TEST THE RECORD, NOT THE IMPUTATION: MODEL CRITICISM OF TEMPORAL POINT PROCESSES FROM AGGREGATED COUNTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Event data such as neural spike trains and activity logs are often recorded only as counts per interval. Exact-time diagnostics for temporal point processes, such as time rescaling, must then impute event times. For history-dependent models, we show that the imputation can decide the verdict: different imputations of the same record can reverse the diagnostic. We instead criticize a model through the distribution it induces on the recorded counts. For continuous-time neural generators this distribution is implicit, so our test, OBSCRITIC, estimates count forecasts by simulation or particle filtering. It then accumulates signed residuals that show which counts are mispredicted after which histories. Calibrating the whole procedure by Monte Carlo keeps the test valid in finite samples, even when the forecasts are approximate. On synthetic data where frequencies and adjacent transitions of count categories stay fixed, OBSCRITIC detects changes in higher-order dynamics that imputation-based tests miss. A jointly calibrated library of lag features detects the same alternative. On this matched alternative, signed count residuals are substantially more powerful than a PIT-based diagnostic using the same forecasts. OBSCRITIC also detects misspecification of a Transformer Hawkes Process and, on real datasets, rejects lower-capacity generators more often. These results support evaluating generators through their distributions on recorded counts.

1 INTRODUCTION

An aggregated event log records counts in time intervals. A continuous-time model generates event times, and a logger maps them to these counts. The discarded times lie outside the observed sigma-field. Nevertheless, a standard diagnostic reconstructs a timestamped path, applies time rescaling, and reports the result as a property of the record (Brown et al., 2002; Gerhard & Gerstner, 2010).

This mismatch is consequential for modern models of temporal point processes (TPPs). An *imputation* (also called a completion) is a timestamped path whose bin counts match the record. In recurrent or self-attentive generators, different imputations induce different latent histories and later forecasts. Figure 1 shows that these choices can reverse the rescaling verdict for the same record.

The correct null is the distribution obtained after passing the generator through the logger. If a latent path $X \sim P_\theta$ and a *logger* $\mathcal{C}$ produces the *record* $Y = \mathcal{C}(X)$, then the scientific target is the pushforward $P_\theta^Y = \mathcal{C}_\# P_\theta$. Forecasts must condition on $\mathcal{F}_{t-1}^Y = \sigma(Y_1, \ldots, Y_{t-1})$, not on an analyst-selected imputation. This is a prequential problem—sequential forecasts are judged by what materializes next (Dawid, 1984)—with a logger that determines the observable experiment (Heitjan & Rubin, 1991).

Existing path critics assume event configurations or evaluable intensities (Yang et al., 2019; Baum et al., 2023; Baars et al., 2025; Kling & Vetter, 2024). Interval-censored Hawkes models fit aggregated counts with specialized approximations (Rizoiu et al., 2022); methods that use SMC in state space and particle-MCMC methods reconstruct Hawkes histories that agree with the recorded counts for parameter inference (Chen et al., 2025; Lambe et al., 2025). Censoring-aware inference integrates missing events for prediction (Boyd et al., 2023; Fernandez et al., 2020). Conditional jitter tests fine timing given coarse counts (Amarasingham et al., 2012). Generic predictive checks compare chosen statistics under simulation (Gelman et al., 1996); they do not construct sequential residuals at the

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

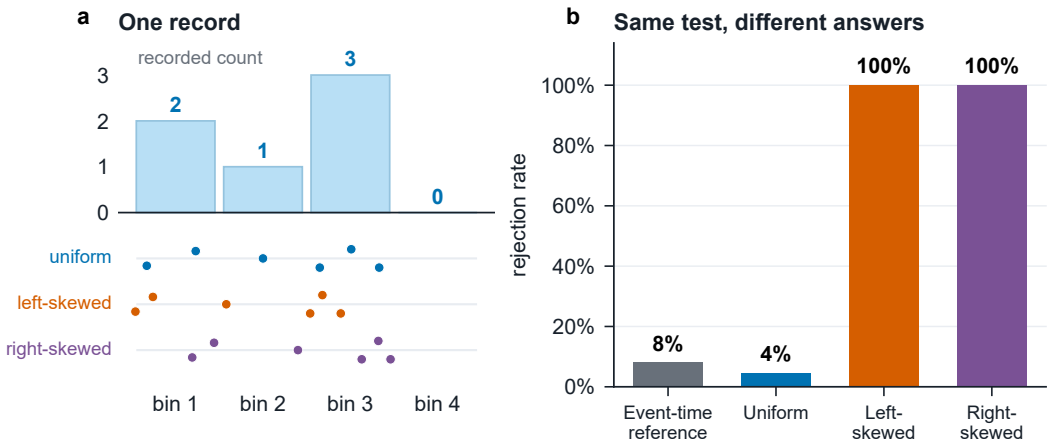

Figure 1: **Imputation changes the diagnostic for the same record.** Within-bin imputations preserve every recorded count but can reverse the rescaling verdict without outer recalibration. Vertical offsets only separate nearby events. Table 1 compares calibrated procedures; Table S4 gives their levels.

logger's filtration. ObsCritic evaluates a supplied generator by identifying mispredicted count categories and calibrating the approximate computation, without evaluating the marginal likelihood of the counts.

Its contributions are:

1. **A conditional discrepancy for the record produced by the logger.** We formulate criticism at the observable pushforward. Signed count residuals indexed by recorded history identify which counts fail after which history states, using either a lookup table estimated from simulations or particle filtering.
2. **Calibration of the implemented critic.** Under the null, exchangeable records processed with the same randomized procedure yield a Monte Carlo p-value valid in finite samples, even with approximate forecasts.
3. **A benchmark for conditional dynamics with matched path summaries.** We preserve each path's frequencies of count categories and transitions between adjacent categories, then alter its second-order organization. Same-feature controls and a joint lag library test the roles of conditional forecasting and feature choice.

2 Criticism on Recorded Counts

The test asks whether the model can generate the counts that remain after the logger discards event times.

Why imputation cannot identify the target. For any record y, the set $\mathcal{C}^{-1}(y)$ of latent paths consistent with that record generally contains many elements. An exact-path statistic $D(x)$ is a function of the record only when it is constant on every such set. Recurrent diagnostics usually fail this condition because changing within-bin times changes the hidden state.

Let $X = (\tau_i)_{i\geq 1}$ be a latent temporal point process from a generator P_θ. The logger maps X to $Y_{1:T}$. In the main setting,

$$Y_t = \#\{i : \tau_i \in (b_{t-1}, b_t]\}, \qquad H_0 : Y_{1:T} \sim \mathcal{C}_{\#} P_\theta. \tag{1}$$

The model is fit on data independent of the tested record, with parameters held fixed during testing. If the same data are used for fitting and testing, the fitting procedure must be repeated inside every null replicate or handled by selective inference.

For counts, define the atom dictionary

$$\phi(Y_t) = \big(\mathbf{1}\{Y_t = 0\}, \mathbf{1}\{Y_t = 1\}, \mathbf{1}\{Y_t = 2\}, \mathbf{1}\{Y_t = 3\}, \mathbf{1}\{Y_t \geq 4\}\big)^\top. \tag{2}$$

Each coordinate indicates a count category, which we call an atom. The forecast given the recorded past and its residual are

$$\mu_t = \mathbb{E}_\theta[\phi(Y_t) \mid \mathcal{F}_{t-1}^Y], \qquad e_t = \phi(Y_t) - \mu_t. \tag{3}$$

Each coordinate has a direct interpretation: a persistent positive residual for $Y_t = 2$ means two-event bins occur more often than predicted after conditioning on the recorded past.

3 ObsCritic

The statistic accumulates forecast errors within states of the recorded history.

3.1 Two implementations of count forecasts

For a history feature h_t with finitely many values, computed from past counts (Section 3.2), *ObsCritic-table*, the simulation-based lookup-table implementation, estimates $P_\theta(\phi(Y_t) \mid h_t)$ from generator simulations passed through the logger. The residuals can be tabulated by history state and count category when the history feature takes few values.

ObsCritic-particle, the particle-filter implementation, handles forecasts that depend on a continuous or model-native latent state. It uses the generator's state transition, future sampler, next-event density, and survival factor. These event-level factors score latent segments that contain the observed number of events in the bin. We call these matching segments. The probability $P_\theta(Y_t = n \mid \mathcal{F}_{t-1}^Y)$ of the recorded count is estimated by simulation. At the start of bin t, particles represent latent histories compatible with $Y_{1:t-1}$.

Prediction precedes assimilation: the implementation predicts the next count category, records its residual, and then conditions the particles on the observed bin. This order keeps Y_t out of its own forecast $\widehat{\mu}_t$.

For an observed count n in a bin of width Δ, the proposal draws and sorts n independent uniform positions,

$$q(u_{1:n} \mid n) = \frac{n!}{\Delta^n} \mathbf{1}\{0 < u_1 < \cdots < u_n < \Delta\}. \tag{4}$$

It has support on every matching segment. This ordered-uniform proposal with event-density and survival weighting follows the interval-count Hawkes SMC construction of Chen et al. (2025) and Lambe et al. (2025), which we reuse to forecast recorded counts. The importance weight is the event density along the proposed gaps times survival to the bin boundary, divided by Eq. (4). Because q is constant across particles conditional on n, its factor cancels during normalization. For $n = 0$, the proposal is a singleton weighted by the survival probability.

Algorithm 1: ObsCritic-particle
Input: generator and logger; observed counts $Y_{1:T}$; atom map $\phi : \mathcal{Y} \to \mathbb{R}^K$; predictable feature $h_t \in \mathbb{R}^d$; N particles; M predictive draws; outer bank size B.
For each observed or model-simulated group: initialize generator states and ages, $S \leftarrow 0$, $V \leftarrow 0$. For $t = 1, \ldots, T$: (1) sample particle ancestors with replacement and simulate M complete next-bin futures. (2) form $\widehat{\mu}_t$ from their atoms, set $\widehat{e}_t = \phi(Y_t) - \widehat{\mu}_t$, and, after burn-in, update $S \leftarrow S + h_t \otimes \widehat{e}_t$ and $V \leftarrow V + (h_t h_t^\top) \otimes \frac{M+1}{M-1}[\mathrm{diag}(\widehat{\mu}_t) - \widehat{\mu}_t \widehat{\mu}_t^\top]$. (3) only now assimilate Y_t: draw sorted-uniform matching segments, compute event-density/survival importance weights, and systematic-resample N particles. Return $Q = S^\top V^+ S$ and the signed coordinates of S.
Calibrate: process the B model-simulated groups specified in Section 3.3 and report the upper rank in Eq. (6).

The factor $(M + 1)/(M - 1)$ in Algorithm 1 corrects the covariance scale when the observation and the M predictive draws are conditionally iid from the same categorical law (Appendix A.1); at a fixed predictive budget, this positive constant scaling leaves rank decisions unchanged.

3.2 PREDICTABLE LOCALIZATION

Let $h_t \in \mathbb{R}^d$ be measurable with respect to $\mathcal{F}_{t-1}^Y$ and let $\widehat{e}_t \in \mathbb{R}^K$ denote the residual for count categories. The accumulated score and quadratic statistic are

$$g_t = h_t \otimes \widehat{e}_t \in \mathbb{R}^{dK}, \qquad S_T = \sum_{t=1}^{T} g_t, \qquad Q_T = S_T^\top \widehat{V}_T^+ S_T, \tag{5}$$

where $\widehat{V}_T = \sum_t (h_t h_t^\top) \otimes \widehat{\Sigma}_t$ is the $dK \times dK$ predictable covariance and $+$ denotes the Moore–Penrose inverse on the identifiable categorical subspace. For one-hot history states, this is the block-diagonal covariance implemented in our experiments. Reshaping S_T into a $d \times K$ matrix gives a signed residual table. Its coordinates identify over- and under-predicted multiplicities in each history state.

3.3 CALIBRATION OF THE IMPLEMENTED ALGORITHM

Let G_0 be the observed group and $G_1, \ldots, G_B$ independent groups drawn from the model and passed through the same logger. Every group is processed by the same randomized implementation: particle count, predictive budget, resampling rule, regularization, and auxiliary-seed distribution are held fixed. The upper-tail p-value is

$$p_{\mathrm{MC}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{Q_b \geq Q_0\}}{B+1}. \tag{6}$$

Under H_0, the $B+1$ statistics are exchangeable, so Eq. (6) is super-uniform (Besag & Clifford, 1989).

4 GUARANTEES AND INFORMATION BOUNDARY

Property (Imputation invariance)**.** *If two latent event sequences map to the same record y, then for the same auxiliary seed* OBSCRITIC *returns the same statistic and signed residual table on both sequences.*

Corollary 1 (Monte Carlo rank validity)**.** *Under Eq.* (1) *and the fitting and simulation conditions of Sections 2 and 3.3,* $\Pr(p_{\mathrm{MC}} \leq \alpha) \leq \alpha$ *for the test as run, including particle approximation.*

Proposition 1 (Observed-filtration residual)**.** *If the exact conditional forecast μ_t is used and h_t is predictable, then $h_t \otimes e_t$ is a martingale difference sequence with respect to $\mathcal{F}_t^Y$.*

The proposition concerns the full-past forecast in Eq. (3). A conditional lookup table targets a coarser state h_t; the test as run is calibrated by Section 3.3.

Property (Information boundary)**.** *If two latent models have the same pushforward law under $\mathcal{C}$, every test measurable from Y has identical size and power for the two models. More generally, the size–power difference is bounded by the total-variation distance between their observed pushforward laws.*

5 EXPERIMENTS

We test conditional dynamics, compare generators on recorded data, and examine the limits of the chosen count categories.

5.1 EVALUATION PROTOCOL

Seeds, alternatives, observation windows, thresholds, decision criteria, and method configurations were fixed before each corresponding evaluation; model parameters remain unchanged during evaluation. The observed statistic comes from a held-out group, and null simulations follow Section 3.3. The matched comparison in Section 5.2 uses the largest set of calibration and evaluation paths; Appendix A.7 lists independent replications. Table S2 and Appendix C.1 specify each design and calibration rule.

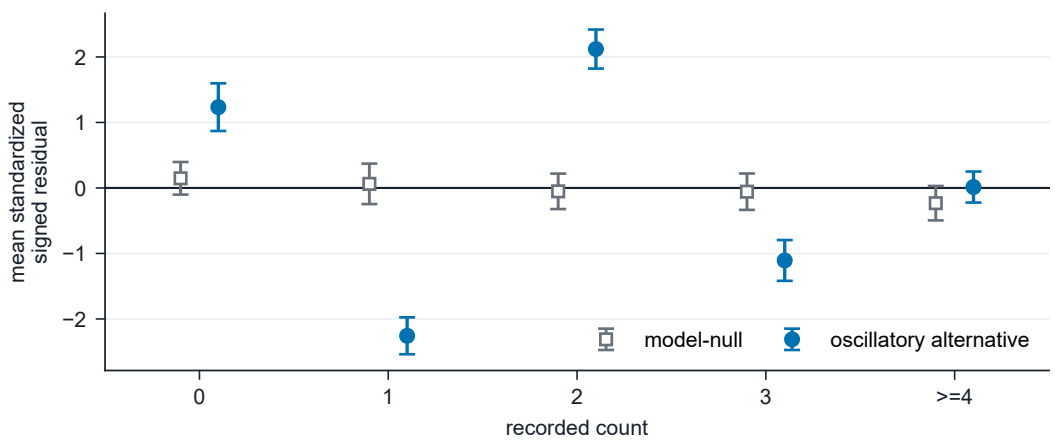

Figure 2: **Signed residuals localize the error.** The oscillatory alternative gives the recurrent generator negative self-connections (Appendix A.8). This alternative produces signed excesses and deficits in the displayed count categories; model-null means remain near zero. Bars show intervals for the mean standardized residual; Table S3 specifies their level and sample size.

Table 1: Matched comparison after bin reordering at the level specified in Table S4. The alternative preserves frequencies of count categories and transitions between adjacent categories. Both implementations use the same lag-2 feature and residual statistic.

Method	Model-null ($n = 300$)	Matched power ($n = 300$)
ObsCritic-particle, lag-2	**12/300 (4.0%)**	**203/300 (67.7%)**
ObsCritic-table, lag-2	**15/300 (5.0%)**	**193/300 (64.3%)**
ObsCritic-particle, intercept	17/300 (5.7%)	19/300 (6.3%)
Empirical atom frequency	18/300 (6.0%)	19/300 (6.3%)
First-order table	16/300 (5.3%)	17/300 (5.7%)
Q20	15/300 (5.0%)	89/300 (29.7%)
Max imputation	16/300 (5.3%)	13/300 (4.3%)

Within each calibrated comparison, all methods receive the same record and are calibrated separately under that study's rule. ObsCritic-table estimates probabilities of count categories given recent count categories from independent model simulations. We compare four kinds of tests. Imputation tests fill in event times uniformly within each bin and apply time-rescaling KS. A randomized probability integral transform (PIT) test using the Ljung–Box statistic (Q20) measures autocorrelation in randomized ranks derived from ObsCritic-particle's forecasts. Frequency and first-order controls use count-category frequencies or transitions between adjacent categories. Path MMD (Gretton et al., 2012) compares whole records. Appendix C.1 gives the statistics and configurations.

The matched study uses the upper rank in Eq. (6). The oscillatory and Transformer studies use an empirical-quantile threshold with a different Monte Carlo rejection level. Table S4 records the exact levels, numbers of calibration paths, and no-ties calculation.

5.2 Conditional dynamics under exact matching

For each null path, we build an alternative by reordering its bins to favor returns of the form $a \to b \to a$. Here a and b denote the categories $0, 1, 2, 3, \geq 4$, encoded by $A_t = \min(Y_t, 4)$. Every raw count value and every consecutive category pair appears exactly as often as before. Bins retain their counts and within-bin event positions; Appendix A.7 gives the construction. Controls based only on category frequencies or adjacent-category transitions therefore cannot separate the paired paths. ObsCritic indexes residuals by the category two bins earlier, $h_t = \mathrm{onehot}(\min(Y_{t-2}, 4))$; ObsCritic-table uses the same lag-2 feature and residual statistic.

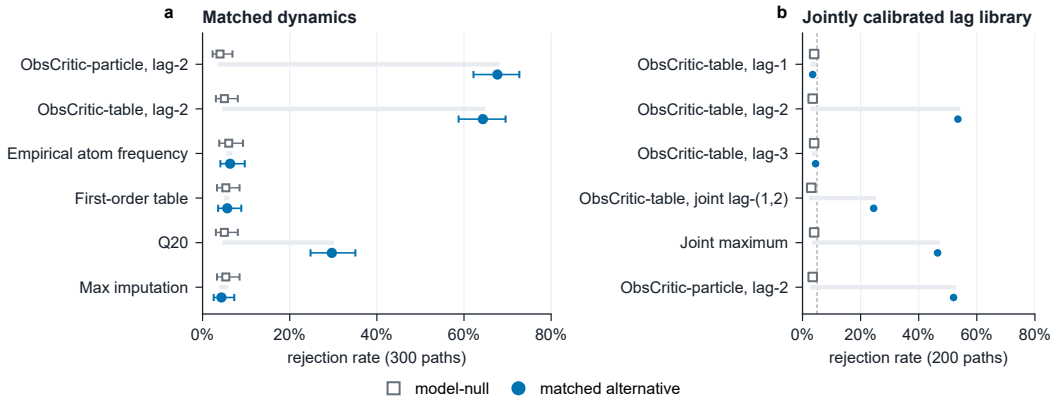

Figure 3: **Conditional dynamics survive matching frequencies and adjacent transitions of count categories.** (a) Model-null rejection and power in the main matched study; Table 1 gives the values and Table S5 the Wilson intervals. (b) A jointly calibrated library of lag features detects the matched alternative; the wrong lag alone remains at size.

ObsCritic-particle attains **67.7%** power and ObsCritic-table **64.3%**, while Max imputation reaches **4.3%** (Table 1). The separation comes from indexing residuals by recorded lag-2 states. Frequency and first-order controls retain null-level rejection because their summaries are unchanged. The same-feature comparison shows that conditional features provide the main separation, with additional detections from particle forecasting. Appendix C.2 reports intervals, paired decisions, matching checks, and runtime.

A jointly calibrated library of lag features detects the same matched alternative. Calibrating the maximum together accounts for searching across the included lag features. Its maximum over the lag and joint-lag tables reaches **46.5%** power; the mismatched lag alone remains at size (Figure 3; Table S6). The qualitative separation also appears in independent replications of the matched comparison. Table S1 reports these replications together with the recurrent oscillatory, sign-flipped, conditional-imputation, and log-score studies.

5.3 Real-data capacity ladder

We use the public TriTPP benchmarks distributed with Add-and-Thin (Shchur et al., 2020b; Lüdke et al., 2023). Taxi records pickups in southern Manhattan. The two Yelp streams record check-ins at McCarran Airport and across Mississauga businesses. PUBG records player deaths within matches. Twitter records one user's posting times.

Increasing generator capacity removes the discrepancies in the chosen count categories across these datasets (Figure 4). The comparison changes the generator while keeping the critic and the chosen count categories the same. The capacity ladder compares a recurrent exponential generator with increasingly expressive LogNormal mixtures. Taxi worsens at the intermediate generator before the strongest model removes the detected discrepancy; PUBG clears the statistic at both mixture levels. Table S7 gives the capacity configurations, dataset counts, and model-null evaluation.

At the strongest capacity, Path MMD still detects global discrepancies for some generators even though ObsCritic finds no mismatch in the chosen count categories. The tests inspect different properties: chosen count categories given history for ObsCritic and global differences between paths for MMD. ObsCritic conditions its forecasts for next-bin count categories on recorded history; MMD uses a kernel on paths. Table S7 and Figure 6 report the comparison. Table S1 also reports particle-budget, irregular-grid, thinning-and-clutter, and saturation checks, including settings with low power.

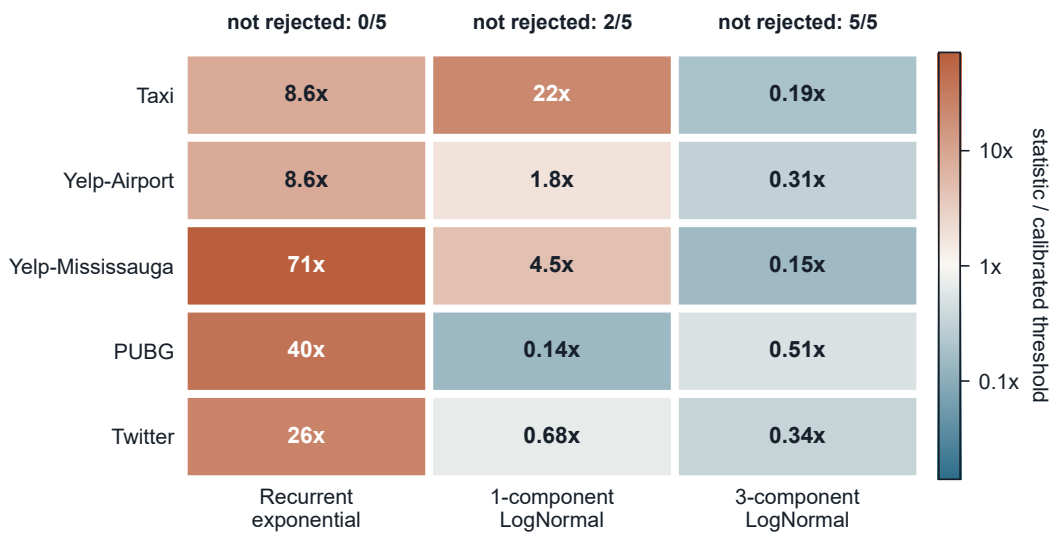

Figure 4: **Calibrated criticism tracks generator capacity.** Each cell is the observed statistic divided by its calibrated threshold; values below one indicate no detected mismatch in the chosen count categories. The same critic is used at every capacity. Taxi and PUBG are not monotone across individual levels; Section 5.3 discusses the intermediate generator.

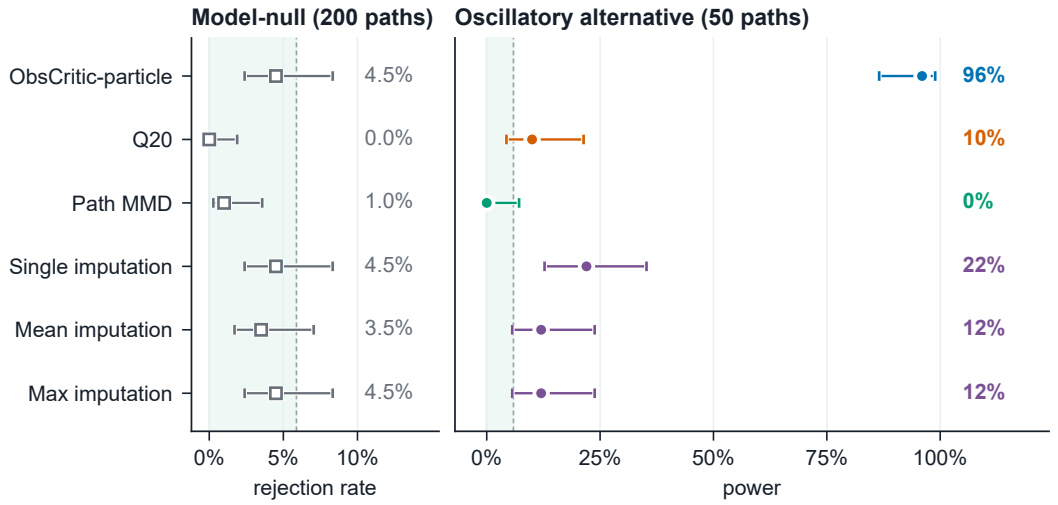

Figure 5: **An oscillatory stress test with an official Transformer generator.** Squares show model-null rejection and circles show power; horizontal bars are Wilson intervals. The shaded band ends at this study's Monte Carlo rejection level. Table S8 gives sample sizes, intervals, and results.

5.4 Transformer Hawkes Process

The official EasyTPP (Xue et al., 2024) implementation of the Transformer Hawkes Process (THP) (Zuo et al., 2020) provides a modern generator for the oscillatory study. The model, thresholds, methods, and set of alternatives are shared across the null evaluations (Appendix C.4).

Under this study's Monte Carlo rejection level, ObsCritic-particle attains **96%** power versus **22%** for the strongest imputation baseline (Figure 5). Here the alternative changes the marginal count law, leaving a discrepancy in the recorded counts. Empirical atom frequency is also effective. Table S8 gives every method's null rejection and power, including the low-power Q20 and Path MMD results, and the reported confidence intervals.

6 Related work

Point-process goodness of fit. Discrete-time rescaling corrects the finite-resolution bias of continuous-time rescaling for discretized spike trains (Haslinger et al., 2010). Time rescaling transforms exact event paths using an evaluable conditional intensity (Brown et al., 2002). Stein–Papangelou tests address event configurations (Yang et al., 2019); sequential kernel Stein tests extend criticism to discrete sequences of variable length when the target law or its probability ratios can be scored (Baum et al., 2023). Generalized-score tests handle misspecification in self-exciting models using a quasi-likelihood score that can be evaluated (Wei et al., 2021), while Residual TPP decomposes event-level residuals (Yuan & Fang, 2025). Bootstrap and innovation-martingale tests add consistency or asymptotic guarantees that are distribution-free for observed paths of point processes (Kling & Vetter, 2024; Baars et al., 2025). ObsCritic addresses the complementary case in which the logger induces a conditional count law that cannot be evaluated directly.

Neural generators. RMTPP learns a recurrent representation of event history for intensity-based prediction (Du et al., 2016). LogNormMix models conditional inter-event times with log-normal mixtures, enabling likelihood evaluation and sampling (Shchur et al., 2020a).

Censoring, coarsening, and latent-time SMC. Interval-censored Hawkes processes build specialized models for fitting aggregate counts (Rizoiu et al., 2022); mark-censored inference marginalizes missing event types for prediction (Boyd et al., 2023); censoring-aware KSDs treat time-to-event data (Fernandez et al., 2020). Most directly, Chen et al. (2025) use a state-space representation and SMC likelihood estimate for discretely observed Hawkes inference, while Lambe et al. (2025) use ordered-uniform latent event times and importance weights in multivariate particle MCMC. ObsCritic turns this inner particle machinery toward evaluation through forecasts of recorded counts, signed residuals by history, and calibration of the entire procedure.

Randomization and predictive checking. Conditional jitter resamples fine spike timing while conditioning on coarse counts (Amarasingham et al., 2012); its null concerns sub-bin temporal structure. Our null is instead the unconditional pushforward of a supplied generator through its logger. Posterior predictive checks and Monte Carlo tests provide the outer simulation logic (Gelman et al., 1996; Besag & Clifford, 1989). The contribution here is the inner discrepancy: signed count residuals indexed by recorded history.

Forecast evaluation and simulation calibration. Prequential analysis evaluates sequential probability forecasts by their realized outcomes (Dawid, 1984). For count data, randomized PITs, marginal calibration diagrams, and proper scores provide a mature toolbox when the one-step predictive mass function is available (Czado et al., 2009; Gneiting & Raftery, 2007). Conditional-density Stein tests provide localized discrepancies for differentiable conditional laws (Jitkrittum et al., 2020). PGF and Pearson-residual tests likewise diagnose specified count time-series laws (Hudecová et al., 2015; Weiß et al., 2020). ObsCritic constructs a forecast through the logger and evaluates its approximation by exchangeable simulation, connecting forecast criticism with simulation calibration (Talts et al., 2018; Besag & Clifford, 1989).

7 Limitations

With a finite set of count categories, an alternative can preserve every selected conditional moment while changing another feature; Path MMD therefore remains a complementary global check. Finite-sample validity applies under the fitting and simulation conditions of Sections 2 and 3.3; the analytic martingale interpretation requires stronger consistency and rank conditions. Power depends on whether the logger retains the distinction being tested. The current particle implementation requires the model's event density and survival function, while leaving the count likelihood implicit. The THP study tests one controlled alternative; evaluation of fitted THPs on real logs remains future work. On real data, non-rejection indicates compatibility with the chosen count categories only.

8 CONCLUSION

OBSCRITIC evaluates a generator through its distribution on recorded counts, using standardized count forecasts, signed count residuals, and a Monte Carlo rank test for the procedure as run. Criticism on the recorded counts replaces an analyst's invented history with the information the record actually contains.

AI USE DISCLOSURE

The authors used generative AI assistants (OpenAI Codex and Claude) in a supporting role, for literature search, assistance with code implementation, checking derivations, figure formatting, language editing, and translation. The research questions, experimental design, proofs, analysis, and conclusions are the authors' own. All numerical results were produced by the authors' experiment code. The authors reviewed and verified all AI-assisted content, including code and citations, and take full responsibility for the paper.

REPRODUCIBILITY STATEMENT

Machine-readable results, configurations, and figure-source hashes accompany the manuscript. The appendices specify generators, observation rules, reference simulations, and proofs. The anonymous artifact README provides commands for reproducing the experiments, figures, and manuscript.

REFERENCES

Asohan Amarasingham, Matthew T. Harrison, Nicholas G. Hatsopoulos, and Stuart Geman. Conditional modeling and the jitter method of spike resampling. *Journal of Neurophysiology*, 107(2): 517–531, 2012.

Justin Baars, Sami Umut Can, and Roger J. A. Laeven. Asymptotically distribution-free goodness-of-fit testing for point processes. *arXiv preprint arXiv:2503.24197*, 2025.

Jerome Baum, Heishiro Kanagawa, and Arthur Gretton. A kernel stein test of goodness of fit for sequential models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1936–1953. PMLR, 2023.

Julian Besag and Peter Clifford. Generalized monte carlo significance tests. *Biometrika*, 76(4): 633–642, 1989.

Alex Boyd, Yuxin Chang, Stephan Mandt, and Padhraic Smyth. Inference for mark-censored temporal point processes. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 226–236. PMLR, 2023.

Emery N. Brown, Riccardo Barbieri, Valérie Ventura, Robert E. Kass, and Loren M. Frank. The time-rescaling theorem and its application to neural spike train data analysis. *Neural Computation*, 14(2):325–346, 2002.

Feng Chen, Tsz-Kit Jeffrey Kwan, and Tom Stindl. Estimating the hawkes process from a discretely observed sample path. *Journal of Computational and Graphical Statistics*, 34(4):1419–1431, 2025. doi: 10.1080/10618600.2025.2450463.

Claudia Czado, Tilmann Gneiting, and Leonhard Held. Predictive model assessment for count data. *Biometrics*, 65(4):1254–1261, 2009. doi: 10.1111/j.1541-0420.2009.01191.x.

A. Philip Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147 (2):278–292, 1984. doi: 10.2307/2981683.

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*, pp. 1555–1564. Association for Computing Machinery, 2016. doi: 10.1145/2939672.2939875.

Tamara Fernandez, Nicolas Rivera, Wenkai Xu, and Arthur Gretton. Kernelized stein discrepancy tests of goodness-of-fit for time-to-event data. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3112–3122. PMLR, 2020.

Christopher A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, 2014. doi: 10.1002/qj.2270.

Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807, 1996.

Felipe Gerhard and Wulfram Gerstner. Rescaling, thinning or complementing? on goodness-of-fit procedures for point process models and generalized linear models. *Neuroinformatics*, 8:93–105, 2010.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Robert Haslinger, Gordon Pipa, and Emery Brown. Discrete time rescaling theorem: Determining goodness of fit for discrete time statistical models of neural spiking. *Neural Computation*, 22(10): 2477–2506, 2010. doi: 10.1162/NECO_a_00015.

Daniel F. Heitjan and Donald B. Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19(4): 2244–2253, 1991. doi: 10.1214/aos/1176348386.

Šárka Hudecová, Marie Hušková, and Simos G. Meintanis. Tests for time series of counts based on the probability-generating function. *Statistics*, 49(2):316–337, 2015. doi: 10.1080/02331888.2014.979826.

Wittawat Jitkrittum, Heishiro Kanagawa, and Bernhard Schölkopf. Testing goodness of fit of conditional density models with kernels. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pp. 221–230. PMLR, 2020.

José C. F. Kling and Mathias Vetter. On goodness-of-fit testing for self-exciting point processes. *arXiv preprint arXiv:2407.09130*, 2024.

Jason J. Lambe, Feng Chen, Tom Stindl, and Tsz-Kit Jeffrey Kwan. Parametric inference for the discretely observed multivariate hawkes process using particle markov chain monte carlo. *arXiv preprint arXiv:2503.18351*, 2025.

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*, volume 36, 2023. URL https://openreview.net/forum?id=tn9Dldam9L.

Marian-Andrei Rizoiu, Alexander Soen, Shidi Li, Pio Calderon, Leanne J. Dong, Aditya Krishna Menon, and Lexing Xie. Interval-censored hawkes processes. *Journal of Machine Learning Research*, 23(338):1–84, 2022.

Oleksandr Shchur, Marin Biloš, and Stephan Günnemann. Intensity-free learning of temporal point processes. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=HygOjhEYDH.

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*, volume 33, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/00ac8ed3b4327bdd4ebbebcb2ba10a00-Abstract.html.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

Song Wei, Shixiang Zhu, Minghe Zhang, and Yao Xie. Goodness-of-fit test for mismatched self-exciting processes. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1243–1251. PMLR, 2021.

Christian H. Weiß, Lukas Scherer, Boris Aleksandrov, and Martin H.-J. M. Feld. Checking model adequacy for count time series by using pearson residuals. *Journal of Time Series Econometrics*, 12(1):20180018, 2020. doi: 10.1515/jtse-2018-0018.

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 *International Conference on Learning Representations*, 2024.

Jiasen Yang, Vinayak Rao, and Jennifer Neville. A stein–papangelou goodness-of-fit test for point processes. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 226–235. PMLR, 2019.

Ruoxin Yuan and Guanhua Fang. Residual TPP: A unified lightweight approach for event stream data analysis. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73455–73477. PMLR, 2025.

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 of *Proceedings of Machine Learning Research*, pp. 11692–11702. PMLR, 2020.

A ADDITIONAL DIAGNOSTICS

Table S1: Supplementary experiment inventory. Null and alternative groups are distinct; finite-rank levels follow Table S4. The same empirical-frequency, first-order, and lag-2 controls retain their definitions across banks.

Study	Bank or configuration	Finite-rank level	Results and role
Recurrent oscillatory	50 calibration; 50 null; 50 alternatives	$3/51 = 5.88\%$	Power: ObsCritic 68%, frequency 64%, Q20 12%, single/mean/max imputation 14/8/8%. 50-path null, ObsCritic and max: 3/50 each.
Recurrent 500-path null bank	500 model-null paths; thresholds from 50 calibration paths	$3/51 = 5.88\%$	ObsCritic 32/500; max-16 imputation 29/500.
Sign-flipped response	50 alternatives	$3/51 = 5.88\%$	ObsCritic and Q20: 50/50; single/mean/max imputation: 43/50, 47/50, 45/50.
Conditional imputation	64-particle lineage; rescaling KS; thresholds from 50 calibration paths	$3/51 = 5.88\%$	50-path null 4/50; oscillatory power 5/50; 500-path null 41/500. The entire lineage-sampling and testing procedure is calibrated.
Prequential log score	Shared particle forecast	$3/51 = 5.88\%$	50-path null 6/50; power 10/50; 500-path null 54/500 under the reused 50-null threshold.
Burst-state feature	$h_t = (1, \mathbf{1}\{Y_{t-1} \geq 2\})$; thresholds from 50 calibration paths	$3/51 = 5.88\%$	Null 5/50; power 33/50; 500-path null 26/500; covariance rank eight.
Matched replication A	99 calibration; 200 null/alternative	$5/100 = 5\%$	Particle lag-2 power 95/200; frequency 6/200; first-order table 11/200.
Matched replication B	99 calibration; 200 null/alternative	$5/100 = 5\%$	Particle lag-2 power 129/200; frequency 8/200; first-order table 18/200.
Particle budget	$N = 16, 32, 64, 128$; 50 per cell	$3/51 = 5.88\%$	Null 4% throughout; power 52/62/60/62%; median runtime below 0.56 s at $N = 128$.
Irregular grid	50 groups per scenario	$3/51 = 5.88\%$	ObsCritic: null 4%, oscillatory power 40%, sign-flipped power 98%; Q20: 8/4/100%.
Thinning and clutter	50 groups per scenario	$3/51 = 5.88\%$	ObsCritic: null 6%, oscillatory power 2%, sign-flipped power 76%; Q20: 6/2/98%.
Counter saturation	50 groups per scenario	$3/51 = 5.88\%$	ObsCritic: null 0%, oscillatory power 32%, sign-flipped power 84%; Q20: 14/10/100%.
Poisson predictive check	Bin means 0.4, 1.5, 3.5	Not applicable	Mean error $\leq 5.33 \times 10^{-4}$; covariance error $\leq 3.84 \times 10^{-4}$; absolute mean residual ≤ 0.0130.
Real-data strongest model	Taxi, two Yelp streams, PUBG, Twitter	5% (real); 5.88% (split null)	Atom p-values 0.75/0.48/0.78/0.32/0.38; pooled null 9/245; ESS ratios 0.970–1.000; runtime 3.43–5.37 s.
Per-path reversal	1,200 bin packages; 50 alternatives	$3/51 = 5.88\%$	No detected shift: ObsCritic 5/50, frequency 3/50, first-order table 5/50, Q20 7/50, strongest imputation 2/50.

A.1 Implementation and predictive scaling

The count-compatible proposal absorbs observed multiplicity before resampling. High effective sample size is expected when that proposal captures most of the observation; ESS measures the remaining mismatch with the generator-weighted target. The channel experiments in Table S1 additionally implement irregular bins, thinning with clutter, and saturation. Marks and other stochastic observation kernels require a corresponding compatible proposal and are not established by these experiments.

Suppose that, conditional on the recorded past, $W_0, W_1, \ldots, W_M$ are iid one-hot draws from the same categorical law with covariance Σ_t. Let S_M be the covariance of $W_{1:M}$ with denominator M, and $\bar{W} = M^{-1} \sum_{m=1}^{M} W_m$. Then

$$\mathbb{E}[S_M] = \frac{M-1}{M}\Sigma_t, \qquad \mathrm{Var}(W_0 - \bar{W}) = \frac{M+1}{M}\Sigma_t. \tag{7}$$

Thus $(M+1)/(M-1)$ corrects the finite-predictive covariance scale under the stated common-law assumption. This is the finite-ensemble setting of fair forecast evaluation (Ferro, 2014). Outer rank calibration applies the entire implementation exchangeably, including finite-particle approximation and auxiliary randomization; it does not make an approximate conditional covariance equal to the population covariance.

Under particle consistency, a conditional Lindeberg condition, covariance consistency, and rank stability, Q_T converges to a chi-square law on the identifiable subspace. This supplies the asymptotic interpretation; the reported decisions use outer rank calibration.

The recurrent oscillatory experiment uses an intercept-only feature. The additional burst-state experiment uses the two-coordinate feature in Table S1; both feature maps and the burst threshold precede evaluation. The model-conditional imputation baseline traces one 64-particle count-compatible lineage and applies time-rescaling KS. Its calibration repeats proposal, resampling, lineage selection, and testing. The prequential categorical log score uses the same particle predictive counts as Q20 and ObsCritic, replacing the signed discrepancy with predictive typicality.

The official THP experiment contains 300 unique, seed-reconstructible group keys: 50 calibration, 200 evaluation-null, and 50 alternative groups. Each imputation preserves observed bin counts and the atom covariance has rank four. Artifacts pin the checkpoint, event bank, and EasyTPP source revision.

A.2 Configuration registry and count regime

The headline, particle, and duration studies use independent seed banks and answer different questions. Table S2 records their distinct configurations and experimental roles.

Table S2: Configuration registry. Predictive draws are count draws per bin; each study uses its own seed bank for the stated purpose.

Study role	History	Particles / draws	Independent bank	Result
Edge-preserving precision	60 s	64 / 128	499 cal.; 300/300	ObsCritic-particle 67.7%; ObsCritic-table 64.3%
Non-oracle lag library	60 s	64 / 128	199 cal.; 200/200	Joint maximum 46.5%; oracle lag-2 53.5%
Imputation comparison	60 s	64 / 128	50 per scenario; bank A	Atom 68%; max-16 8%
Particle ablation	60 s	16, 32, 64, 128 / 128	50 per cell; bank B	52, 62, 60, 62%
Duration curve	60, 120, 240 s	64 / 128	50 per cell; bank C	72, 86, 100%

The main recurrent null evaluation contains 60,000 post-burn-in bins. Counts $0, 1, 2, 3, \geq 4$ occur with frequencies 21.3%, 36.4%, 26.7%, 11.4%, and 4.2%; the mean is 1.419 events/bin and the

maximum is 9. Across all four scenarios the maximum ranges from 8 to 10. The study therefore covers multi-event bins up to 10, with average occupancy below 1.5 events/bin; denser regimes lie outside the reported range.

A.3 Finite-history null replication

A 30-calibration/30-evaluation study across five finite-history generators produced 21/150 atom rejections (14.0%). Across 20,000 exchangeable relabelings of the same 300 null statistics, the mean is 6.47%, close to the finite-rank level $2/31 = 6.45\%$. A split at least as large as 14.0% occurs with probability 0.013. Calibration and evaluation use the same generator, initialization, burn-in, binning, proposal, auxiliary-randomness distribution, and scoring path; only their disjoint seed namespaces differ. A 100-calibration/100-evaluation replication per dataset gives 19/500 atom rejections (3.8%; Wilson 95% interval 2.45–5.86%) and 16/500 Q20 rejections (3.2%; 1.98–5.13%). Thus the high-precision experiment recovers the predicted finite-rank behavior across generators.

In the oscillatory study, the prequential log score rejects 54/500 model-null paths (10.8%) under its reused 50-null threshold. This rate is conditional on the reused threshold; the exact rank guarantee averages over calibration-bank and evaluation draws.

A.4 Closed-form predictive validation

For homogeneous Poisson generators with 50-ms bin means 0.4, 1.5, and 3.5, the exact five-atom count law is available analytically. Over 4,096 sequential bins per regime, the maximum absolute error of the 128-draw $\widehat{\mu}_t$ is 5.33×10^{-4}; the maximum covariance-entry error relative to the exact finite-ensemble target is 3.84×10^{-4}; and the largest absolute mean residual is 0.0130. Median ESS is one in every regime. This closed-form study validates the predictive and covariance implementation across sparse and dense count regimes.

A.5 Predictive budget and logger resolution

On an independent 50/50/50 recurrent bank at the common $3/51$ rank level, predictive budgets $M = 8, 16, 32, 128$ give model-null rates $4, 6, 10, 10\%$ and oscillatory power $70, 74, 80, 84\%$. Corrected and raw covariance versions make identical rank decisions at every M: within a given budget, the correction derived from Eq. (7) is a positive scalar rescaling. It changes the quadratic scale and asymptotic interpretation without changing a separately calibrated rank test. Median runtime rises from 0.45 to 0.65 s.

Changing only the logger width to 25, 50, and 100 ms gives model-null rates $0, 4, 8\%$ and power $24, 46, 60\%$ on a separate shared-path bank. Mean null counts per bin are 0.714, 1.428, and 2.855. The fixed five-atom dictionary is sparse at 25 ms and concentrates the oscillatory rate shift at 100 ms; the result demonstrates that observation resolution controls retained power. All covariance ranks are four, and the same dictionary and configurations are used at every width.

A.6 Exact marginal matching

The reversal control in Table S1 preserves each count and within-bin position while reversing 1,200 bin packages; it produced no detectable shift.

A stricter long-memory control preserves every four-gram, every lag-1/2/3 pair table, the joint lag-(1,2) table, and all exact-count packages, then maximizes eight-step returns over 512 Euler trails. Both the particle predictive-state critic (8/200) and the jointly calibrated low-order library (6/200) remain at the 5% rank level. Recorded distinctions outside the selected predictable features need not move the critic, even when the reordered path has substantially different longer organization.

A.7 Edge-preserving conditional alternative

The 128 reference paths for the precision study use simulation seeds 1,010,012,400–1,010,012,527. Calibration uses 1,020,012,528–1,020,013,026; evaluation-null uses 1,021,013,027–1,021,013,326; alternative source paths use 1,022,013,327–1,022,013,626. These sets are disjoint. The lookup-table and particle-filter implementations intentionally score the same 499+300+300 paths for paired

comparison. For the feature-library study, the 64 reference seeds are 1,010,012,100–1,010,012,163, disjoint from its calibration and evaluation namespaces. Matched replications A and B each use 99 calibration statistics and contain particle lag-2, frequency, and first-order controls, not the ObsCritic-table lag-2 implementation.

For an atom sequence $y_{0:T}$, define the directed multigraph with one edge $y_{t-1} \to y_t$ for every adjacent pair. Every Euler trail with the same endpoints uses precisely this edge multiset and therefore preserves the full 5×5 transition table, which also fixes the atom histogram. The alternative enumerates 256 seeded trails and selects the one maximizing $\sum_{t=2}^{T} \mathbf{1}\{y_t = y_{t-2}\}$. Exact-count packages—including their normalized within-bin positions—are then reassigned within each unchanged atom class. Across all 300 precision alternatives, the maximum atom-count error and maximum transition-count error are both zero; packages, endpoints, and within-bin positions are preserved exactly.

The lag-2 statistic has rank 20, the intercept statistic rank four, and both reference transition statistics rank 20. All statistics are finite. Matched replications A and B use independent banks with 99 calibration statistics and give particle lag-2 power 95/200 and 129/200; the 499-calibration precision bank gives 203/300. The same-feature particle-free control gives 193/300 on the precision alternative and 15/300 on model-null. Their paired alternative decisions are 188 both, 92 neither, 15 particle-only, and 5 table-only. The corresponding empirical-frequency powers are 6/200, 8/200, and 19/300, while first-order-table powers are 11/200, 18/200, and 17/300. Particle detections missed by frequency versus the reverse are 185–1; paired counts against the first-order table, Q20, and max-16 are 187–1, 132–18, and 190–0.

A.8 Synthetic generator and oscillatory alternative

For inter-event gap Δ_i and elapsed time s after event i, the recurrent generator is

$$h_i = \tanh(Rh_{i-1} + w \log(1 + 20\Delta_i) + b), \qquad \lambda(t_i + s \mid \mathcal{H}_{t_i}) = \exp(v^\top h_i + c + 4s), \tag{8}$$

with $h_0 = 0$. The null uses

$$R_0 = \begin{bmatrix} .55 & .10 & 0 \\ -.12 & .48 & .08 \\ .05 & -.10 & .52 \end{bmatrix}, \quad w = \begin{bmatrix} .35 \\ -.25 \\ .20 \end{bmatrix}, \quad b = \begin{bmatrix} .05 \\ -.02 \\ .03 \end{bmatrix}, \quad v_0 = \begin{bmatrix} .45 \\ -.30 \\ .25 \end{bmatrix}, \quad c_0 = \log 18. \tag{9}$$

The oscillatory alternative changes only

$$R_1 = \begin{bmatrix} -.62 & .18 & 0 \\ -.10 & -.55 & .16 \\ .15 & 0 & -.58 \end{bmatrix}, \quad v_1 = \begin{bmatrix} .90 \\ -.62 \\ .52 \end{bmatrix}, \quad c_1 = 3.03182266108032, \tag{10}$$

leaving w, b, and the elapsed-time slope fixed. Paths last 70 s, the first 10 s are burn-in, and the logger records counts in 50-ms bins. These equations, all seeds, and the event simulator are included in the anonymous supplement.

B Proof sketches

Imputation invariance. The algorithm receives only y, the supplied model, the logger, and an auxiliary seed. Two latent representatives in $\mathcal{C}^{-1}(y)$ therefore induce identical calls, resampling decisions, predictive draws, residuals, and output.

Finite-algorithm validity. Under $H_0, G_0, G_1, \ldots, G_B$ are iid after the same logger. Independent auxiliary seeds with the same law make the implemented statistics exchangeable. The upper rank in Eq. (6), including ties, is therefore super-uniform.

Observed-filtration martingale. By Eq. (3), $\mathbb{E}[e_t \mid \mathcal{F}^Y_{t-1}] = 0$. Predictability of h_t gives $\mathbb{E}[h_t \otimes e_t \mid \mathcal{F}^Y_{t-1}] = h_t \otimes \mathbb{E}[e_t \mid \mathcal{F}^Y_{t-1}] = 0$.

Information boundary. Any test measurable from Y has expectation determined by the pushforward distribution. Equal pushforwards give equal rejection probabilities. For unequal pushforwards, the difference in expectation of a $[0, 1]$-valued test is bounded by total variation.

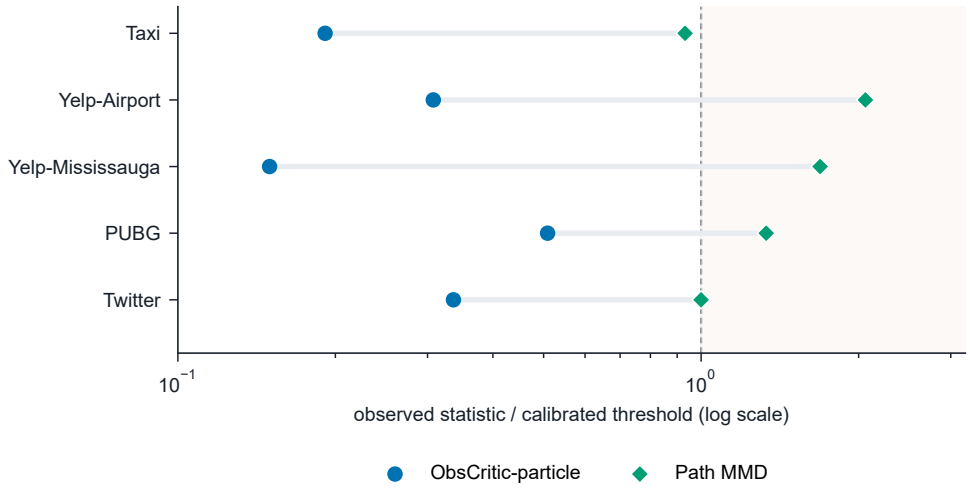

Figure 6: **Local and global checks examine different model properties.** The strongest generators have atom-critic ratios below one (circles); Path MMD detects global discrepancies where its ratio exceeds one (diamonds). Path MMD ratios are 0.93 for Taxi and 0.9999 for Twitter; neither is rejected.

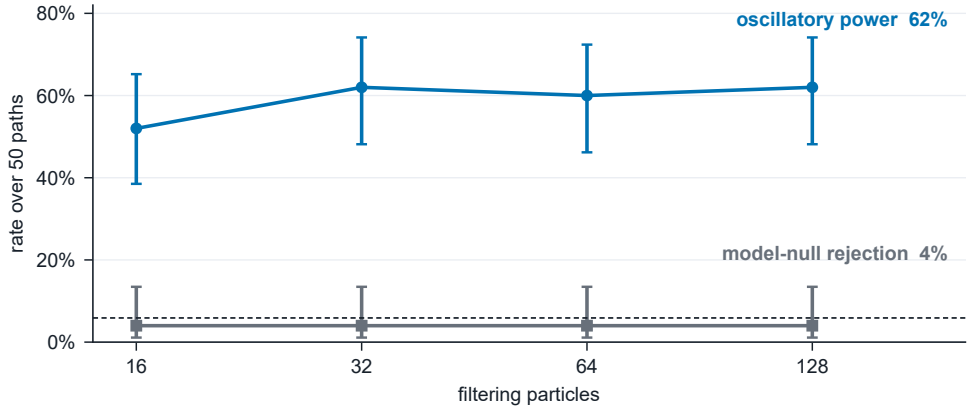

Figure 7: **Particle ablation.** Model-null rejection remains at 4%; power changes little beyond 32 particles. Error bars are 95% Wilson intervals over 50 paths.

C Complete Experimental Configurations and Results

C.1 Baselines and Calibration Levels

All baselines receive the same record. The uniform-imputation baseline draws 16 independent within-bin imputations and applies time-rescaling KS to each. The first, mean, and maximum statistics each undergo whole-procedure calibration. Path MMD (Gretton et al., 2012) operates on the records. The lag-2 lookup-table implementation estimates $P(A_t \mid A_{t-2})$ from independent null simulations and applies the same $h_t \otimes e_t$ quadratic geometry without particles. Q20 and ObsCritic-particle share the predictive count distribution and observed filtration: Q20 aggregates randomized-rank autocorrelations, whereas ObsCritic retains signed count-atom errors. For n scored bins, predictive counts $C_t^{(1:M)}$, and $V_t \sim \mathrm{Unif}(0,1)$, Q20 uses the Ljung–Box statistic with 20 lags,

$$u_t = \frac{\sum_m \mathbf{1}\{C_t^{(m)} < Y_t\} + V_t\big(1 + \sum_m \mathbf{1}\{C_t^{(m)} = Y_t\}\big)}{M+1}, \quad Q_{20} = n(n+2)\sum_{\ell=1}^{20} \frac{\widehat{\rho}_\ell^2}{n-\ell}, \tag{11}$$

where $\widehat{\rho}_\ell$ is the lag-ℓ autocorrelation of $\sqrt{12}(u_t - 1/2)$. The empirical-frequency critic is the unconditional endpoint of the same logger-aligned atom family. It replaces the history-conditioned forecast with a rank-four pooled atom law estimated from 16 independent model paths in the oscillatory study, retaining the same five-atom geometry and outer calibration bank. The main lag-2 table uses 128 reference model paths disjoint from calibration and evaluation; the feature library uses 64. Reference probabilities, with 0.5 smoothing per cell, remain unchanged throughout scoring.

Table S3: Baseline and display configurations.

Component	Configuration or result
Imputation instability	50 paths, 50-ms bins, nominal 5% time-rescaling KS without outer recalibration. Uniform/left-skewed/right-skewed imputations give 4/100/100% rejection; the event-time reference shown in Figure 1 is 8%.
Imputation ensembles	16 independent uniform imputations; Single imputation, Mean imputation, and Max imputation each receive whole-procedure calibration. Model-conditional imputation traces one 64-particle compatible lineage.
Reference forecasts	Oscillatory empirical atom frequency: 16 independent model paths, rank-four atom law. Main lag-2 table: 128 reference paths; feature library: 64. Reference probabilities use 0.5 smoothing per cell and are not updated while scoring.
Localization figure	50 oscillatory alternatives; excess 0- and 2-atoms and deficits at 1 and 3. Bars are 95% intervals for mean standardized residuals; model-null means remain near zero.

Table S4: Finite-rank calibration levels; no ties are assumed in the exact rank counts.

Study	Calibration rule
Matched precision study	499 calibration statistics; 25 of 500 exchangeable ranks reject at $\alpha = .05$, an exact 5% level, using Eq. (6).
Matched replications A/B	99 calibration statistics; upper-rank Monte Carlo rule; $5/100 = 5\%$ with no ties.
Feature library and long-memory control	199 calibration statistics; upper-rank Monte Carlo rule; $10/200 = 5\%$ with no ties.
Finite-history null replications	30 calibration statistics: $2/31 = 6.45\%$; 100 calibration statistics: $5/101 \approx 4.95\%$, using the `higher` 0.95 quantile and strict exceedance.
Strongest-model real-data check	Real-data Monte Carlo p-values use 99 null simulations ($5/100 = 5\%$); the split-null check uses 50 calibration and 49 evaluation groups ($3/51 = 5.88\%$).
Oscillatory and THP studies	50 calibration statistics; NumPy-style `higher` empirical 0.95 quantile; three of 51 ranks reject, giving $3/51 = 5.88\%$.

C.2 Matched dynamics and feature library

Table S5: Complete matched precision results (300 model-null and 300 alternative paths, 499 calibration statistics). Wilson intervals are 95%.

Method or check	Result
ObsCritic-particle, lag-2	Null 12/300 (4.0%); power 203/300 (67.7%, interval 62.2–72.7%).
ObsCritic-table, lag-2	Null 15/300 (5.0%); power 193/300 (64.3%, interval 58.8–69.5%).

Method or check	Result
ObsCritic-particle, intercept	Null 17/300 (5.7%); power 19/300 (6.3%).
Empirical atom frequency	Null 18/300 (6.0%); power 19/300 (6.3%).
First-order table	Null 16/300 (5.3%); power 17/300 (5.7%).
Q20	Null 15/300 (5.0%); power 89/300 (29.7%).
Max imputation	Null 16/300 (5.3%); power 13/300 (4.3%); maximum over 16 imputations.
Paired particle/table detections	188 both, 92 neither, 15 particle-only, 5 table-only.
Construction	Choose among 256 Euler trails; median two-step returns rise from 309.5 to 360. Atom-count and transition-count matching errors are both zero.
Resource use	Median end-to-end particle runtime 2.61 s per path group.

Table S6: Non-oracle feature-library results (199 calibration, 200 model-null, 200 alternative paths).

Method	Result
Joint maximum	8/200 null rejections (4.0%); 93/200 alternatives (46.5%). Joint calibration covers ObsCritic-table lag-1, lag-2, lag-3, and joint lag-(1,2).
ObsCritic-table, lag-2	Oracle feature: 107/200 alternatives (53.5%).
ObsCritic-particle, lag-2	104/200 alternatives (52.0%).
ObsCritic-table, lag-3	Mismatched feature: 9/200 alternatives (4.5%).
ObsCritic-table, lag-1	Power 3.5%, as displayed in Figure 3.
ObsCritic-table, joint lag-(1,2)	Power 24.5%, as displayed in Figure 3.

C.3 Real-data capacity and global checks

Table S7: Real-data capacity ladder and complementary global diagnostics.

Endpoint	Result
Benchmarks	Taxi, Yelp-Airport, Yelp-Mississauga, PUBG, Twitter.
Capacity levels	Recurrent exponential, one-component LogNormal mixture, three-component LogNormal mixture. No selected-atom mismatch on 0/5, 2/5, and 5/5 datasets respectively.
Non-monotone cases	Taxi worsens at the intermediate model; PUBG passes both mixture levels. Figure 4 retains every dataset-level statistic/threshold ratio.
Strongest-model null study	99 fitted-model simulation groups per dataset; pooled split-null rejection 9/245 (3.67%).
Local versus global	Path MMD finds no discrepancy for 2/5 strongest generators; ObsCritic finds no selected-atom mismatch for 5/5. These are distinct diagnostic targets.

C.4 Transformer Hawkes Process

The THP experiment uses 64 filtering particles, 128 predictive draws, 16-imputation ensembles, 50 calibration groups, and 50 alternatives. The 50- and 200-group model-null banks share the model, thresholds, methods, and alternative bank. Its finite-rank level is $3/51 = 5.88\%$.

Table S8: THP results: 200 model-null groups and 50 alternatives.

Method	Null rejection	Power
ObsCritic-particle	9/200 (4.5%)	48/50 (96%)
Empirical atom frequency	6/200 (3.0%)	47/50 (94%)
Q20	0/200 (0%)	10%
Path MMD	2/200 (1.0%)	0%
Single imputation	9/200 (4.5%)	22%
Mean imputation	7/200 (3.5%)	12%
Max imputation	9/200 (4.5%)	12%

The reported Wilson 95% interval for ObsCritic power is 86.5–98.9%; for the 9/200 null rejection of ObsCritic and Single imputation it is 2.39–8.33%. The empirical-frequency result demonstrates a marginal-law change in this alternative.

C.5 Negative controls and low-power regimes

Table S9: Negative controls, null anomaly, and low-power observation regimes. Finite-rank levels follow the study-specific rules in Table S4.

Study	Finite-rank level	Numerical result and interpretation
Finite-history null replication (Appendix A.3)	$2/31 = 6.45\%$ (30/30); $5/101 \approx 4.95\%$ (100/100)	30-calibration/30-evaluation replication: 21/150 rejections (14.0%). Across 20,000 exchangeable relabelings, mean 6.47%, versus $2/31 = 6.45\%$; probability of a split at least as large is 0.013. 100-calibration/100-evaluation replication: 19/500 (3.8%, Wilson interval 2.45–5.86%) for the atom critic and 16/500 (3.2%, 1.98–5.13%) for Q20.
Long-memory control (Appendix A.6)	5%	Particle predictive-state critic 8/200; Joint maximum 6/200 at the 5% rank level. The alternative preserves four-grams, lag-1/2/3 pair tables, joint lag-(1,2), and exact-count packages while maximizing eight-step returns over 512 Euler trails.
Reversal control (Table S1)	$3/51 = 5.88\%$	1,200 bin packages, 50 alternatives; ObsCritic 5/50, frequency 3/50, First-order table 5/50, Q20 7/50, strongest imputation 2/50. No detected shift.
Thinning and clutter	$3/51 = 5.88\%$	50 groups per scenario: ObsCritic null/oscillatory/sign-flipped rejection 6/2/76%; Q20 6/2/98%. Oscillatory power is low for both.
Counter saturation	$3/51 = 5.88\%$	50 groups per scenario: ObsCritic null/oscillatory/sign-flipped rejection 0/32/84%; Q20 14/10/100%. The low oscillatory powers remain part of the observation-channel boundary.