Microlensing Detection and Inference via Learned Bayes Factors

Nolan Smyth [1 2 3] **Laurence Perreault-Levasseur** [1 2 3 4 5 6] **Yashar Hezaveh** [1 2 3 4 5 6]

Abstract

We present a unified framework for gravitational microlensing event detection and parameter inference. Traditional pipelines use deterministic hard cuts on photometric statistics, systematically missing low-magnification events in the finite-source regime. We instead frame detection as Bayesian model comparison using Evidence Networks, which learn calibrated Bayes factors from binary-labeled simulations, and combine this with Neural Posterior Estimation (NPE) for amortized parameter inference. Both share a transformer encoder that handles irregularly-sampled time series without imputation. On simulated Roman Space Telescope data, our Evidence Network achieves 99.9% detection efficiency with a false-positive rate below 6×10^{-4} on simulated data with augmentation and noise, but no astrophysical confounders. Gains are most dramatic in the extreme finite-source regime ($\rho \gtrsim 5$), where detection rates reach $\sim$95% versus $\sim$65% for hard cuts, precisely the short-duration free-floating planet events most constraining for formation scenarios. Our NPE provides calibrated posteriors, working towards real-time analysis at survey scale.

1. Introduction

The Nancy Grace Roman Space Telescope will conduct an unprecedented Galactic Exoplanet Survey targeting over 200 million stars toward the Galactic Bulge (Penny et al., 2019). Observing at 15-minute cadence over five 72-day seasons, Roman will revolutionize our understanding of exoplanet demographics through gravitational microlensing, a technique uniquely sensitive to low-mass planets, including terrestrial-mass worlds and free-floating planets (FFPs) unbound to any host star. Yield predictions suggest Roman will detect $\sim$1,400 bound exoplanets (Penny et al., 2019) and $\sim$250 FFPs (Johnson et al., 2020), with lens mass and distance measurements achievable for a significant fraction of detected hosts (Terry et al., 2025), enabling direct tests of planet formation and dynamical ejection scenarios (Zwart, 2024; Coleman & DeRocco, 2024). FFPs are expected to outnumber bound terrestrial planets by a factor of 20 or more (Sumi et al., 2023), and the first direct mass measurement of a free-floating planet via microlensing parallax (Dong et al., 2026) was recently achieved. Characterizing the low-mass end of the FFP mass function requires sensitivity to short-duration low-magnification events, making performance in this regime crucial to key science drivers.

Realizing this science return requires solving two coupled challenges: *detection* (identifying microlensing signals amid billions of light curves dominated by photometric noise and astrophysical contaminants) and *inference* (estimating model parameters corresponding to those signals). Traditional approaches address these problems sequentially and with fundamentally different methodologies. Detection often relies on deterministic *hard cuts* that apply threshold-based criteria to photometric summary statistics. Representative examples include requiring $\Delta\chi^2 \geq 300$ (difference between microlensing model and flat baseline) combined with $N_{\text{consec}} \geq 6$ consecutive points exceeding 3σ above baseline (Johnson et al., 2020), or variants using the sum of significances over consecutive high-SNR points as a metric (e.g. Sumi et al., 2011; Mroz et al., 2017). While computationally efficient, hard cuts have limitations. In particular, they are ad-hoc, provide no probabilistic interpretation (a light curve either passes or fails, with no uncertainty quantification), and they can exhibit strong parameter-dependence, missing low-magnification or short-duration events.

For detected events, characterization traditionally employs Markov Chain Monte Carlo (MCMC) sampling to explore the posterior distribution over physical parameters. MCMC provides asymptotically exact Bayesian inference, but requires inference-time computation. This means running MCMCs from scratch for each light curve scales poorly with number of detections. Moreover, MCMC characteriza-

[1]Ciela Institute, Montreal Institute for Astrophysics and Machine Learning, Montréal QC, Canada [2]Mila - Quebec AI Institute, Montréal QC, Canada [3]Department of Physics, Université de Montréal, Montréal QC, Canada [4]Center for Computational Astrophysics, Flatiron Institute, New York, USA [5]Perimeter Institute for Theoretical Physics, Waterloo, Canada [6]Trottier Space Institute, McGill University, Montréal, Canada. Correspondence to: Nolan Smyth <nolan.smyth@umontreal.ca>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

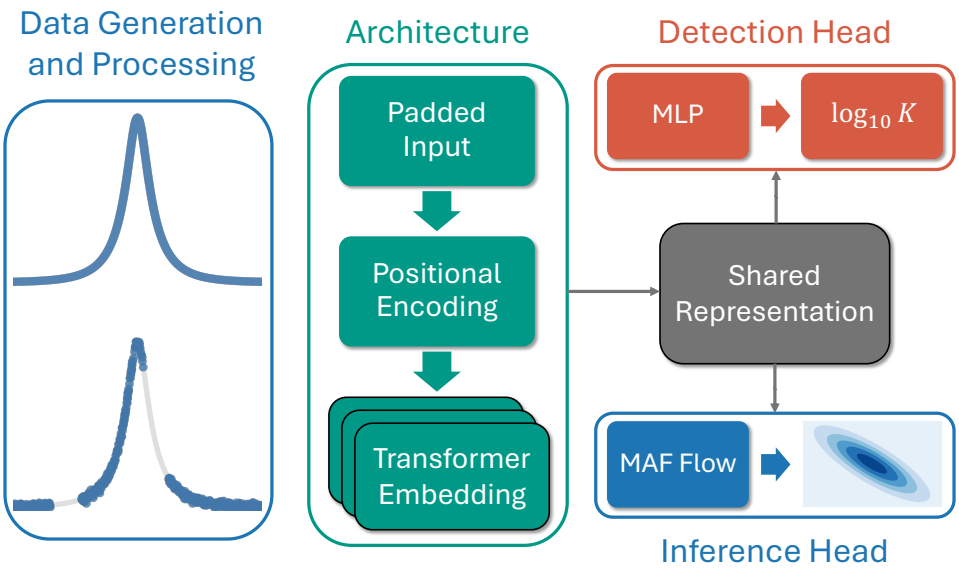


Figure 1. Joint architecture for microlensing detection and parameter inference. Simulated light curves undergo data augmentation and padding. A shared transformer encoder with multi-head self-attention processes the irregularly-sampled time series into a fixed-dimensional summary statistic, the shared representation $\mathbf{s}$. Two specialized heads operate on $\mathbf{s}$. The detection branch outputs calibrated Bayes factors $\log_{10} K$, while the inference branch uses a masked-autoregressive normalizing flow to estimate the full posterior $p(\theta|x)$. This unified framework enables both tasks to benefit from shared features learned by the transformer.

tion occurs *after* detection, decoupling model comparison (signal vs. noise) from parameter estimation in a way that discards information: the Bayes factor for signal detection naturally arises from the same marginal likelihood integrals computed during MCMC.

This work presents a unified neural framework that addresses both challenges within a single probabilistic formalism. We combine *Evidence Networks* (Jeffrey & Wandelt, 2024) for fast Bayes factor estimation with *Neural Posterior Estimation* (NPE) (Papamakarios & Murray, 2018; Greenberg et al., 2019) for amortized posterior inference, both sharing a transformer-based encoder for irregularly-sampled time series. Evidence Networks frame detection as supervised Bayesian model comparison: given labeled training data $\{(\mathbf{x}_i, m_i)\}$ with $m_i \in \{0, 1\}$ indicating noise ($\mathcal{M}_0$) vs. signal ($\mathcal{M}_1$), a network learns the Bayes factor $K \equiv p(\mathbf{x}|\mathcal{M}_1)/p(\mathbf{x}|\mathcal{M}_0)$. Crucially, this requires *only binary labels*, making the approach applicable to scenarios where model evidences are slow or intractable.

Prior work has applied NPE to binary microlensing (Zhang et al., 2021), transformer embeddings for single-lens inference on irregular time-series data (Smyth et al., 2025), and normalizing flows with reversible-jump MCMC for microlensing model selection (Keehan et al., 2022). Our framework extends these by unifying detection and characterization within a single amortized pipeline (Section 2.3).

Our contributions are threefold: (i) a unified probabilistic framework for detection and characterization that enables targeted scientific questions with calibrated uncertainty (e.g. the probability a light curve contains an FFP with $t_E < 1$ day); (ii) the first application of Evidence Networks to irregularly-sampled astrophysical time series, recovering calibrated Bayes factors from binary labels alone and replacing ad hoc $\Delta\chi^2$ thresholds; and (iii) on simulated Roman-like data, improved detection completeness in the extreme finite-source regime ($\rho \gtrsim 5$) relative to literature hard cuts (Section 3).[1]

2. Methods

Our framework consists of two complementary neural networks operating on the same input representation: an Evidence Network for signal detection via Bayes factor estimation, and a Neural Posterior Estimator for inference.

2.1. Data and Modeling

We model Finite-Source Point-Lens (FSPL) events using `ESPLMag2` as implemented in `VBMicrolensing` (Bozza et al., 2018), which uses a pre-computed table, the entries of which are calculated via contour integration. Note that this is also the backend behind `FSPLargemodel` in `pyLIMA`, another microlensing modeling package (Bachelet et al., 2017).

[1]The code used in this work is available at: https://github.com/NolanSmyth/evidence_networks_sbi_microlensing.

The five microlensing parameters we consider are $\boldsymbol{\theta} = (t_0, u_0, \log_{10} t_E, \log_{10} \rho, f_s)$: the time of closest lens-source alignment (days), the impact parameter in units of the Einstein radius θ_E, the Einstein crossing time (days), the normalized source radius $\rho \equiv \theta_s/\theta_E$ (with θ_s the angular source size), and the source flux fraction $f_s \in [0, 1]$. The observed flux at time t, normalized to the baseline, is

$$F(t) = f_s A(t, \boldsymbol{\theta}) + (1 - f_s) \tag{1}$$

where $A(t, \boldsymbol{\theta})$ is the magnification factor computed accounting for finite size of the source, and $1-f_s$ represents blended light from unresolved stars common in crowded fields.

For simplicity, we adopt box-uniform priors: $t_0 \in [0, 20]$ days, $u_0 \in [0, 1.5]$, $\log_{10}(t_E) \in [-1, 1.3]$ (0.1–20 days, spanning short-duration FFP candidate events), $\log_{10}(\rho) \in [-2, 1]$ (source radii from $0.01\theta_E$ to $10\theta_E$), and $f_s \in [0.1, 1.0]$. We stress that for a production-ready pipeline, the priors should be informed by galactic and population models. For example, if FFPs follow a power-law mass function (Sumi et al., 2023), one would expect a much higher density of short-duration events whose duration is determined by the angular size of the source, rather than a peaked mass function (DeRocco et al., 2025) which would predict fewer events with $\rho \gg 1$. Ideally, these factors should also be incorporated into a self-consistent hierarchical model, rather than being chosen ad hoc.

Each light curve is represented as a padded sequence of length $L = 1000$ with three channels per timestep: $\mathbf{x}_i = (t_i^{\text{norm}}, F_i, \sigma_i)$ for $i = 1, \ldots, L$. Times are normalized to $[-1, 1]$ over the 20-day window via $t^{\text{norm}} = 2t/T_{\text{total}} - 1$, allowing the network to learn position-invariant features. F_i is the normalized flux relative to the baseline (taken to be unity), and σ_i is the per-point photometric uncertainty. Padded positions take a value of -2 in all channels, enabling the transformer to identify valid data via attention masking.

2.2. Data Augmentation

To demonstrate performance on some realistic observing conditions, we employ on-the-fly data augmentation during training. For each parameter sample $\boldsymbol{\theta}$ drawn from the prior:

1. **Base simulation**: Generate a dense, noiseless light curve using a cadence of 15 minutes (1920 points uniformly spaced over 20 days)
2. **Seasonal gaps**: Introduce 0–3 gaps of length 1–10 days each, simulating seasonal observability windows or downtime
3. **Random dropout**: Remove a random fraction (0–60%) of remaining points, mimicking bad readings or cadence variations
4. **Photometric noise**: Add Gaussian noise with per-point standard deviation σ drawn uniformly from $[0.001, 0.02]$ in relative flux units (corresponding to $\sim$0.1–2%, covering the typical expected photometric precision of Roman)

We apply recoverability filters to exclude training samples with insufficient coverage. Events must satisfy: at least 5 points within $1.5t_E$ of t_0 (peak region coverage), at least 5 points beyond $3.0t_E$ from t_0 (baseline establishment), and *peak* signal-to-noise ratio SNR > 5, where SNR is computed as $(A_{\text{peak}} - 1)/\sigma_{\text{baseline}}$. These criteria ensure the network trains on physically recoverable events while learning to handle realistic data quality.

2.3. Transformer Embedding Network

The key challenge is converting variable-length, irregularly-sampled light curves into fixed-dimensional summaries that preserve the temporal structure needed for both detection and parameter inference. Both the Evidence Network and Neural Posterior Estimator share a common embedding architecture based on the transformer encoder (Vaswani et al., 2023). This design offers several advantages. First, self-attention learns long-range dependencies: a point at $t = 5$ days can directly attend to baseline measurements at $t = 15$ days, capturing the full event morphology. Second, attention weights are permutation-invariant but position-aware through encodings; the network learns what matters about the temporal structure, not fixed grid positions. Third, the shared embedding serves both detection and inference, enabling transfer learning: features useful for computing Bayes factors also help constrain parameters.

The network processes 3-channel input $\mathbf{X} \in \mathbb{R}^{B \times L \times 3}$ (batch size B, sequence length L) to produce a summary $\mathbf{s} \in \mathbb{R}^{B \times d_{\text{model}}}$. Each timestep $\mathbf{x}_i = (t_i^{\text{norm}}, F_i, \sigma_i)$ is first projected from 3 dimensions to $d_{\text{model}} = 256$ via a learned linear layer $\mathbf{h}_i^{(0)} = \mathbf{W}_{\text{proj}}\mathbf{x}_i + \mathbf{b}_{\text{proj}}$, scaled by $\sqrt{d_{\text{model}}}$ for variance stabilization. This projection creates a high-dimensional representation where the network can learn complex relationships between flux, time, and uncertainty.

We append a learnable classification token $\mathbf{c} \in \mathbb{R}^{d_{\text{model}}}$ to the sequence. This token serves as an information aggregator: through self-attention, it attends to all timesteps and accumulates evidence for detection or parameter constraints. Its final hidden state becomes our summary $\mathbf{s}$.

We add sinusoidal positional encodings (Vaswani et al., 2023); our normalized time channel already provides explicit positional cues, so we empirically observed that these encodings offer only marginal benefit, but speed up convergence somewhat. The core architecture is 6 transformer encoder layers with 8-head self-attention and position-wise feedforward networks ($d_{\text{ff}} = 512$), with standard resid-

ual connections and layer normalization. Padded positions (marked by time $= -2$) are masked in the attention computation, so the network handles variable-length sequences without interpolation or imputation. We extract the CLS token's final hidden state and apply layer normalization.

2.4. Evidence Networks for Signal Detection

2.4.1. Bayesian Framework

Unlike $\Delta\chi^2$ statistics, which have no rigorous Bayesian interpretation for the non-nested model comparison, Bayes factors have a clear meaning. We treat microlensing event detection as a model selection problem: given observed data $\mathbf{x}$ (a light curve), we compare the signal hypothesis $\mathcal{M}_1$ (microlensing event present) against the noise-only hypothesis $\mathcal{M}_0$. The posterior probability of the event model relative to the noise-only model is

$$\frac{p(\mathcal{M}_1|\mathbf{x})}{p(\mathcal{M}_0|\mathbf{x})} = \frac{p(\mathbf{x}|\mathcal{M}_1)}{p(\mathbf{x}|\mathcal{M}_0)}\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)} = K\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)}, \quad (2)$$

where $p(\mathcal{M}_1)/p(\mathcal{M}_0)$ is the prior probability of any given light curve containing a true event.

The Bayes factor K quantifies the relative evidence, or how much to update your model prior given the data:

$$K = \frac{p(\mathbf{x}|\mathcal{M}_1)}{p(\mathbf{x}|\mathcal{M}_0)} = \frac{\int p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M}_1)p(\boldsymbol{\theta}|\mathcal{M}_1)\, d\boldsymbol{\theta}}{p(\mathbf{x}|\mathcal{M}_0)} \quad (3)$$

where the numerator marginalizes over the prior $p(\boldsymbol{\theta}|\mathcal{M}_1)$ to compute the model evidence. Unlike $\Delta\chi^2$ evaluated at maximum-likelihood parameters, this marginalization automatically penalizes model complexity via the Occam factor: a larger parameter space must spread its predictive probability over more possibilities, reducing the evidence unless the data strongly favor the signal model. Computing this integral analytically or via sampling is often slow or intractable in practice. However, an Evidence Network trained on binary labeled training data alone can learn to quickly and accurately predict $\log K$ directly from the data.

2.4.2. Architecture and Training

We train the Evidence Network using the leaky parity-odd power (l-POP) exponential loss with exponent $\alpha = 2$, following the recommendations of Jeffrey & Wandelt (2024):

$$\mathcal{L}(f(\mathbf{x}), m) = \exp\left[\left(\frac{1}{2} - m\right) J_\alpha(f(\mathbf{x}))\right] \quad (4)$$

where $m \in \{0, 1\}$ is the binary label (1 for signal, 0 for noise), $f(\mathbf{x})$ is the network output, and $J_\alpha(z) = z + z|z|^{\alpha-1}$ is the l-POP transform. For the optimal network f^* minimizing the expected loss, we have:

$$\log K = J_\alpha(f^*(\mathbf{x})). \quad (5)$$

The l-POP loss is not the only choice that leads to an estimate of the Bayes factor, but it does spread the dynamic range of Bayes factors across many orders of magnitude, preventing saturation for strong signals while maintaining sensitivity to weaker signals.

The Evidence Network itself consists of the transformer embedding followed by a 3-layer multilayer perceptron with hidden dimensions $[256, 128, 1]$ and ReLU activations. The final layer outputs a scalar $f(\mathbf{x})$ transformed to $\log_{10} K$ via l-POP.

We train for 200 epochs using the Adam optimizer (Kingma & Ba, 2017) with learning rate 10^{-4}, batch size 1024, and early stopping based on validation loss (patience 20 epochs). Training data consists of 100,000 simulated light curves with equal signal/noise fractions, generated on-the-fly with augmentation as described in Section 2.2.

The network is trained using *only binary labels*, not the true Bayes factors. In real microlensing applications, ground-truth model evidences are unavailable; labels typically come from human classification, conventional detection algorithms, or simulations. The analytic $\log K$ values in our toy model (Appendix A) serve solely for post-hoc validation. This constraint ensures the method is applicable to realistic scenarios where Bayes factors cannot be computed.

2.5. Neural Posterior Estimation for Parameter Inference

For detected events, we estimate the posterior $p(\boldsymbol{\theta}|\mathbf{x})$ using simulation-based inference. We employ Neural Posterior Estimation, which directly learns the posterior via conditional density estimation rather than explicitly modeling the likelihood. The posterior estimator combines the transformer embedding network with a Masked Autoregressive Flow (MAF) normalizing flow (Papamakarios et al., 2018). The transformer produces a summary statistic $\mathbf{s} = g_\phi(\mathbf{x}) \in \mathbb{R}^{256}$, which conditions the flow, so that $p_\psi(\boldsymbol{\theta}|\mathbf{x}) \approx p_\psi(\boldsymbol{\theta}|\mathbf{s})$, with

$$p_\psi(\boldsymbol{\theta}|\mathbf{s}) = p_\mathbf{z}\left(T_\psi^{-1}(\boldsymbol{\theta}|\mathbf{s})\right)\left|\det\frac{\partial T_\psi^{-1}}{\partial\boldsymbol{\theta}}\right|, \quad (6)$$

where $T_\psi : \mathbb{R}^5 \times \mathbb{R}^{256} \to \mathbb{R}^5$ is the learned invertible transformation (flow), $p_\mathbf{z} = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a standard Gaussian base distribution, and the Jacobian determinant accounts for the change of variables. The MAF architecture uses 10 autoregressive layers, each a masked affine coupling layer with context embedding $\mathbf{s}$ injected at every layer.

The autoregressive property ensures efficient density evaluation: the i-th parameter dimension is generated conditioned

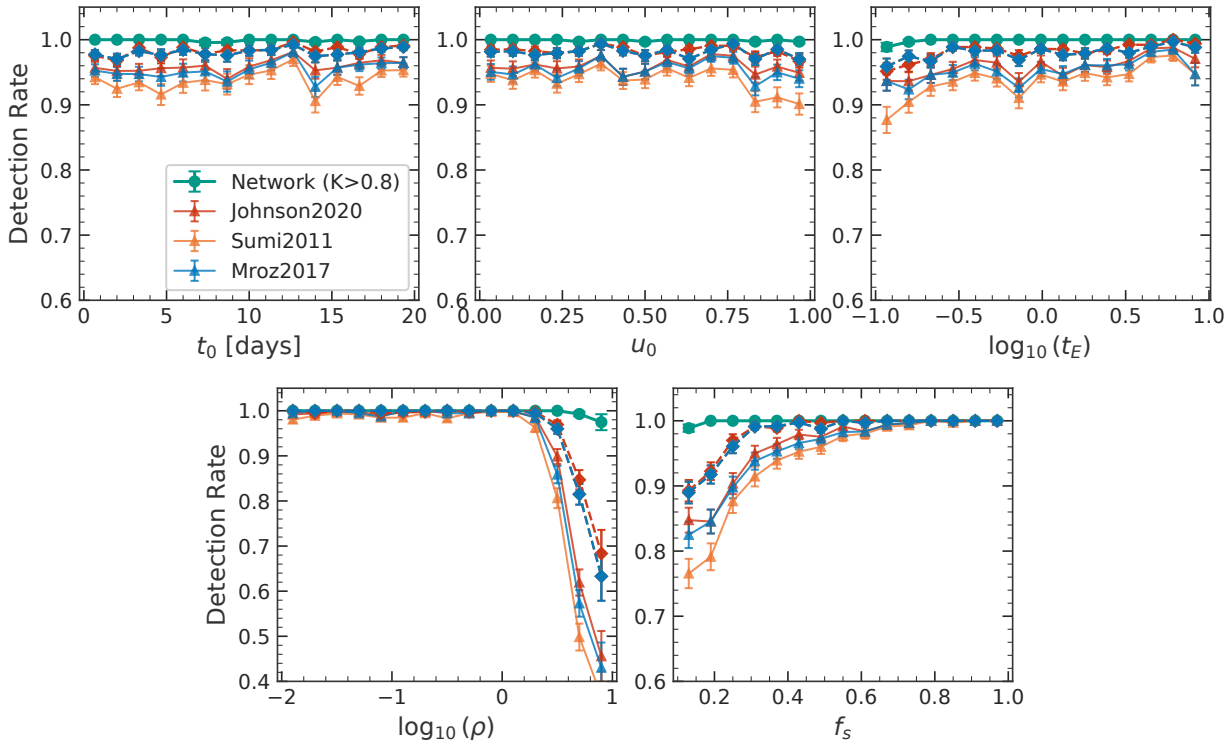


Figure 2. Detection efficiency as a function of FSPL parameters for the Evidence Network and literature hard cuts. The Evidence Network maintains high efficiency in the extreme finite-source regime ($\rho \gtrsim 5$) where hard cuts systematically fail. Both as-published and re-tuned hard cuts are shown with re-tuned shown as dashed lines; hard-cut statistics use oracle (true) parameters, so the comparison is conservative. Error bars denote per-bin binomial 68% confidence intervals.

on previous dimensions $\theta_{1:i-1}$ and the summary $\mathbf{s}$, enabling tractable likelihood computation via:

$$\log p_\psi(\boldsymbol{\theta}|\mathbf{s}) = \log p_{\mathbf{z}}(T_\psi^{-1}(\boldsymbol{\theta}|\mathbf{s})) + \log \left| \det \frac{\partial T_\psi^{-1}}{\partial \boldsymbol{\theta}} \right| \quad (7)$$

Intuitively, the log-posterior at $\boldsymbol{\theta}$ is the log-density of the base Gaussian evaluated at the point that maps to $\boldsymbol{\theta}$ under the flow, plus a Jacobian correction for the change of variables.

We train via maximum likelihood on 80,000 simulated pairs $\{(\boldsymbol{\theta}_i, \mathbf{x}_i)\}_{i=1}^{80000}$, with an additional 20,000 held out for validation, using the `sbi` library (Boelts et al., 2025). We jointly optimize the transformer parameters ϕ and flow parameters ψ so that simulated parameter-data pairs receive high posterior density under the model:

$$\max_{\phi,\psi} \mathbb{E}_{(\boldsymbol{\theta},\mathbf{x})\sim p(\boldsymbol{\theta},\mathbf{x})} \left[\log p_\psi(\boldsymbol{\theta}|g_\phi(\mathbf{x}))\right] \quad (8)$$

Training uses the same data augmentation pipeline (Section 2.2), ensuring the network learns to handle gaps, dropout, and variable noise. We employ the Adam optimizer with learning rate 10^{-4}, batch size 512, and train for up to 400 epochs with early stopping (patience 30 epochs) based on validation negative log-likelihood. Training requires approximately 18 hours on a single NVIDIA H100 GPU.

At test time, drawing posterior samples for a new observation requires only a single forward pass through the transformer to compute $\mathbf{s} = g_\phi(\mathbf{x})$, followed by sampling from the flow $\boldsymbol{\theta}^{(i)} \sim p_\psi(\cdot|\mathbf{s})$ by transforming base samples $\mathbf{z}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ via $\boldsymbol{\theta}^{(i)} = T_\psi(\mathbf{z}^{(i)}|\mathbf{s})$. Generating 10,000 posterior samples takes $\sim$ 21 ms per event, a $\sim$16,000$\times$ speedup over a comparably-tuned `emcee` sampler on the same hardware (NVIDIA H100); see Appendix D for details.

3. Application to Microlensing Detection and Characterization

Before applying the method to microlensing, we validate it on a toy Gaussian-bump detection problem where the Bayes factor admits a closed-form solution. The Evidence Network recovers near-oracle ROC performance (AUC = 0.725 vs. 0.726) and produces well-calibrated posteriors, confirming that binary labels alone suffice to recover calibrated Bayes factor estimates; full details and figures are in Appendix A.

Having validated the methodology, we apply the complete framework to simulated microlensing light curves. This section presents results for both signal detection (via Bayes factors) and parameter characterization (via SBI posteriors).

3.1. Detection Performance vs. Literature Hard Cuts

Traditional microlensing surveys employ deterministic detection criteria that apply threshold-based cuts to photometric summary statistics. As noted in Section 1, these "hard cuts" lack probabilistic interpretation and exhibit strong parameter-dependence that systematically misses certain event types. To quantify these limitations, we compare our Evidence Network against methods from the literature.

The first method, used in Johnson et al. (2020) for the Nancy Grace Roman Space Telescope yield predictions, requires both $\Delta\chi^2 \geq 300$ (where $\Delta\chi^2$ measures the improvement in fit between a microlensing model and a flat baseline) and $N_{\rm consec} \geq 6$ consecutive photometric points exceeding 3σ above baseline. The second criterion, used in Sumi et al. (2011) for the MOA-II Galactic Bulge survey, requires $\chi_{3+} \geq 80$ (the sum of significances over all consecutive high-SNR points) combined with $N_{\rm consec} \geq 3$. Finally, Mroz et al. (2017) adopted a more permissive threshold of $\chi_{3+} \geq 32$ with $N_{\rm consec} \geq 3$ for the OGLE survey.

For a fair comparison, we compute these statistics on the same test set of 5,000 simulated events with data augmentations as described in Section 2.2. Critically, when evaluating the hard cuts, we calculate $\Delta\chi^2$ using the *true* injected FSPL parameters; this is an oracle upper bound, since in practice one must fit the model, incurring optimization failures and local minima that would further degrade detection rates. We apply the literature thresholds as published: the Johnson et al. (2020) cuts are Roman-tuned at the cadence level ($N_{\rm consec} = 6$ at 15-minute sampling) but were calibrated against Roman's full multi-season observing strategy rather than a 20-day window, while the Sumi et al. (2011); Mroz et al. (2017) thresholds inherit from ground-based surveys. To ensure the comparison does not merely penalize thresholds calibrated for a different observing setup, we additionally *re-tune* each hard cut (jointly sweeping the statistic thresholds) on a validation set drawn from the same generator used to train the network (Appendix C, Table 2). The Evidence Network threshold is likewise selected on validation data.

For the Evidence Network, we adopt a detection threshold of $\log_{10} K > 0.8$ (equivalently, $K > 6.3$, or approximately 1 in a million chance of misclassification under ideal conditions). This threshold was selected to achieve zero false positives on our validation set, ensuring high purity for candidate detections; on the held-out test set the network likewise yields no false positives among 5,000 Gaussian-noise events, corresponding to a 95% upper limit on the false-positive rate of 6×10^{-4}. With this conservative criterion, the Evidence Network detects **4,996 of 5,000 signals (99.9%)**, outperforming all hard cuts: Johnson et al. (2020) achieves 95.9%, Sumi et al. (2011) 93.9%, and Mroz et al. (2017) 95.4%. Even after re-tuning, the evidence network significantly outperforms the hard cuts, despite the cuts retaining the advantage of using oracle (true) parameters.

The network's advantage is most pronounced in the extreme finite-source regime ($\rho \gtrsim 5$), where detection efficiency remains $\sim 95\%$ compared to $\sim 65\%$ for hard cuts. This regime dominates short-duration events, precisely the population of greatest interest for constraining the FFP mass function. Hard cuts fail here because finite-source events have lower peak magnification despite distinctive morphological signatures; the Evidence Network extracts evidence from the full temporal structure rather than relying on peak flux alone.

Figure 2 breaks this down by FSPL parameter: the Evidence Network maintains consistently high detection rates across the parameter space, while hard cuts degrade significantly in the finite-source regime. The hard cuts also show modest degradation for highly-blended events (low f_s), another important case for Roman's densely populated fields of view toward the Galactic Bulge. Representative events detected only by the network (Appendix B) illustrate the failure modes: low peak magnification, observation gaps interrupting consecutive high-SNR points, and ongoing events without an established baseline.

3.2. Posterior Estimation: Injected vs. Recovered Parameters

For events classified as likely signals by the Evidence Network, we estimate posterior distributions $p(\boldsymbol{\theta}|\boldsymbol{x})$ over the five physical parameters using the transformer-based Neural Posterior Estimator (Section 2.5). The key question for amortized inference is whether the network produces posteriors that are both *accurate* (centered on the true parameter values) and *calibrated* (credible intervals reflect genuine uncertainty rather than overconfident or underconfident predictions). We assess both properties on the 4,996 events detected by the Evidence Network at our chosen threshold ($\log_{10} K > 0.8$).

Figure 3 shows injected versus recovered parameter values for all detected events. The correlation between injected and median recovered values is high for all five parameters. Binned median trends (shaded regions) closely track the identity line (dashed gray), with only mild deviations at the edges.

Physical degeneracies. When the source is much smaller than the impact parameter ($\rho \ll u_0$), finite-source effects are negligible and ρ is poorly constrained. Conversely, when

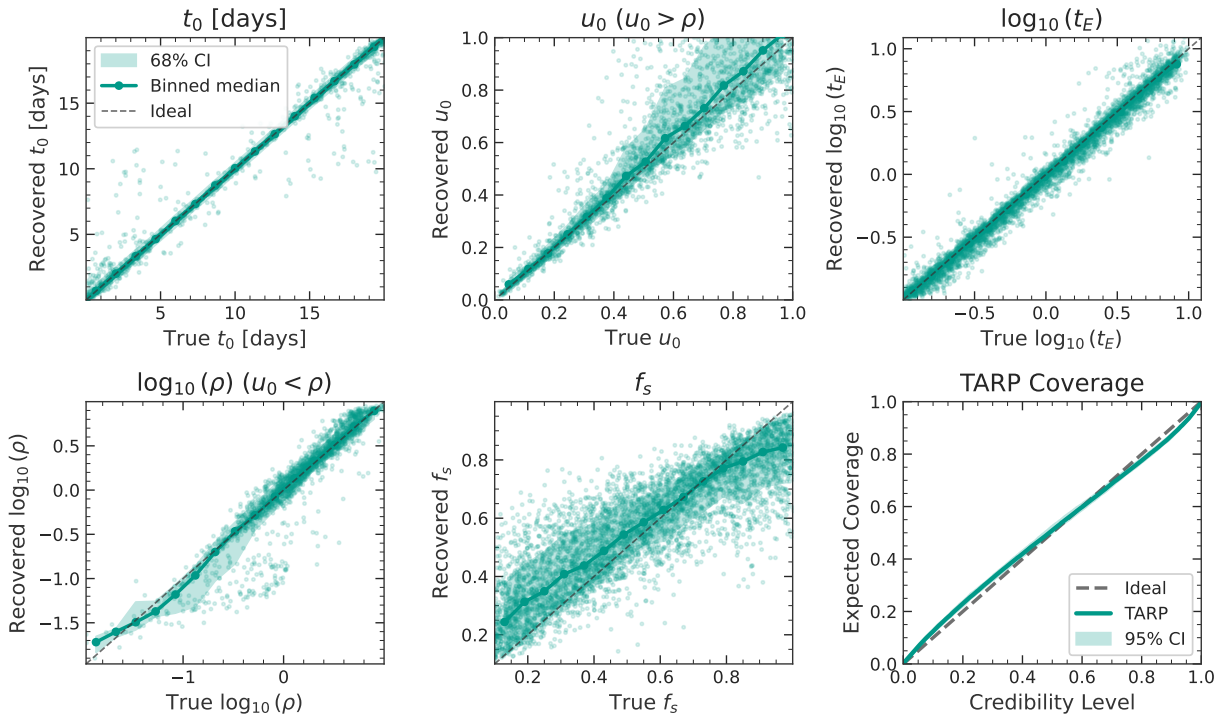


Figure 3. Posterior calibration assessment for the events detected by the Evidence Network. **Top row and middle-left:** Injected vs. recovered parameter values show strong correlation with ideal 1:1 line (gray dashed). Shaded regions show 68% credible intervals. Note physical degeneracies: u_0 is only constrained when $u_0 > \rho$ (impact parameter exceeds source size), while ρ is only constrained when $u_0 < \rho$ (finite-source effects dominate). **Bottom right:** TARP coverage test demonstrates good calibration and coverage.

$u_0 < \rho$ (source larger than impact parameter), the event is dominated by finite-source effects and u_0 becomes degenerate. We therefore only plot a subset of events for these two parameters corresponding to the non-degenerate regions. The blend parameter shows the largest scatter as expected: blending primarily affects the amplitude normalization. Nonetheless, the posteriors correctly identify highly-blended vs. minimally-blended events. We quantify the residual biases and degeneracy axes in Appendix E.

Calibration and Coverage To rigorously validate posterior calibration, we apply the Test of Accuracy with Random Points (TARP) (Lemos et al., 2023). The bottom right panel of Figure 3 shows that the expected coverage closely tracks the credibility level of the posteriors, demonstrating that they are not biased and are reasonably well-calibrated.

4. Discussion

Our unified framework replaces ad hoc hard cuts with calibrated Bayes factors via Evidence Networks, combined with amortized posterior inference via NPE.

4.1. Advantages of the Unified Framework

Our approach combines signal detection and parameter characterization within a single probabilistic framework, offering several advantages over traditional microlensing analysis pipelines:

Transferable representations. The two heads share an encoder but are trained separately, so the framework is "unified" at the level of the learned representation rather than a single joint objective. We show in Appendix C that this sharing is beneficial, not merely architectural: an Evidence head trained on the frozen, SBI-trained encoder transfers from inference to detection at essentially no cost, matching the headline detection rate at a fraction of the training budget and wall-clock cost.

Amortized inference. Both the Evidence Network and SBI posterior estimator require upfront computational investment for training (18–20 hours total on a single GPU). However, once trained, inference for new observations is nearly instantaneous: detecting signals via Bayes factors requires $\sim$10 milliseconds per light curve, while drawing 10,000 posterior samples requires $\sim$100 milliseconds. This one-time training cost amortizes over billions of light curves, enabling real-time analysis at survey scale.

Natural handling of irregular data. The transformer architecture processes light curves as unordered sets of observations with attention masking, requiring no interpolation to fixed grids or imputation of missing data. This is particularly valuable for microlensing surveys, which may exhibit seasonal gaps, variable cadence driven by observing strategy, and detector-specific systematics.

Principled model comparison. The Evidence Network outputs calibrated model posterior probabilities $p(\mathcal{M}_1|\mathbf{x})$, enabling principled decision-making under uncertainty: observers can set detection thresholds based on desired false positive rates or incorporate prior odds to account for event rarity, naturally trading off completeness against purity.

4.2. Current Limitations and Future Extensions

False positives and non-Gaussian noise. Our training data contains only microlensing signals and Gaussian noise, so the framework does not currently model astrophysical false positives such as stellar flares, eclipsing binaries, cataclysmic variables, pulsating variables, or instrumental systematics. The reported false-positive rate is therefore an upper bound for the signal-detection problem in isolation, not a survey-level false-positive rate, which in practice contains contaminants that are challenging to distinguish from true positives. For instance, Mroz (2024) argued on prior grounds that TESS event TIC-107150013 is a stellar flare from a magnetically active giant, despite the light curve strongly favoring a finite-source microlensing fit over flare models ($\Delta\mathrm{BIC} \gtrsim 13$) (Kunimoto et al., 2025); resolving such tensions demands combining calibrated likelihoods with realistic astrophysical priors, not light curve vetting alone. Real survey data also exhibit correlated noise, cosmic rays, and bad pixels. Both limitations are in principle addressable: contaminants can be injected into training, and the noise distribution can be learned from real observations (Legin et al., 2023; 2024). More generally, SBI is sensitive to distributional shifts (Filipp et al., 2024), so production priors should be informed by galactic population models rather than the box-uniform priors used here.

Fixed observing window. The network is trained on 20-day windows with $t_E \sim 0.1$–20 days, limiting applicability to longer-duration events ($t_E \sim 30$–200 days) from more massive lenses. Extension requires variable-length windows with adaptive padding or sliding-window tiling of long light curves.

Binary lens models. Binary lenses produce complex caustic morphologies; Zhang et al. (2021) demonstrated NPE for binary parameter estimation but not detection. Extending our framework requires higher-dimensional parameter spaces and potentially hierarchical model comparison.

Population-level inference. Our framework performs event-level inference, whereas scientific questions (e.g., the FFP mass function) require hierarchical models $p(\boldsymbol{\phi}|\{\mathbf{x}_i\})$. Recent hierarchical SBI work (Erickson et al., 2024; Perkins et al., 2024) points toward natural extensions.

4.3. Comparison with Prior Work

Evidence Networks. Jeffrey & Wandelt (2024) introduced Evidence Networks for cosmological model comparison using regularly-sampled CMB power spectra and galaxy survey statistics. Our work represents the first application to irregularly-sampled astrophysical time series, demonstrating that the l-POP-Exponential loss generalizes to variable-length sequential data when combined with transformer embeddings. The key architectural innovation is replacing fixed-length dense networks with attention-based aggregation over unordered observations.

SBI for microlensing. Zhang et al. (2021) applied NPE to binary microlensing parameter estimation but did not address signal detection or single-lens events. Our contribution is twofold: (1) extending to the full single-lens FSPL parameter space with rigorous TARP validation, and (2) unifying detection and characterization within a single framework via Evidence Networks.

Differentiable microlensing. Recent work has introduced differentiable microlensing simulators that enable gradient-based inference. `microJAX` (Miyazaki & Kawahara, 2025) provides a fully differentiable JAX-based ray-shooting implementation, while `microlux` (Ren & Zhu, 2025) implements differentiable contour integration for binary lenses, with Ren et al. (2026) demonstrating its application to real events via Hamiltonian Monte Carlo. These approaches are complementary to ours: differentiable simulators enable gradient-based samplers but require a differentiable forward model, while our SBI framework operates with any black-box simulator and amortizes inference across all observations.

Transformer time-series models. Salinas et al. (2023) employed transformers to classify *TESS* transits vs. false positives using attention maps to identify diagnostic features. Our approach differs in using transformers as summary statistic extractors for downstream detection and density estimation, enabling calibrated probabilistic inference with explicit uncertainty quantification.

4.4. Broader Implications

The combination of Evidence Networks and transformer-based SBI addresses a general challenge in time-domain astronomy: automated, scalable, and principled analysis of irregularly-sampled transient signals. Beyond microlensing, this framework may be applicable to domains such as supernova classification, exoplanet transit detection, variable

star labeling, and more. The key requirement is simply a forward simulator capable of generating sufficient training data.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Bachelet, E., Norbury, M., Bozza, V., and Street, R. pyLIMA : An open source package for microlensing modeling. I. presentation of the software and analysis on single lens models, September 2017.

Boelts, J., Deistler, M., Gloeckler, M., Tejero-Cantero, Á., Lueckmann, J.-M., Moss, G., Steinbach, P., Moreau, T., Muratore, F., Linhart, J., Durkan, C., Vetter, J., Miller, B. K., Herold, M., Ziaeemehr, A., Pals, M., Gruner, T., Bischoff, S., Krouglova, N., Gao, R., Lappalainen, J. K., Mucsányi, B., Pei, F., Schulz, A., Stefanidi, Z., Rodrigues, P., Schröder, C., Zaid, F. A., Beck, J., Kapoor, J., Greenberg, D. S., Gonçalves, P. J., and Macke, J. H. Sbi reloaded: A toolkit for simulation-based inference workflows, August 2025.

Bozza, V., Bachelet, E., Bartolić, F., Heintz, T., Hoag, A., and Hundertmark, M. VBBinaryLensing: A public package for microlensing light curve computation. *Monthly Notices of the Royal Astronomical Society*, 479(4):5157–5167, October 2018. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/sty1791.

Coleman, G. A. L. and DeRocco, W. Predicting the Galactic population of free-floating planets from realistic initial conditions, July 2024.

DeRocco, W., Penny, M. T., Johnson, S. A., and McGill, P. Reconstructing the Free-floating Planet Mass Function with the Nancy Grace Roman Space Telescope, April 2025.

Dong, S., Wu, Z., Ryu, Y.-H., Udalski, A., Mroz, P., Rybicki, K. A., Hodgkin, S. T., Wyrzykowski, L., Eyer, L., Bensby, T., Chen, P., Wang, S. X., Gould, A., Yang, H., Albrow, M. D., Chung, S.-J., Han, C., Hwang, K.-H., Jung, Y. K., Shin, I.-G., Shvartzvald, Y., Yee, J. C., Zang, W., Kim, D.-J., Lee, C.-U., Park, B.-G., Poleski, R., Skowron, J., Szymanski, M. K., Soszynski, I., Pietrukowicz, P., Kozlowski, S., Skowron, D. M., Ulaczyk, K., Gromadzki, M., Ratajczak, M., Iwanek, P., Wrona, M., Mroz, M. J., Rixon, G., Harrison, D. L., and Breedt, E. A free-floating-planet microlensing event caused by a Saturn-mass object. *Science*, 391(6780):96–99, January 2026. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adv9266.

Erickson, S., Wagner-Carena, S., Marshall, P., Millon, M., Birrer, S., Roodman, A., Schmidt, T., Treu, T., Schuldt, S., Shajib, A., Venkatraman, P., and Collaboration, T. L. D. E. S. Lens Modeling of STRIDES Strongly Lensed Quasars using Neural Posterior Estimation, October 2024.

Filipp, A., Hezaveh, Y., and Perreault-Levasseur, L. Robustness of Neural Ratio and Posterior Estimators to Distributional Shifts for Population-Level Dark Matter Analysis in Strong Gravitational Lensing, November 2024.

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J. Emcee: The MCMC Hammer. *Publications of the Astronomical Society of the Pacific*, 125(925):306–312, March 2013. ISSN 00046280, 15383873. doi: 10.1086/670067.

Greenberg, D. S., Nonnenmacher, M., and Macke, J. H. Automatic Posterior Transformation for Likelihood-Free Inference, May 2019.

Jeffrey, N. and Wandelt, B. D. Evidence Networks: Simple losses for fast, amortized, neural Bayesian model comparison, January 2024.

Johnson, S. A., Penny, M. T., Gaudi, B. S., Kerins, E., Rattenbury, N. J., Robin, A. C., Novati, S. C., and Henderson, C. B. Predictions of the Nancy Grace Roman Space Telescope Galactic Exoplanet Survey II: Free-Floating Planet Detection Rates. *The Astronomical Journal*, 160(3):123, August 2020. ISSN 1538-3881. doi: 10.3847/1538-3881/aba75b.

Keehan, D., Yarndley, J., and Rattenbury, N. Microlensing model inference with normalising flows and reversible jump MCMC. *Astronomy and Computing*, 41:100657, October 2022. ISSN 22131337. doi: 10.1016/j.ascom.2022.100657.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization, January 2017.

Kunimoto, M., DeRocco, W., Smyth, N., Bryson, S., and Gaudi, B. S. Searching for Free-Floating Planets with TESS: Results from Sectors 61-65, October 2025.

Legin, R., Adam, A., Hezaveh, Y., and Levasseur, L. P. Beyond Gaussian Noise: A Generalized Approach to Likelihood Analysis with non-Gaussian Noise. *The Astrophysical Journal Letters*, 949(2):L41, June 2023. ISSN 2041-8205, 2041-8213. doi: 10.3847/2041-8213/acd645.

Legin, R., Isi, M., Wong, K. W. K., Hezaveh, Y., and Perreault-Levasseur, L. Gravitational-Wave Parameter Estimation in non-Gaussian noise using Score-Based Likelihood Characterization, October 2024.

Lemos, P., Coogan, A., Hezaveh, Y., and Perreault-Levasseur, L. Sampling-Based Accuracy Testing of Posterior Estimators for General Inference, June 2023.

Miyazaki, S. and Kawahara, H. microJAX: A Differentiable Framework for Microlensing Modeling with GPU-Accelerated Image-Centered Ray Shooting, October 2025.

Mroz, P. TESS Free-floating Planet Candidate Is Likely a Stellar Flare, May 2024.

Mroz, P., Udalski, A., Skowron, J., Poleski, R., Kozlowski, S., Szymanski, M. K., Soszynski, I., Wyrzykowski, L., Pietrukowicz, P., Ulaczyk, K., Skowron, D., and Pawlak, M. No large population of unbound or wide-orbit Jupiter-mass planets. *Nature*, 548(7666):183–186, August 2017. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature23276.

Papamakarios, G. and Murray, I. Fast ε-free Inference of Simulation Models with Bayesian Conditional Density Estimation, April 2018.

Papamakarios, G., Pavlakou, T., and Murray, I. Masked Autoregressive Flow for Density Estimation, June 2018.

Penny, M. T., Gaudi, B. S., Kerins, E., Rattenbury, N. J., Mao, S., Robin, A. C., and Novati, S. C. Predictions of the WFIRST Microlensing Survey I: Bound Planet Detection Rates. *The Astrophysical Journal Supplement Series*, 241(1):3, February 2019. ISSN 1538-4365. doi: 10.3847/1538-4365/aafb69.

Perkins, S. E., McGill, P., Dawson, W., Abrams, N. S., Lam, C. Y., Ho, M.-F., Lu, J. R., Bird, S., Pruett, K., Golovich, N., and Chapline, G. Disentangling the Black Hole Mass Spectrum with Photometric Microlensing Surveys. *The Astrophysical Journal*, 961(2):179, February 2024. ISSN 0004-637X, 1538-4357. doi: 10.3847/1538-4357/ad09bf.

Ren, H. and Zhu, W. A differentiable binary microlensing model using adaptive contour integration method. *The Astronomical Journal*, 169(3):170, March 2025. ISSN 0004-6256, 1538-3881. doi: 10.3847/1538-3881/adb1b2.

Ren, H., Zang, W., Zhu, W., Ryu, Y.-H., Tang, Y., Zhang, J., Albrow, M. D., Chung, S.-J., Gould, A., Han, C., Hwang, K.-H., Jung, Y. K., Shin, I.-G., Shvartzvald, Y., Yang, H., Yee, J. C., Kim, D.-J., Lee, C.-U., Park, B.-G., Tang, Y., Maoz, D., Mao, S., and Qian, Q. KMT-2025-BLG-1314 and KMT-2025-BLG-1392: Two microlensing planetary/brown-dwarf candidates analyzed with differentiable code, March 2026.

Salinas, H., Pichara, K., Brahm, R., Pérez-Galarce, F., and Mery, D. Distinguishing a planetary transit from false positives: A Transformer-based classification for planetary transit signals. *Monthly Notices of the Royal Astronomical Society*, 522(3):3201–3216, July 2023. ISSN 0035-8711. doi: 10.1093/mnras/stad1173.

Smyth, N., Perreault-Levasseur, L., and Hezaveh, Y. Transformer Embeddings for Fast Microlensing Inference, December 2025.

Sumi, T., Kamiya, K., Udalski, A., Bennett, D. P., Bond, I. A., Abe, F., Botzler, C. S., Fukui, A., Furusawa, K., Hearnshaw, J. B., Itow, Y., Kilmartin, P. M., Korpela, A., Lin, W., Ling, C. H., Masuda, K., Matsubara, Y., Miyake, N., Motomura, M., Muraki, Y., Nagaya, M., Nakamura, S., Ohnishi, K., Okumura, T., Perrott, Y. C., Rattenbury, N., Saito, T., Sako, T., Sullivan, D. J., Sweatman, W. L., Tristram, P. J., Yock, P. C. M., Szymanski, M. K., Kubiak, M., Pietrzynski, G., Poleski, R., Soszynski, I., Wyrzykowski, L., and Ulaczyk, K. Unbound or Distant Planetary Mass Population Detected by Gravitational Microlensing. *Nature*, 473(7347):349–352, May 2011. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature10092.

Sumi, T., koshimoto, N., Bennett, D. P., Rattenbury, N. J., Abe, F., Barry, R., Bhattacharya, A., Bond, I. A., Fujii, H., Fukui, A., Hamada, R., Hirao, Y., Silva, S. I., Itow, Y., Kirikawa, R., Kondo, I., Matsubara, Y., Miyazaki, S., Muraki, Y., Olmschenk, G., Ranc, C., Satoh, Y., Suzuki, D., Tomoyoshi, M., Tristram, P. J., Vandorou, A., Yama, H., and Yamashita, K. Free-Floating planet Mass Function from MOA-II 9-year survey towards the Galactic Bulge, July 2023.

Terry, S. K., Bachelet, E., Zohrabi, F., Verma, H., Crisp, A., Huston, M., McGee, C., Penny, M., Abrams, N. S., Albrow, M. D., Anderson, J., Bagheri, F., Beaulieu, J.-P., Bellini, A., Bennett, D. P., Bergsten, G., Bhadra, T. D., Bhattacharya, A., Bond, I. A., Bozza, V., Brandon, C., Novati, S. C., Carey, S., Christiansen, J., DeRocco, W., Gaudi, B. S., Hulberg, J., Silva, S. I., Jones, S. E., Kerins, E., Khakpash, S., Kruszynska, K., Lam, C., Lu, J. R., Malpas, A., Miyazaki, S., Mroz, P., Murlidhar, A., Nataf, D., Newman, M., Olmschenk, G., Poleski, R., Ranc, C., Rattenbury, N. J., Rybicki, K., Saggese, V., Sobeck, J., Stassun, K. G., Stephan, A. P., Street, R. A., Sumi, T., Suzuki, D., Vandorou, A., Vyas, M., Yee, J. C., Zang, W., and Zhang, K. Predictions of the Nancy Grace Roman Space Telescope Galactic Exoplanet Survey. IV. Lens Mass and Distance Measurements, October 2025.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention Is All You Need, August 2023.

Zhang, K., Bloom, J. S., Gaudi, B. S., Lanusse, F., Lam, C., and Lu, J. R. Real-Time Likelihood-Free Inference of Roman Binary Microlensing Events with Amortized Neural Posterior Estimation. *The Astronomical Journal*, 161(6):262, June 2021. ISSN 0004-6256, 1538-3881. doi: 10.3847/1538-3881/abf42e.

Zwart, S. F. P. The origin of free-floating objects in the Galaxy, September 2024.

A. Validation on Toy Model with Analytic Bayes Factors

Before applying Evidence Networks to microlensing data, we validate the methodology on a simplified problem where ground-truth Bayes factors can be computed analytically. This toy model demonstrates that transformer embedding networks trained on binary labels alone can recover calibrated Bayes factor predictions for low SNR bump-hunt timeseries problems.

A.1. Toy Model Setup

We construct a signal detection problem using Gaussian bump templates embedded in white noise. The observed time series $\mathbf{d} \in \mathbb{R}^T$ follows one of two models:

- **Signal hypothesis** $\mathcal{M}_1$: $\mathbf{d} = A\mathbf{g} + \mathbf{n}$, where $A \sim \mathcal{N}(\mu_A, \sigma_A^2)$ is a random amplitude, $\mathbf{g}$ is a unit-norm Gaussian template, and $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2\mathbf{I})$ is white noise
- **Noise-only hypothesis** $\mathcal{M}_0$: $\mathbf{d} = \mathbf{n}$

For this linear-Gaussian model, the Bayes factor admits a closed-form solution. Defining the matched-filter statistic $x \equiv \mathbf{d}^\top\mathbf{g}$, the template norm $G \equiv \mathbf{g}^\top\mathbf{g}$, and the parameter $\alpha \equiv \sigma_A^2 G/\sigma_n^2$, the natural logarithm of the Bayes factor is:

$$\log K = -\frac{1}{2}\log(1+\alpha) + \frac{1}{2}\frac{\alpha}{1+\alpha}\rho^2 \tag{9}$$

where $\rho^2 \equiv x^2/(\sigma_n^2 G)$ is the matched-filter signal-to-noise ratio squared. This analytic expression serves as ground truth for validation.

We generate toy data with $T = 200$ time points over a 20-day window with noise standard deviation $\sigma_n = 0.1$. Gaussian templates are centered and have widths of 1.5 days, with amplitudes drawn from the prior $A \sim \mathcal{N}(0.1, 0.1^2)$. We train the Evidence Network on 50,000 simulated light curves with equal signal/noise fractions, using binary labels only. The analytic $\log K$ values are reserved strictly for post-training evaluation.

The Evidence Network uses the same transformer encoder described in Section 2.3 (6 layers, 8 heads, $d_{\text{model}} = 64$), followed by a 2-layer MLP classifier with hidden dimension 64. We train using the l-POP-Exponential loss with $\alpha = 2$ for 60 epochs, employing an Adam optimizer with learning rate 10^{-4} and a step learning rate scheduler (factor 0.5 every 20 epochs without improvement).

A.2. Toy Model Performance

Figure 5a shows the receiver operating characteristic (ROC) curve for the trained Evidence Network compared against an "oracle" classifier that uses the analytic $\log_{10} K$ directly. The network achieves near-oracle discriminative performance (AUC = 0.725 vs. oracle 0.726), demonstrating that Evidence Networks recover calibrated Bayes factors from binary labels alone without requiring explicit likelihood computation. The near-perfect overlap of the ROC curves shows that the network has learned to extract essentially all available statistical information for discriminating signal from noise. The network's true positive rate at 1% false positive rate is 0.32, compared to the oracle's 0.33. At higher false positive rates (5% and 10%), the network achieves values within 0.01 of the oracle, confirming robust performance across the full operating range.

While ROC curves measure discriminative performance, calibration assesses whether predicted probabilities match empirical frequencies, a crucial requirement for Bayesian model comparison. We apply the coverage test of Jeffrey & Wandelt (2024): events are binned by predicted model posterior $p(\mathcal{M}_1|\mathbf{x}) = K/(1+K)$, and well-calibrated predictions satisfy $p(\mathcal{M}_1|\mathbf{x}) \approx f_{\text{signal}}$ within binomial uncertainty $\sigma_{\text{binom}} = \sqrt{f_{\text{signal}}(1-f_{\text{signal}})/N_{\text{bin}}}$, where f_{signal} is the empirical fraction of true signals in each bin. Figure 5b shows the resulting calibration curve. The points lie almost exactly on the identity line, with deviations consistent with binomial sampling uncertainty.

These results motivate application of Evidence Networks to microlensing detection, where we similarly have access to labeled training data (from simulations) but lack analytic expressions for Bayes factors.

B. Representative Events Detected by the Evidence Network

To make the failure modes of hard cuts concrete, Figure 6 shows six finite-source events that the Evidence Network assigns strong detections ($\log_{10} K > 14$) but that are missed by the Johnson et al. (2020) criteria. All six lie in the short-duration, large-ρ regime ($t_E \lesssim 1$ day, $\rho \gtrsim 2$) that dominates the low-mass FFP population.

Three distinct failure modes appear. First, finite-source broadening spreads the magnification signal over several days, so even with clear morphological structure the peak

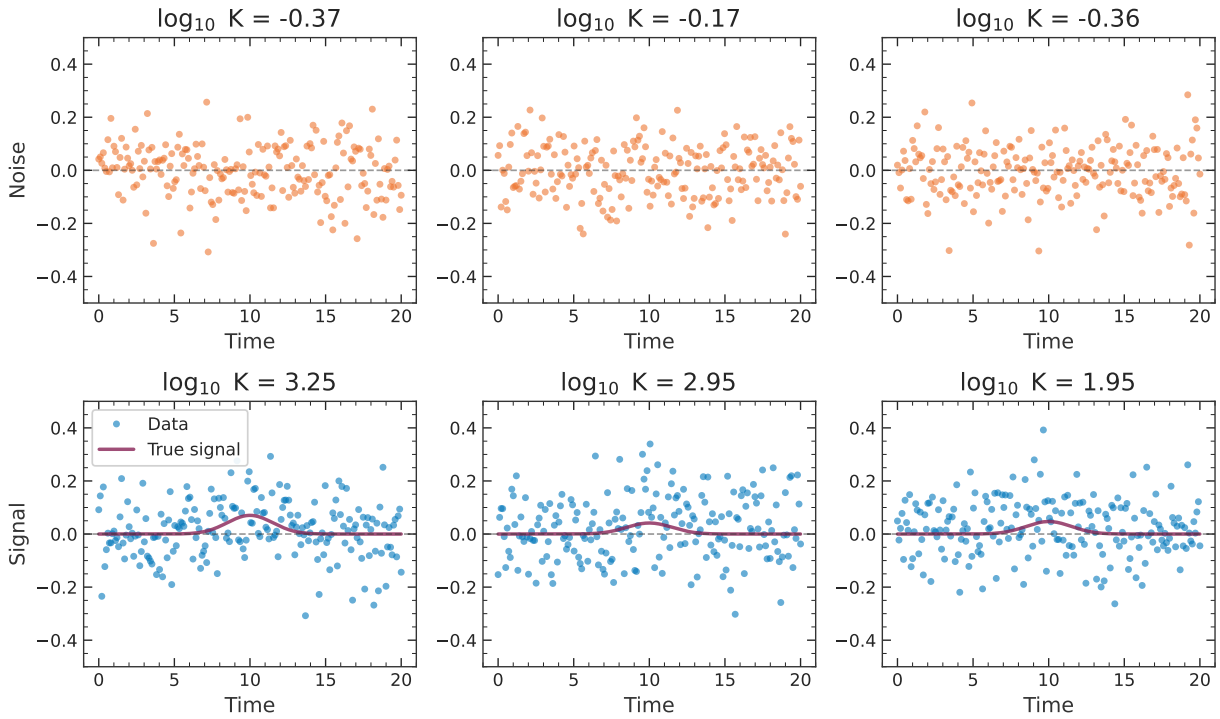
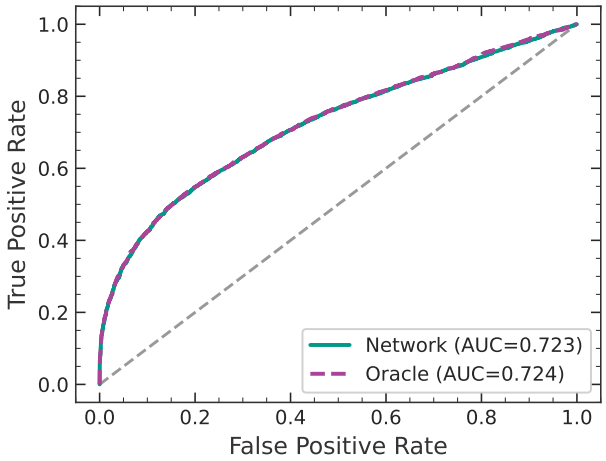
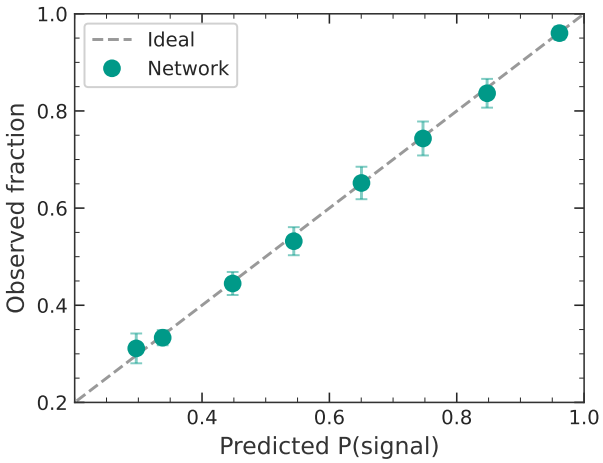


Figure 4. Toy Data Examples. The top row are noise-only and the bottom row are high Bayes-factor signals with the true underlying signal shown in red.



(a) ROC curve comparing Evidence Network to oracle classifier



(b) Calibration: predicted vs. observed signal fraction

Figure 5. Toy model validation with analytic Bayes factors. **(a)** The Evidence Network achieves near-oracle ROC performance (AUC=0.725 vs 0.726), demonstrating it extracts essentially all discriminative information. **(b)** Calibration curve shows predicted model posteriors match empirical signal fractions, confirming the network produces well-calibrated Bayes factor estimates from binary labels alone.

rarely clears the 3σ threshold needed for $N_{\text{consec}} \geq 6$ consecutive points (top left, top middle). Second, observation gaps coincident with the peak interrupt the run of high-SNR points, so a detectable event fails the consecutivity requirement despite an unambiguous signal in the ungapped data (top right, bottom left). Third, events whose peak falls near the edge of the observing window never establish a clean pre- or post-event baseline, causing the $\Delta\chi^2$ reference fit to be biased and the cut to reject the event (bottom middle, bottom right).

In all three cases the signal remains extractable because

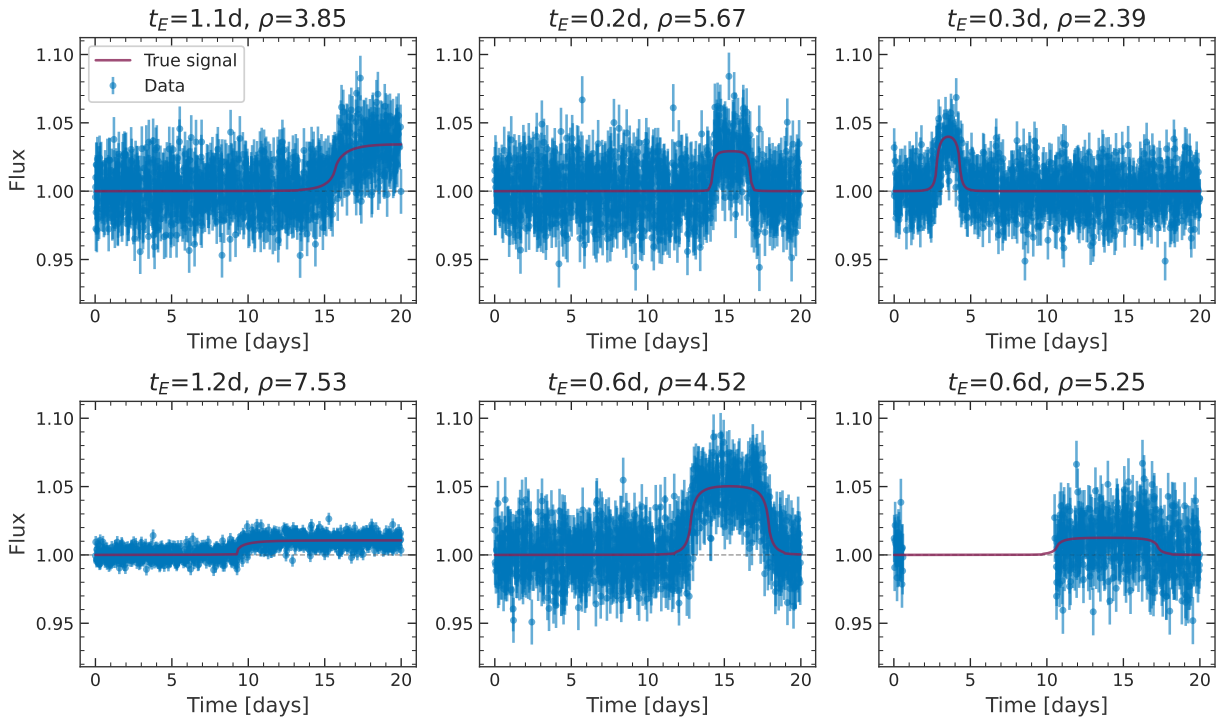


Figure 6. Representative events detected by the Evidence Network (with $\log_{10} K > 14$, strong detections) but missed by hard cuts from Johnson et al. (2020). These examples illustrate systematic failures of threshold-based detection: low peak magnification (top left, middle), observation gaps interrupting consecutive high-SNR points (top right, bottom left), and ongoing events without established baseline (bottom middle, right). The Evidence Network extracts signal from the full temporal morphology rather than requiring specific summary statistics to exceed fixed thresholds.

the transformer attends to the full light curve rather than to a fixed summary statistic. This is why the Evidence Network's advantage is concentrated precisely in the $\rho \gtrsim 5$ regime most relevant to the FFP mass function.

C. Shared- vs. Separate-Encoder Ablation and Re-Tuned Hard Cuts

Does sharing the encoder help? To test whether the shared transformer encoder is beneficial or merely architectural, we train an Evidence detection head on top of the *frozen* encoder learned by the NPE (inference) model, and compare it head-to-head against a dedicated Evidence model trained from scratch. Both are evaluated on the same seeded test set of 5,000 signals and 5,000 Gaussian-noise events, with the detection threshold on $\log_{10} K$ chosen per model so that the validation false-positive rate is zero. We run the comparison at two training budgets: 20K-samples and 100K-samples. Table 1 reports the result.

Table 1. Transfer vs. dedicated encoder detection performance at two training budgets. At both budgets the transfer model reaches 99.9%; the from-scratch detector matches it only approximately, and only at the larger budget.

Budget	Model	$\log_{10} K$ thr.	TPR
20K	Dedicated	18.2	93.0%
20K	Transfer	1.20	99.9%
100K	Dedicated	1.80	98.9%
100K	Transfer	0.11	99.9%

Two readings follow. First, at the small 20K budget the transfer model already matches the headline 99.9% detection rate while the matched from-scratch detector lags at 93.0%: the SBI-trained embedding does the heavy lifting and the head needs only a small classification budget. Second, at the matched 100K budget the from-scratch detector closes most of the gap (98.9% vs. 99.9%), but the transfer model still wins. The transfer model also reaches FPR $= 0$ at a smaller threshold; its noise $\log_{10} K$ distribution is consistent with the inference-trained encoder having absorbed information about the full microlensing manifold that a classification-only loss does not capture. Finally, on identical hardware and hyperparameters, the transfer model trains $3.8\times$ faster than the from-scratch model.

Re-tuned hard cuts. We also re-tune each hard cut on a validation set drawn from the same generator used to train the network, jointly sweeping the statistic thresholds. Table 2 reports both the literature and re-tuned criteria. Re-tuning raises every cut's TPR to $\sim$97% (from 91–94%), but all remain below the Evidence Network's 99.9%. The hard cuts retain the advantage of oracle (true) lensing parameters throughout.

Table 2. Literature vs. re-tuned hard-cut detectors on the seeded test set. "Stat" / "Consec" are the statistic and consecutive-point thresholds. TPR is the detection rate; FPR is on the Gaussian-noise test set. Hard cuts use oracle parameters.

Method	Lit. stat / consec	Tuned stat / consec	Lit. TPR	Tuned TPR (FPR)
Johnson et al. (2020)	$\Delta\chi^2 \geq 300$ / 6	300 / 2	95.9%	97.0% (0)
Sumi et al. (2011)	$\chi_{3+} \geq 80$ / 3	6.98 / 2	93.9%	97.2% (2×10^{-4})
Mroz et al. (2017)	$\chi_{3+} \geq 32$ / 3	6.98 / 2	95.4%	97.2% (2×10^{-4})

D. NPE vs. MCMC Wall-Clock Benchmark

We measure wall-clock cost directly. We compare amortized NPE sampling against an `emcee` (Foreman-Mackey et al., 2013) sampler on a single NVIDIA H100. The MCMC budget is 5,000 retained samples per event (32 walkers, 1,000-step burn-in). Table 3 reports per-event timings. The median speedup is $\sim$16,000×.

Table 3. Per-event wall-clock cost of NPE sampling vs. `emcee` MCMC on the same forward model (NVIDIA H100, 20 events, 10,000 NPE samples / 5,000 MCMC samples per event).

Quantity	Median	Mean
NPE per event	21 ms	44 ms
MCMC per event	346 s	333 s
Speedup (MCMC/NPE)	16,400×	14,500×

We report wall-clock cost only. The MCMC time is for single-event sampling with a modest walker count and a CPU-side likelihood evaluator. Optimizing the MCMC configuration could shrink this gap, but amortized inference will continue to be much faster.

E. Posterior Bias Decomposition

Table 4 summarizes the per-parameter posterior accuracy and calibration over all 5,000 test events. The well-constrained parameters (t_0, $\log_{10} t_E$) have small bias; the degeneracy-affected parameters (u_0, $\log_{10} \rho$, f_s) carry larger residuals, with 95% credible-interval coverage remaining between 0.90 and 0.97 throughout.

Table 4. Per-parameter posterior diagnostics over 5,000 test events: median bias, RMSE, Pearson correlation between injected and recovered medians, and empirical coverage of the 68% and 95% credible intervals.

Parameter	Bias	RMSE	r	68%	95%
t_0	+0.010	0.445	0.997	0.826	0.972
u_0	+0.084	0.222	0.770	0.636	0.925
$\log_{10} t_E$	−0.018	0.064	0.994	0.661	0.931
$\log_{10} \rho$	+0.056	0.436	0.824	0.553	0.901
f_s	+0.029	0.131	0.881	0.643	0.916

Figure 7 decomposes these residuals against the physically relevant axes. The $\log_{10} \rho$ bias is consistent with zero in the well-resolved finite-source regime ($u_0 \gg \rho$) and develops a structured bias only as $u_0 \sim \rho$, exactly where the $u_0 \leftrightarrow \rho$ degeneracy folds two physical solutions onto one light curve. The f_s bias shrinks with peak magnification, as expected from the $(f_s, A_{\rm peak})$ degeneracy. Finally, u_0 recovery is essentially unbiased for $u_0 > 0.2$ and, near the detection boundary ($u_0 \to 1$, $u_0 \sim \rho$), inflates its credible interval rather than producing a point bias. The network widens its posterior where the data are uninformative rather than becoming overconfident.

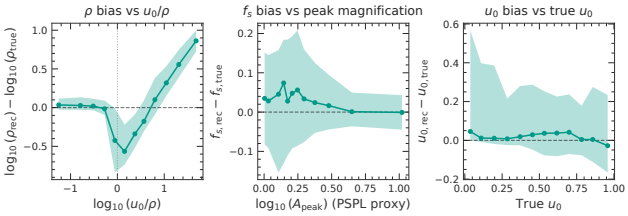


Figure 7. Posterior bias decomposition. **Left:** $\log_{10} \rho$ residual vs. u_0/ρ. **Middle:** f_s residual vs. peak magnification. **Right:** u_0 residual vs. true u_0. Bands show 16–84th percentile residuals per bin.