Transformer Embeddings for Fast Microlensing Inference

Nolan Smyth[1,2,3]
Laurence Perreault-Levasseur[1,2,3,4] **Yashar Hezaveh**[1,2,3,4]
[1]Université de Montréal [2]Mila [3]Ciela Institute [4]CCA, Flatiron Institute
{nolan.smyth, laurence.perreault.levasseur, yashar.hezaveh}@umontreal.ca

Abstract

The search for free-floating planets (FFPs) is a key science driver for upcoming microlensing surveys like the Nancy Grace Roman Galactic Exoplanet Survey. These rogue worlds are typically detected via short-duration microlensing events, the characterization of which often requires analyzing noisy, irregularly-sampled observations. We present a pipeline for this task using simulation-based inference. We use a Transformer encoder to learn a compressed summary representation of the raw time-series data, which in turn conditions a neural posterior estimator. We demonstrate that our method produces accurate and well-calibrated posteriors over three orders of magnitude faster than traditional methods. We also demonstrate its performance on KMT-BLG-2019-2073, a short-duration FFP candidate event.

1 Introduction

Free-floating planets (FFPs) may be the most ubiquitous type of terrestrial-mass exoplanet, potentially outnumbering their bound counterparts by a factor of more than 20 [1]. Low-mass FFPs primarily form in planetary disks and are subsequently ejected from their system of origin [2, 3]. Due to their negligible electromagnetic radiation and lack of host star, FFPs are extremely difficult to detect. Gravitational microlensing is the most promising technique to find these rogue worlds [4]. Microlensing occurs when a compact foreground mass passes near the line of sight to a background star, warping the light around the lens. The multiple images produced at the observer are unresolved, resulting in a temporary smooth, achromatic magnification of the background star.

The Nancy Grace Roman Space Telescope is expected to detect thousands of FFPs [5]. This will usher in a new era of exoplanet demographics, illuminating the origins of these elusive objects [6, 7]. Crucial to this effort is rapid characterization of microlensing signals. Traditional methods like Markov Chain Monte Carlo (MCMC) are computationally expensive and do not scale well to the billions of light curves Roman will deliver. While anomaly detection pipelines are expected to significantly filter this dataset, a significant proportion will still require detailed characterization. Simulation-Based Inference (SBI) provides a powerful framework for efficient posterior estimation by amortizing the cost of simulation.

Neural Posterior Estimation (NPE) is a SBI approach where a neural network is trained to learn the Bayesian Posterior $p(\theta|x)$ over the model parameters θ given the observed data x, bypassing the need for likelihood evaluations. A key benefit of this amortized approach is that training is a one-time cost, after which inference for new observations is extremely fast, requiring only a single forward pass through the trained network. Previous work has used amortized SBI for binary microlensing events, motivated by the computational cost of the forward model [8]. This work used a 1D ResNet with a Gated Recurrent Unit as an embedding network for fixed-length, regularly-sampled data. We found that a similar, well-calibrated recurrent network trained on regularly-sampled light-curves

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

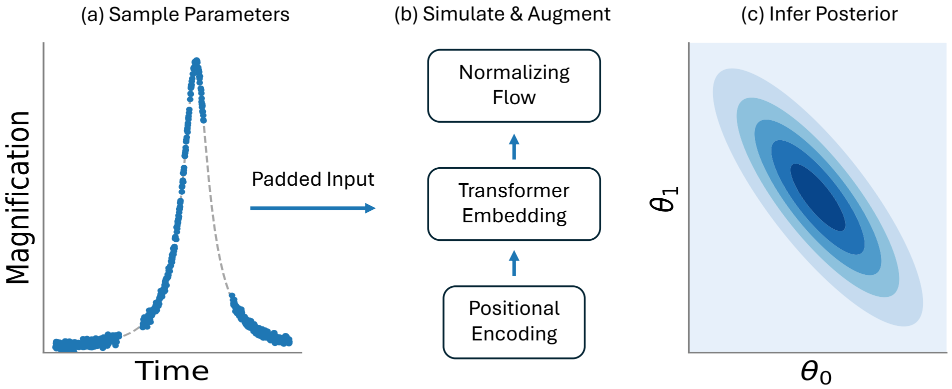


Figure 1: Our SBI pipeline. Parameters θ are drawn to simulate a light curve, which is then augmented with data gaps, dropout, and noise. The resulting observation x is fed into a Transformer encoder to produce an embedding that conditions a normalizing flow approximating the posterior $p(\theta|x)$.

failed catastrophically when even minor data gaps were introduced. This is a classic example of the well-known difficulty of handling distributional shifts in time-series data [9, 10].

In this work, we develop an end-to-end SBI pipeline for microlensing parameter estimation that naturally handles variable length, sparse, noisy, and irregularly sampled time-series data.[1] While previous work has applied transformer encoders for astrophysical time series data (e.g. [11, 12]), the primary contribution of this work is to demonstrate a robust and scalable solution to this specific challenge, enabling fast and accurate inference directly from raw, time-series data without the need for imputation or complex pre-processing. We use a Transformer encoder [13], as the self-attention mechanism is particularly well-suited to this type of data. A schematic of this pipeline is shown in Figure 1. We validate our pipeline, showing it produces accurate and well-calibrated posteriors across a wide range of event morphologies. We also demonstrate the pipeline's performance on a real-world FFP candidate, KMT-2019-BLG-2073 [14], recovering posterior distributions consistent with the publicly available data [14, 15].

2 Method

We simulate microlensing events using a finite-source, point-lens (FSPL) model implemented with the `VBMicrolensing` package [16]. The parameters, collectively denoted by θ, are: the time of closest approach t_0; the minimum impact parameter in units of the Einstein radius u_0; the Einstein crossing time t_E; the normalized source radius $\rho \equiv \theta_\star/\theta_E$; and the source flux fraction $f_s \equiv \frac{F_{\rm source}}{F_{\rm source}+F_{\rm blend}}$. The total observed flux at time t, normalized to the baseline, is $F(t) = f_s A(t, \theta) + (1 - f_s)$, where A is the magnification. All timescales are in units of days.

To train a robust network, we employ on-the-fly data augmentation. For each training sample θ_i drawn from the prior shown in Table 1a, we generate a dense, noiseless light-curve. We then apply a sequence of random augmentations: **Seasonal Gaps:** Introduce 0 to 3 gaps, each with a length of 1 to 10 days. **Random Dropout:** Remove a random fraction (0%-60%) of the remaining points to simulate data quality cuts. **Noise Injection:** Add Gaussian photometric noise, with σ for each light-curve drawn uniformly from $[0.001, 0.02]$ in relative flux units.

Each light-curve is represented as a padded sequence of length $L = 1000$. Each timestep has three channels: $x_i = (t_i^{\rm norm}, F_i, \sigma_i), i = 1, ..., L$. Times are normalized to $[-1, 1]$ over a window of $T = 20$ days and σ_i is the per-point photometric uncertainty. Padded timesteps are masked with a value of -2. To ensure the network trains on meaningful signals, we filter these augmented light-curves using a set of recoverability criteria: at least 5 data points must lie within $t_E/2$ of the peak, t_0; at least 5 points must lie more than $2t_E$ from the peak to establish a baseline magnitude; the peak magnification must be at least $5\times$ larger than the mean per-point flux error, $F(t_0)/\sigma > 5$. In a full analysis pipeline, there would be a process for identifying the most promising light-curves - these criteria are meant to serve as a quality assurance check to avoid the network being trained or evaluated on light-curves that would not have a recoverable microlensing signal.

[1] All the code used in this work is available at: https://github.com/NolanSmyth/sbi_microlensing_transformers.

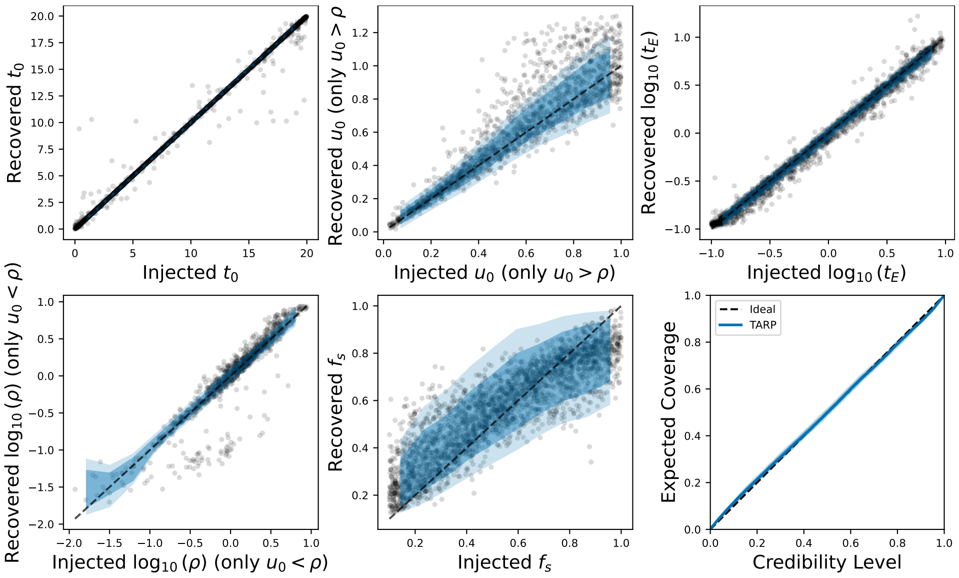


Figure 2: Injected–recovered plots show recovered median posterior parameters. The shaded regions show the 16-84 and 5-95 percentile ranges, averaged per bin. The TARP diagnostic is shown in the bottom right, demonstrating excellent calibration.

Our NPE pipeline uses a standard Transformer encoder [13] to map the input sequence $x \in \mathbb{R}^{L\times 3}$ to a summary vector $z \in \mathbb{R}^d$. The network consists of an input projection layer, sinusoidal positional encoding, and a stack of 6 Transformer layers with 8 attention heads, a model dimension of 256, and a feed-forward dimension of 512. To aggregate the variable-length sequence output, we perform normalized average pooling over the unmasked timesteps. We then use a Masked Autoregressive Flow, as implemented by the `sbi` package [17, 18], to model the posterior distribution $p(\theta|x)$, conditioned on the Transformer embedding. The network is trained by minimizing the negative log-likelihood of the posterior on pairs of simulations $\{(\theta_i, x_i)\}$. We trained our network on 80,000 simulated events, plus an additional 20,000 for validation during training. We use the Adam optimizer [19] with an initial learning rate of 10^{-4} and a `ReduceLROnPlateau` scheduler that reduces the learning rate by a factor of 0.5 with a patience of 10 epochs. All training and evaluation was conducted in $\sim$ 20 hours on a single Nvidia H100 GPU using 16GB of memory.

3 Results

We assess the calibration of our trained network on a test set of 5,000 simulated events, drawing 5,000 posterior samples from each event. Figure 2 shows the injected-vs-recovered values for each parameter and the overall model calibration. Each panel, except the bottom right, shows the posterior median recovered values against the true injected values. For the impact parameter u_0 and source size ρ, we plot only the regimes where they are identifiable. For point-source-like events $(u_0 > \rho)$, u_0 is well-recovered, while for events with strong finite-source effects $(u_0 < \rho)$, ρ is well-recovered. The bottom right panel shows a TARP (Tests of Accuracy with Random Points) diagnostic [20]. Calibrated TARP is both a necessary and sufficient condition for an accurate posterior estimator. Crucially, the NPE yields a more than 10^4 factor speedup compared to ensemble sampling as implemented in `emcee`[21] to generate the same number of posterior samples (see Appendix A for more details).

We also applied the pipeline to KMT-2019-BLG-2073 [14], a short-duration microlensing event with pronounced finite-source effects, classified as a "likely FFP candidate". We used publicly available I-band data from the KMTNet survey's pySIS pipeline [15]. [2] We convert the raw magnitudes to an absolute flux scale with a reported zero point of 18.15, $F_{\text{abs}} = 10^{-0.4(I-18.15)}$, and then normalize the light-curve by the baseline flux. Since the fit reported in [14] was performed on data from KMTNet's TLC pipeline, it does not provide a direct comparison for the publicly available pySIS data we analyze here. Such differences between different photometric extraction pipelines is important and should be considered depending on the application. The recovered posteriors are shown in Table 1b.

[2] https://kmtnet.kasi.re.kr/ulens/event/2019/view.php?event=KMT-2019-BLG-2073.

Table 1: Priors and recovered posteriors.

(a) Priors

Parameter	Prior
t_0	$\mathrm{Uniform}(0,\ 20)$
u_0	$\mathrm{Uniform}(0,\ 1.5)$
t_E	$\mathrm{LogUniform}(0.1,\ 20)$
ρ	$\mathrm{LogUniform}(10^{-2},\ 10)$
f_s	$\mathrm{Uniform}(0.1,\ 1)$

(b) Recovered posteriors for KMT-2019-BLG-2073

Parameter	Recovered value	Reported Fit
t_0	$8708.60^{+0.02}_{-0.02}$	8708.58
u_0	$0.20^{+0.11}_{-0.10}$	0.32
t_E	$0.355^{+0.034}_{-0.031}$	0.50
ρ	$0.832^{+0.080}_{-0.090}$	N/A
f_s	$0.82^{+0.11}_{-0.13}$	0.61

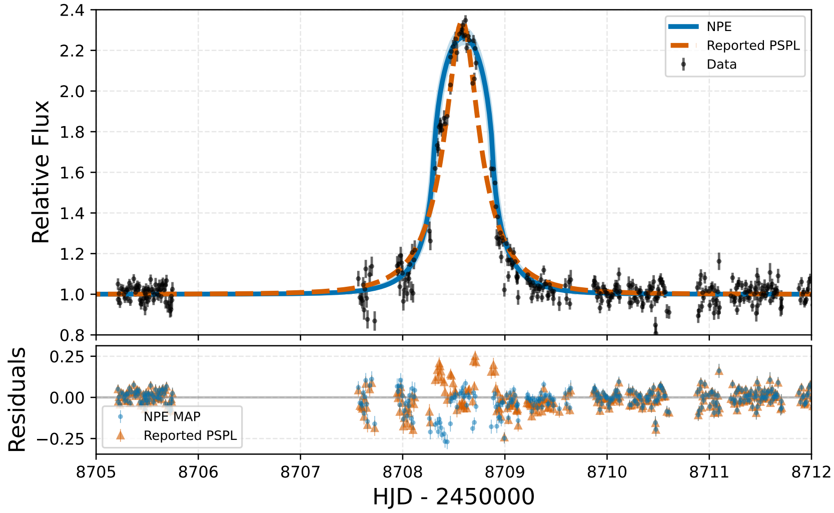


Figure 3: Data and model fits to KMT-2019-BLG-2073 using pySIS (I-band). Time is shown as Heliocentric Julian Date (HJD). The shaded band shows the 5-95 percentile range recovered by the NPE, while the dashed line is the reported PSPL fit.

Figure 3 shows the data, light-curve recovered by the NPE, and the best-fit point-source-point-lens (PSPL) model. The FSPL model provides an excellent fit around the peak where finite-source effects dominate, with residuals comparable to or smaller than those of the PSPL fit.

4 Discussion and Limitations

Our SBI approach with a Transformer embedding demonstrates a powerful, automated method for microlensing characterization. In deployment, the priors should be informed by the expected distribution of detectable events, given by a hierarchical galactic and lens population model [22]. This is beyond the scope of the current work, but important as SBI methods can be sensitive to distributional shifts [23].

Furthermore, the model is trained using Gaussian noise, which is a good approximation of the fundamental Poisson measurement uncertainty but does not capture more complex, systematic noise sources present in real detectors or insidious false-positives. False-positive signals are ubiquitous due to stellar variability, magnetic activity, and other transients that can mimic temporary increased flux in a source. This issue also arises with traditional methods (see e.g. [24, 25, 26]). It will be crucial to also model these false positives and either include them in the detection pipeline, or model them for accurate inference. Additionally, the model is trained on a fixed 20-day window, tailored to short-duration events associated with low-mass FFPs, but limiting its applicability to long-duration events from more massive lenses like brown dwarfs or intermediate mass black holes [27, 28, 29]. This is straightforward to address with an extended prior, depending on the use-case. Our application to KMT-2019-BLG-2073 serves as a promising proof-of-concept. However, a more extensive validation on a larger, diverse set of real-world events is required to fully establish the method's reliability and is a key direction for future work.

There are several avenues for future architectural exploration. Ablation studies would help determine the optimal model size and embedding dimension. We utilized sinusoidal positional encoding added to the input. While the timestamps of the inputs should contain all the necessary positional information, we noted that this encoding resulted in faster convergence during training. Quantifying this improvement and comparing encoding schemes could potentially offer significant benefits. Also, while our use of normalized average pooling over unmasked tokens is simple and effective, other pooling or aggregation mechanisms may better capture information from the features of the microlensing peak.

Lastly, we note that our pipeline is complimentary to automated anomaly detection pipelines (see e.g. [30]) that are intended to reduce the number of light-curves to be manually analyzed.

5 Conclusion

We have presented a robust and calibrated SBI pipeline for microlensing analysis. By leveraging a Transformer encoder, our method can directly process sparse, noisy, and irregularly-sampled time-series data, overcoming a major limitation of amortized inference approaches in this domain. This work provides a powerful, scalable tool for analyzing the large datasets. This will be crucial for upcoming surveys like the Roman Space Telescope, helping to accelerate the discovery and characterization of free-floating planets.

References

[1] Takahiro Sumi, Naoki koshimoto, David P. Bennett, Nicholas J. Rattenbury, Fumio Abe, Richard Barry, Aparna Bhattacharya, Ian A. Bond, Hirosane Fujii, Akihiko Fukui, Ryusei Hamada, Yuki Hirao, Stela Ishitani Silva, Yoshitaka Itow, Rintaro Kirikawa, Iona Kondo, Yutaka Matsubara, Shota Miyazaki, Yasushi Muraki, Greg Olmschenk, Clement Ranc, Yuki Satoh, Daisuke Suzuki, Mio Tomoyoshi, Paul J. Tristram, Aikaterini Vandorou, Hibiki Yama, and Kansuke Yamashita. Free-Floating planet Mass Function from MOA-II 9-year survey towards the Galactic Bulge, July 2023.

[2] Simon F. Portegies Zwart. The origin of free-floating objects in the Galaxy, September 2024.

[3] Gavin A. L. Coleman and William DeRocco. Predicting the Galactic population of free-floating planets from realistic initial conditions, July 2024.

[4] B. Scott Gaudi. Microlensing Surveys for Exoplanets. *Annual Review of Astronomy and Astrophysics*, 50(1):411–453, September 2012.

[5] Samson A. Johnson, Matthew T. Penny, B. Scott Gaudi, Eamonn Kerins, Nicholas J. Rattenbury, Annie C. Robin, Sebastiano Calchi Novati, and Calen B. Henderson. 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.

[6] Andrew Gould, Youn Kil Jung, Kyu-Ha Hwang, Subo Dong, Michael D. Albrow, Sun-Ju Chung, Cheongho Han, Yoon-Hyun Ryu, In-Gu Shin, Yossi Shvartzvald, Hongjing Yang, Jennifer C. Yee, Weicheng Zang, Sang-Mok Cha, Dong-Jin Kim, Seung-Lee Kim, Chung-Uk Lee, Dong-Joo Lee, Yongseok Lee, Byeong-Gon Park, and Richard W. Pogge. Free-Floating Planets, the Einstein Desert, and 'Oumuamua. *Journal of The Korean Astronomical Society*, 55(5):173–194, October 2022.

[7] William DeRocco, Matthew T. Penny, Samson A. Johnson, and Peter McGill. Reconstructing the Free-floating Planet Mass Function with the Nancy Grace Roman Space Telescope, April 2025.

[8] Keming Zhang, Joshua S. Bloom, B. Scott Gaudi, Francois Lanusse, Casey Lam, and Jessica R. Lu. Real-Time Likelihood-Free Inference of Roman Binary Microlensing Events with Amortized Neural Posterior Estimation. *The Astronomical Journal*, 161(6):262, June 2021.

[9] Jean-Christophe Gagnon-Audet, Kartik Ahuja, Mohammad-Javad Darvishi-Bayazi, Pooneh Mousavi, Guillaume Dumas, and Irina Rish. WOODS: Benchmarks for Out-of-Distribution Generalization in Time Series, April 2023.

[10] Satya Narayan Shukla and Benjamin M. Marlin. A Survey on Principles, Models and Methods for Learning from Irregularly Sampled Time Series, January 2021.

[11] Nolan Koblischke and Jo Bovy. SpectraFM: Tuning into Stellar Foundation Models, November 2024.

[12] Gemma Zhang, Thomas Helfer, Alexander T. Gagliano, Siddharth Mishra-Sharma, and V. Ashley Villar. Maven: A Multimodal Foundation Model for Supernova Science, August 2024.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023.

[14] Hyoun-Woo Kim, Kyu-Ha Hwang, Andrew Gould, Jennifer C. Yee, Yoon-Hyun Ryu, Michael D. Albrow, Sun-Ju Chung, Cheongho Han, Youn Kil Jung, Chung-Uk Lee, In-Gu Shin, Yossi Shvartzvald, Weicheng Zang, Sang-Mok Cha, Dong-Jin Kim, Seung-Lee Kim, Dong-Joo Lee, Yongseok Lee, Byeong-Gon Park, and Richard W. Pogge. KMT-2019-BLG-2073: Fourth Free-Floating-Planet Candidate with $ϑ_\rm E < 10 \rmμas$. *The Astronomical Journal*, 162(1):15, July 2021.

[15] M. D. Albrow, K. Horne, D. M. Bramich, P. Fouqué, V. R. Miller, J.-P. Beaulieu, C. Coutures, J. Menzies, A. Williams, V. Batista, D. P. Bennett, S. Brillant, A. Cassan, S. Dieters, D. Dominis Prester, J. Donatowicz, J. Greenhill, N. Kains, S. R. Kane, D. Kubas, J.-B. Marquette, K. R. Pollard, K. C. Sahu, Y. Tsapras, J. Wambsganss, and M. Zub. Difference imaging photometry of blended gravitational microlensing events with a numerical kernel. *Monthly Notices of the Royal Astronomical Society*, 397(4):2099–2105, August 2009.

[16] V. Bozza, E. Bachelet, F. Bartolić, T. Heintz, A. Hoag, and M. Hundertmark. VBBinaryLensing: A public package for microlensing light curve computation. *Monthly Notices of the Royal Astronomical Society*, 479(4):5157–5167, October 2018.

[17] David S. Greenberg, Marcel Nonnenmacher, and Jakob H. Macke. Automatic Posterior Transformation for Likelihood-Free Inference, May 2019.

[18] Jan Boelts, Michael Deistler, Manuel Gloeckler, Álvaro Tejero-Cantero, Jan-Matthis Lueckmann, Guy Moss, Peter Steinbach, Thomas Moreau, Fabio Muratore, Julia Linhart, Conor Durkan, Julius Vetter, Benjamin Kurt Miller, Maternus Herold, Abolfazl Ziaeemehr, Matthijs Pals, Theo Gruner, Sebastian Bischoff, Nastya Krouglova, Richard Gao, Janne K. Lappalainen, Bálint Mucsányi, Felix Pei, Auguste Schulz, Zinovia Stefanidi, Pedro Rodrigues, Cornelius Schröder, Faried Abu Zaid, Jonas Beck, Jaivardhan Kapoor, David S. Greenberg, Pedro J. Gonçalves, and Jakob H. Macke. Sbi reloaded: A toolkit for simulation-based inference workflows, August 2025.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017.

[20] Pablo Lemos, Adam Coogan, Yashar Hezaveh, and Laurence Perreault-Levasseur. Sampling-Based Accuracy Testing of Posterior Estimators for General Inference, June 2023.

[21] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman. Emcee: The MCMC Hammer. *Publications of the Astronomical Society of the Pacific*, 125(925):306–312, March 2013.

[22] Naoki Koshimoto, Junichi Baba, and David P. Bennett. A Parametric Galactic Model toward the Galactic Bulge Based on Gaia and Microlensing Data. *The Astrophysical Journal*, 917(2):78, August 2021.

[23] Andreas Filipp, Yashar Hezaveh, and Laurence Perreault-Levasseur. Robustness of Neural Ratio and Posterior Estimators to Distributional Shifts for Population-Level Dark Matter Analysis in Strong Gravitational Lensing, November 2024.

[24] Michelle Kunimoto, William DeRocco, Nolan Smyth, and Steve Bryson. Searching for Free-Floating Planets with TESS: I. Discovery of a First Terrestrial-Mass Candidate, April 2024.

[25] Przemek Mroz. TESS Free-floating Planet Candidate Is Likely a Stellar Flare, May 2024.

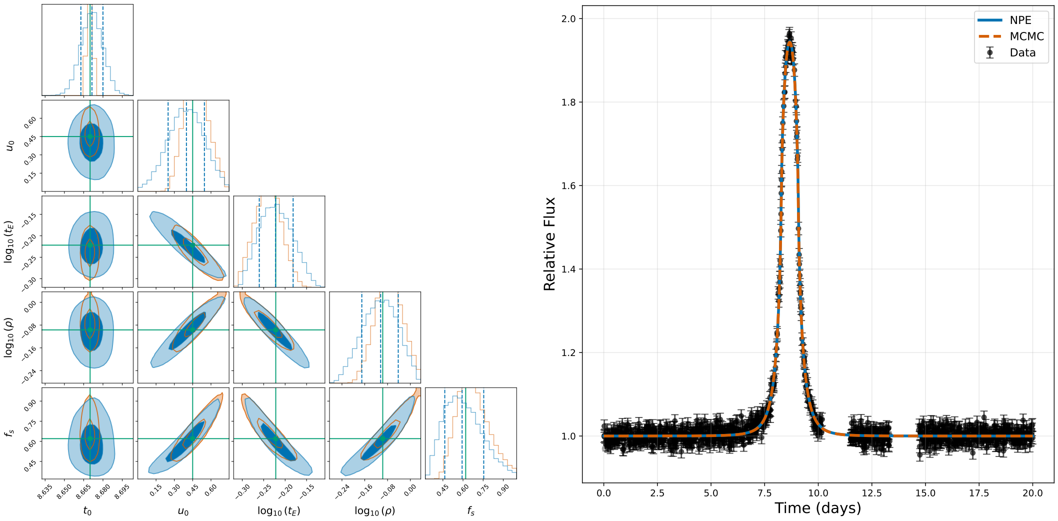
[26] Hongjing Yang, Weicheng Zang, Tianjun Gan, Renkun Kuang, Andrew Gould, and Shude Mao. How Rare are TESS Free-Floating Planets?, August 2024.

[27] Scott E. Perkins, Peter McGill, William A. Dawson, Ming-Feng Ho, Natasha S. Abrams, Simeon Bird, and Jessica R. Lu. Hints of an Anomalous Lens Population towards the Galactic Bulge, March 2025.

[28] Scott Ellis Perkins, Peter McGill, William Dawson, Natasha S. Abrams, Casey Y. Lam, Ming-Feng Ho, Jessica R. Lu, Simeon Bird, Kerianne Pruett, Nathan Golovich, and George Chapline. Disentangling the Black Hole Mass Spectrum with Photometric Microlensing Surveys. *The Astrophysical Journal*, 961(2):179, February 2024.

[29] William DeRocco, Evan Frangipane, Nick Hamer, Stefano Profumo, and Nolan Smyth. Rogue worlds meet the dark side: Revealing terrestrial-mass primordial black holes with the Nancy Grace Roman Space Telescope. *Physical Review D*, 109(2):023013, January 2024.

[30] Javier Viaña, Kyu-Ha Hwang, Zoë de Beurs, Jennifer C. Yee, Andrew Vanderburg, Michael D. Albrow, Sun-Ju Chung, Andrew Gould, Cheongho Han, Youn Kil Jung, Yoon-Hyun Ryu, In-Gu Shin, Yossi Shvartzvald, Hongjing Yang, Weicheng Zang, Sang-Mok Cha, Dong-Jin Kim, Seung-Lee Kim, Chung-Uk Lee, Dong-Joo Lee, Yongseok Lee, Byeong-Gon Park, and Richard W. Pogge. LensNet: Enhancing Real-time Microlensing Event Discovery with Recurrent Neural Networks in the Korea Microlensing Telescope Network. *The Astronomical Journal*, 169(3):159, March 2025.



(a) Posterior comparison. MCMC results are displayed in orange and NPE results in blue. The true parameter values are shown in green.

(b) Simulated light-curve and recovered models for the MCMC and NPE. The 16-84 percentile confidence intervals are shown, but are practically negligible as both models fit the data extremely well.

Figure 4: NPE and MCMC posteriors on the same simulated event.

A Benchmarks

We show the results of the NPE and a traditional MCMC for a simulated lightcurve in Figure 4. The MCMC uses 32 walkers, 5,000 burn-in steps, followed by 10,000 samples. We directly generate 15,000 samples from the NPE. Drawing 15,000 samples from the trained NPE takes just 0.08 seconds on a GPU and 0.82 seconds on a CPU. For comparison, running the MCMC sampler for 15,000 steps took 959 seconds on a CPU, resulting in an inference speedup factor of about 1.2×10^3 for a single light curve on the same hardware and a factor of more than 10^4 when GPU accelerated.

As shown in Figure 4a, the NPE posteriors are in excellent agreement with the MCMC results, with a slight broadening visible in some parameters. This is consequence of the amortized inference. The network has learned a approximation to ensure robust calibration across the entire range of possible light curves, but is well-calibrated across the entire prior space, as revealed by TARP diagnostic. The posterior estimate for a single event could quickly be refined using importance sampling or as seeding to a down-stream MCMC if necessary. The light-curves recovered by both the NPE and MCMC fit the data extremely well, as shown in Figure 4b. The extremely narrow bands corresponding to the 16-84 percentile light-curves also demonstrates the inherent degeneracies of microlensing recovered by the model.