Research
A SCIENCE PARTNER JOURNAL

RESEARCH ARTICLE

Toward Efficient and Accurate EMRI Parameter Estimation: A Machine Learning-Enhanced MCMC Framework

Bo Liang[1,2†], Chang Liu[1,2,3†], Hanlin Song[4], Zhenwei Lyu[5], Minghui Du[1*], Peng Xu[1,2,6*], Ziren Luo[1,2,7], Sensen He[8], Haohao Gu[8], Tianyu Zhao[1], Manjia Liang[1], Yuxiang Xu[1], Li-e Qiang[3], Mingming Sun[9], and Wei-Liang Qian[10]

[1]Center for Gravitational Wave Experiment, National Microgravity Laboratory, Institute of Mechanics, Chinese Academy of Sciences, Beijing 100190, China. [2]Taiji Laboratory for Gravitational Wave Universe (Beijing/Hangzhou), University of Chinese Academy of Sciences (UCAS), Beijing 100049, China. [3]National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China. [4]School of Physics, Peking University, Beijing 100871, China. [5]Leicester International Institute, Dalian University of Technology, Panjin 124221, China. [6]Lanzhou Center of Theoretical Physics, Lanzhou University, Lanzhou 730000, China. [7]Key Laboratory of Gravitational Wave Precision Measurement of Zhejiang Province, Hangzhou Institute for Advanced Study, UCAS, Hangzhou 310024, China. [8]Baidu Inc., Beijing 100085, China. [9]AGI Lab, Beijing Institute of Mathematical Sciences and Applications, Beijing, China. [10]Escola de Engenharia de Lorena, Universidade de São Paulo, 12602-810 Lorena, SP, Brazil.

*Address correspondence to: duminghui@imech.ac.cn (M.D.); xupeng@imech.ac.cn (P.X.)
†These authors contributed equally to this work.

Citation: Liang B, Liu C, Song H, Lyu Z, Du M, Xu P, Luo Z, He S, Gu H, Zhao T, et al. Toward Efficient and Accurate EMRI Parameter Estimation: A Machine Learning-Enhanced MCMC Framework. *Research* 2026;9:Article 1055. https://doi.org/10.34133/research.1055

Submitted 23 August 2025
Revised 2 December 2025
Accepted 4 December 2025
Published 8 January 2026

Copyright © 2026 Bo Liang et al. Exclusive licensee Science and Technology Review Publishing House. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution License (CC BY 4.0).

The detection of gravitational waves from extreme-mass-ratio inspirals (EMRIs) in space-based antennas like Taiji and Laser Interferometer Space Antenna promises deep insights into strong-field gravity and black hole physics. However, the complex, highly degenerate, and nonconvex likelihood landscapes characteristic of EMRI parameter spaces pose severe challenges for conventional Markov chain Monte Carlo (MCMC) methods. Under realistic instrumental noise and broad priors, these methods demand impractical computational costs but are prone to becoming trapped in local maxima, leading to biased and unreliable parameter estimates. To address these challenges, we introduce flow-matching MCMC (FM-MCMC), a novel Bayesian framework that integrates continuous normalizing flows (CNFs) with parallel tempering MCMC (PTMCMC). By generating high-likelihood regions via CNFs and refining them through PTMCMC, FM-MCMC enables robust exploration of the nontrivial parameter spaces, achieves orders-of-magnitude improvement in computational efficiency, and, more importantly, ensures statistically unbiased inference. By enabling real-time, unbiased parameter inference, FM-MCMC could unlock the full scientific potential of EMRI observations, and would serve as a scalable pipeline for precision gravitational-wave astronomy.

Introduction

Gravitational wave (GW) observation has gradually revolutionized our view of the Universe since the first detection of GW150914 by the LIGO-Virgo Collaboration [1,2]. Over the past decade, ground-based detectors such as LIGO, Virgo, and KAGRA have cataloged over 100 merger events of compact binaries, including binary black holes, binary neutron stars, and black hole–neutron star systems [2–4]. These observatories, limited mainly by terrestrial noises, operate in the 10 to 10^3 Hz frequency band and probe astrophysical systems with masses up to hundreds of solar masses [5–7]. However, the millihertz GW universe encompassing sources like extreme-mass-ratio inspirals (EMRIs), massive black hole binaries (MBHBs), and galactic binaries (GBs) remains largely unexplored. The planned space-based interferometers, such as Taiji, TianQin, and Laser Interferometer Space Antenna (LISA) [8–10], promise to open a new window into such low-frequency GW universe, in which EMRIs are among the most promising sources for probing strong-field gravity and black hole physics [11,12].

EMRIs, in which a stellar-mass compact object inspirals into a supermassive black hole through more than 10^5 orbital cycles, encode a wealth of information about the spacetime geometry and accretion dynamics in the vicinity of the supermassive black hole [13]. These systems are expected to retain substantial orbital eccentricities in the final stages before plunge [11]. The

combination of high eccentricity and extreme mass ratio produces GW signals with a rich harmonic structure and tens of thousands of orbital cycles within the detector's sensitive frequency band. Such unique waveform characteristics make EMRIs powerful probes of the nature of gravity (including, for example, the direct test of the "no-hair" theorem) [14–18] and of the astrophysical environments (including dark matter distributions) surrounding the supermassive black holes [19–23], while unlocking EMRI's full scientific objectives and implications will require precise measurements of key physical parameters, such as the central black hole's mass M and spin a, the mass of the stellar-mass compact object μ, and the orbital eccentricity e.

However, the application of traditional Bayesian methods like Markov chain Monte Carlo (MCMC) to EMRI signal parameter estimation faces significant challenges, primarily due to prohibitive computational costs and inadequate exploration of the parameter space [24,25]. MCMC relies on intensive waveform evaluations to explore the parameter space, as the Bayesian posterior occupies only a minuscule region within the high-dimensional space [26], necessitating an enormous number of iterations to locate the global optimum corresponding to the true parameters. Such challenges are further exacerbated by the long duration of EMRI signals (lasting months to years [24,25]), which significantly increases the computational cost of each likelihood evaluation. Another fundamental challenge in EMRI searches stems from the highly nonconvex structure of the likelihood surface, which is riddled with numerous local maxima [27–29]. These local peaks correspond to nonlocal degeneracies in the parameter space, that distinct regions could yield waveforms nearly indistinguishable from the true signal. In other words, widely separated parameter regions can produce highly similar gravitational waveforms, leading to multiple strong but spurious likelihood peaks. Crucially, the global maximum associated with the true parameters often lies far from these suboptimal solutions in parameter space, making it extremely difficult to identify without exhaustive exploration.

Traditional methods like grid searches or MCMC are highly susceptible to becoming trapped in these local maxima during the parameter estimation process [26].

In data analysis scenarios, where instrumental noise is present (as illustrated in the bottom panel of Fig. 1), the performance of traditional inference methods is further compromised [28,30–32]. Instrumental noise introduces additional complexity to the likelihood surface, substantially increasing the number of local maxima. These maxima are not only more numerous but also sharper and more densely packed, making the parameter space highly rugged [32]. As a result, inference algorithms such as MCMC or hybrid methods are prone to becoming trapped in these spurious local maxima, particularly when degeneracies in EMRI signals are present [28]. This dramatically increases the difficulty of converging to the true parameters.

Key scientific objectives, such as testing strong-field gravity with EMRI systems or inferring black hole physical parameters, remain beyond reach with existing analysis techniques [17,33–37]. Biased parameter estimation introduces significant systematic errors and may lead to incorrect astrophysical interpretations, e.g., misestimation of spin or orbital dynamics [33,34,36,38]. Therefore, the development of novel algorithms capable of efficiently exploring high-dimensional, noise-corrupted parameter spaces while circumventing local optima is essential for fully unlocking the scientific opportunities offered by EMRI GW observations [39–41].

To address these challenges, we propose flow-matching MCMC (FM-MCMC): a hybrid framework that integrates continuous normalizing flows (CNFs) [42,43] with parallel tempering MCMC (PTMCMC) [44,45]. The FM-MCMC framework begins with flow matching, which rapidly and coarsely explores the parameter space through gradient-based trajectory learning. This approach effectively mitigates initialization sensitivity by identifying high-likelihood regions in a computationally

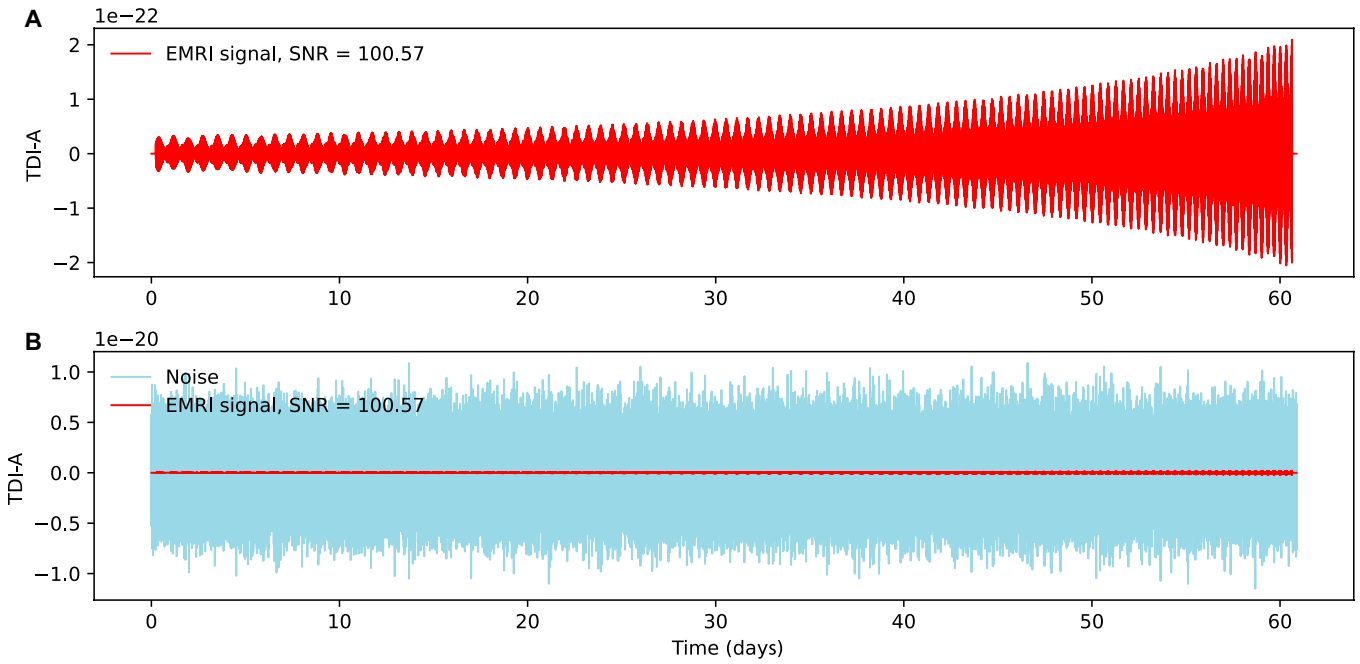


Fig. 1. (A) The EMRI signal under instrumental noise conditions. The EMRI waveform (red curve) is generated with the augmented analytic kludge model, showing the time-delay interferometry channel A observable over a 60-day observation period. (B) The same EMRI signal combined with simulated Taiji instrumental noise (blue shading), illustrating the challenge of parameter recovery in realistic noise environments.

efficient manner. Subsequently, PTMCMC is initialized using samples from these high-likelihood regions, enabling fine-grained exploration and leveraging its robust capability for local precision posterior estimation.

Our framework is illustrated in Fig. 2, with methodological details provided in the "Accelerating MCMC sampling with CNFs" section.

For EMRI signals with signal-to-noise ratios (SNRs) exceeding 60—which defines typical "bright" sources prioritized in space-based detector observations—our method achieves reliable parameter recovery (especially for intrinsic parameters) with true values falling within the 1σ posterior credible intervals even under realistic instrumental noise conditions. In contrast, under identical noise conditions, traditional Bayesian methods such as standard MCMC sampling become trapped in local maxima of the likelihood function, resulting in substantial biases across all parameter estimates. This methodological advance enables efficient and unbiased parameter estimation for EMRI systems under near-realistic observational conditions, addressing a long-standing bottleneck for conventional Bayesian inference approaches in EMRI analysis. With this breakthrough, our approach would pave the way for unlocking the scientific landscape revealed by EMRI observations.

The paper is structured as follows. Results reports the results based on numerical simulations and compares the performance of our method with that of standard MCMC. In Discussion, we conclude with a discussion of implications and future research directions. Methods details the methodologies of data generation and preprocessing, which support model training and validation. We then present the machine learning framework designed for Bayesian posterior estimation of EMRI signals.

Results

We present parameter estimation results for injected EMRI signals, with the source parameters drawn from the prior distributions detailed in Table 1. The waveform model employed in this study is based on EMRIs around a Kerr black hole, implemented within the FEW framework [26,46]. Detector response is modeled for the Taiji mission (as a representative of LISA-like missions), and the science data streams incorporate Gaussian stationary noise generated according to Taiji's noise budget.

In this study, we focused on a subset of the parameters, especially the mass M and spin a of the central black hole, as well as the mass of the stellar-mass compact object μ and the orbital eccentricity e, which are crucial for understanding the fundamental properties of the system and hold significant implications for strong-field gravity and astrophysics [47]. We have simplified the analysis to an 8-dimensional parameter space to demonstrate the capabilities of our method. The full 14-dimensional parameter space, which includes orbital semi-latus rectum p_0 and luminosity distance d, will be addressed in future work. As a representative example, the injected EMRI signal analyzed in this section has the following parameters: primary mass $M = 1.0593 \times 10^6 M_\odot$, secondary mass $\mu = 8.6556 M_\odot$, dimensionless spin $a = 0.1931$, initial semi-latus rectum $p_0 = 10.0$, initial eccentricity $e_0 = 0.4870$, luminosity distance = 2.0 Gpc, sky and spin orientations $(\theta_S, \phi_S, \theta_K, \phi_K) = (1.672, 1.0372, 0.4040, 3.6169)$ radians, and the initial orbital phase is set to zero. The waveform is sampled at 25-s intervals over a 2-month observation period. The resulting SNR is 87.9.

To benchmark against our machine learning approach, we perform traditional Bayesian inference using Eryn [44], an advanced MCMC sampler built upon emcee [45].

Eryn has emerged as a popular tool for space-based GW data analysis, particularly in multisource and high-dimensional inference problems (e.g., Ref. [48]). We apply both Eryn and our method to the same injected EMRI signal, enabling a direct comparison of their sampling efficiencies, convergence properties, and parameter estimation accuracies. For such comparison, each method employs two distinct initialization strategies:

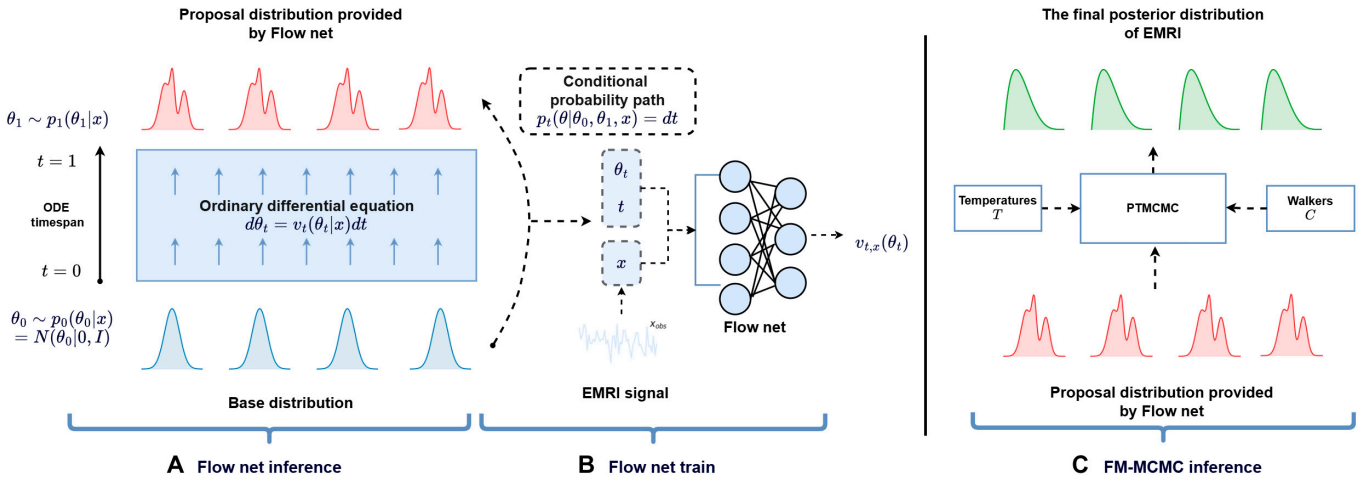


Fig. 2. (A) Flow net inference stage. It demonstrates the transformation from a baseline Gaussian distribution $\theta_0 \sim p_{\theta_0}(\theta_0|x) = \mathcal{N}(\theta_0|0, I)$ (blue) to the time-evolving EMRI posterior distribution $\theta_t \sim p_{\theta_t}(\theta_t|x)$ (pink) through an ordinary differential equation (ODE). Here, $\theta_0 \sim \mathcal{N}(0, I)$ denotes the baseline distribution, x represents preprocessed EMRI data (details explained in Methods), and θ_t models the continuously interpolated transformation state at normalized time $t \in [0,1]$. The conditional probability path of the entire ODE is modeled by Flow net. The theoretical foundation and implementation details are comprehensively described in the "Continuous normalizing flows" section. (B) Flow net training stage. It illustrates the Flow net training process, where the Flow net learns the velocity field v_θ by minimizing the flow matching [42] loss function. (C) FM-MCMC inference stage. During FM-MCMC inference initialization, PTMCMC determines starting distribution points through its temperature parameter T and walks C. This initialization distribution is then generated by Flow net's proposal distribution. Final posterior sampling is executed exclusively via PTMCMC's likelihood evaluations, completing the Bayesian inference cycle for EMRI parameter estimation.

Table 1. Prior distributions used in this work. For each parameter, the table lists its lower and upper bounds, assuming a uniform distribution between them.

Parameter	Description	Prior lower bound	Prior upper bound	Units
M	Central black hole mass	9×10^5	1.1×10^6	$M_\odot$
μ	Secondary object mass	50	100	$M_\odot$
a	Central black hole spin	0.1	0.9	–
e_0	Orbital eccentricity	0.1	0.6	–
θ_S	Sky location polar angle	0 rad	π rad	rad
ϕ_S	Sky location azimuthal angle	0 rad	2π rad	rad
θ_K	Initial black hole spin polar angle	0 rad	π rad	rad
ϕ_K	Initial black hole spin azimuthal angle	0 rad	2π rad	rad

1. Informative initialization setting: Both intrinsic parameters (M, μ, a, e_0) and extrinsic parameters ($\theta_S, \phi_S, \theta_K, \phi_K$) are initialized within a narrow neighborhood of the true values, with absolute deviations less than 10^{-7} times the corresponding true parameter values. This configuration provides favorable initial conditions to assess convergence under idealized, truth-proximal starting points.
2. Practical initialization setting: The intrinsic parameters (M, μ, a, e_0) are initialized by randomly drawing samples from their respective broad prior distributions, as defined in Table 1. In contrast, the extrinsic parameters (θ_S, ϕ_S, θ_K, ϕ_K) are initialized identically to setting 1, within a tight range ($<10^{-7}\times$ true values) around the true values. This setup evaluates the sampler's ability to recover the correct posterior mode when intrinsic parameters are initialized under realistic (i.e., uninformed) priors.

Both settings employed identical parallel tempering configurations with 20 walkers and 20 temperature ladder levels. For FM-MCMC, 20,000 iterations were used, while PT-MCMC was run with 50,000 iterations in order to provide sufficient sampling. Despite this increased iteration count, PT-MCMC still struggled to achieve convergence, highlighting the efficiency advantage of FM-MCMC. For FM-MCMC, the priors are fixed and generated by FMPE, which also provides initial sampling points. This setup reflects the different strategies of the two approaches: PT-MCMC depends on broad predefined priors, while FM-MCMC leverages informed priors from FMPE to improve efficiency.

Result of parallel tempering Markov chain Monte Carlo

Running MCMC under setting 1 where all parameters are initialized close to the true values, we successfully recover the posterior distributions of the source parameters (see Fig. 3). This confirms that MCMC can accurately reconstruct EMRI parameters when starting from favorable initial conditions near the global maximum of the likelihood. However, in realistic data analysis, the true parameters of an EMRI are unknown, and such informed initialization is not feasible. To assess the practical performance of the MCMC sampler, we perform inference under setting 2, sampling intrinsic parameters drawn from broad priors.

The resulting posteriors exhibit significant biases across all intrinsic parameters, with the true values lying outside the 3σ credible regions (see Fig. 4). This behavior is due to the inherent multimodality and strong degeneracies in the EMRI data analysis, which cause MCMC chains to become trapped in local maxima. As a result, the sampler oscillates around suboptimal solutions without achieving convergence to the global optimum. Therefore, under current settings, using PTMCMC alone is insufficient for robust scientific analysis of EMRI data, as it lacks the efficiency and global exploration capability required for such complex parameter spaces. To the best of our knowledge, no existing study has demonstrated that MCMC methods can reliably recover the correct posterior distribution for EMRIs when initialized from broad and realistic priors. This limitation highlights the severe challenges posed by the high-dimensional, multimodal, and degenerate structure of the EMRI likelihood surface.

Result of CNFs

To evaluate the performance of our CNF model (named flow matching posterior estimation, FMPE for short [42,43]), we first tested it on 1,000 EMRI signals with source parameters randomly drawn from the priors defined in Table 1. The resulting posterior samples were then used to generate the P–P (probability–probability) plot, as shown in Fig. 5.

The P–P plot shows that the cumulative distribution functions (CDFs) of the posterior percentile values for the intrinsic parameters closely align with the diagonal line. This validates the statistical reliability and unbiasedness of our CNF-based posterior estimation across the full extent of the prior volume.

We further compared the performance of FMPE and MCMC under the more realistic setting 2, in which intrinsic parameters are initialized from broad priors. As shown in Fig. 6, FMPE successfully recovered the posterior distributions of intrinsic parameters (e.g., M and a) with high accuracy, while clearly identifying the true values (red lines). In stark contrast, MCMC under the same broad initialization (Fig. 4) failed to converge to the global maximum of the likelihood due to parameter degeneracies and local maxima.

Notably, FMPE achieves accurate posterior inference in minutes per inference, an improvement of several orders of magnitude over MCMC, which requires days to reach thermalization. This remarkable computational efficiency demonstrates

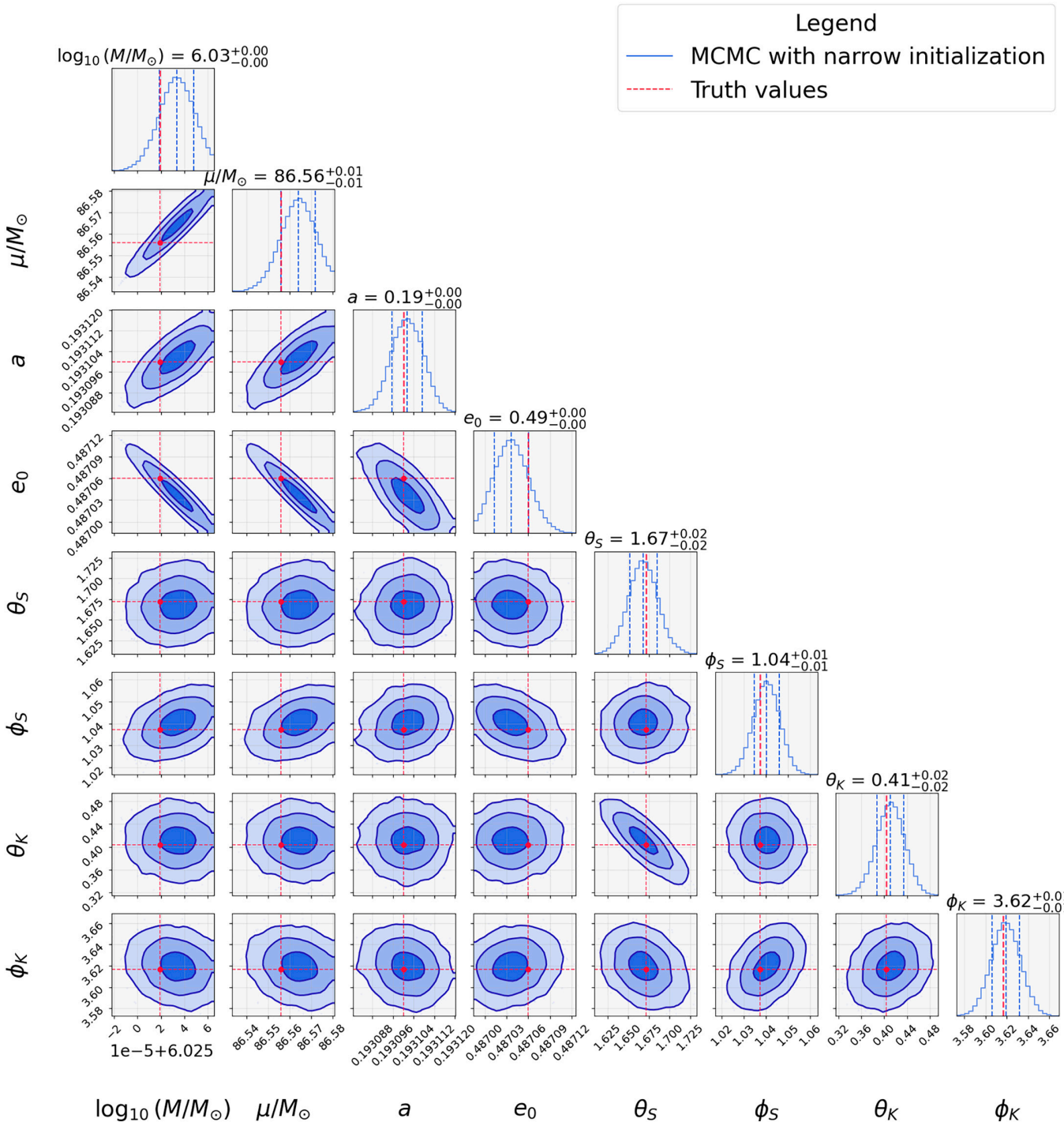


Fig. 3. MCMC parameter recovery with narrow initialization. Posterior distributions from MCMC chains initialized near true EMRI parameters (red lines) using setting 1. In the figure, the blue line represents the posterior probability distribution of the EMRI parameters. The 3 contours represent the 1σ, 2σ, and 3σ confidence regions of the posterior distribution. This validates the MCMC reliability with proper initialization, demonstrating that the chains converge accurately when started within a reasonable proximity to the true parameters.

that our pure machine learning-based approach enables rapid and unbiased parameter estimation for EMRIs, even in the challenging noise-dominated situations.

Moreover, the inferred posterior distributions significantly narrow the initial Bayesian prior ranges. This prior refinement can inform and improve subsequent analyses for both machine learning models and traditional MCMC pipelines. In particular, as shown in the next subsection, the high-fidelity posteriors generated by our FMPE model provide an ideal starting distribution for MCMC samplers. By initializing chains within the high-probability region of parameter space, FMPE enables a "hot start" for MCMC sampling that mitigates the risk of trapping in local maxima, addressing the major bottleneck in traditional EMRI parameter estimation.

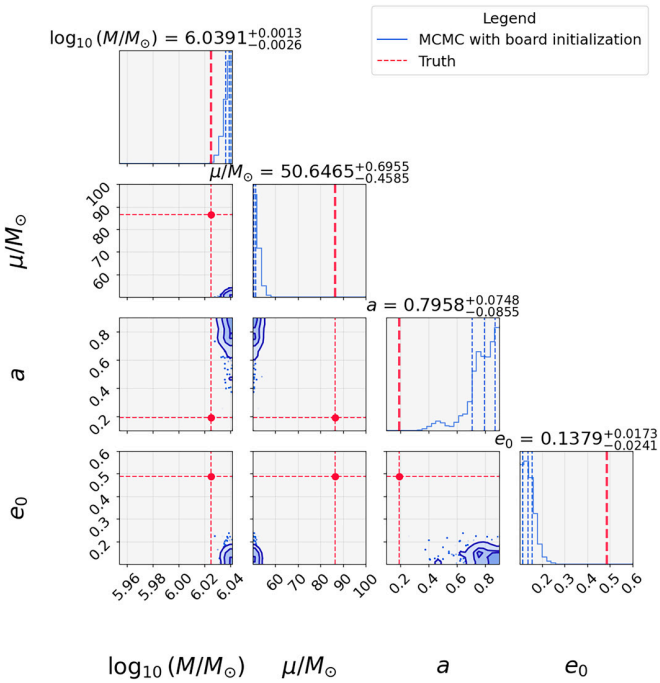


Fig. 4. MCMC parameter estimation with broad initialization. Posterior distributions from MCMC chains initialized with wide priors. In the figure, the blue line represents the posterior probability distribution of the EMRI parameters. The dashed red lines indicate the true values of the parameters. The 3 contours represent the 1σ, 2σ, and 3σ confidence regions of the posterior distribution.

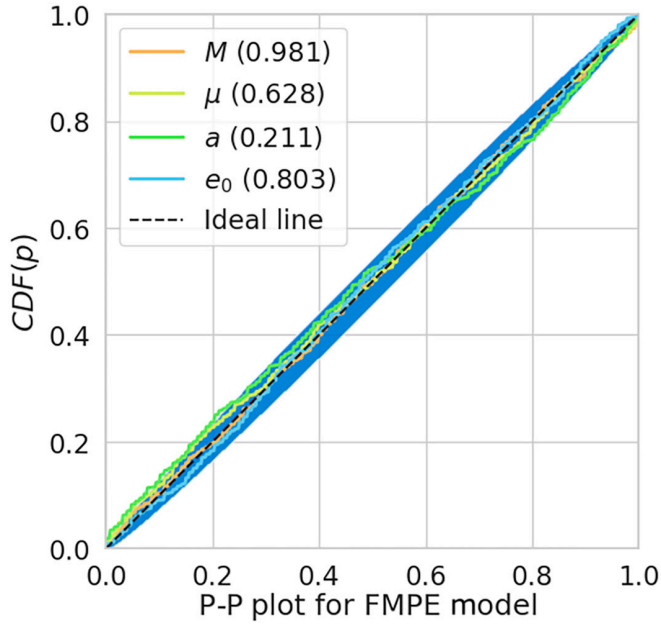


Fig. 5. P–P plot for the FMPE model, based on 500 simulated EMRI signals. The dashed line represents the ideal case, while the colored lines represent the empirical CDFs of the parameters. The shaded blue region indicates the 95% confidence bands. All parameters lie well within the confidence region, demonstrating that the CNF captures posterior uncertainty with high fidelity.

Result of FM-MCMC

Building on the flow-based acceleration techniques, we propose in this work the pioneering FM-MCMC framework: a novel framework that seamlessly fuses the posterior sampling dynamics of CNF with the chain and temperature dynamics of PTMCMC.

Figure 7 demonstrates the parameter recovery of FM-MCMC for the same EMRI signal analyzed in the previous subsection under the broad priors listed in Table 1. The full inference was completed within 48 h on a single NVIDIA RTX 4090 GPU, demonstrating the feasibility of FM-MCMC for realistic EMRI data analysis tasks. Crucially, unlike the case for stand-alone MCMC (Fig. 4), FM-MCMC converges unambiguously to the true values (red lines in Fig. 7) across all intrinsic parameters. Moreover, as quantified in Figs. 6 and 7, FM-MCMC also achieves significant improvements over FMPE. This enhanced performance arises from the synergistic integration of the rapid exploration of machine learning with the refined sampling of MCMC, combining the speed of machine learning with the statistical reliability of Bayesian sampling. We note that while normalizing flows are sometimes reported to under-represent small modes, in our framework, the CNF does not serve as the final posterior estimator. Instead, CNF proposals are used to accelerate initialization, and the subsequent PTMCMC stage ensures more precise Bayesian inference. As shown in Fig. 7 and in the additional walk results in Fig. S2, FM-MCMC successfully identifies and explores multiple modes in the likelihood surface, including subdominant ones. This confirms that the framework does not collapse to a single dominant mode but robustly preserves the multimodal structure of the EMRI parameter space.

Clear evidence presented in Table 2 offers direct quantitative comparisons between the injected signal parameters and those recovered by FM-MCMC and stand-alone MCMC. This rigorously confirms the unprecedented precision of the FM-MCMC framework, which, to the best of our knowledge, marks the first

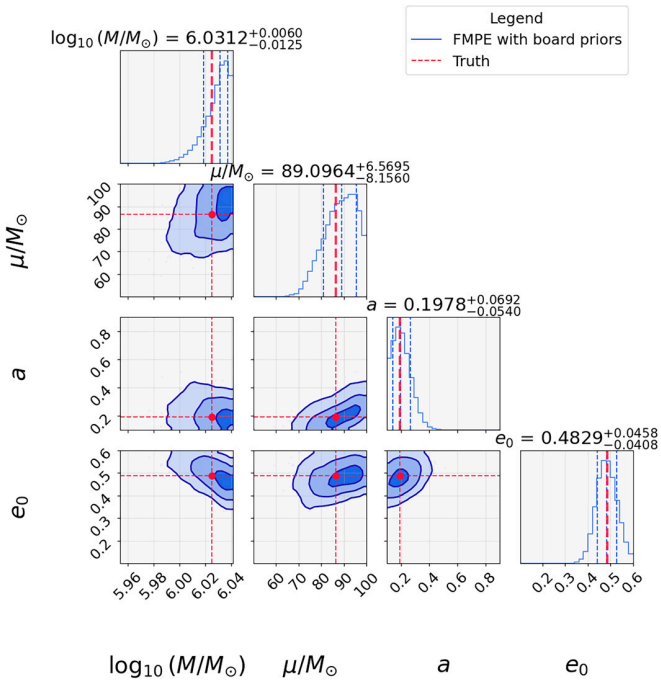


Fig. 6. EMRI parameter estimation using FMPE under broad priors. Posterior distributions from FMPE initialized with wide priors. The blue shading represents the posterior probability distribution of EMRI parameters. Red dashed lines indicate injected true parameter values. The 3 contours represent the 1σ, 2σ, and 3σ confidence regions of the posterior distribution.

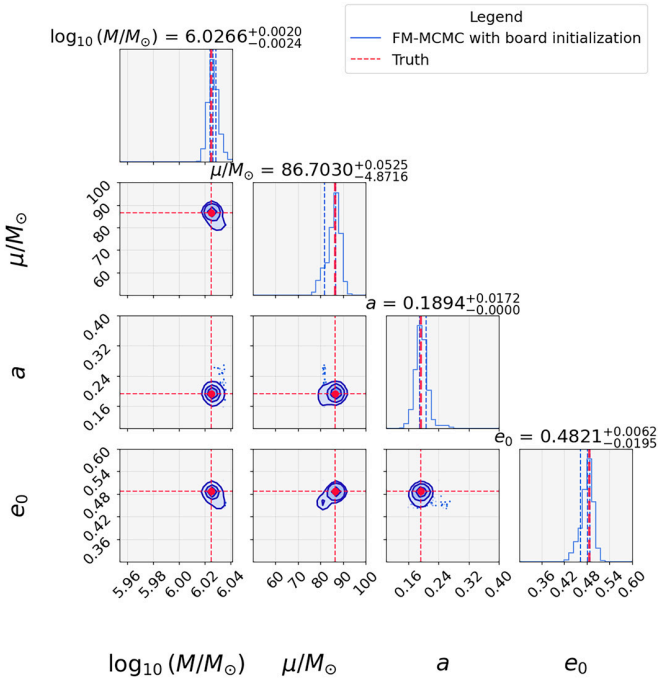


Fig. 7. The posterior distributions from the FM-MCMC framework initialized with wide priors. The blue contours show the posterior probability distribution of the EMRI parameters, while the red dashed lines indicate the true parameter values. The contours represent the 1σ, 2σ, and 3σ confidence regions of the posterior distribution, highlighting the presence of multiple modes in the likelihood surface. The walkers shown in the figure correspond to the 0th walker, which has the largest negative log-likelihood among the 20 walkers.

Table 2. The table compares the parameters injected with those recovered by FM-MCMC and Eryn. The recovered values correspond to the median estimates, and the 1σ confidence intervals are shown for each parameter.

Parameter	Injected value	FM-MCMC	Eryn
$\log_{10}M[M_\odot]$	6.0250	$6.0266^{+0.0020}_{-0.0024}$	$6.0391^{+0.0013}_{-0.0026}$
$\mu[M_\odot]$	86.5560	$86.7030^{+0.0525}_{-4.8716}$	$50.6465^{+0.6955}_{-0.4585}$
a	0.19310	$0.1894^{+0.0172}_{-0.0000}$	$0.7958^{+0.0748}_{-0.0855}$
e_0	0.48706	$0.4821^{+0.0062}_{-0.0195}$	$0.1379^{+0.0173}_{-0.0241}$

successful estimation of the intrinsic parameters for Kerr EMRIs under realistic noise conditions of LISA-like missions with fully unrestricted priors. This advancement finally brings within reach a deep understanding of the astrophysical formation channels [49,50], unprecedented precision tests of GR including direct validation of the "no hair" theorem [51,52], and a novel pathway for probing potential dark matter signatures around massive black holes [49,53,54] with EMRI observations.

Discussion

In this work, we have proposed and demonstrated the FM-MCMC framework for EMRI parameter estimation. Compared with traditional methods, FM-MCMC significantly improves computational efficiency while maintaining high accuracy, thereby making EMRI parameter estimation more tractable. Our results show that this approach can effectively overcome the difficulties associated with highly multimodal likelihood surfaces and efficiently recover the intrinsic parameters of EMRIs. While our current focus has been on demonstrating the feasibility and efficiency of FM-MCMC, realistic EMRI data analysis will face further challenges. In addition to instrumental noise, one must also consider uncertainties in the noise PSD, astrophysical foregrounds such as galactic white dwarf binaries, and GW backgrounds from other astrophysical sources. Developing methods that remain effective under such complex noise and confusion conditions will be a key direction of future research.

We also note that our current framework, similar to standard MCMC methods, assumes that the signal parameters lie within the predefined prior distribution used to train the CNF model. Inference cannot be guaranteed if the data fall completely outside this defined parameter range, and the development of robust out-of-distribution handling, potentially using methods such as transfer learning or fine-tuning to adapt the model to new data regimes, remains an important direction for future work. So far, the so-called "global fitting" problem—the joint analysis of all signals and noises—remains an active research area. Several prototype designs for global fit pipeline (e.g., Refs. [39, 48, 55]) have been proposed, which generally adopt a modular structure. Namely, the overall workflow is divided into modules, such as the characterization of noise, foreground and background, as well as the parameter estimation of MBHBs, GBs, and EMRIs, with exchange of information and residual data between them. Among these target signals, MBHB mergers are high-SNR (up to $\mathcal{O}(10^3)$ to $\mathcal{O}(10^4)$) transient signals, while GBs are persistent signals highly localized in the frequency domain. It is therefore natural to analyze them in the pipeline's early stages, and noise characterization modules should also be invoked for the construction of likelihoods. Besides, verification sources (such as verification GBs) and auxiliary data of the spacecrafts (such as the readouts of gravitational reference sensors, fluctuations of temperature and residual magnetic fields near core payloads, and platform attitude jitters) can also provide supplementary information for calibrating the noise profile.

Consequently, it is reasonable to focus here on the analysis of EMRI, under the condition that other bright signals have been extracted and the spectra of noise/foreground/background are known in the prior stages. Crucially, the key challenge for EMRI parameter estimation, and the core contribution of this work, lies in overcoming local optima and computational complexity to enable inference across relatively wide priors.

We show that compared with traditional methods, FM-MCMC can efficiently achieve global maxima under highly multimodal likelihood surfaces, making EMRI parameter estimation more tractable. The proposed algorithm is designed to ultimately function as the "EMRI parameter estimation module" in the full global fitting pipeline. In future work, we will also explore the potential interaction of EMRI signals with instrumental noises and GW foreground/background, upon integration into the full global fitting pipeline. For example, the nonstationarity in instrumental noise and astrophysical foreground can be handled through methods such as whitening in the time-frequency domain [41,56], an approach fully compatible with our framework.

This work establishes FM-MCMC, the first machine learning-enhanced Bayesian framework integrating CNFs with PTMCMC, which resolves the long-standing challenge in EMRI data analysis, that of global convergence in high-dimensional, multimodal parameter spaces. Conventional MCMC methods are prone to become trapped in local likelihood maxima, leading to systematic biases in the recovery of intrinsic EMRI parameters under realistic instrumental noise conditions and initial priors. In contrast, given flow matching to learn and deploy globally informed proposal distributions, that by dynamically generating high-likelihood regions via CNFs and refining them through PTMCMC, FM-MCMC enables robust exploration of degenerate, high-dimensional parameter spaces, while achieving order-of-magnitude improvements in computational efficiency and, more importantly, ensuring statistically reliable inference. For EMRI analysis, FM-MCMC alleviates parameter-estimation degeneracies arising from both instrumental noise and the complex harmonic structure of AAK waveforms with higher modes. Our results demonstrate that FM-MCMC can successfully recover EMRI parameters even under broad prior assumptions, highlighting its robustness compared to traditional methods. Looking forward, the extension of this framework to realistic data conditions, including uncertain noise PSDs, astrophysical foregrounds, and potential GW backgrounds, will be an essential step. Ultimately, such developments will make EMRI parameter estimation a feasible and powerful tool for unlocking the rich scientific opportunities enabled by future space-based GW observations.

Last but not least, FM-MCMC enables real-time analysis of EMRI signals, reducing inference times from days to hours on a single GPU, and establishing a scalable pipeline for the future planned space-based antennas like Taiji, TianQin, and LISA. Future work will extend this hybrid paradigm to multisource inferences and continue to advance GW astronomy into the era of intelligent data exploitation.

Methods

The construction of the datasets for EMRI analysis in this work rests on 4 key components: (a) the numerical synthesis of EMRI waveforms, (b) the specification of the detector orbit configuration, (c) the application of time-delay interferometry (TDI), and (d) the preprocessing of data for model training. As illustrated in Fig. 8, this methodological framework sequentially integrates these components to ensure physical fidelity and computational tractability. The following subsections provide detailed descriptions of each operational stage within this pipeline.

Extreme mass ratio inspiral waveforms

The accurate waveform modeling for EMRI systems constitutes a critical challenge, requiring high-fidelity implementations of gravitational self-force (GSF) calculations and black hole perturbation theory to achieve the sub-radian phase precision for meaningful scientific inference [57,58]. Perturbative expansions of the binary metric have been applied at the first order for solving general orbits around Kerr black holes [59], with significant progress also made in second-order calculations [60,61]. However, the computational efficiency of numerical GSF calculations remains insufficient to meet the demands of data analysis. Recent developments in effective-one-body models [62,63] synthesize GSF data into resumed potentials and try to balance between physical self-consistency and computational efficiency. On the other hand, rapid generation of approximate kludge models has been developed [64–67]. Early analytic kludge (AK) models [64], while sacrificing some accuracy, offer significantly improved computational efficiency. Numerical kludge (NK) models [65,68], on the other hand, achieve high-fidelity EMRI waveforms by fitting the perturbative calculations. The augmented analytic kludge (AAK) models [26,66,67] incorporate the NK model, enabling the generation of high-precision EMRI waveforms without significantly increasing computational costs. To date, the most advanced computational framework FastEMRIWaveforms (FEW) [26,46] can rapidly compute EMRI waveforms in the time domain. For data preparation, we employ the AAK model implemented in FEW to generate the EMRI waveforms. The model's sub-radian phase accuracy satisfies the precision criteria required for reliable parameter inference in this study.

Detector orbit configuration

Both the Taiji and LISA antennas consist of 3 spacecrafts (S/Cs) forming near-equilateral triangular constellations in heliocentric orbits. By continuously monitoring variations in the optical path lengths between S/Cs using laser interferometers, these detectors can measure the imprints induced by incident GWs. In this paper, we model the single-arm GW response in terms of laser frequency modulations as defined in [69] and [70], which explicitly depends on the detector's orbit configuration. As mentioned, we take Taiji as the representative of LISA-like missions, and employ an equal-arm analytic model to

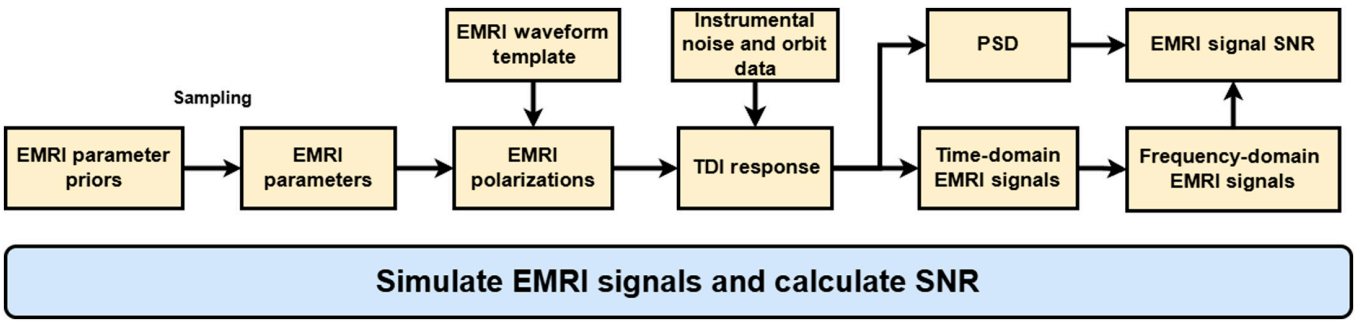


Fig. 8. The framework for EMRI signal generation and preprocessing pipeline. This framework begins with parameter priors and waveform generation, followed by orbital data and instrumental noise input, TDI response calculation, noise spectral analysis, and time-to-frequency domain transformation through preprocessing steps.

describe its orbit, where the coordinates of S/C$_i$ ($i \in \{1,2,3\}$) in the solar system barycentric frame read [71,72]:

$$\boldsymbol{R}_i(t) = \left\{ R\cos\beta + \frac{L}{2\sqrt{3}}\left[\frac{1}{2}\sin 2\beta \sin\gamma_i - \left(1+\sin^2\beta\right)\cos\gamma_i\right], \right.$$
$$R\sin\beta + \frac{L}{2\sqrt{3}}\left[\frac{1}{2}\sin\beta\cos\gamma_i - \left(1+\cos^2\beta\right)\sin\gamma_i\right], \quad (1)$$
$$\left. -\frac{L}{2}\cos\left[\beta-\gamma_i\right] \right\},$$

where

$$\beta(t) \equiv \frac{2\pi}{T}t + \beta_0, \quad \gamma_i \equiv \frac{2\pi(i-1)}{3} + \gamma_0. \quad (2)$$

Related parameters are specified as follows: the nominal arm length of Taiji is $L = 3\times10^9$ m, the orbital radius and period of Taiji's guiding center are $R = 1$ AU and $T = 1$ year, respectively, and the initial phase angles are set to $\beta_0 = \gamma_0 = 0$, without loss of generality. Taiji has a nominal arm length of 3 million km (compared to LISA's 2.5 million km) and a readout noise budget of 8 pm/$\sqrt{\text{Hz}}$ (versus LISA's 15 pm/$\sqrt{\text{Hz}}$). Thus, Taiji is expected to have slightly better sensitivity to GW signals (e.g., approximately 1.3 times better at 1 mHz); see Ref. [73] for the overall design of Taiji, and Fig. 1 of Ref. [74] for the comparison among Taiji, LISA, and TianQin's sensitivities. As a result, Taiji may detect a slightly larger number of observable EMRIs, which, in turn, would lead to greater demand for the analysis of these signals. However, due to the current significant uncertainties in EMRI population models, it is not yet feasible to provide a quantitative comparison.

Time-delay interferometry

For space-based GW detection, the inequality of arm and the instability of lasers causes the laser frequency to be orders of magnitudes larger than the target GW signals [75]. To achieve the designed sensitivity, TDI is proposed effectively to suppress laser frequency noise in the data processing stage [76,77]. The basic principle of TDI is to apply appropriate time delays to the single-arm measurements and combine them to synthesize an effective equal-arm interferometer. First-generation TDI combinations perform well for static unequal-arm configurations, while second-generation TDI further extends this capability under situations with time-varying arm lengths and, hence, is more suitable for the realistic Taiji and LISA orbits [78]. For example, the second-generation Michelson TDI combination $X(t)$ is defined as [79]:

$$\begin{aligned} X(t) =& y_{1'} + y_{3,2'} + y_{1,22'} + y_{2',322'} \\ &+ y_{1,3'322'} + y_{2',33'322'} + y_{1',3'33'322'} \\ &+ y_{3,2'3'33'322'} - y_1 - y_{2',3} - y_{1',3'3} \\ &- y_{3,2'3'3} - y_{1',22'3'3} - y_{3,2'22'3'3} \\ &- y_{1,22'22'3'3} - y_{2',322'22'3'3}. \end{aligned} \quad (3)$$

where y denotes the single-arm measurement, and the digits following the comma represent time-delay operations applied according to the corresponding arm lengths (see Ref. [79] for details). The Y and Z channels are obtained by cyclically permuting the indices according to the rule $1 \to 2, 2 \to 3$, and $3 \to 1$. We further derive the widely adopted quasi-noise-orthogonal $\{A, E, T\}$ channels from $\{X, Y, Z\}$ as follows:

$$\begin{aligned} A &= \frac{1}{\sqrt{2}}(Z - X), \\ E &= \frac{1}{\sqrt{6}}(X - 2Y + Z), \\ T &= \frac{1}{\sqrt{3}}(X + Y + Z). \end{aligned} \quad (4)$$

Since data analysis is performed in the form of TDI variables, it would be beneficial to provide an explanation on the impacts of detector orbit and TDI on the EMRI waveforms. Firstly, the rotation of the detector constellation causes time-varying amplitude modulation. Secondly, the translation motion of the detector induces a Doppler frequency shift. Thirdly, TDI combination involves the summation of waveforms at different times. For the impact of second-generation TDI, an intuitive illustration to the effect of TDI can be given under the equal-arm approximation. Taking the second-generation Michelson-X channel as an example, the Fourier transform of TDI result can be expressed as

$$\tilde{X}(f) \approx -4e^{-i3u}\sin 2u \sin u \tilde{X}_0(f), \quad (5)$$

with $u \equiv 2\pi f L/c$, $\tilde{X}_0(f)$ being the response of a Michelson interferometer, and the preceding factor represents the additional correction introduced by the TDI combination (relative to $\tilde{X}_0(f)$; see also the result for the first generation in Ref. [71]). The GW responses in the A and E channels are equivalent to two Michelson interferometers oriented at 45° relative to each other [80], whereas the T channel is insensitive to signals and hence referred to as the "noise" channel or "null" channel.

In this study, to leverage their complementary sensitivity, we utilize both the A and E channels for the likelihood calculation in the final MCMC parameter estimation. For the training of the CNF model (described in the "Accelerating MCMC sampling with CNFs" section), however, we used data exclusively from the A channel. This choice was made for computational efficiency, as the A channel alone proved sufficient to generate an effective proposal distribution for initializing the MCMC sampling. The T channel, which is typically used for noise characterization, was not included as this study operates under the assumption of known noise PSDs; see also the Discussion section. The TDI response calculations for all channels were implemented using the open-source tool FastLISAResponse (https://github.com/mikekatz04/lisa-on-gpu/tree/master). Each data sample (used for training, as detailed in the "Dataset generation" section) comprises 2-month-duration TDI-A channel measurements, simulated at a sampling interval of 25 s. The GPU-accelerated heterogeneous computing architecture employed for both waveform generation and response emulation [69] enables the generation of single waveforms with latency under 1 s, while maintaining the required numerical precision [81].

Data preprocessing

Although GPU acceleration significantly improves the efficiency of waveform generation, the high dimensionality of the raw time-domain data still poses difficulties for machine learning models. Frequency domain analysis [81–83] has proven to be an exceptionally effective method for estimating parameters in EMRI analysis. Therefore, we transform the time-domain samples to the frequency domain using the Fast Fourier Transform (FFT) algorithm.

For discrete time-series $s^A_{EMRI}(t_n)$ with N samples, the spectral components can be derived as

$$S^A_{\mathrm{EMRI}}(f_k) = \mathcal{F}\{w(t_n)\cdot s^A_{\mathrm{EMRI}}(t_n)\} = \sum_{n=0}^{N-1} w(t_n) s^A_{\mathrm{EMRI}}(t_n) e^{-i2\pi kn/N}, \quad (6)$$

where $t_n = n\Delta t$ defines the sampling times and Δt is the sampling interval, $f_k = k/(N\Delta t)$ denotes frequency bins, and $w(t_n)$ is the Tukey window ($\alpha = 0.05$). $\mathcal{F}$ denotes the FFT algorithm. The windowed signal satisfies

$$s^{\mathrm{window}}_{\mathrm{EMRI}}(t_n) = \begin{cases} w(t_n)\cdot s^A_{\mathrm{EMRI}}(t_n), & 0 \le n < N_{\mathrm{orig}} \\ 0, & N_{\mathrm{orig}} \le n < N_{\mathrm{pad}}, \end{cases} \quad (7)$$

with N_{orig} being the original signal length and N_{pad} being the zero-padded length.

In the EMRI analysis, conventional min–max normalization may induce information degradation. Therefore, we propose a preprocessing scheme that combines spectral whitening with standardization. For the TDI-*A* channel, we first compute its noise power spectral density (PSD) in terms of the test-mass acceleration noise $S_{\mathrm{acc}}(f)$ and the optical metrology system noise $S_{\mathrm{oms}}(f)$,

$$\mathrm{PSD}(f) = 32\sin^2(4u)\sin^2(2u)\left[S_{\mathrm{oms}}(f)(2+\cos(2u)) + 2S_{\mathrm{acc}}(f)(3+2\cos(2u)+\cos(4u))\right], \quad (8)$$

where $u \equiv \pi L f/c$ is the normalized frequency, and $S_{\mathrm{oms}}(f)$, $S_{\mathrm{acc}}(f)$ take the designed spectral profiles of Taiji (see, e.g., Ref. [84]).

The amplitude spectral density (ASD) is then derived as $\mathrm{ASD} \equiv \sqrt{\mathrm{PSD}}$. To ensure that the SNR remains consistent before and after whitening, we introduce a scale factor:

$$\kappa = \sqrt{\frac{N_t}{4\Delta t}}, \quad (9)$$

with N_t being the number of time samples. The whitening process is ultimately implemented through spectral normalization:

$$\tilde{s}_{\mathrm{whitened}}(f) = \frac{s^{\mathrm{window}}_{\mathrm{EMRI}}(t_n)}{\mathrm{ASD}\cdot\kappa}, \quad (10)$$

This pipeline effectively decouples noise correlations while maintaining the phase coherence and amplitude characteristics of EMRI GW signals.

We then establish a physically consistent noise injection framework based on whitened signals. The complex Gaussian noise in the frequency domain is modeled with independent real and imaginary components,

$$n(f) = n_{\mathrm{real}}(f) + i\cdot n_{\mathrm{imag}}(f), \quad n_{\mathrm{real}}, n_{\mathrm{imag}} \sim \mathcal{N}(0,\sigma^2), \quad (11)$$

where the variance $\sigma^2 = 1$ is guaranteed by whitening normalization. The noise injection is implemented through spectral superposition,

$$s'(f) = \tilde{s}_{\mathrm{whitened}}(f) + n(f). \quad (12)$$

This process preserves the phase coherence of the signals while ensuring consistency with Taiji's noise PSD characteristics. This completes the full pipeline for generating individual EMRI GW signals.

Dataset generation

EMRI data are inherently challenging to process due to their large size [85], and this difficulty is further amplified in machine learning applications, where models are trained on batched data. After the FFT, each sample can reach lengths of up to millions. While dimensionality reduction techniques like image-based downsampling or adaptive pooling could theoretically compress the data volume, these methods risk losing vital relativistic features present in EMRI waveforms. To preserve the full harmonic coherence and orbital features essential for robust parameter inference, the TDI-*A* channel EMRI signals were kept in their pristine, noise-free form within the training dataset. During each training iteration, distinct realizations of instrumental noise were dynamically generated and injected into the clean signal data, which accurately emulate the stochastic noise characteristics encountered in real observations. Specifically, we first perform random sampling within the prior parameter ranges listed in Table 1, selecting 20,000 EMRI signal parameters with SNRs exceeding 60 to form the training dataset.

An additional 2,000 parameter sets meeting the same SNR criterion were selected to constitute the test set.

Accelerating MCMC sampling with CNFs

This study introduces an innovative method, FM-MCMC, which accelerates MCMC sampling through CNFs. By integrating CNFs into the widely used Eryn [44] MCMC framework for space-based GW data analysis, the proposed approach significantly enhances the efficiency of parameter estimation. Specifically, we first train a CNF model to rapidly generate posterior distributions for EMRI's intrinsic parameters. Through a dynamic matching mechanism, the sample size generated by the model automatically adapts to Eryn MCMC's walker count and temperature ladder configuration. Simultaneously, based on the posterior distribution characteristics output by CNFs, the system intelligently optimizes Eryn MCMC's initial parameter space to achieve automated contraction of prior distribution ranges.

Continuous normalizing flows

While current machine learning applications in natural sciences frequently employ neural posterior estimation [33,72,86–88] with discrete normalizing flows [89,90], recent advances suggest that flow matching with CNFs [91]—termed FMPE [42,43]—offers superior training efficiency and accuracy [33,42,88]. We examine 2 probability distributions over $\mathbb{R}^d$ characterized by densities $q(\theta_0)$ (base distribution q_0) and $q(\theta_1|x)$ (target posterior q_1), where x represents observational data. The fundamental objective of CNFs is to construct a diffeomorphism $f: \mathbb{R}^d \to \mathbb{R}^d$ satisfying the transport condition: if $\theta_0 \sim q_0$, then $f(\theta_0) \sim q_1$. CNFs realize this transformation via a continuous process governed by a temporal parameter $t \in [0,1]$, where the dynamics are defined by a neural velocity field $v_{t,x}: \mathbb{R}^d \to \mathbb{R}^d$,

$$\frac{d}{dt}\phi_{t,x}(\theta) = v_{t,x}(\phi_{t,x}(\theta)), \quad \phi_{0,x}(\theta) = \theta_0, \quad \phi_{1,x}(\theta) = \theta_1, \quad (13)$$

where $\phi_{t,x}$ represents the flow map transporting samples from q_0 to q_1 under the conditioning variable x. In this context, x corresponds to the EMRI signal in the presence of noise. Target distribution samples are generated by numerically integrating Eq. 13 with initial conditions drawn from q_0. The associated probability density evolution follows the continuity equation,

$$\frac{\partial}{\partial t} q_{t,x}(\theta) + \nabla_\theta \cdot \left(q_{t,x}(\theta) v_{t,x}(\theta) \right) = 0, \tag{14}$$

yielding the instantaneous density via

$$q(\theta|x) = q_0(\theta_0) \exp\left(- \int_0^1 \nabla_\theta \cdot v_{t,x}(\theta_t)\, dt \right). \tag{15}$$

This formulation enables simultaneous density evaluation and sampling through adaptive ordinary differential equation (ODE) solvers.

Following flow matching principles, we parameterize the velocity field through regression on conditional vector fields,

$$u_t(\theta|\theta_1) = \theta_1 - \theta, \tag{16}$$

which induce Gaussian probability paths

$$p_t(\theta|\theta_1) = \mathcal{N}\left(\theta_1 t, (1-t)^2 I_d\right). \tag{17}$$

These paths interpolate between a standard normal distribution at $t = 0$ and a posterior distribution centered at θ_1 as $t \to 1$. The training objective is to minimize the expected deviation between learned and target vector fields,

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{t\sim\mathcal{U}(0,1)} \mathbb{E}_{\theta_1\sim p(\theta)} \mathbb{E}_{x\sim p(x|\theta_1)} \mathbb{E}_{\theta_t\sim p_t(\theta_t|\theta_1)} \|r\|^2, \tag{18}$$

$$r = v_{t,x}(\theta_t) - u_t(\theta_t|\theta_1), \tag{19}$$

where $v_{t,x}$ learns to transport samples while preserving density consistency through the divergence term in Eq. 15.

High-precision Bayesian inference

Our framework employs CNFs, trained through flow matching, to enable rapid posterior estimation for EMRI. The trained model generates 10^3 posterior samples in just a few minutes, providing high-quality, optimized initial proposals for PTMCMC sampling. The CNF sample generation dynamically adapts to PTMCMC's chain configurations and temperature ladder parameters. Final posterior distributions are derived from thermally equilibrated PTMCMC chains.

Our training set comprises 20,000 clean EMRI waveform samples, each waveform representing a 2-month Taiji observation window sampled at fixed intervals, totaling 200,000 data points. The neural architecture comprises two core components: (a) a 1-dimensional convolutional dimensionality reduction encoder that compresses raw GW signals from 200,000 to 2,048 dimensions through nonlinear operations while preserving critical phase information, and (b) a CNF network with 21 residual blocks, achieving progressive feature compression from 3,072 to 4 dimensions.

This hierarchical architecture achieves synergistic optimization of EMRI signal embeddings x, temporal evolution parameter t, and dynamic parameter states θ_t through a multiphase feature interaction mechanism. By progressively integrating the 3 critical components, the framework establishes an accurate mapping to the vector field $v_{t,x}(\theta_t)$. The training was conducted on an NVIDIA RTX 4090 GPU, utilizing the Adam optimizer (initial learning rate 0.0001) with a 2,000 epoch cosine annealing schedule. The full training cycle required approximately 48 h.

Acknowledgments

We sincerely thank the PaddlePaddle team (https://www.paddlepaddle.org.cn/) of Baidu for providing the arithmetic support in this study.

Funding: This study is supported by the National Key Research and Development Program of China (Grant Nos. 2021YFC2201901, 2021YFC2203004, 2020YFC2200100, and 2021YFC2201903) and the International Partnership Program of the Chinese Academy of Sciences (Grant No. 025GJHZ2023106GC). We also gratefully acknowledge the financial support from Brazilian agencies Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS), Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). This work is supported by the High-Performance Computing Platform of Peking University.

Author contributions: B.L. led the core methodological development of the FM-MCMC framework, performed the main numerical simulations and data analysis, and wrote the initial draft of the manuscript. M.D. and P.X. supervised the overall project, provided critical theoretical guidance on the research direction, secured necessary funding, and were responsible for the final review and editing of the manuscript. C.L., H.S., Z. Lyu, Z. Luo, S.H., H.G., T.Z., M.L., Y.X., L.Q., M.S., and W.-L.Q. participated in the core topic discussion and provided valuable scientific insights. All authors reviewed and approved the final manuscript.

Competing interests: The authors declare that they have no competing interests.

Data Availability

The datasets generated and analyzed during the current study were produced using specified open-source tools and simulated according to the described methodological framework. The EMRI waveforms were generated using the https://github.com/emri-wg/FEWFastEMRIWaveforms (FEW) framework, utilizing the AAK model. The detector response and instrumental noise budget for the https://github.com/TriangleDataCenter/Triangle-Simulator Taiji mission were modeled using the open-source tool https://github.com/mikekatz04/lisa-on-gpu/FastLISAResponse. The source code for the FM-MCMC framework and the data used for generating the figures in this paper are publicly available on GitHub at https://github.com/aspirinliang/FM-MCMChttps://github.com/aspirinliang/FM-MCMC.

Supplementary Materials

Notes S1 to S4
Figs. S1 to S6

References

1. Aasi J, Abbott BP, Abbott R, Abbott T, Abernathy MR, Ackley K, Adams C, Adams T, Addesso P, Adhikari RX, et al.

Advanced LIGO. *Class Quant Grav*. 2015;32(7):Article 074001.
2. Abbott BP, Abbott R, Abbott TD, Abernathy MR, Acernese F, Ackley K, Adams C, Adams T, Addesso P, Adhikari RX, et al. Observation of gravitational waves from a binary black hole merger. *Phys Rev Lett*. 2016;116(6):Article 061102.
3. LIGO-Virgo-KAGRA Collaboration. LIGO-Virgo-KAGRA Announce the 200th Gravitational Wave Detection of O4! 2025. [accessed 19 July 2025] https://www.ligo.caltech.edu/news/ligo20250320
4. Abbott R, Abbott TD, Abraham S, Acernese F, Ackley K, Adams A, Adams C, Adhikari RX, Adya VB, Affeldt C, et al. GWTC-2: Compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run. *Phys Rev X*. 2021;11(2):Article 021053.
5. Abbott R, Abbott TD, Acernese F, Ackley K, Adams C, Adhikari N, Adhikari RX, Adya VB, Affeldt C, Agarwal D, et al. GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run. *Phys Rev X*. 2023;13(4):Article 041039.
6. Ruan WH, Liu C, Guo ZK, Wu YL, Cai RG. The LISA-Taiji network: Precision localization of coalescing massive black hole binaries. *Research*. 2021;2021:Article 6014164.
7. Xu X, Liu H, Tan Y. Verification of laser heterodyne interferometric bench for Chinese spaceborne gravitational wave detection missions. *Research*. 2024;7:0302.
8. Hu WR, Wu YL. The Taiji program in space for gravitational wave physics and the nature of gravity. *Natl Sci Rev*. 2017;4(5):685–686.
9. Luo J, Chen LS, Duan HZ, Gong YG, Hu S, Ji J, Liu Q, Mei J, Milyukov V, Sazhin M, et al. TianQin: A space-borne gravitational wave detector. *Class Quant Grav*. 2016;33(3):Article 035010.
10. Amaro-Seoane P, Audley H, Babak S, Baker J, Barausse E, Bender P, Berti E, Binetruy P, Born M, Bortoluzzi D, et al. Laser interferometer space antenna. arXiv. 2017. https://doi.org/10.48550/arXiv.1702.00786.
11. Babak S, Gair J, Sesana A, Barausse E, Sopuerta CF, Berry CP, Berti E, Amaro-Seoane P, Petiteau A, Klein A. Science with the space-based interferometer LISA. V: Extreme mass-ratio inspirals. *Phys Rev D*. 2017;95(10):Article 103012.
12. Gair JR, Vallisneri M, Larson SL, Baker JG. Testing general relativity with low-frequency. *Living Rev Relativ*. 2013;16(1):7.
13. Laghi D, Tamanini N, Del Pozzo W, Sesana A, Gair J, Babak S, Izquierdo-Villalba D. Gravitational-wave cosmology with extreme mass ratio inspirals. *Mon Not Roy Astron Soc*. 2021;508(3):4512–4531.
14. Sopuerta CF, Yunes N. Extreme- and intermediate-mass ratio inspirals in dynamical Chern Simons modified gravity. *Phys Rev D*. 2009;80(6):Article 064006.
15. Barack L, Cutler C. Using LISA extreme-mass-ratio inspiral sources to test off-Kerr deviations in the geometry of massive black holes. *Phys Rev D*. 2007;75(4):Article 042003.
16. Speri L, Barsanti S, Maselli A, Sotiriou TP, Warburton N, van de Meent M, Chua AJ, Burke O, Gair J. Probing fundamental physics with Extreme Mass Ratio Inspirals: A full Bayesian inference for scalar charge. arXiv. 2024. https://doi.org/10.48550/arXiv.2406.07607
17. Zou X, Zhong X, Han WB, Mohanty SD. Constraint on the deviation of Kerr metric via bumpy parameterization and particle swarm optimization in extreme mass-ratio inspirals. arXiv. 2025. https://doi.org/10.48550/arXiv.2506.21955.
18. Xin S, Han WB, Yang SC. Gravitational waves from extreme-mass-ratio inspirals using general parametrized metrics. *Phys Rev D*. 2019;100(8):Article 084055.
19. Gair JR, Tang C, Volonteri M. LISA extreme-mass-ratio inspiral events as probes of the black hole mass function. *Phys Rev D*. 2010;81(10):Article 104014.
20. Speri L, Antonelli A, Sberna L, Babak S, Barausse E, Gair JR, Katz ML. Probing accretion physics with gravitational waves. *Phys Rev X*. 2023;13(2):Article 021035.
21. Cole PS, Bertone G, Coogan A, Gaggero D, Karydas T, Kavanagh BJ, Spieksma TF, Tomaselli GM. Disks, spikes, and clouds: Distinguishing environmental effects on BBH gravitational waveforms. arXiv. 2022. https://doi.org/10.48550/arXiv.2211.01362.
22. Hannuksela OA, Ng KCY, TGF L. Extreme dark matter tests with extreme mass ratio inspirals. *Phys Rev D*. 2020;102(10):Article 103022.
23. Yue XJ, Han WB, Chen X. Dark matter: An efficient catalyst for intermediate-mass ratio-inspiral events. *Astrophys J*. 2019;874(1):34.
24. Gair JR, Barack L, Creighton T, Cutler C, Larson SL, Phinney ES, Vallisneri M. Event rate estimates for LISA extreme mass ratio capture sources. *Class Quant Grav*. 2004;21(20):S1595–S1606.
25. Saltas ID, Oliveri R. EMRI MC: A GPU-based code for Bayesian inference of EMRI wave nforms. arXiv. 2023. https://doi.org/10.48550/arXiv.2311.17174
26. Chua AJK, Katz ML, Warburton N, Hughes SA. Rapid generation of fully relativistic extreme-mass-ratio-inspiral waveform templates for LISA data analysis. *Phys Rev Lett*. 2021;126(5):Article 051102.
27. Babak S, Baker JG, Benacquista MJ, Cornish NJ, Larson SL, Mandel I, McWilliams ST, Petiteau A, Porter EK, Robinson EL, et al. The mock LISA data challenges: From challenge 3 to challenge 4. *Class Quant Grav*. 2010;27(8):Article 084009.
28. Chua AJK, Cutler CJ. Nonlocal parameter degeneracy in the intrinsic space of gravitational wave signals from extreme-mass-ratio inspirals. *Phys Rev D*. 2022;106(12):Article 124046.
29. Zou XB, Mohanty SD, Luo HG, Liu YX. Swarm intelligence methods for extreme mass ratio Inspiral search: First application of particle swarm optimization. *Universe*. 2024;10(2):Article 96.
30. Cornish NJ. Detection strategies for extreme mass ratio inspirals. *Class Quant Grav*. 2011;28(9):Article 094016.
31. Babak S, Gair JR, Porter EK. An algorithm for the detection of extreme mass ratio inspirals in LISA data. *Class Quant Grav*. 2009;26(13):Article 135004.
32. Zou X, Soumya M, Xie Q, Chen X, Luo H, Liu Y, Han W, Jiao J, Zhang X, Zhao S. An overview of EMRI data analysis. *Acta Sci Nat Univ Pekin*. 2025;61(181):181–194.
33. Liang B, Du M, Wang H, Xu Y, Liu C, Wei X, Xu P, Qiang LE, Luo Z. Rapid parameter estimation for merging massive black hole binaries using ODE-based generative models. arXiv. 2024. https://doi.org/10.48550/arXiv.2407.07125
34. Yun Q, Han WB, Guo YY, Wang H, Du M. The detection, extraction and parameter estimation of extreme-mass-ratio inspirals with deep learning. *Sci Chin Phys Mech Astron*. 2025;68(1):Article 210413.
35. Zou XB, Mohanty SD, Luo HG, Liu YX. Search for extreme mass ratio inspirals using particle swarm optimization and reduced dimensionality likelihoods. *Universe*. 2024;10(4):171.

36. Ye CQ, Fan HM, Torres-Orjuela A, Zhang J, Hu YM. Identification of gravitational waves from extreme-mass-ratio inspirals. *Phys Rev D*. 2024;109(12):Article 124034.
37. Shen P, Cui Q, Han WB. Assessing the systematic errors of extreme-mass-ratio inspirals waveforms for testing general relativity. *Phys Rev D*. 2025;111(2):Article 024004.
38. Zou X, Zhong X, Han WB, Mohanty SD. Constraining deviations from the Kerr metric via a bumpy parametrization and particle swarm optimization in extreme mass-ratio inspirals. *Phys Rev D*. 2025;112(8):Article 084075.
39. Littenberg TB, Cornish NJ. Prototype global analysis of LISA data with multiple source types. *Phys Rev D*. 2023;107(6):Article 063004.
40. Deng S, Babak S, Le Jeune M, Marsat S, Plagnol E, Sartirana A. Modular global-fit pipeline for LISA data analysis. *Phys Rev D*. 2025;111(10):Article 103014.
41. Du M, Luo Z, Xu P. Enhancing Taiji's parameter estimation under nonstationarity: A time frequency domain framework for galactic binaries and instrumental noises. *Phys Rev D*. 2025;112(8):Article 083036.
42. Dax M, Wildberger J, Buchholz S, Green SR, Macke JH, Scholkopf B. Flow matching for scalable simulation-based inference. arXiv. 2023. https://doi.org/10.48550/arXiv.2305.17161
43. Lipman Y, Chen RT, Ben-Hamu H, Nickel M, Le M. Flow matching for generative modeling. arXiv. 2022. https://doi.org/10.48550/arXiv.2210.02747
44. Karnesis N, Katz ML, Korsakova N, Gair JR, Stergioulas N. Eryn: A multipurpose sampler for Bayesian inference. *Mon Not Roy Astron Soc*. 2023;526(4):4814–4830.
45. Foreman-Mackey D, Hogg DW, Lang D, Goodman J. Emcee: The MCMC hammer. *Publ Astron Soc Pac*. 2013;125(925):306–312.
46. Katz ML, Chua AJK, Speri L, Warburton N, Hughes SA. Fast extreme-mass-ratio-inspiral waveforms: New tools for millihertz gravitational-wave data analysis. *Phys Rev D*. 2021;104(6):Article 064047.
47. LISA Consortium Waveform Working Group. Waveform modelling for the laser interferometer space antenna. arXiv. 2023. https://doi.org/10.48550/arXiv.2311.01300
48. Katz ML, Karnesis N, Korsakova N, Gair JR, Stergioulas N. Efficient GPU-accelerated multisource global fit pipeline for LISA data analysis. *Phys Rev D*. 2025;111(2):Article 024060.
49. Abbott BP, Abbott R, Abbott TD, Abernathy MR, Acernese F, Ackley K, Adams C, Adams T, Addesso P, Adhikari RX, et al. Binary black hole mergers in the first advanced LIGO observing run. *Phys Rev X*. 2016;6(4):Article 041015.
50. Abbott BP, Abbott R, Abbott TD, Abernathy MR, Acernese F, Ackley K, Adams C, Adams T, Addesso P, Adhikari RX, et al. Astrophysical implications of the binary black hole merger GW150914. *Astrophys J Lett*. 2016;818(2):L22.
51. Gair JR, Vallisneri M, Larson SL, Baker JG. Testing general relativity with low-frequency, space-based gravitational-wave detectors. *Living Rev Relativ*. 2013;16(1):Article 7.
52. Wardell B, Pound A, Warburton N, Miller J, Durkan L, Le Tiec A. Gravitational wave forms for compact binaries from second-order self-force theory. *Phys Rev Lett*. 2023;130(24):Article 241402.
53. Yunes N, Yagi K, Pretorius F. Theoretical physics implications of the binary black-hole mergers GW150914 and GW151226. *Phys Rev D*. 2016;94(8):Article 084002.
54. Abbott BP, Abbott R, Abbott TD, Abernathy MR, Acernese F, Ackley K, Adams C, Adams T, Addesso P, Adhikari RX, et al. Tests of general relativity with GW150914. *Phys Rev Lett*. 2016;116(22):Article 221101.
55. Deng S, Babak S, Jeune ML, Marsat S, Plagnol E, Sartirana A. Modular global-fit pipeline for LISA data analysis. arXiv. 2025. https://doi.org/10.48550/arXiv.2501.10277
56. Cornish NJ. Time-frequency analysis of gravitational wave data. *Phys Rev D*. 2020;102(12):Article 124038.
57. Barack L, Pound A. Self-force and radiation reaction in general relativity. *Rept Prog Phys*. 2019;82(1):Article 016904.
58. Pound A, Wardell B. Black hole perturbation theory and gravitational self-force. arXiv. 2021. https://doi.org/10.48550/arXiv.2101.04592
59. van de Meent M. Gravitational self-force on generic bound geodesics in Kerr spacetime. *Phys Rev D*. 2018;97(10):Article 104033.
60. Pound A, Wardell B, Warburton N, Miller J. Second-order self-force calculation of gravitational binding energy in compact binaries. *Phys Rev Lett*. 2020;124(2):Article 021101.
61. Warburton N, Pound A, Wardell B, Miller J, Durkan L. Gravitational-wave energy flux for compact binaries through second order in the mass ratio. *Phys Rev Lett*. 2021;127(15):Article 151102.
62. Albertini A, Gamba R, Nagar A, Bernuzzi S. Effective-one-body waveforms for extreme mass-ratio binaries: Consistency with second-order gravitational self-force quasicircular results and extension to nonprecessing spins and eccentricity. *Phys Rev D*. 2024;109(4):Article 044022.
63. Shen P, Han WB, Zhang C, Yang SC, Zhong XY, Jiang Y, Cui Q. Influence of mass-ratio corrections in extreme-mass-ratio inspirals for testing general relativity. *Phys Rev D*. 2023;108(6):Article 064015.
64. Barack L, Cutler C. LISA capture sources: Approximate waveforms, signal-to-noise ratios, and parameter estimation accuracy. *Phys Rev D*. 2004;69(8):Article 082005.
65. Babak S, Fang H, Gair JR, Glampedakis K, Hughes SA. "Kludge" gravitational waveforms for a test-body orbiting a Kerr black hole. *Phys Rev D*. 2007;75(2):Article 024005.
66. Chua AJK, Gair JR. Improved analytic extreme-mass-ratio inspiral model for scoping out eLISA data analysis. *Class Quant Grav*. 2015;32(23):Article 232002.
67. Chua AJK, Moore CJ, Gair JR. Augmented kludge waveforms for detecting extreme-mass ratio inspirals. *Phys Rev D*. 2017;96(4):Article 044005.
68. Gair JR, Glampedakis K. Improved approximate inspirals of test-bodies into Kerr black holes. *Phys Rev D*. 2006;73(6):Article 064037.
69. Katz ML, Bayle JB, Chua AJK, Vallisneri M. Assessing the data-analysis impact of LISA orbit approximations using a GPU-accelerated response model. *Phys Rev D*. 2022;106(10):Article 103001.
70. Baghi Q, Karnesis N, Bayle JB, Besançon M, and Inchauspé H. Uncovering gravitational-wave backgrounds from noises of unknown shape with LISA. Journal of Cosmology and Astroparticle Physics. 2023: 066. DOI: 10.1088/1475-7516/2023/04/066
71. WG L. LISA Data Challenge Manual. https://lisa-ldc.lal.in2p3.fr/static/data/pdf/LDC-manual-002.pdf
72. Du M, Liang BH, Wang H, Xu P, Luo Z, Wu Y. Advancing space-based gravitational wave astronomy: Rapid detection and parameter estimation using normalizing flows. 2023. https://api.semanticscholar.org/CorpusID:260775495

73. Luo Z, Wang Y, Wu Y, Hu W, Jin G. The Taiji program: A concise overview. *Prog Theor Exp Phys*. 2021;2021(5):05A108.
74. Wang H, Du M, Xu P, Zhou Y-F. Challenges in space-based gravitational wave data analysis and applications of artificia intelligence. *Sci Sin Phys Mech Astron*. 2024;54(7): Article 270403.
75. Tinto M, Dhurandhar S. Time-delay interferometry. *Living Rev Relativ*. 2021;24(1):Article 1.
76. Armstrong JW, Estabrook FB, Tinto M. Time-delay interferometry for space-based gravitational wave searches. *Astrophys J*. 1999;527(2):814.
77. Ni WT, Shy JT, Tseng SM, Xu X, Yeh HC, Hsu WY, Liu WL, Tzeng SD, Fridelance P, Samain E, et al. Progress in mission concept study and laboratory development for the ASTROD (Astrodynamical Space Test of Relativity using Optical Devices). In: Allah dadi FA, Casani EK, Maclay TD, editors. *Small spacecraft, space environments, and instrumentation technologies*. Vol. 3116. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. SPIE; 1997. p. 105–116.
78. Tinto M, Dhurandhar S, Malakar D. Second-generation time-delay interferometry. *Phys Rev D*. 2023;107(8):Article 082001.
79. Otto M. Time-delay interferometry simulations for the laser interferometer space antenna [thesis]. Leibniz University Hannover; 2015.
80. Wang G, Ni WT. Revisiting time delay interferometry for unequal-arm LISA and TAIJI. *Phys Scr*. 2023;98(7): Article 075005.
81. Burke O, Piovano GA, Warburton N, Lynch P, Speri L, Kavanagh C, Wardell B, Pound A, Durkan L, Miller J. Accuracy requirements: Assessing the importance of first post-adiabatic terms for small-mass-ratio binaries. arXiv. 2024. https://doi.org/10.48550/arXiv.2310.08927
82. Bayle JB, Bonga B, Caprini C, Doneva D, Muratore M, Petiteau A, Rossi E, Shao L. Overview and progress on the laser interferometer space antenna mission. *Nat Astron*. 2022;6(12):1334–1338.
83. Speri L, Karnesis N, Renzini AI, Gair JR. A roadmap of gravitational wave data analysis. *Nat Astron*. 2022;6(12): 1356–1363.
84. Wang G, Ni WT, Han WB, Yang SC, Zhong XY. Numerical simulation of sky localization for LISA-TAIJI joint observation. *Phys Rev D*. 2020;102(2):Article 024089.
85. Babak S, Gair JR, Cole RH. Extreme mass ratio inspirals: Perspectives for their detection. In: *Equations of motion in relativistic gravity*. Springer International Publishing; 2015. p. 783–812.
86. Papamakarios G, Murray I. Fast ϵ-free inference of simulation models with Bayesian conditional density estimation. arXiv. 2018. https://doi.org/10.48550/arXiv.1605.06376
87. Dax M, Green SR, Gair J, Macke JH, Buonanno A, Scholkopf B. Real-time gravitational wave science with neural posterior estimation. *Phys Rev Lett*. 2021;127(24):Article 241103.
88. Gebhard TD, Wildberger J, Dax M, Angerhausen D, Quanz SP, Schölkopf B. Inferring atmospheric properties of exoplanets with flow matching and neural importance sampling. arXiv. 2023. https://doi.org/10.48550/arXiv.2312.08295
89. Rezende D, Mohamed S. Variational inference with normalizing flows. In: *International conference on machine learning*. PMLR; 2015. p. 1530–1538.
90. Papamakarios G, Nalisnick E, Rezende DJ, Mohamed S, Lakshminarayanan B. Normalizing flows for probabilistic modeling and inference. *J Mach Learn Res*. 2021;22(57):1–64.
91. Chen RTQ, Rubanova Y, Bettencourt J, and Duvenaud D. Neural ordinary differential equations. arXiv. 2019. https://doi.org/10.48550/arXiv.1806.07366