression. Values report standard error in parentheses over 100 bootstrap replicates. Lower is better (↓). Best in **bold**, second best underlined. Metric definitions in Appendix D.

Metric	λ	EQ	SO	TB	TX	YLP	YLA
$\mathcal{E}$ ($\times 10^{-3}$)	0	**31.3** (0.58)	57.8 (10.3)	**512** (23.6)	**77.3** (2.94)	54.3 (10.4)	8.87 (5.68)
	0.1	33.4 (0.63)	14.4 (4.78)	534 (25.6)	79.9 (2.98)	89.9 (19.2)	14.8 (8.28)
	1	31.9 (0.61)	**14.1** (3.25)	529 (26.0)	79.1 (2.61)	**29.1** (13.3)	**8.79** (5.22)
W_1 ($\times 10^{-2}$)	0	**2.95** (0.05)	14.3 (1.16)	**57.3** (2.29)	**7.98** (0.19)	18.0 (1.43)	**8.46** (0.90)
	0.1	3.00 (0.05)	8.39 (0.99)	59.1 (2.45)	8.29 (0.22)	20.6 (1.64)	9.56 (0.95)
	1	2.95 (0.05)	**7.79** (0.73)	59.2 (2.65)	8.05 (0.18)	**14.7** (1.83)	8.71 (0.69)
Sig-W_1 ($\times 10^{-5}$)	0	790 (50.2)	**0.19** (0.17)	2.16 (0.89)	3.13 (0.96)	2.34 (1.72)	**5.79** (4.92)
	0.1	**457** (38.4)	0.51 (0.29)	**0.17** (0.19)	3.10 (1.00)	1.99 (1.68)	24.6 (11.0)
	1	566 (43.3)	0.22 (0.18)	0.21 (0.20)	**0.56** (0.38)	**0.69** (0.85)	8.20 (6.05)
CRPS ($\times 10^{-1}$)	0	5.85 (0.11)	13.3 (0.54)	0.37 (0.01)	1.14 (0.02)	4.80 (0.43)	4.82 (0.13)
	0.1	5.58 (0.10)	5.28 (0.09)	0.37 (0.01)	1.15 (0.02)	3.27 (0.12)	4.79 (0.14)
	1	**5.39** (0.11)	**5.13** (0.09)	**0.36** (0.01)	**1.11** (0.02)	**2.83** (0.12)	**4.76** (0.14)
L_λ ($\times 10^{-2}$)	0	22.7 (0.89)	**5.98** (0.41)	9.94 (1.15)	6.32 (1.03)	14.0 (1.15)	**13.7** (1.20)
	0.1	23.9 (1.14)	7.16 (0.64)	7.94 (0.96)	8.15 (0.96)	16.2 (0.91)	16.7 (1.17)
	1	**20.4** (0.99)	7.42 (0.53)	**4.29** (0.83)	**4.10** (0.85)	**9.40** (0.94)	14.6 (1.22)
$L_{\log(\tau)}$ ($\times 10^{-2}$)	0	16.7 (0.84)	8.17 (0.61)	13.4 (0.68)	4.18 (0.42)	**8.54** (1.73)	7.96 (1.09)
	0.1	14.4 (0.86)	**6.62** (0.81)	**10.3** (0.44)	**4.00** (0.47)	18.4 (1.41)	9.68 (1.20)
	1	**11.3** (0.90)	7.13 (0.86)	13.8 (0.48)	7.27 (0.58)	18.9 (0.70)	**7.28** (1.34)
ACD ($\times 10^{-2}$)	0	**5.42** (0.76)	13.0 (0.51)	2.13 (0.39)	3.72 (0.47)	**1.63** (0.33)	**2.23** (0.58)
	0.1	6.27 (0.76)	**6.47** (0.51)	1.10 (0.27)	2.73 (0.41)	3.80 (0.62)	2.83 (0.62)
	1	7.86 (0.76)	8.77 (0.51)	**0.97** (0.25)	**1.48** (0.36)	2.95 (0.60)	2.37 (0.56)
PCD ($\times 10^{-2}$)	0	8.42 (0.62)	12.6 (0.28)	5.85 (0.27)	6.25 (0.23)	13.7 (0.70)	12.4 (0.72)
	0.1	**7.84** (0.64)	13.8 (0.45)	**5.84** (0.26)	6.16 (0.24)	13.2 (0.69)	12.4 (0.73)
	1	10.1 (0.70)	**10.9** (0.32)	5.85 (0.25)	**5.80** (0.23)	**12.6** (0.69)	**12.3** (0.73)

Table S11: Dataset statistics. K denotes the number of event types (marks).

Dataset	K	# Events			# Sequences			Seq. Length		
		Train	Val	Test	Train	Val	Test	Min	Mean	Max
PS	3	14293	4796	4837	1200	400	400	2	12.0	24
IP	3	45232	14916	14944	3000	1000	1000	2	15.0	31
H1	1	11720	3638	3833	1200	400	400	2	9.6	36
H3	3	39357	12915	13061	1200	400	400	11	32.7	70
SC	1	24369	8138	8100	1200	400	400	18	20.3	23
SR	1	31368	10982	11081	1200	400	400	1	26.7	75
EQ	7	46358	6212	13845	3000	400	900	1	15.4	17
TB	17	72183	11272	27955	1300	200	500	27	55.7	63
TX	10	50454	7204	14420	1400	200	400	35	36.0	37
SO	22	88766	25266	26015	1401	401	401	38	63.6	100
YLA	1	5803	1991	1923	191	63	65	9	30.5	55
YLP	1	10542	3550	3531	191	63	65	3	55.2	107

F Information about Datasets

The benchmark contains 6 synthetic and 6 real-world datasets. Together they span memoryless, time-varying, self-exciting, renewal, and self-correcting generative regimes. All synthetic datasets are split 60%/20%/20% into training, validation, and test sets. Real-world datasets use the splits described below.

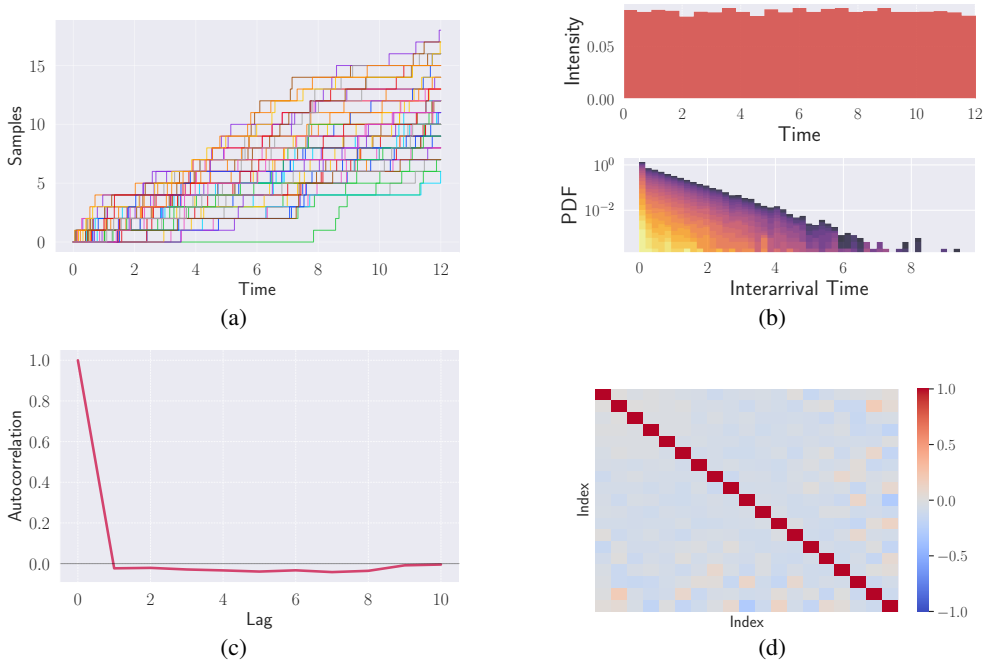


Figure S13: Diagnostics for the Poisson dataset (PS). (a) Representative sample paths from the dataset. (b) Empirical temporal summaries. Top: intensity function. Bottom: interarrival times PDF. (c) Empirical autocorrelation of interarrival times. (d) Correlation structure between interarrival times.

Homogeneous Poisson (PS). We generate 2,000 event sequences from a homogeneous Poisson process on $[0, T_{\max}]$ with horizon $T_{\max} = 12$ and constant rate $\lambda = 1$. Equivalently, interarrival times are sampled i.i.d. from $\mathrm{Exp}(1)$.

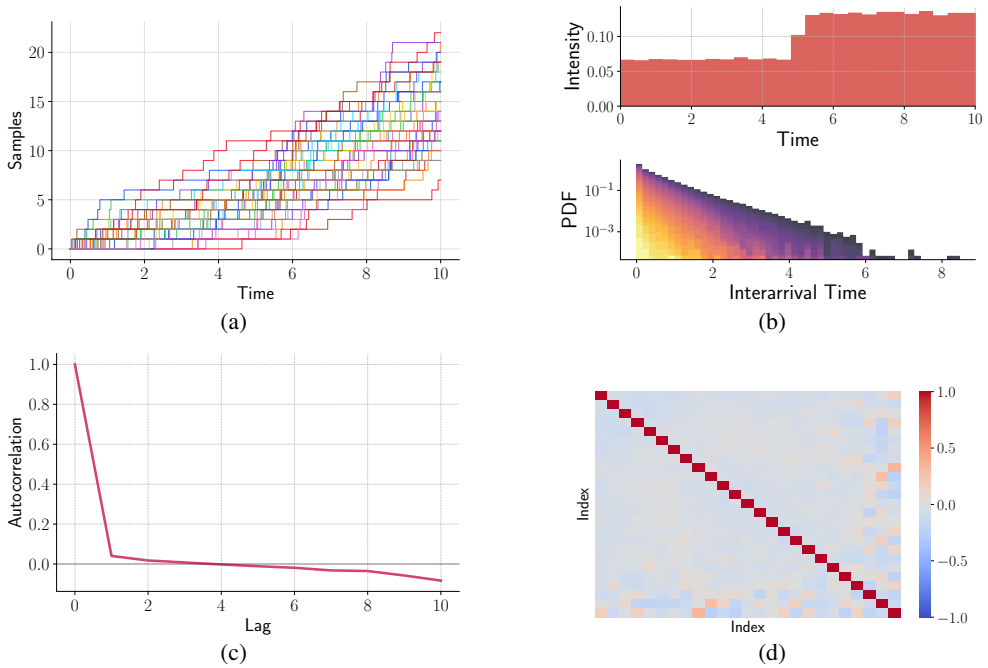


Figure S14: Diagnostics for the inhomogeneous Poisson dataset (IP). (a) Representative sample paths from the dataset. (b) Empirical temporal summaries. Top: intensity function. Bottom: interarrival times PDF. (c) Empirical autocorrelation of interarrival times. (d) Correlation structure between interarrival times.

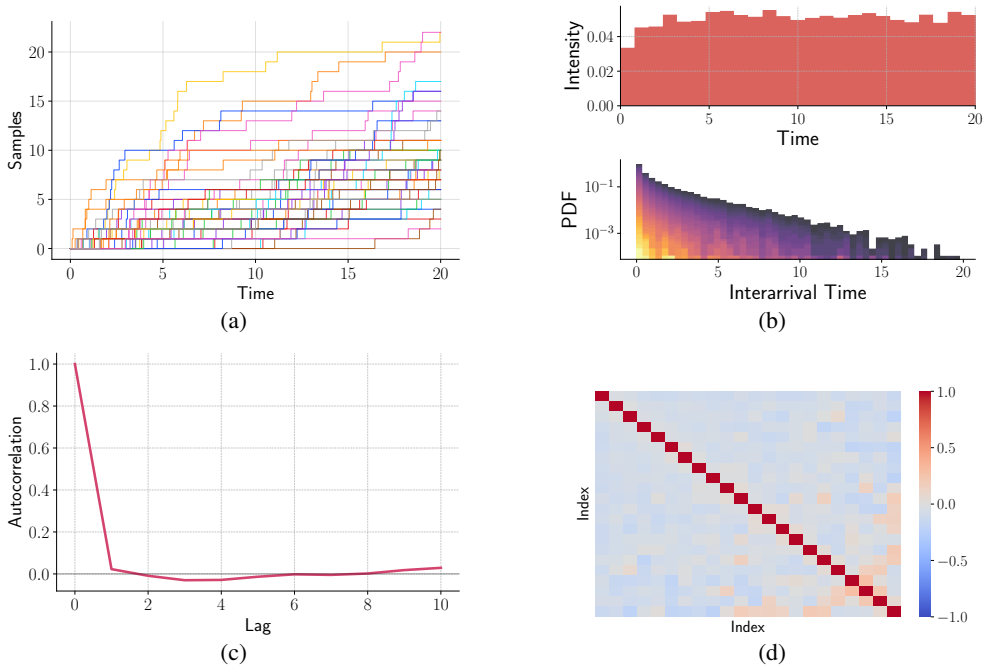


Figure S15: Diagnostics for the Hawkes dataset (H1). (a) Representative sample paths from the dataset. (b) Empirical temporal summaries. Top: intensity function. Bottom: interarrival times PDF. (c) Empirical autocorrelation of interarrival times. (d) Correlation structure between interarrival times.

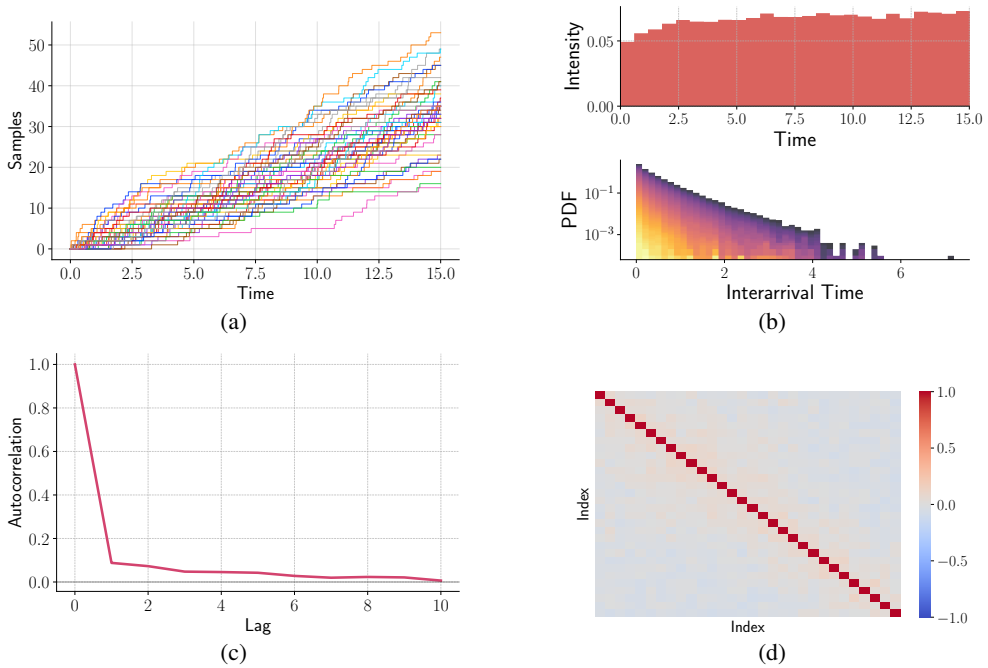


Figure S16: Diagnostics for the 3-dimensional Hawkes dataset (H3). (a) Representative sample paths from the dataset. (b) Empirical temporal summaries. Top: intensity function. Bottom: interarrival times PDF. (c) Empirical autocorrelation of interarrival times. (d) Correlation structure between interarrival times.

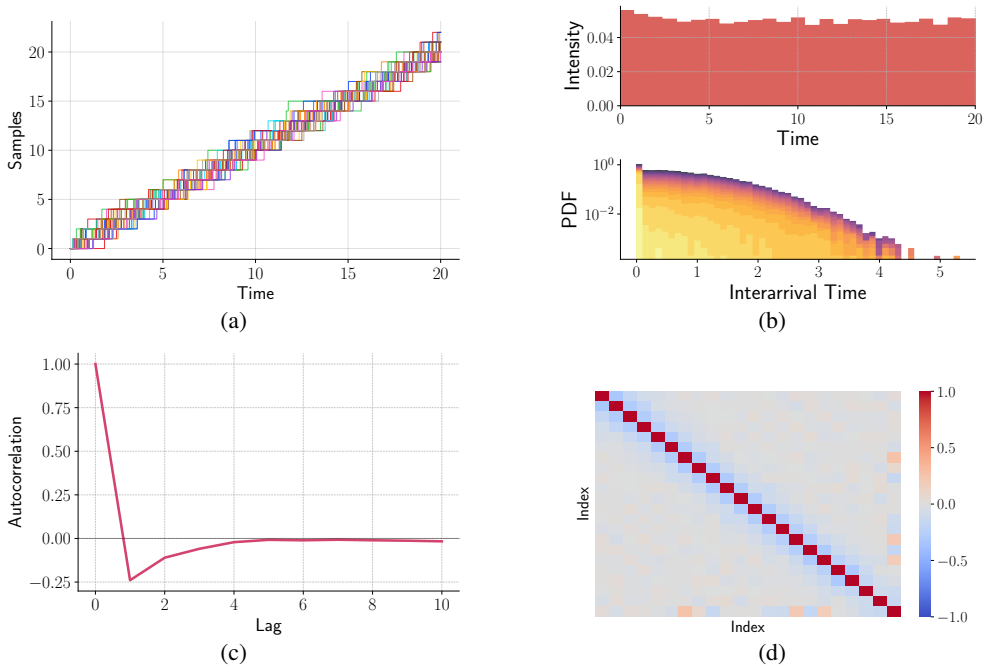


Figure S17: Diagnostics for the self-correcting dataset (SC). (a) Representative sample paths from the dataset. (b) Empirical temporal summaries. Top: intensity function. Bottom: interarrival times PDF. (c) Empirical autocorrelation of interarrival times. (d) Correlation structure between interarrival times.

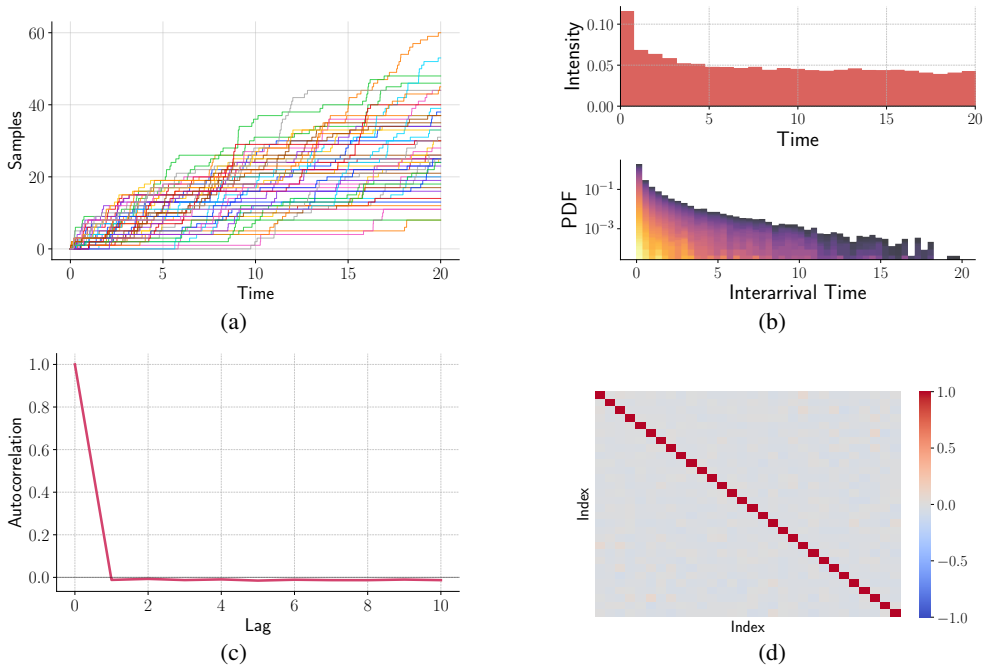


Figure S18: Diagnostics for the stationary renewal dataset (SR). (a) Representative sample paths from the dataset. (b) Empirical temporal summaries. Top: intensity function. Bottom: interarrival times PDF. (c) Empirical autocorrelation of interarrival times. (d) Correlation structure between interarrival times.

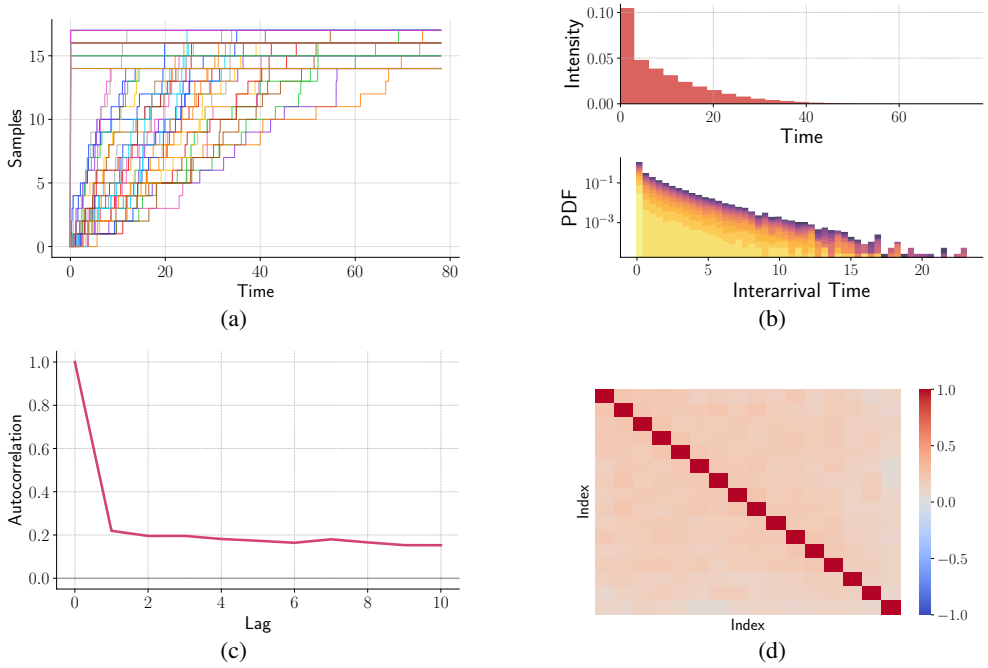


Figure S19: Diagnostics for the earthquake dataset (EQ). (a) Representative sample paths from the dataset. (b) Empirical temporal summaries. Top: intensity function. Bottom: interarrival times PDF. (c) Empirical autocorrelation of interarrival times. (d) Correlation structure between interarrival times.

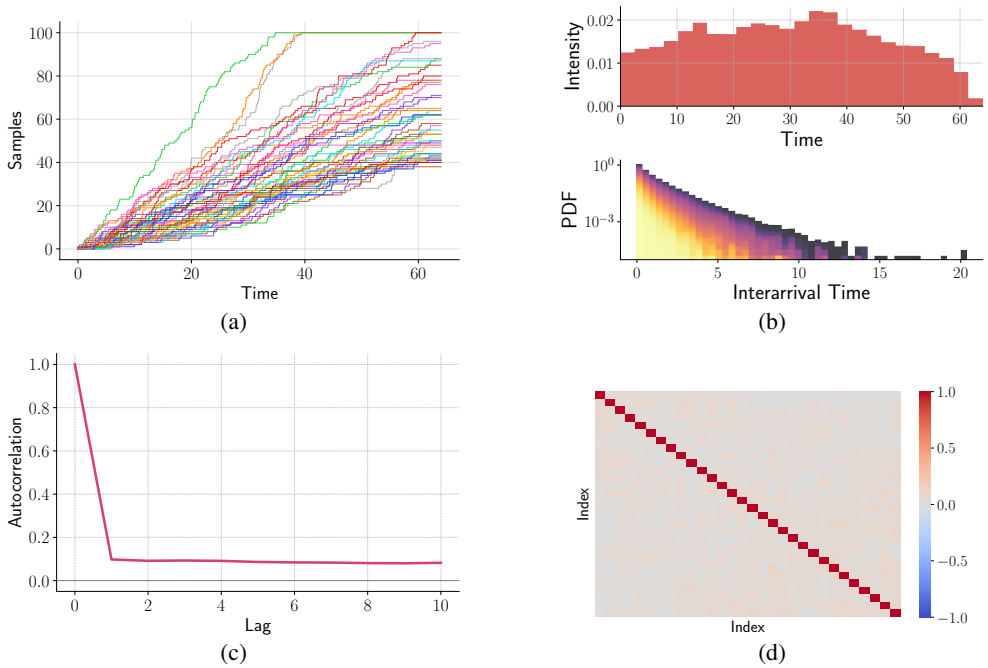


Figure S20: Diagnostics for the Stack Overflow dataset (SO). (a) Representative sample paths from the dataset. (b) Empirical temporal summaries. Top: intensity function. Bottom: interarrival times PDF. (c) Empirical autocorrelation of interarrival times. (d) Correlation structure between interarrival times.

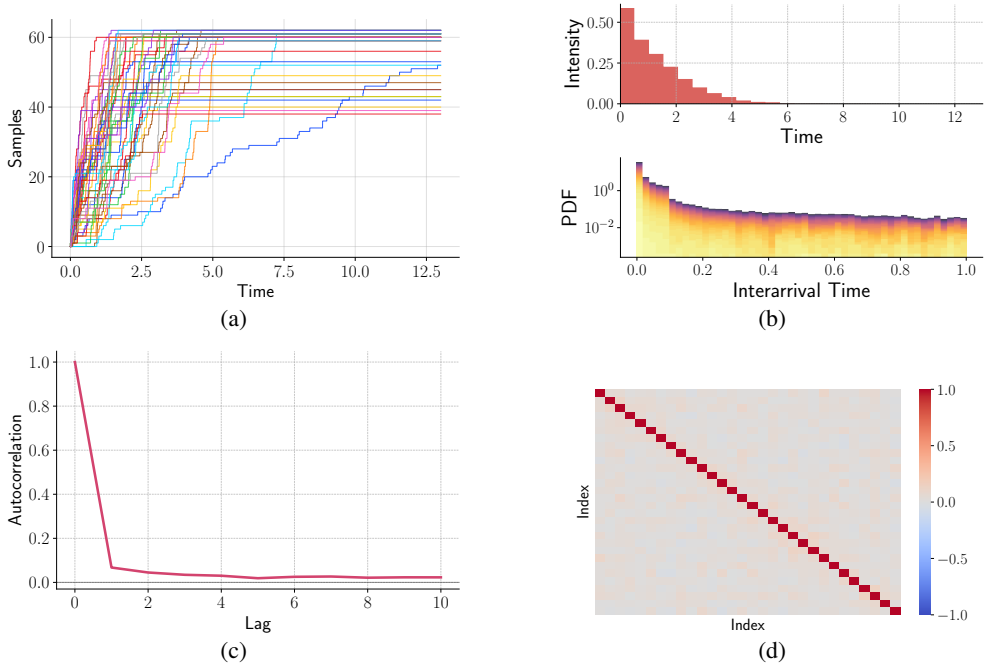


Figure S21: Diagnostics for the Taobao dataset (TB). (a) Representative sample paths from the dataset. (b) Empirical temporal summaries. Top: intensity function. Bottom: interarrival times PDF. (c) Empirical autocorrelation of interarrival times. (d) Correlation structure between interarrival times.

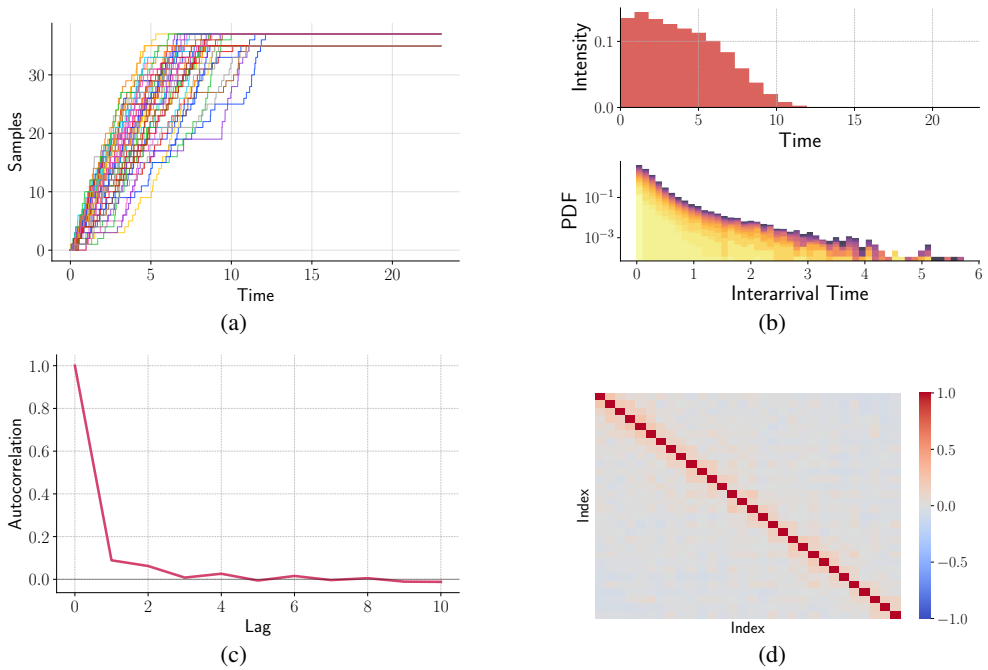


Figure S22: Diagnostics for the Taxi dataset (TX). (a) Representative sample paths from the dataset. (b) Empirical temporal summaries. Top: intensity function. Bottom: interarrival times PDF. (c) Empirical autocorrelation of interarrival times. (d) Correlation structure between interarrival times.

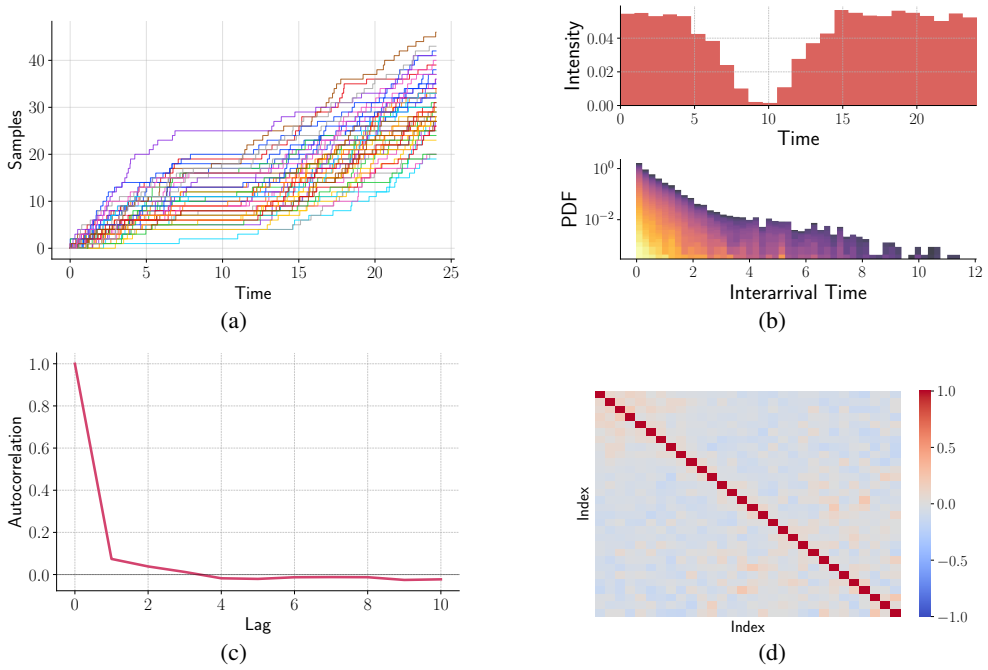


Figure S23: Diagnostics for the Yelp-Airport dataset (YLA). (a) Representative sample paths from the dataset. (b) Empirical temporal summaries. Top: intensity function. Bottom: interarrival times PDF. (c) Empirical autocorrelation of interarrival times. (d) Correlation structure between interarrival times.

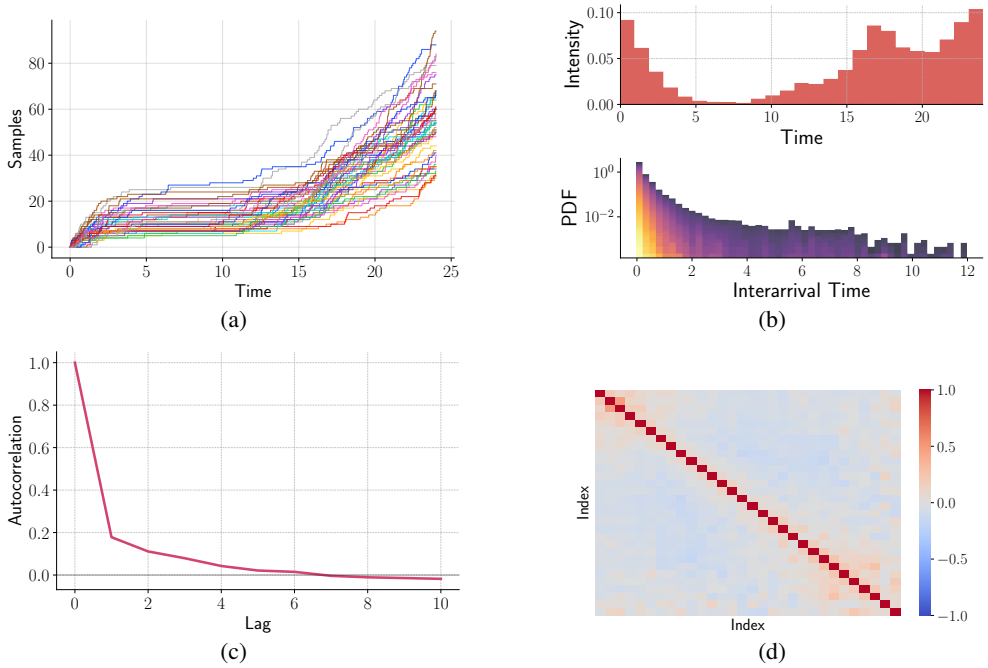


Figure S24: Diagnostics for the Yelp dataset (YLP). (a) Representative sample paths from the dataset. (b) Empirical temporal summaries. Top: intensity function. Bottom: interarrival times PDF. (c) Empirical autocorrelation of interarrival times. (d) Correlation structure between interarrival times.

Inhomogeneous Poisson (IP). We generate 5,000 sequences from an inhomogeneous Poisson process on $[0, T_{\max}]$ with $T_{\max} = 10$ and piecewise-constant intensity

$$\lambda(t) = 1 + \mathbb{1}_{t \geq T_{\max}/2}\,. \tag{S106}$$

Thus, the rate is 1 on the first half of the window and 2 on the second half. Sequences are simulated using the Lewis-Ogata thinning algorithm with upper bound $\bar{\lambda} = 2$.

Univariate Hawkes (H1). We generate 2,000 sequences from a univariate Hawkes process with exponential kernel,

$$\lambda(t \mid \mathcal{H}_t) = \mu + \alpha \sum_{t_i < t} \beta e^{-\beta(t - t_i)}, \tag{S107}$$

using $\mu = 0.3$, $\alpha = 0.4$, $\beta = 1.0$, and horizon $T_{\max} = 20$. Simulation is performed with the `tick` library. Since $\alpha < 1$, the process is stationary. This setting introduces temporal clustering through self-excitation and tests whether a model can capture history-dependent intensities.

Multivariate Hawkes (H3). We generate 2,000 sequences from a 3-dimensional Hawkes process with exponential kernels using `tick`. The baseline intensities are $\mu_k = 0.5$ for all k, the decay parameters are $\beta_{k\ell} = 1$ for all pairs (k, ℓ), and the excitation matrix is

$$A = \begin{pmatrix} 0.5 & 0.1 & 0.0 \\ 0.1 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.1 \end{pmatrix}, \tag{S108}$$

with horizon $T_{\max} = 15$. The 3 component streams are merged into a single time-ordered event sequence, and the source dimension is used as a categorical mark ($K = 3$). This dataset evaluates a model's ability to capture both cross-excitation across dimensions and mark-dependent intensity structure.

Self-correcting (SC). We generate 2,000 sequences on $[0, T_{\max})$ with $T_{\max} = 20$ from a self-correcting process with conditional intensity

$$\lambda(t \mid \mathcal{H}_t) = \exp\big(\mu t - \alpha N_t\big), \tag{S109}$$

using $\mu = 1$ and $\alpha = 1$. The intensity grows between events, and each event scales it by $e^{-\alpha}$. Sequences are therefore more regular than under a Poisson process: lengths range from 18 to 23 events, and consecutive interarrival times are negatively correlated.

Stationary renewal (SR). We generate 2,000 sequences from a renewal process on $[0, T_{\max})$ with $T_{\max} = 20$. Interarrival times τ_k are i.i.d. lognormal with mean 1 and standard deviation 6, that is $\log \tau_k$ is Gaussian with mean $-s^2/2$ and variance $s^2 = \log 37$. The heavy tail produces bursts of short gaps separated by long silences, and sequence lengths range from 1 to 75 events. Since each sequence starts with a fresh renewal at $t = 0$, the empirical intensity is highest near $t = 0$ before it settles. Interarrival times are independent by construction, so their autocorrelation vanishes at every nonzero lag.

SR and SC follow the synthetic processes of Omi et al. [44]. We simulate them with the samplers of Shchur et al. [50], which also generated the SR and SC data of the benchmark of Lüdke et al. [34].

Taxi (TX). Time-stamped taxi pick-up and drop-off events in New York City [56], partitioned by borough and action type into $K = 10$ event categories (5 boroughs $\times$ 2 actions). The subset comprises 2,000 randomly sampled drivers with a mean sequence length of 36 events. The standard EasyTPP train/validation/test split is used.

Stack Overflow. A dataset of user-interaction events on Stack Overflow, where event types correspond to distinct awarded-badge categories [56]. The standard EasyTPP train/validation/test split is used.

Taobao. User click sequences on the Taobao shopping platform, collected between 25 November and 3 December 2017 [56]. Item categories are ranked by frequency. The most frequent are retained and all remaining categories are merged into a single *other* class, yielding $K = 17$ event types. The subset comprises the 2,000 most active users with a mean sequence length of 56 events. The standard EasyTPP train/validation/test split is used.

Earthquake. Time-stamped seismic events recorded across the Conterminous United States from 1996 to 2023, derived from the USGS catalogue [56]. The standard EasyTPP train/validation/test split is used.

Yelp-Mississauga (YLP). This real-world dataset, taken from the benchmark of Lüdke et al. [34], contains user check-in event sequences for businesses in the city of Mississauga during 2018. As in that benchmark, it is unmarked ($K = 1$) and each sequence contains only ordered check-in times.

Yelp-Airport (YLA). This dataset comes from the same benchmark and Yelp source as YLP. It contains check-in times at McCarran International Airport during 2018, with 319 sequences on a 24-hour window. It is unmarked ($K = 1$) and split as YLP.

G Dataset Training Details

Categorical event types (marks). For datasets with $K > 1$ event types, event i carries a categorical mark $m_i \in \{1, \dots, K\}$, where Table S11 reports K. We concatenate a learnable embedding of m_i with the log-interarrival encoding before passing the result to the LSTM (with hidden state $h_i \in \mathbb{R}^H, H \in \mathbb{N}$), so h_i encodes both times and marks. A categorical head on h_i predicts the next mark and is trained by cross entropy against the observed m_{i+1} using observed histories (teacher forcing). During free-running generation, the head samples the next mark, whose embedding enters the next recurrent update. For unmarked datasets ($K = 1$), we omit the mark embedding, head, and loss.

Optimisation. All experiments were optimised using the Adam optimiser [29], with learning rates tailored to each specific task. The search implementation selects checkpoints by the mean of the position-wise validation interarrival and event-time histogram losses (`val_epdf`). It ranks configurations by the sum of minimum ranks on 8 validation diagnostics: Sig-W_1, the 2 position-wise histogram losses, $\mathcal{E}$, W_1, ACD, PCD, and CRPS. Pathwise diagnostics use free-running generation, and CRPS uses predictions conditional on observed histories.

Source Code Availability. The complete source code to reproduce the experiments of the paper is available at `github.com/Code-Cornelius/sigtpp`.

Software Environment. All numerical experiments were conducted on a system running Ubuntu 24.04.4 LTS with Python 3.8 and PyTorch 1.12.1+cu116. CUDA was available through PyTorch, with cuDNN version 8302.

Hardware Infrastructure. The machine is equipped with 3 NVIDIA L40S GPUs, each with 48 GB

Table S12: Hyperparameter search grid for all models. †Fixed value, not searched.

Model	Hyperparameter	Values
WGAN	Generator LR η_G	10^{-3}, 10^{-4}, 10^{-5}
	Discriminator LR η_D	10^{-2}, 10^{-3}, 10^{-4}
	Hidden size	16, 32
	Lipschitz reg. λ	10^{-2}, 10^{-3}, 10^{-4}
DETER	Learning rate η	10^{-2}, 10^{-4}
	Hidden size	8, 16, 32
GAMMA	Learning rate η	1, 10^{-2}, 10^{-4}
VAE	Learning rate η	10^{-3}, 10^{-4}, 10^{-5}
	Hidden size	16, 32
	Latent dimension	8, 16
	Reconstruction weight	0.01, 0.1, 1, 10
	KL annealing epochs†	50
	Free bits†	0.1
DDPM	Learning rate η	10^{-3}, 10^{-4}
	Hidden size	16, 32
	Diffusion steps T	10, 50, 100
	Batch size	512, 2048, 4096, 30000
SIGTPP	Generator LR η_G	10^{-3}, $5\cdot10^{-4}$, 10^{-4}
	Hidden size	16, 32
	Log-interarrival embedding dimension†	64
	Signature degree†	8
	Teacher forcing	True, False
	Terminal anchor	free, residual
	Detach cumulative channel	True, False
SIGREG	Generator LR η_G	10^{-3}, 10^{-4}
	Regression loss	L^1, L^2
	Noise mode	existing, fixed
	Conditional-loss weight λ	0, 0.1, 1

of VRAM, using NVIDIA driver version 580.126.09. The system is powered by dual AMD EPYC 9354 32-core processors, providing 128 CPU threads in total, with a maximum CPU frequency of 3.80 GHz. It has 768 GB of RAM and NVMe SSD storage.

Notes on SIGTPP hyperparameters. Table S12 lists the hyperparameters of SIGTPP. The terminal anchor determines how each path's endpoint is handled before computing its signature. The theoretical analysis uses the Φ-embedded representation on a fixed compact domain, which is the setting covered by the stability results. In the implementation, we evaluate two path representations during validation. The "residual" setting uses the Φ-based terminal anchor, while the "free" setting computes signatures directly over the observed variable-length interval ending at the last cumulative time. The "detach cumulative channel" flag changes only the backward pass. Detachment leaves the forward path, signature, and numerical value of Sig-W_1 unchanged, but drops the derivative through the cumulative-time coordinate.

Theorem 6 concerns the fixed-horizon embedding Φ of Definition 4. The residual option implements this embedding by appending the terminal knot $(T_{\max}, T_{\max} - t_m)$ in time-augmented coordinates, which records the final censored interval. The "free" option omits this knot and ends at the last event. The stated stability bounds do not cover this variant.

Lastly, training uses standardised degree-8 signature coordinates, whereas evaluation uses unscaled degree-3 coordinates.

Degree ablation. Table 3 uses a separate SIGTPP grid. It crosses the signature degree $M \in \{2,3,4,5,6,8\}$ with generator learning rate $\eta_G \in \{10^{-2}, 10^{-3}, 10^{-4}\}$, hidden size $\{16, 32\}$, and

teacher forcing in the signature branch (True, False), giving $3 \times 6 \times 2 \times 2 = 72$ configurations per dataset. It fixes the free terminal anchor, and no cumulative-channel detachment.

SIGREG ablation. For the ablation in Appendix E.3, the generator predicts each next event from the observed prefix in a separate teacher-forced pass and averages absolute (L^1) or squared (L^2) errors over all valid next-event positions in the batch. With noise mode equal to existing, the error acts on individual noisy outputs. With noise mode equal to fixed, it acts on the output with decoder noise fixed at its mean. The latter is a deterministic prediction, not an average over generated outputs. All runs fix hidden size 32, signature degree 8, the free terminal anchor and no cumulative-channel detachment.

H Limitations

A primary limitation of our approach is the computational cost of the truncated signature transform, whose feature dimension scales exponentially with the signature degree M and polynomially with the path dimension d. This may hinder direct extension to high-dimensional or marked TPPs. Second, our current embedding relies on a linear interpolation lift. While effective in practice, this lift introduces an artificial continuous transition at jump times and is therefore not fully faithful to the jump structure of the process. Third, our theoretical analysis relies on a determinacy property of the pushforward law of the temporal point process under the embedding, but such determinacy conditions are already difficult to verify in practice even for continuous processes, and we do not investigate them further here. Finally, our theoretical bounds are preliminary and likely conservative: the constants are loose, and a sharper characterisation of the embedding remains open. A better understanding of how the required truncation degree depends on the nature of the temporal point process, together with the relationship between moment conditions and the law of the process, is promising directions for future work.