

Residual neural likelihood estimation and its application to gravitational-wave astronomy

Mattia Emma¹**Department of Physics, Royal Holloway University of London, Egham, TW20 0EX, United Kingdom*Gregory Ashton²*Department of Physics, Royal Holloway University of London, Egham, TW20 0EX, United Kingdom and Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, United Kingdom*

(Received 21 January 2026; accepted 20 May 2026; published 22 June 2026)

Simulation-based inference provides a powerful framework for Bayesian inference when the likelihood is analytically intractable or computationally prohibitive. By leveraging machine-learning techniques and neural density estimators, it enables flexible likelihood or posterior modeling directly from simulations. We introduce residual neural likelihood estimation (RNLE), a modification of neural likelihood estimation (NLE) that learns the likelihood of non-Gaussian noise in gravitational-wave detector data. Exploiting the additive structure of the signal and noise generation processes, RNLE directly models the noise distribution, substantially reducing the number of simulations required for accurate parameter estimation and improving robustness to realistic noise artifacts. The performance of RNLE is demonstrated using a toy model, simulated gravitational-wave signals, and real detector noise from ground based interferometers. In the presence of loud non-Gaussian transients, we find that RNLE can improve robustness of parameter recovery when trained on appropriately constructed datasets. We further assess the stability of the method by quantifying the variability introduced by retraining the conditional density estimator on statistically identical datasets with different optimization seeds, referred to as training noise, and investigate evidence-weighted combinations of independently trained RNLE realizations as a diagnostic of this variability. An implementation of RNLE is publicly available in the SBILBY package, enabling its deployment within gravitational-wave astronomy and a broad range of scientific applications requiring flexible, simulation-based likelihood estimation.

DOI: [10.1103/PhysRevD.113.124064](https://doi.org/10.1103/PhysRevD.113.124064)

I. INTRODUCTION

The first direct detection of a gravitational wave (GW) signal by the Laser Interferometer Gravitational-Wave Observatory (LIGO) occurred on the September 14, 2015 [1]. Since then, the LIGO-Virgo-KAGRA (LVK) Collaboration [2–4] has reported the observation of hundreds of additional compact binary coalescence events [5]. The analysis of these GW signals has enabled unprecedented tests of Einstein’s theory of general relativity [6–9], insights into the evolutionary history of the Universe [10,11], and the nature of compact objects and matter at extreme densities [12–15]. With the increasing number of events, it will become possible to infer the properties of the

cosmological population of black holes and the formation pathways of compact binary systems with increasing precision [16]. To robustly carry out these studies, it is crucial to accurately infer the parameters of the GW sources [17].

The first algorithms for parameter estimation were developed in the late 1980s and employed maximum likelihood estimation techniques [18]. With the application of the Metropolis-Hastings algorithm [19,20] to the calculation of posterior probabilities and the increased performance of hardware and computing power, in the mid 1990s it became feasible to use Bayesian inference in the multidimensional parameter space of GWs. The first to apply Markov chain Monte Carlo (MCMC) to GWs from compact binary mergers were Christensen and Meyer [21]. The strength of the Bayesian approach lies in its ability to incorporate our theoretical knowledge of the GW waveform model to obtain both a point estimate and a credible interval for each inferred parameter. MCMC algorithms were further developed by the LVK collaboration, leading to the LALInference adaptive MCMC routine [22]. This made

*Contact author: mattia.emma@ligo.org

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

it possible to estimate all 15 parameters describing the first GW signal from a spinning binary black hole merger [23,24]. Skilling's work [25,26] independently introduced nested sampling as a general Bayesian inference algorithm. This method was adapted to gravitational-wave parameter estimation [22], where posterior samples are obtained as a byproduct of evidence evaluation. This method and further refined MCMC techniques were merged into the Bilby software [27,28]. This is now the main tool for the production of parameter estimation results in the LVK observing runs [4,29]. An alternative algorithm, RIFT [30–32], uses Gaussian process interpolation with a highly parallelizable grid-based parameter estimation approach [33], and has also been successfully employed for parameter estimation in the third observing run (O3) [34,35].

While Bayesian parameter estimation algorithms have been instrumental in extracting the physical properties of the first hundreds of detected GWs events and continue to be actively refined [36–40], their accuracy is limited by a number of underlying assumptions. In frequency-domain analyses, one such limitation arises from the use of the Whittle likelihood [41–44], which approximates the exact likelihood by assuming that the detector strain noise is stationary and Gaussian. These assumptions are violated in the presence of transient noise artifacts, commonly referred to as glitches [45–47]. Due to their frequent occurrence and potentially high signal-to-noise ratio (SNR), glitches have been shown to contaminate nearly 20% of detected binary black hole signals and essentially all binary neutron star signals [48–51]. More broadly, glitches represent only one manifestation of the nonstationary behavior of interferometric gravitational-wave detectors, which also arises from time-dependent variations in the noise spectrum and environmental disturbances. Such nonstationarities introduce correlations between Fourier components and temporal variations in the noise power spectral density, leading to biased posterior distributions, underestimated uncertainties, and inconsistencies across data segments [52].

The most prominent signal affected by a glitch is GW170817 [53], the first observed binary neutron star (BNS) merger. The simultaneous detection of an electromagnetic counterpart made it stand out as the first, and to date only, confirmed multimessenger event [54–56]. The coincidence of GW170817 with an extremely loud glitch in the Livingston detector [57] demonstrated the importance of developing methods to perform accurate and fast parameter estimation in the presence of transient noise artifacts. Different techniques have been developed to achieve this goal [48]. Some aim to subtract known and modeled nontransient noise directly from the strain data using knowledge provided by the interferometer's auxiliary channels, e.g., GWSUBTRACT [58–60]. Other algorithms aim to model and subtract glitches from the data before analyzing the signal with the Whittle likelihood [61,62] or to model glitch and signal simultaneously [63–65]. Among

these, the BayesWave algorithm is the primary method used by LVK [66]. BayesWave employs linear combinations of wavelets to model non-Gaussianities and also allows for the simultaneous modeling of glitches and signals [67–69]. It is used by the LVK collaboration to model and subtract glitches from data containing GW signals [70]. However, Udall *et al.* [71] showed that the glitch subtraction process is inherently limited, leaving behind residual glitch power with an SNR of 3–7 as an unavoidable consequence. They further demonstrated that this residual power is enough to introduce biases in parameter estimation results.

Concurrently, different machine learning (ML) algorithms [72] have also been developed for the classification of nontransient noise [73–75], the deglitching of gravitational wave data, and to distinguish between glitch and signal power [76–78]. While these methods introduce significant improvements with respect to the Whittle likelihood and have been successfully applied in the analysis of gravitational wave events [51,79–81], none of them can yet be reliably applied in every setting to fully unbiased GW parameter estimation results [82]. More recently, novel approaches have been proposed to directly address this limitation by reducing or bypassing assumptions about the noise realization. For instance, Chatterjee and Jani [83] introduce a framework that reconstructs the gravitational wave signal independently of the underlying noise, while Legin *et al.* [84] present a score-based diffusion model that learns the non-Gaussian noise distribution directly from detector data, enabling accurate parameter recovery even in the presence of loud glitches. These developments represent important steps toward robust inference in glitch-contaminated data.

simulation based inference (SBI) has established itself as a powerful ML-based framework to tackle complex inference problems that are computationally difficult for traditional Bayesian methods or for which an analytical expression of the likelihood function is unavailable [85–88]. The central idea of SBI is to exploit a forward model capable of generating synthetic data given some model parameters. These simulations are then used to train a conditional density estimator that approximates either the likelihood, the posterior, or a related quantity, thereby enabling efficient Bayesian inference without requiring an explicit likelihood function. Modern SBI methods are typically classified into three main categories, depending on the quantity that is learned: neural likelihood estimation (NLE) [89–92], neural posterior estimation (NPE) [93,94], and neural ratio estimation (NRE) [95,96]. In NLE, a neural network is trained to approximate the likelihood function, and inference is performed via standard Bayesian techniques using the learned likelihood. In contrast, NPE directly models the posterior distribution, and so samples can be directly drawn. Finally, NRE estimates the likelihood-to-evidence ratio, which can be used to construct the posterior via importance weighting or rejection sampling.

Each approach offers distinct advantages in terms of training stability, computational efficiency, and flexibility, and the choice between them depends on the structure of the simulator and the inference task at hand. In the past decade, SBI has found wider application in GW astronomy and astrophysics [97–102]. It has been employed for the rapid generation of sky maps for public alerts [103], and algorithms have been developed for the current [104,105] and next generation of GW detectors [106–108]. A specific application of SBI to gravitational-wave parameter estimation is the DINGO software package [109–111]. DINGO employs real detector data with simulated gravitational-wave signal injections to train a NPE model, yielding an approximate posterior distribution. This approximation is subsequently refined through rejection sampling [43,112], and has been shown to reproduce the results obtained with standard Bayesian inference pipelines such as Bilby [110]. However, because the rejection sampling step relies on a Gaussian likelihood approximation, the resulting inference inherits the associated assumptions of stationarity and Gaussianity, and may therefore be biased in the presence of non-Gaussian noise transients.

In this work, we employ a modification of NLE to address the challenges posed by non-Gaussian noise in gravitational-wave data analysis. We adopt NLE rather than NPE for two key reasons. First, NLE has been shown to exhibit higher simulation efficiency, enabling it to learn more complex distributions with a fixed number of simulations [113]. Second, unlike NPE, NLE allows the conditional density estimator to be trained solely on detector noise while still yielding posterior distributions for the signal parameters, a property that is essential for our approach. We introduce residual neural likelihood estimation (RNLE), which combines the NLE algorithm implemented in the SBI package [89,114] with the additive structure of gravitational-wave signal and noise, with the goal of learning a likelihood for real detector data that captures nonstationarities and transient non-Gaussianities. The likelihood is obtained by training the conditional density estimator exclusively on noise realizations. For each likelihood evaluation, a waveform approximant realization is subtracted from the observed data, and the neural likelihood is evaluated on the resulting residuals. This strategy is conceptually related to the approach introduced in Legin *et al.* [84], which also exploits the additivity of signal and noise, and learns the noise distribution directly from detector data to produce an unbiased estimate of the likelihood function. In contrast to Legin *et al.* [84], which employ score-based diffusion models and learns the score of the noise distribution, our method uses neural likelihood estimation within the SBI framework and learns the likelihood directly from the noise distribution. We can therefore integrate the learned likelihood into the Bilby framework, and perform Bayesian parameter estimation using the nested sampling implementation of DYNesty [115].

The paper is structured as follows. In Sec. II, we give a summary of Bayesian parameter estimation and discuss the theoretical validity of the RNLE approach. In Sec. II A, we explain the setup of the RNLE algorithm and the specifics of the employed estimator. In Sec. III we compare results obtained with NLE, RNLE, and the Whittle likelihood for a simple toy model. In Secs. IV and V, we benchmark RNLE against the Whittle likelihood using simulated detector noise for the training and a signal from a binary black hole (BBH) merger injected into simulated noise as an observation. After demonstrating the validity of our algorithm, in Sec. VI, we analyze BBH merger signals injected in quasi-Gaussian detector noise. In Sec. VII, we analyze BBH injections in the proximity of very loud glitches to study the capability of our algorithm to recover accurate posteriors when the Whittle likelihood is biased. To follow up on this analysis, in Sec. VIII we analyze BBH injections around a specific glitch with RNLE likelihoods trained on varying datasets, and in Sec. IX we assess the training noise by retraining on the same dataset with a different machine learning seed. We conclude in Sec. X, summarizing our results and outlining future development and usage possibilities for the SBILBY software package [116].

II. METHODOLOGY

Given observation data $\vec{d}$ and a predictive model M with associated parameters θ , Bayesian inference centers around constructing a likelihood $\mathcal{L}(\vec{d}|\theta, M)$ and prior $\pi(\theta|M)$, then seeking to estimate the posterior probability density

$$p(\theta|\vec{d}, M) \propto \mathcal{L}(\vec{d}|\theta, M)\pi(\theta|M), \quad (1)$$

and the normalizing evidence

$$\mathcal{Z}(M|\vec{d}) = \int \mathcal{L}(\vec{d}|\theta, M)\pi(\theta|M)d\theta. \quad (2)$$

Calculation of the posterior and evidence can only be done in closed form for a limited set of problems (typically, those with so-called conjugate priors [117]). Therefore, computational Bayesian inference approaches have been developed¹ enabling approximation with induced numerical error.

Computational approaches typically require an explicit construction of the likelihood in closed form. This is suitable for many astrophysical data analysis problems where the likelihood takes a simple form following standard probability distributions. Most often, this is the Gaussian distribution thanks to the central limit theorem [118]. However, in some instances, the likelihood is intractable (i.e., difficult or impossible to write down in

¹Grid-based approaches can be applied in low-dimensional problems, but quickly become intractable for more than a few dimensions, therefore stochastic sampling methods such as MCMC and nested sampling are generally applied in higher-dimensional problems.

closed form) or computationally costly [17,119–121]. Over the last few years, a new approach has been developed to address such problems: SBI. While there are multiple variants of SBI, in this work we consider NLE in which a neural density estimator q_ϕ , with weights ϕ , is trained to approximate $\mathcal{L}(\vec{d}|\theta)$ using training data consisting of N pairs $\{\theta_i, \vec{d}_i\}$. To construct the training data, samples θ_i are drawn from a generating distribution (generally the prior). The samples are then passed into the simulator $s(\theta_i)$ to produce a realization of the $\vec{d}_i$. Note that the simulator may, in general, also depend on a set of latent variables z that are randomly sampled during the simulation.

Let us now restrict ourselves to the set of cases where the forward model described by the simulator consists of a deterministic model $m(\vartheta)$ and an additive noise process $n(\lambda)$ where we have defined two subsets ϑ , the model parameters and λ , the parameters describing the noise process with $\theta = \vartheta \cup \lambda$. That is, we consider only models where $s_i = m_i + n_i$ where $m_i = m(\vartheta)$ is deterministic while $n_i = n(\lambda)$ is a random sample drawn from a noise distribution.

Standard NLE proceeds by taking as input the full simulator $s(\theta)$ and proceeding to approximate $\mathcal{L}(\vec{d}|\theta)$. However, a wide variety of astrophysical problems are well described by a deterministic model with additive noise. Under this assumption, it is appropriate to instead approximate the residual likelihood $\mathcal{L}(\vec{d} - m(\vartheta)|\lambda)$ which is simpler due to the lower dimensionality.

To better understand the motivation let us consider cases where the likelihood is known in closed form. For additive white Gaussian noise with standard deviation σ , for a single observation d_i , the log-likelihood is given by

$$\log \mathcal{L}(d_i|\theta) = -\frac{1}{2} \left(\frac{(d_i - m(\vartheta))^2}{\sigma^2} - \log(2\pi\sigma^2) \right), \quad (3)$$

and if the set of data $\vec{d}$ is independent and identically distributed, then

$$\log \mathcal{L}(\vec{d}|\theta) = \sum_i \log \mathcal{L}(d_i|\theta). \quad (4)$$

However, given that the data only enters the likelihood through the residual $r_i = d_i - m(\vartheta)$, we can introduce the *residual likelihood* as

$$\log \mathcal{L}(r_i|\lambda) = -\frac{1}{2} \left(\frac{(r_i)^2}{\sigma^2} - \log(2\pi\sigma^2) \right), \quad (5)$$

which is identical to the standard likelihood function. For standard applications of this likelihood, it would be odd to differentiate between the two likelihoods, but if one were to apply NLE, learning the residual likelihood is easier as the density estimate that must be learned is only that of the

noise distribution and not the full joint distribution of the data and θ .

More related to GW is the case of colored Gaussian noise. For interferometric gravitational-wave data, in the absence of a signal (which we refer to as M_n), the detector data is well described by stationary coloured Gaussian noise. Assuming we know *a priori* the power spectral density (PSD) that describes the noise, if $\vec{d}$ is a regular real-valued time-series with a sampling frequency f_s and duration T , then the likelihood is known and can be approximated by the so-called Whittle likelihood,

$$\log \mathcal{L}(\vec{d}|M_n) \propto -\frac{1}{2} \langle \tilde{d}, \tilde{d} \rangle, \quad (6)$$

where $\tilde{d} = \text{fft}(\vec{d})/f_s$ is the scaled complex-valued Fourier-transformed data, the noise-weighted inner product is defined as,

$$\langle a, b \rangle := \frac{4}{T} \sum_j \mathcal{R} \left(\frac{a_j^* b_j}{P_j} \right), \quad (7)$$

and P_j is the real-valued PSD. If a signal is present in the data, then the generate model M_s predicts an additive sum of the signal and noise such that

$$\log \mathcal{L}(\vec{d}|\vartheta, M_s) = -\frac{1}{2} \langle \tilde{d} - m(\vartheta), \tilde{d} - m(\vartheta) \rangle - \sum_j \log(2\pi P_j). \quad (8)$$

Again, we can rewrite this as a residual likelihood

$$\log \mathcal{L}(\vec{r}|M_s) \propto -\frac{1}{2} \langle \tilde{r}, \tilde{r} \rangle - \sum_j \log(2\pi P_j). \quad (9)$$

In this instance, the noise has no parameters (i.e. λ is the empty set). However, one could also consider parameterizing the PSD $P_j \rightarrow P_j(\lambda)$ in which case

$$\log \mathcal{L}(\vec{r}|\lambda, M_s) = -\frac{1}{2} \langle \tilde{r}, \tilde{r} \rangle_\lambda - \sum_i \log(2\pi P_j(\lambda)), \quad (10)$$

where we add the subscript λ to the noise-weighted inner product to denote that these parameters are required by the PSD. This can straightforwardly be generalized for a multi-detector analysis. Since the detector noise is assumed to be independent, the joint likelihood is obtained by combining the individual detector likelihoods multiplicatively, which corresponds to summing their logarithms. Thus, for detectors $d \in \{1, \dots, D\}$, the total log-likelihood is

$$\log \mathcal{L}_{\text{tot}}(\vec{r}|\lambda, M_s) = \sum_{d=1}^D \log \mathcal{L}_d(\lambda), \quad (11)$$

where $\mathcal{L}_d(\vec{r}|\lambda, M_s)$ denotes the RNLE-learned likelihood for detector d evaluated at parameters λ .

For these cases, where the likelihood is known in closed form, NLE is not required. However, our hypothesis is that when the assumptions underlying the likelihood construction are no longer valid (e.g. the noise is non-Gaussian, nonstationary, or contains transient noise artifacts), provided the noise and signal remain additive, the residual likelihood is easier to approximate with NLE than the full likelihood.

A. Simulation-based inference setup

We employ the SBI package's implementation of NLE to construct the RNLE algorithm, which we implement within the newly developed SBILBY framework [116]. Following the approach described in Sec. II, the package comprises two main components for the simulation and analysis of GW data.

The first component is specific to gravitational-wave parameter estimation and provides functionality to simulate Gaussian noise realizations from prescribed power spectral densities, as well as to access real detector noise. In this work, we always use pieces of data of 4 s, and the analysis is performed in the time domain. The pieces of data are whitened in the frequency domain and subsequently cut in the time domain to obtain training data for the conditional density estimators (see Sec. IV for the details). This training data, consisting of pairs of noise parameters and data $\{\lambda_i, \vec{d}_i\}$, is passed to the SBI package to train the conditional density estimator through the second component of our algorithm. This second component is independent of the particular inference problem and can be employed in any application for which separate generative models of the noise and signal processes are implemented. If the residual noise model is not explicitly parametrized, an unconditional density estimator would in principle be sufficient. In this work, however, we employ a conditional estimator by adopting the SBI implementation of NLE, thereby leveraging its well-established framework for conditional density estimation based on parameter-data pairs. This choice enables a direct integration of RNLE within the existing SBI and Bilby ecosystems, avoiding the need for a separate custom implementation. While this introduces a modest increase in computational cost, it provides a practical proof-of-concept realization of RNLE within a standard inference pipeline. At the same time, this formulation naturally accommodates future extensions in which the noise model is explicitly parametrized through additional hyperparameters. Therefore, throughout this work, we employ the masked autoregressive flow (MAF) as the conditional density estimator [122].

A MAF models the joint probability density of the parameters by decomposing it into a product of one-dimensional conditional densities, each parametrized by an autoregressive neural network (ARNN) [123,124].

The ARNNs provide a flexible and tractable parametrization of conditional probability distributions. In ARNNs, the joint probability of a multivariate variable $\mathbf{x} = (x_1, \dots, x_D)$ is factorized as

$$p(\mathbf{x}) = \prod_{i=1}^D p(x_i|x_{1:i-1}), \quad (12)$$

where each conditional distribution $p(x_i|x_{1:i-1})$ is represented by a neural network whose inputs depend only on the preceding components $x_{1:i-1}$. This autoregressive structure enables the model to efficiently capture statistical dependencies among parameters while maintaining exact likelihood evaluation. In our framework, the ARNN sequentially transforms samples from a base distribution into complex likelihood surfaces $\mathcal{L}(\vec{d} - m(\vartheta)|\lambda, M_s)$. To ensure reproducibility, we employ the default architecture provided in SBI v0.22.0. In this configuration, the likelihood $p(x|\theta)$ is modeled using a stack of five autoregressive flow transforms, each parametrized by a masked autoencoder for distribution estimation network [125] with two hidden layers of 50 units and rectified linear unit activation functions [126].

The resulting conditional density estimator integrates directly with the Bilby framework, as the NLE implementation is designed to be fully compatible with Bilby's inference interface. For sampling, we employ DYNESTY. At each likelihood evaluation, we subtract a realization of the model $m(\vartheta)$, with ϑ drawn from the prior, from the analyzed data, and evaluate the likelihood on the resulting residuals $r(\theta) = \vec{d}(\theta) - m(\vartheta)$. Although the likelihood directly depends only on the residual parameters λ , we obtain samples for the model parameters as a natural byproduct of this subtraction process. The modular design of our algorithm allows the user to specify custom signal and noise generation functions, ensuring its applicability to a broad class of inference problems beyond gravitational-wave analyses.

Unless otherwise stated, we use the default MAF architecture and training settings provided by SBI v0.22.0, including a training batch size of 50 and a validation fraction of 0.1. The number of epochs is not fixed in advance, and training is terminated automatically by the package's early-stopping criterion when the validation loss fails to improve for 20 epochs.

III. RESULTS: TOY MODEL

To test the validity of RNLE, we begin with the analysis of a simple timeseries problem consisting of a sine-Gaussian toy model following the equation,

$$\mathcal{F}(t, \alpha, f) = A \exp(-(t/\alpha)^2) \times \sin(2\pi f t), \quad (13)$$

where t is the time, and f , A , and α are free parameters that we will refer to as frequency, amplitude, and standard

deviation. We simulate observational data by adding white Gaussian noise sampled from a normal Gaussian distribution with standard deviation σ to a realization of Eq. (13) for specific values of the free parameters. The value of σ is drawn from a uniform prior $\mathcal{U}[0, 2]$. We aim to analyze this observation using likelihoods generated with both the canonical NLE algorithm and our newly implemented RNLE algorithm, and compare the results with those obtained using the correct Gaussian likelihood, following Eqs. (3) and (4). We employ the DYNESTY nested sampling package implemented in Bilby for the posterior sampling. For the training of the NLE conditional density estimator, we employ simulations of white Gaussian noise added to simulations of the signal, employing the priors in Table II in the Appendix. To train RNLE, we employ simulations of white Gaussian noise only, varying the σ parameter.

We perform benchmarking against the Gaussian likelihood first, using a restricted parameter space where f and A are fixed, and then sampling in the full three-dimensional signal space. To assess how well the conditional density estimator recovers the true likelihood, we compute the Jensen-Shannon (JS) divergence between the posterior probability density functions obtained with the (R)NLE and the Gaussian likelihood [127]. For each comparison, we draw 3000 samples from both posteriors, fit a kernel density estimator to each sample set, and evaluate the corresponding probability densities. The JS divergence is computed using the SCIPY implementation [128], which returns the JS distance measured in bits, i.e., using a base-2 logarithm so that $\max(\text{JS}) = 1$. For our analysis, we use the corresponding JS divergence, obtained by squaring the returned distance. This procedure is repeated one hundred times, and the reported values are averages over these trials. We consider two posteriors indistinguishable if their JS divergence is smaller than $10/N$, where N is the number of samples used for the evaluation, e.g., with $N = 3000$ the threshold is 0.003 [28].

The results of the benchmarking are shown in Fig. 1, for two dimensions in the left panel and four dimensions in the right panel. The plots show the variation of the JS divergence value with the number of simulations used in the training for NLE and RNLE against the Gaussian likelihood in green and orange, respectively. From the plots, we see that the JS values for the signal parameter posteriors recovered with RNLE are systematically lower than the NLE ones. The same is not true for the noise parameter σ , for which the values are consistent in the two approaches. The better performance in the RNLE case is related to the lower dimensionality of the training parameter space, one for RNLE and two (four in the right panel) for NLE. We would expect the JS values for the NLE results to eventually converge to the RNLE ones when further increasing the number of simulations. Those results show that RNLE is a viable method to produce posterior probabilities and requires fewer simulations to achieve the same accuracy as NLE. The results also confirm the statistical equivalence of the posterior distributions obtained with the Gaussian likelihood and the one obtained with the RNLE likelihood.

In Fig. 2, we show the time required for a single likelihood evaluation for the NLE, in green, and the RNLE, in orange, trained likelihoods. Similarly to Fig. 1, for the left panel we employed two free parameters, and for the right panel we employed four. The likelihood evaluation time is expected to be independent of the number of simulations used to train the density estimator, since the trained models have the same architecture. The small differences visible in Fig. 2 are due to run-to-run fluctuations and do not indicate a systematic dependence on training-set size. We observe that the computational cost of RNLE likelihood evaluations is comparable to that of standard NLE. While RNLE involves an additional operation, namely the explicit subtraction of a signal realization from the data prior to likelihood evaluation, this overhead

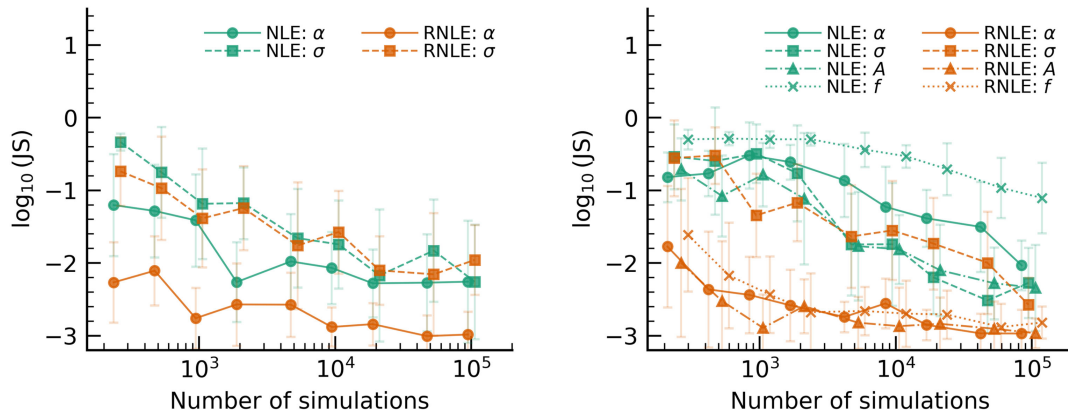


FIG. 1. Logarithm of the Jensen Shannon divergence value for the posterior distributions obtained with the likelihoods trained with NLE, in green, and RNLE, in orange. The squares stand for the noise related parameter σ , the circles, triangles, and crosses for the variance, frequency, and amplitude of the sine-Gaussian signal, respectively. The two-dimensional results are shown in the left panel, while the right panel shows the four-dimensional results.

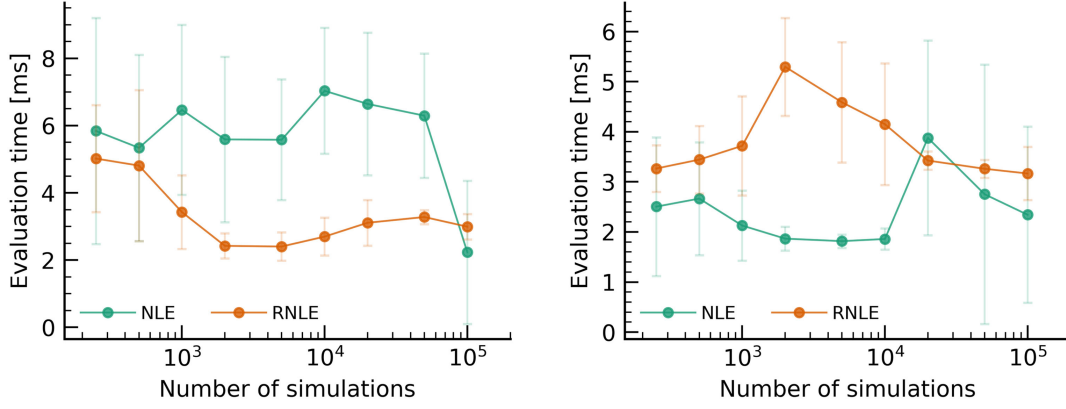


FIG. 2. Likelihood evaluation time in milliseconds for NLE (green) and RNLE (orange) versus the number of simulations used in training. We employ two free parameters in the left panel and four in the right panel. As expected for models with fixed architecture, the likelihood evaluation time shows no systematic dependence on the number of training simulations; the small differences are consistent with run-to-run fluctuations.

does not translate into a systematic increase in total runtime for the toy model considered here, where the forward model evaluation is computationally inexpensive. For gravitational-wave analyses, however, the model evaluation, i.e., the waveform generation, typically constitutes the dominant contribution to the computational cost in standard Bayesian sampling. In this context, RNLE retains the cost of waveform evaluation and introduces an additional overhead associated with the neural likelihood evaluation, making it more expensive than regular sampling. In contrast, once trained, NLE-based inference does not require waveform evaluations during sampling and therefore can be substantially faster. In practice, for our toy-model the overall computational performance depends on several factors, including the hardware architecture, the total number of likelihood evaluations performed by the sampler, the effective sampling efficiency, and the average evaluation time per likelihood call. When accounting for these effects, we find that the total wall time of RNLE-based analyses is broadly similar to that of NLE, with run-to-run fluctuations that are comparable in magnitude for the two methods. As a result, the implementation of RNLE does not incur a significant computational penalty.

IV. CONFIGURING RNLE FOR EFFICIENT GW ANALYSIS

To apply the RNLE algorithm to the analysis of real gravitational wave events, we start by testing it on synthetic data. We create an observation by combining coloured-Gaussian noise, generated from a fixed PSD, and an injected GW waveform produced with the IMRPhenomPv2 waveform approximant [129–132]. The injection values for the signal parameters are listed in Table III. We simulate the Gaussian noise to ensure that the assumptions for the validity of the Whittle likelihood approximation are

satisfied, using the publicly available high-sensitivity PSD curve presented in Abbott *et al.* [133] with a horizon distance of 180 Mpc. To effectively compare the results between our RNLE implementation and Bilby, we ensure that we train the conditional density estimator on data processed as done internally by the Bilby algorithm [134]. In practice, this means that we ensure that our Gaussian noise simulations are properly windowed. Different from a standard Bilby analysis, we are operating in the time domain and therefore need to whiten the data simulated in the frequency domain too. We fix the sampling frequency of the data to 4096 Hz. Since we are operating in the time domain, the whitening procedure for each simulation comprises three main steps [45]:

- (1) A Fast-Fourier transform to obtain the frequency domain representation, $h(f)$, of our time domain data, $h(t)$, after properly windowing it using a Tukey window [135] given by Eq. (11) in Talbot *et al.* [134]. For a data length of 4 s, we employ a roll-off of 0.2 s and a tapering parameter value $\alpha = 0.1$ following the Bilby default in version 2.6.0;
- (2) Whitening the frequency domain signal, after band-passing the data with a frequency mask f_{mask} , between 20 and 2048 Hz, that selects only the frequencies retained in the analysis,

$$\tilde{h}(f) = \frac{h(f)}{\sqrt{P_j \times \frac{1}{N} \sum_{n=0}^{N-1} w[n]^2 \times \frac{d}{4}}}, \quad (14)$$

where d is the duration of the whitened data in seconds, N is the number of data samples, the index n runs over the elements of the window, and P is the PSD.

- (3) An inverse Fast-Fourier transform is used to recover the whitened time domain signal,

$$h_w(t) = \text{IFFT}(\tilde{h}(f)) \times \frac{N_f}{\sqrt{\sum f_{\text{mask}}}}, \quad (15)$$

where N_f is the number of frequency bins. The normalization factor $(\sum f_{\text{mask}})^{-1/2}$ compensates for the fact that the mask sets excluded frequencies to zero, thereby ensuring that the overall amplitude of the whitened time-domain signal remains properly normalized.

The whitened time-domain strain $h_w(t)$ is multiplied by an auxiliary scale parameter σ . This parameter is introduced for implementation reasons as the NLE interface used in this work requires each simulated dataset to be paired with at least one latent parameter drawn from a prior, whereas the noise-only simulations used to train RNLE do not otherwise carry an explicit parameter label. In this sense, σ should be regarded as an auxiliary nuisance parameter rather than as a physical parameter of the detector noise model. We choose a broad prior $\sigma \sim U[0, 2]$ so that the training procedure remains well posed while modifying the noise realizations in a minimal way. Rescaling by σ preserves the morphology and relative structure of the residual time series, including transient features, and therefore provides a simple conditioning variable that does not alter the basic RNLE construction. The use of σ is thus a pragmatic consequence of the present conditional-flow implementation, not a conceptual requirement of residual likelihood estimation itself. In future work, it would be interesting to replace this auxiliary parameter either with an unconditional density estimator or with physically motivated noise hyperparameters. To mitigate numerical artifacts introduced by the whitening procedure at the edges of the time series and to reduce data dimensionality, we

truncate the data. As we have to accordingly crop the observation when analyzing it with the RNLE likelihood, we make sure that we are not cutting away any part of the time series that includes the signal and contributes information to the inference. In Fig. 3, we present a schematic overview of the workflow used to obtain posterior distributions with RNLE from a set of training data realizations—either simulations or real detector data—and a given observation. For validation purposes, the observation consists of a simulated GW signal injected into colored Gaussian noise, whereas in the following sections it corresponds to a GW signal injected into real detector noise. Both the training data and the observations are preprocessed using an identical procedure. The training data realizations are passed to the conditional density estimator, with the settings described in Sec. II A, yielding a likelihood model. For each evaluation of the RNLE likelihood, the GW parameters are sampled from the priors listed in Table IV, and a corresponding frequency-domain waveform is generated using the same waveform model employed for the observation generation. The waveform is whitened and transformed into the time domain, and subsequently subtracted from the whitened time-domain observation. The resulting residual time series is then passed to the likelihood for evaluation. This procedure is repeated at each likelihood evaluation within the nested sampling algorithm, which ultimately produces posterior samples for the signal parameters and the noise residual parameter σ .

A. Analysis of 1D GW signal

To validate the RNLE framework for gravitational-wave parameter estimation, we first perform a simplified

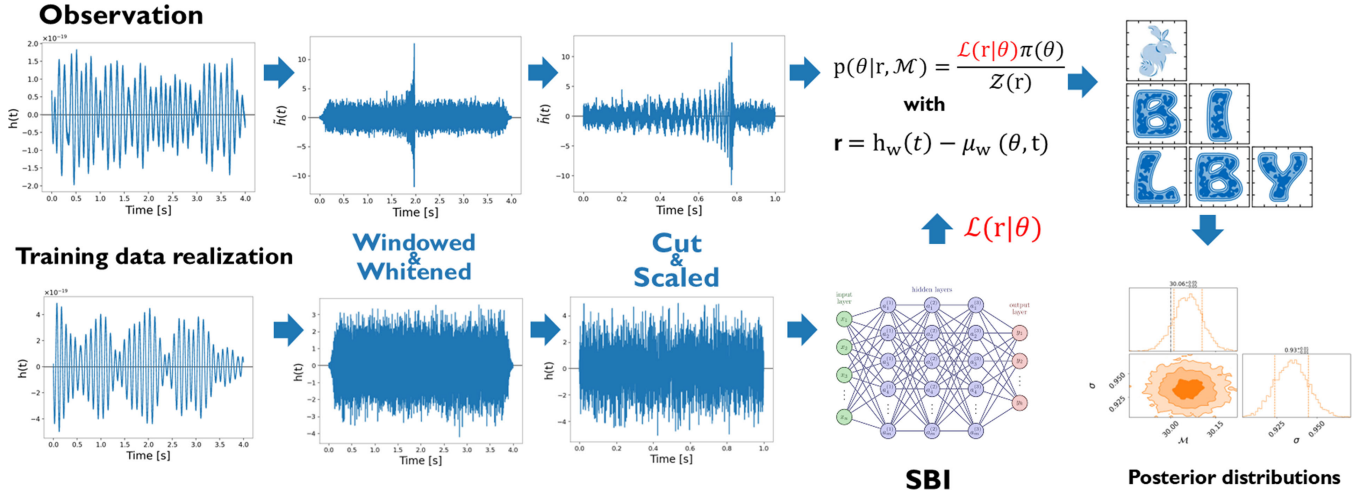


FIG. 3. Schematic overview of the workflow we employ to produce posterior distributions from a set of training data realizations and an observation using RNLE. The preprocessing for the observations, top row, and the training data, bottom row, is the same. We input the training data into the SBI conditional density estimator, which trains a likelihood. This is used in SBILBY with the DYNASTY sampler to perform inference and obtain posterior distributions. At every evaluation of the likelihood, before passing the data to the RNLE likelihood object we subtract a whitened GW signal realization from it.

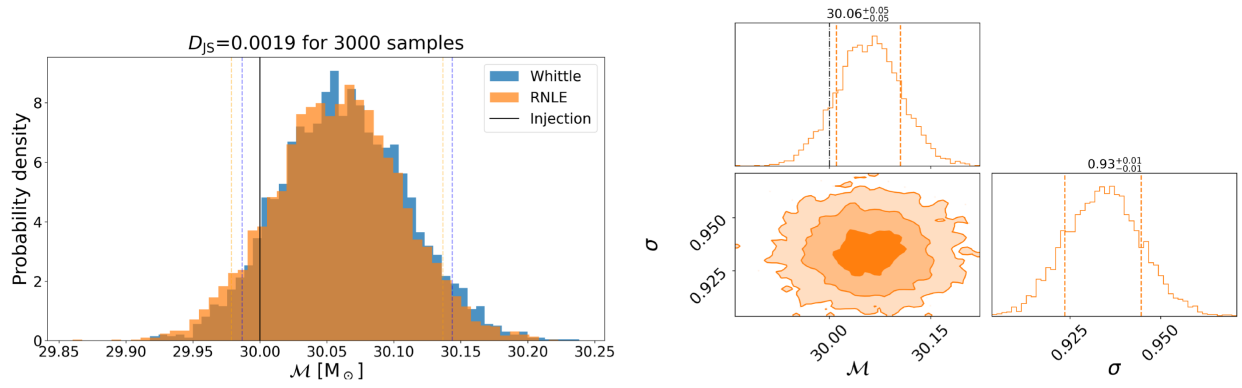


FIG. 4. Left: Comparison of the posterior distribution of the chirp mass $\mathcal{M}$ obtained using the RNLE likelihood trained using 3000 simulations of coloured-Gaussian noise, in orange, and the Whittle likelihood, in blue. The JS divergence value is reported in the title, confirming the agreement between the two results. Right: Full posterior of the RNLE analysis. The histograms show the posterior distributions for the noise scaling parameter σ and the chirp mass. The dash-dotted black line indicates the injected values.

one-dimensional analysis, focusing on a single signal parameter. In this controlled setup, we test whether RNLE can accurately recover posterior distributions consistent with those obtained using the standard Whittle likelihood. Restricting the analysis to the detector-frame chirp mass allows us to systematically explore the algorithm's performance, identify optimal configuration settings, and significantly reduce computational cost.

In the left panel of Fig. 4, we show the comparison between the posterior distribution of the chirp mass [136] obtained with RNLE and that obtained using the Whittle likelihood under identical sampler settings, observation, and priors. We perform a JS divergence test, following the specifics outlined in Sec. III. The resulting value is below the chosen threshold, confirming that the two posterior distributions are statistically consistent.

In the right panel of Fig. 4, we show a corner plot for the RNLE likelihood analysis. The posterior distributions for the auxiliary scaling parameter σ and the chirp mass are approximately uncorrelated. Here, $\sigma = 1$ corresponds to the nominal unrescaled residuals used in the data generation. Since σ is an auxiliary parameter introduced in the present conditional-density implementation, rather than a physical parameter of the signal or detector model, small offsets from $\sigma = 1$ can reflect finite-training effects and approximation error in the learned likelihood.

B. Dependence on training dataset size

To assess how the accuracy of the RNLE likelihood depends on the size of the training dataset, we investigate the convergence of the inferred posterior distributions as a function of the number of training data realizations. In particular, we quantify how the agreement between the RNLE and Whittle likelihood results improves as additional training data are provided to the conditional density estimator.

In Fig. 5, we show the JS divergence between the chirp mass posterior distributions obtained with RNLE and with the Whittle likelihood as a function of the number of simulations used for training. The blue dotted line indicates the JS divergence threshold corresponding to statistical equivalence between the two distributions. The black squares denote the training time required for the RNLE likelihood as a function of the number of simulations. A linear fit in log-log space is performed to illustrate the scaling behavior, showing that the training time approaches linear scaling for $N_{\text{sim}} \gtrsim 10^4$.

The reported training times were obtained using eight CPU cores with the settings described in Sec. II A. From these results, we observe that the JS divergence stabilizes below the equivalence threshold once more than approximately 2000 simulations are used for training, indicating that the RNLE likelihood can reliably reproduce the Whittle likelihood results beyond this point. At the same time, the training time increases rapidly with the size of the dataset, ranging from below one hour for ~ 2000 simulations to several hours for 10^5 simulations. Based on this trade-off between accuracy and computational cost, we adopt 3000 simulations for all subsequent analyses involving a one-dimensional signal prior, unless stated otherwise.

C. Analysis of training noise

To investigate the stability of the conditional density estimator and to quantify the training noise, i.e., the uncertainty associated with the training process itself, we perform an ensemble analysis in which multiple RNLE likelihoods are trained independently. For each chosen number of simulations, we construct three training datasets by drawing independent realizations of coloured Gaussian noise from the same PSD. Each dataset is used to retrain the conditional density estimator with different optimization seeds, allowing us to isolate variability arising from both

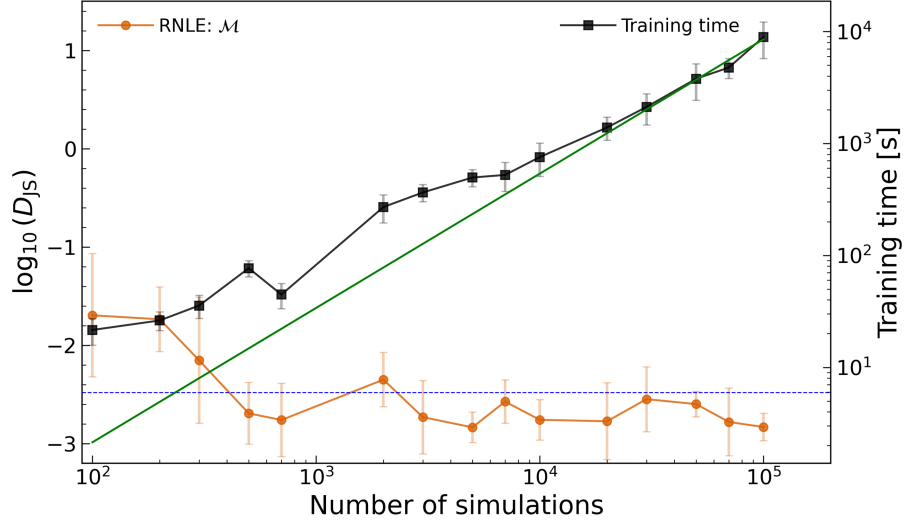


FIG. 5. In orange, the logarithm base ten of the JS divergence value for the chirp mass posteriors, computed between the Whittle and RNLE results, against the number of simulations employed to train the conditional density estimator. The blue dotted line indicates the JS divergence threshold signifying the statistical equivalence between the two distributions. The black squares show the RNLE likelihood training time against the number of simulations. We use a logarithmic scale for the x axis and the right y axis. The error bars are computed over five realizations of retraining and sampling. The green line is obtained using the optimal values from a linear fit in log-log space for the training time versus the number of simulations.

the finite training set and the stochastic nature of the training procedure. Since all training datasets are drawn from the same underlying noise-generating process, we expect that, in the limit of sufficiently large training sets, variations in the specific realizations should not affect the learned likelihood in a systematic way. Any residual variability can therefore be attributed to the stochasticity of the training procedure itself. In Fig. 6, we show violin plots comparing the posterior distribution of the chirp mass obtained with the Whittle

likelihood to those obtained from the different realizations of the RNLE likelihood, for increasing numbers of training simulations. The red dotted line indicates the true chirp mass value of the injected GW signal. Each set of violins corresponds to three independently trained RNLE likelihoods using the same number of simulations.

To quantify the spread among the ensemble members, we compute the log-Bayes factor for each RNLE likelihood with respect to the Whittle likelihood when analyzing the

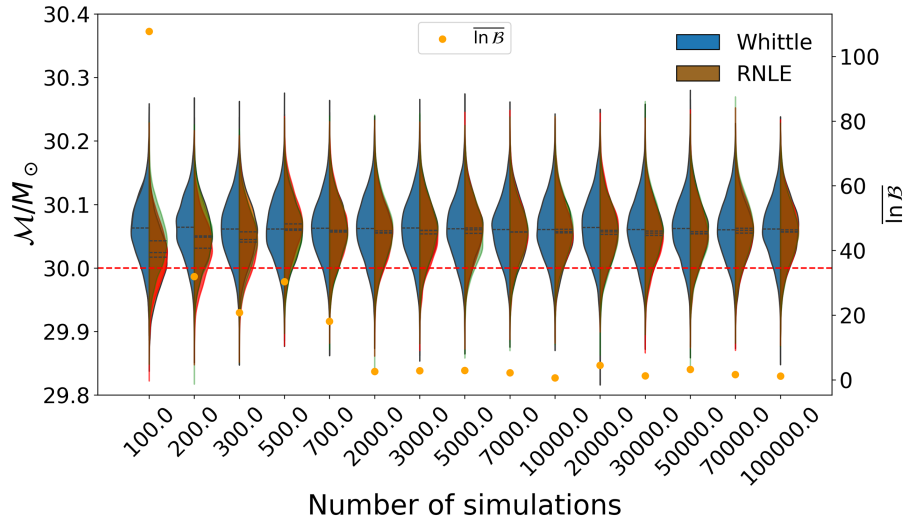


FIG. 6. Violin plot showing the posterior distributions for the chirp mass varying the number of simulations used to train the conditional density estimator. For each violin, the left side in blue, shows the posterior obtained with the Whittle likelihood, while the right side shows the RNLE posteriors obtained with different likelihoods trained on the same number of simulations. Each likelihood realization is trained with a different optimization seed and a different, but statistically equivalent, set of simulations. The red dashed line indicates the true value of the chirp mass for the injected signal. The orange dots represent the mean of the log-Bayes factor computed for the different realizations of the RNLE likelihood against the number of simulations used in the training.

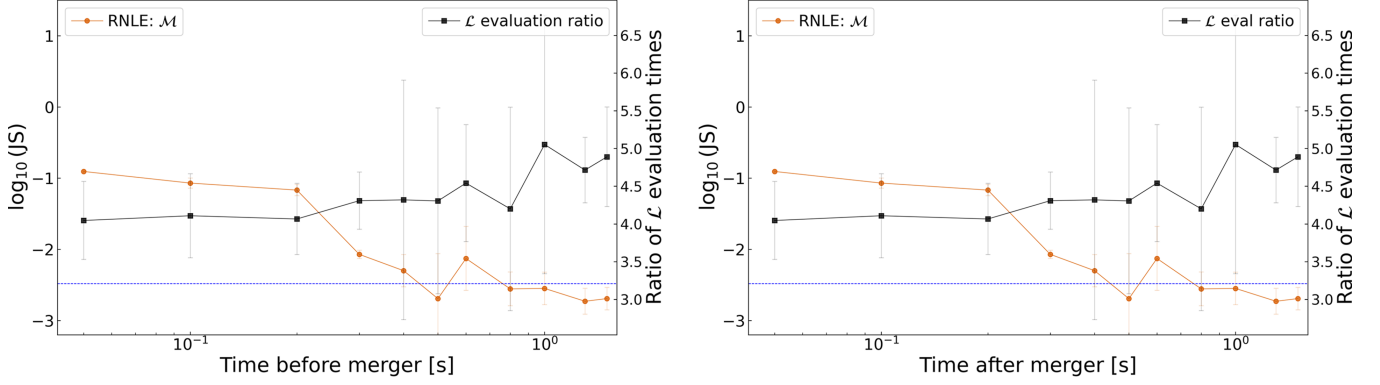


FIG. 7. In orange, the logarithm base ten of the JS divergence value for the chirp mass posteriors against the time included before the merger, left panel, and after the merger, right panel. The blue dotted line indicates the JS divergence threshold signifying the statistical equivalence between the two distributions. The black squares show the ratio between the time needed for a single likelihood evaluation with RNLE and the Whittle likelihood. We use a logarithmic scale for the x axis.

same observation. The orange points in Fig. 6 show the mean of the log-Bayes factors across the ensemble for each training size. We observe that this quantity decreases as the number of simulations increases and stabilizes for training sets with more than approximately 2000 simulations. This behavior suggests that the variability associated with training noise decreases as the size of the training dataset increases, indicating a progressive stabilization of the learned likelihood. This trend is consistent with the results shown in Fig. 5, supporting the choice of $\mathcal{O}(10^3)$ simulations as a reasonable compromise between accuracy and computational cost. A complementary discussion of training noise is presented in Sec. IX, where we also explore a strategy to combine the posteriors obtained from different likelihood realizations, weighting each posterior by its associated Bayesian evidence.

D. Dependence on analysis window size

To quantify how the accuracy of the RNLE likelihood depends on the size of the analyzed data segment, we study the sensitivity of the recovered posterior distributions to variations in the analysis window. In particular, we assess how reducing the amount of data before or after the merger affects the agreement between RNLE and the Whittle likelihood, as measured by the JS divergence of the chirp mass posteriors.

In Fig. 7, we show the JS divergence as a function of the time interval retained before the merger (left panel) and after the merger (right panel) for the RNLE analysis. In each case, the complementary portion of the data segment is held fixed: when varying the premerger duration, we include a fixed 0.3 s of postmerger data, and conversely, when varying the postmerger duration, we include 0.8 s of premerger data. The benchmarking is performed against Whittle likelihood results obtained using a four-second data segment, corresponding to two seconds before and two seconds after the merger. The dotted horizontal line indicates the JS divergence threshold below which the

two posterior distributions are considered statistically equivalent. The square markers denote the ratio between the computational time required for a single likelihood evaluation using RNLE and that required by the Whittle likelihood. While the per-evaluation cost of the RNLE likelihood is only moderately higher—by a factor of ~ 3 – 5 —this does not fully capture the total computational overhead of an end-to-end inference run. In practice, RNLE-based analyses require a larger number of effective likelihood evaluations, owing to the increased dimensionality of the parameter space resulting from the inclusion of an additional noise parameter, which in turn impacts sampling efficiency and convergence. When accounting for both the per-evaluation cost and the total number of likelihood calls performed by the sampler, we find that the overall wall time of RNLE analyses is approximately an order of magnitude larger than that of corresponding Whittle-likelihood runs. We emphasize, however, that this increased computational cost reflects the added flexibility and robustness of the learned likelihood, rather than an inherent inefficiency of the evaluation itself. From the left panel of Fig. 7, we observe that truncating the data too close to the merger removes essential signal information, preventing recovery of posterior distributions consistent with those obtained using the Whittle likelihood. In particular, retaining at least 0.8 s of data prior to the merger is required to recover statistically equivalent posteriors for this injection. This requirement depends on the signal parameters, most notably the chirp mass, as the time spent by the GW signal in the ground based interferometers’ detection band scales as $t_c - t(f) \propto \mathcal{M}^{-5/3}$, where t_c denotes the coalescence time [137].

V. ANALYZING SIMULATED BBHs IN COLORED GAUSSIAN NOISE

Having identified empirically well-performing settings for the training of the conditional density estimator in the one-dimensional case, we now extend our analysis to the

full parameter space of gravitational-wave waveform injections. This step allows us to assess the performance and robustness of the RNLE framework in realistic, high-dimensional inference scenarios. We adopt the injection parameters and prior distributions listed in Tables III and IV throughout this section. For all RNLE analyses presented here, we use a one-second data segment comprising 0.8 s preceding the trigger time and 0.2 s following it, extracted from the four-second stretch employed for whitening. To improve the numerical stability of the learned likelihood in

these higher-dimensional settings, we increase the number of training simulations to 3×10^4 . We emphasize that this choice has not been systematically optimized: while a smaller training set may be sufficient, we have not explored the minimum number of simulations required to ensure stable performance. Importantly, the training procedure remains fully independent of the underlying signal parameters. Figure 8 summarizes the results of three benchmark configurations with increasing complexity. These include: (i) a 10D parameter space for a precessing binary black hole

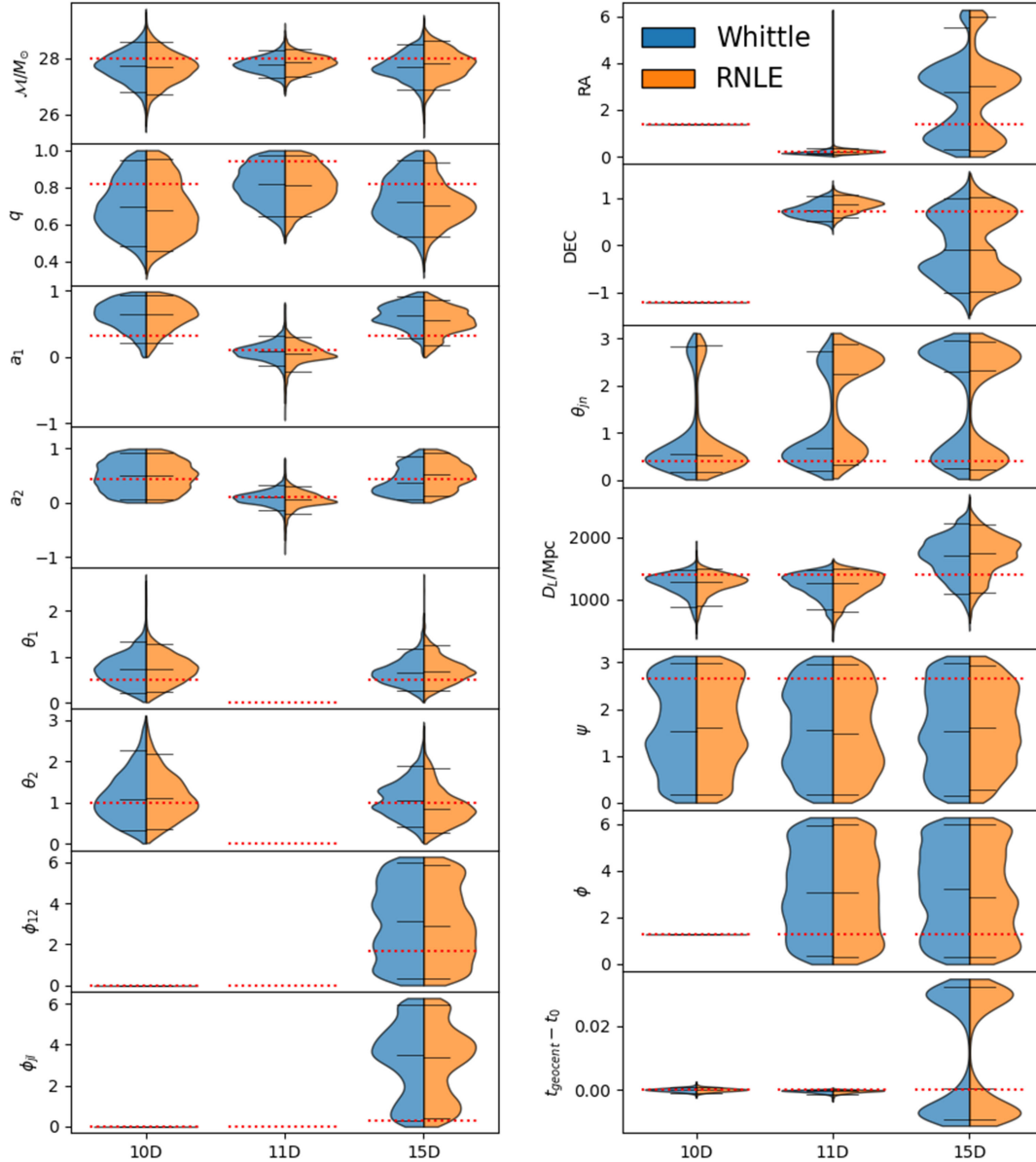


FIG. 8. Comparison of violin plots for the 10-, 11-, and 15-dimensional posterior distributions obtained with the RNLE and Whittle likelihood, in orange and blue, respectively. These were obtained from a precessing BBH merger with a restricted sampling parameter space, an aligned-spin BBH merger signal, and a precessing BBH signal sampling in the full parameter space, respectively. In all cases the GW signal was injected into colored Gaussian noise. The injection values of the signal parameters can be found in Table III and are overplotted as red dotted lines on the violins. JS divergence values for all the parameters are listed in the same table. We show the mean and 90% credibility intervals as black lines for each violin half. For fixed parameters we plot the posterior as a Dirac delta on the true value. The left panel shows the posteriors for the intrinsic parameters, while the right panel shows the extrinsic parameters.

(BBH) signal with a subset of extrinsic parameters fixed; (ii) an 11D aligned-spin BBH signal analyzed using data from two detectors; and (iii) a 15D precessing BBH signal. Each panel in Fig. 8 shows violin plots of the posterior distributions for the sampled parameters. For each parameter, we compare the posterior obtained using the Whittle likelihood, with the posterior obtained with RNLE. From left to right, the violins represent the 10D, 11D, and 15D analyses. The left panel of the figure displays the intrinsic source parameters: the detector-frame chirp mass $\mathcal{M}$, the mass ratio q , the spin magnitudes a_1 and a_2 , the tilt angles θ_1 and θ_2 between each black hole's spin and the orbital angular momentum, the azimuthal angle between the spin vectors ϕ_{12} , and the azimuthal angle between the total and orbital angular momenta ϕ_{jl} [22,23]. The right panel shows the extrinsic parameters: right ascension RA, declination DEC, the viewing angle θ_{JN} , the luminosity distance D_L , the geocenter time t_{geocent} , the polarization angle ψ , and the phase ϕ . The JS divergence values for all the parameters and analyses are reported in Table III in Appendix A. In the 10D analysis, the posterior distributions obtained with RNLE are statistically identical from those obtained with the Whittle likelihood across all parameters, as demonstrated by the JS values. This strong agreement demonstrates that, in moderately high-dimensional settings, RNLE can accurately reproduce the results of traditional likelihood-based inference. We additionally verified that the auxiliary parameter σ does not show strong correlations with the other sampled GW source parameters in the higher-dimensional analyses (a corresponding plot is included in the data release [138]). For the 15D precessing BBH system, small discrepancies are visible in a subset of parameters. These differences are consistent with stochastic sampling variability and reflect the well-known degeneracies and complex posterior structure characteristic of high-dimensional precessing BBH parameter spaces [139,140], rather than a systematic bias introduced by the RNLE method. For comparison, in Fig. 18 we show the JS divergence values between the two 15-dimensional analyses shown in Fig. 8 and values obtained comparing Whittle likelihood analyses employing the same data and settings but different sampling seeds.

To further disentangle potential algorithmic effects from model complexity, we repeat the analysis for an 11D aligned-spin BBH system using simulated data from the Hanford and Livingston detectors. In this configuration, separate conditional density estimators are trained for each detector using disjoint simulation sets, while identical PSDs are employed to ensure consistency in the whitening procedure. The total likelihood is constructed following Eq. (11). This modular approach allows the RNLE framework to scale naturally to larger detector configurations. As shown by the 11D violins, the agreement between the Whittle and RNLE likelihoods remains strong for the intrinsic parameters. This further confirms that the

deviations observed in the 15D case are not indicative of a deficiency in the RNLE framework. Some residual differences persist in the extrinsic parameters, most notably in right ascension, declination, and the viewing angle θ_{JN} , which exhibit multimodal posterior structures. These features are consistent with the sampling variability in complex, multimodal inference problems, and do not suggest a systematic bias, as underlined by the results shown in Fig. 18.

Overall, as is expected, these results validate the accuracy and reliability of the RNLE algorithm across a range of signal dimensionalities and detector configurations.

VI. ANALYZING SIMULATED BBHS IN QUASIGAUSSIAN OBSERVED DATA

Having verified that the RNLE algorithm can reproduce the results obtained with the Whittle likelihood under the assumptions of stationarity and Gaussianity, we next extend the method to data observed by the GW detectors. In this regime, however, agreement with the Whittle likelihood can no longer be interpreted as a validation criterion, since the Whittle likelihood is itself only an approximate model of the noise. The goal of the analyses in this and the following sections is therefore not to establish exact likelihood correctness through comparison with the Whittle likelihood, but rather to assess whether the RNLE-based inference pipeline remains calibrated and robust when applied to observational data. For clarity, experimental control, and to keep the analyses computationally tractable, we restrict the observational-data studies presented below to a single varying parameter. This allows us to isolate the impact of non-Gaussian noise and training-dataset design without conflating these effects with the additional degeneracies and sampling variability characteristic of high-dimensional GW parameter estimation. This choice should therefore be understood as a proof-of-principle simplification, rather than as a limitation of the RNLE framework itself.

We construct a pipeline that downloads publicly available strain data from the GW detectors [141,142], whitens it, and partitions it into one-second segments, following the procedure described in Sec. V and shown in Fig. 3. For each segment, whitening is performed with a PSD estimated from the preceding 16 seconds of data using a Hann² window and employing the median method with the GWPY package [143]. These segments are used as input data to train the conditional density estimator. When evaluating the RNLE likelihood, at each sampling step, the subtracted signal realization is whitened using the same PSD used to whiten the observation.

²We acknowledge that, for consistency with the data's whitening procedure, a Tukey window should be used. As this inconsistency was found in an advanced stage of the analysis, we tested that the effects on the results are negligible, see Appendix C, without further modifications (Fig. 19).

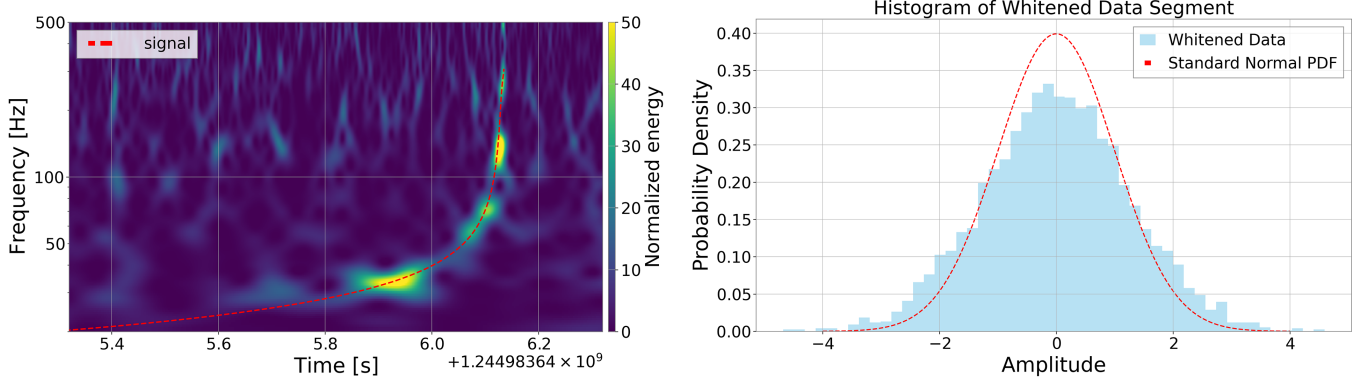


FIG. 9. Left: Time-frequency plot of 1 seconds of data from the Hanford detector around 2:00 a.m. UTC on 10 August 2019 used as the analysis window for RNLE. The data was selected for its quasi-Gaussian noise properties. A red-dashed trace is overplotted to the injected GW signal. The parameters of the injection are recorded in Table III. Right: Gaussianity test for the selected data. We compare the amplitude of the whitened data segment without the signal to a standard normal distribution, red dashed line.

A. Identification of quasi-Gaussian data segment

As an initial test, we consider a segment of data for which the noise is approximately Gaussian: the BNS range [144,145] is stable, the Omicron trigger rate is low [146,147], and no glitches with $\text{SNR} > 5$ are present. We arbitrarily select a 20-second data segment centered on the well-behaved data around 2:00 a.m. Universal Coordinated Time (UTC) on 10 August 2019, which is used to construct the observation and to estimate the PSD for whitening. To generate the training data realizations, we then use 30 minutes of detector data on either side of this central segment, explicitly excluding the 20-second window used for the observation. The data is divided into consecutive one-second segments, each paired with its corresponding PSD. Segments are randomly sampled to construct the training set, ensuring that each segment is used only once.

For the test observation, we inject the same aligned-spin BBH merger used in Sec. V into a segment excluded from

the training set. The resulting time-frequency representation is shown in the left panel of Fig. 9, where the injected signal track is overplotted as a red dotted line. The right panel displays a Gaussianity check of the data segment without the injection by comparing the histogram of whitened amplitudes to a normal distribution.

B. Analysing a BBH injection into a quasi-Gaussian data segment

After training the RNLE conditional density estimator on 3000 data segments, we compare the resulting likelihood with the Whittle likelihood evaluated on the observation shown in Fig. 9. The left panel of Fig. 10 presents the two likelihoods as a function of the chirp mass, holding all other parameters fixed. The RNLE likelihood is broadly consistent with the Whittle likelihood, with minor differences expected given the inherent non-Gaussian and non-stationary features of observational data. This behavior is mirrored in the one-dimensional posteriors obtained using

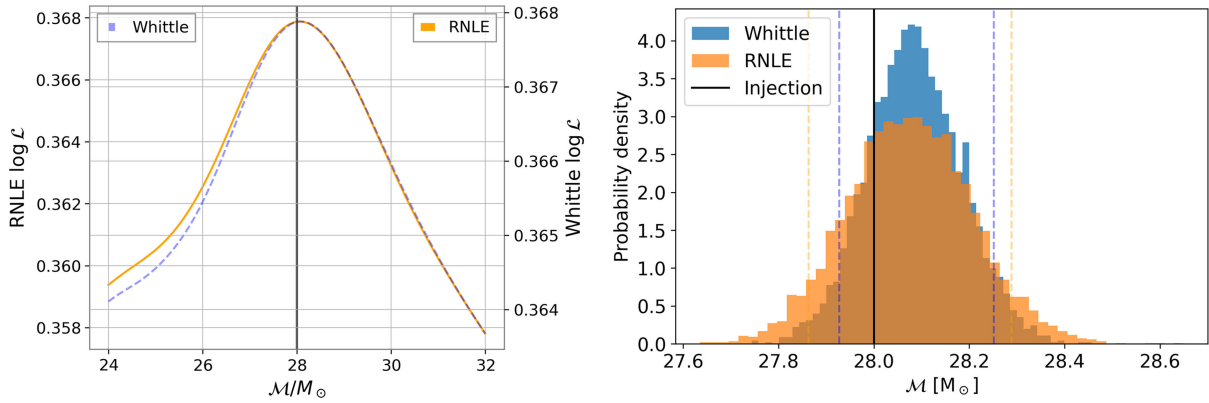


FIG. 10. Left: Comparison between the Whittle likelihood, in blue, and the learnt RNLE likelihood, in orange, for the selected data. Right: Comparison of the posterior probability distributions obtained with DYNesty using the Whittle likelihood, in blue, and the RNLE likelihood, in orange. The black vertical line indicates the injection value, while the dotted vertical lines represent the 90% credible intervals.

DYNesty within Bilby: the RNLE posterior is marginally broader, and the Jensen-Shannon divergence remains above the statistical-equivalence threshold, taking a value of 0.01. Nonetheless, both methods recover comparable median parameter values, demonstrating that the RNLE training and evaluation pipeline can operate reliably on observational data.

To further assess the robustness and calibration of our implementation, we perform a probability-probability (PP) test [28]. We carry out one hundred independent analyses in which both the data segment used for the injection and the injected chirp-mass value are varied, while keeping the trained RNLE likelihood fixed. The chirp mass values are drawn from a uniform prior $\mathcal{U}[26, 30]$, and each injection is analyzed using the same inference settings. For this test, the RNLE likelihood is first trained on 3000 noise realizations drawn from data surrounding the segment used for the observation analyzed in Fig. 4. Figure 11 shows the resulting PP plot, where the x axis denotes the nominal credibility level and the y axis indicates the fraction of events for which the true parameter value lies within the corresponding posterior credible interval. For a statistically well-calibrated inference procedure, the PP curve is expected to follow the diagonal.

The PP plot obtained with 3000 training realizations passes the calibration test, yielding a p value of $p_{\mathcal{M}} = 0.113$. However, the curve exhibits a mild S-shaped deviation from the diagonal, indicative of slightly under-constrained posteriors, matching the visibly broader 90% contours in the right panel of Fig. 10.

We view calibration studies of this kind as the most relevant practical validation strategy, especially in regimes where the data, the true likelihood is unknown. For a deployable RNLE analysis in real detector noise, one

would ideally perform a substantially larger injection campaign using representative data from a given observing run and assess calibration across a broad range of noise realizations, including a statistically meaningful subset containing strong transient artifacts.

VII. ANALYZING BBH INJECTIONS IN EXTREMELY NON-GAUSSIAN DATA SEGMENTS

We have demonstrated that RNLE recovers results consistent with the Whittle likelihood in all scenarios where the assumptions of stationarity and Gaussianity remain valid. We now turn to regimes in which the Whittle likelihood is expected to be biased. In this setting, comparison with the Whittle likelihood is no longer a validation benchmark for correctness, but it remains a useful baseline against which to assess the robustness of RNLE-based inference. The analyses in the remainder of this work should therefore be interpreted as proof-of-principle tests of performance in strongly non-Gaussian data, rather than as a complete validation of the learned likelihood. To this end, we select two segments of detector data containing loud glitches ($\text{SNR} > 20$), inject a gravitational-wave signal on top of each glitch, and attempt to recover the injected chirp mass using DYNesty in combination with a likelihood obtained via RNLE, while keeping all other analysis settings fixed. As in the previous sections, we restrict the inference to a single varying parameter in order to isolate the effect of non-Gaussian noise and clearly demonstrate the viability of the RNLE approach. Although we focus here on a one-dimensional parameter space for clarity, the method is in principle applicable to higher-dimensional analyses. This expectation is motivated by the fact that the RNLE training procedure depends only on the noise properties of the data and is independent of the signal model, as discussed in the previous section. As a result, extending the inference to additional signal parameters would not require retraining the likelihood. Nevertheless, demonstrating the practical performance and scalability of the method in higher-dimensional signal parameter spaces remains an important direction for future work.

A. Identification of a highly non-Gaussian data segment

To identify suitable periods of elevated glitch activity, we analyze data from the LIGO Hanford observatory (LHO) during the second half of the third observing run (O3b) using a 24-hour sliding window, shifted forward by one hour at each step. For each window, we quantify glitch occurrence by extracting the number of entries recorded in the GravitySpy database [148–150]. Figure 12 shows the resulting glitch counts as a function of time across the O3b.

From this distribution, we investigate the properties of time windows associated with elevated glitch rates. Our goal is to analyze an extreme scenario in which the inferred chirp mass is severely biased. For this reason, the time

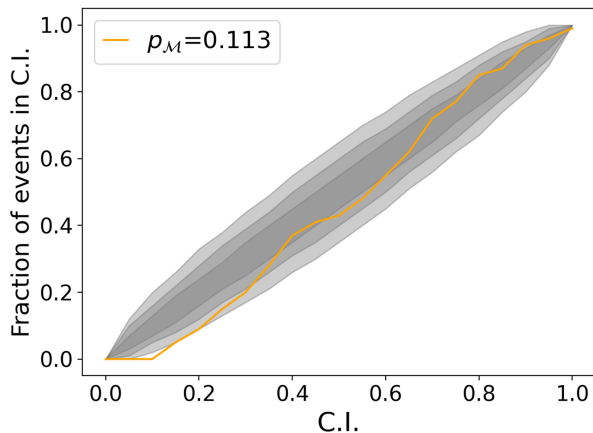


FIG. 11. Probability–robability plot for BBH injections into observational data over 100 realizations. We use one RNLE likelihood model for all runs, drawing the chirp mass uniformly from $\mathcal{U}[26, 30]$ for each injection and embedding each signal in a different detector-data segment. The results shown in the left panel are obtained with an RNLE likelihood trained on 3000 data realizations.

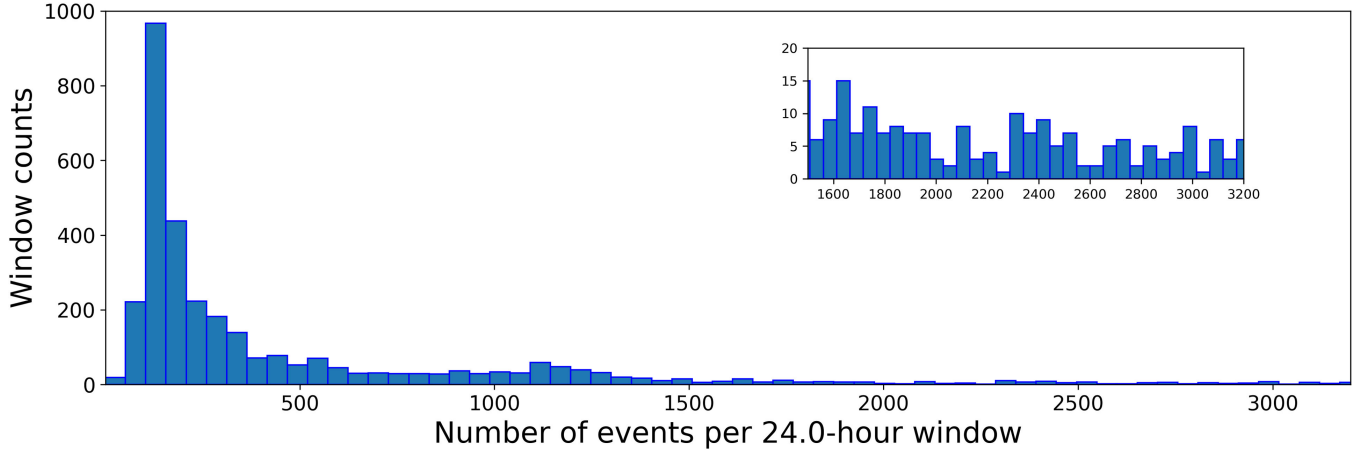


FIG. 12. Histogram of the number of 24-hour windows for each glitch count value in LHO during O3b. Glitch counts are computed using a 24-hour sliding window with a one-hour step, based on the GravitySpy database. The horizontal axis shows the number of glitches within a 24-hour window, while the vertical axis reports the number of windows exhibiting a given glitch count. The inset highlights the windows with the highest glitch rates, from which the analysis window used in this work is selected.

window used in our study is selected based on both the overall glitch rate and the tendency of individual glitches within that period to bias the inference, ensuring that the sample is representative of a strongly non-Gaussian noise environment. We identify the period of November 9–12, 2019 as fitting our requirements. The characteristics of the glitches identified by GravitySpy during this interval are shown in the top panels of Fig. 13.

Within this window, we concentrate on two particularly loud glitches occurring at GPS times 1257456281.4 and 1257541955.8, corresponding to 21:24:23 UTC on November 10, 2019 and 21:12:17 UTC on November 11, 2019, respectively. For clarity, we refer to these as Glitch A and Glitch B. Glitch A has an SNR of 39.94, a duration of 2 s, and a central frequency of 360 Hz, while Glitch B has an SNR of 44.37, a duration of 1.5 s, and a central frequency of 448 Hz. Time–frequency representations of both glitches are shown in Fig. 14. Although GravitySpy classifies both glitches as *Koi fish*, their morphology is more complex than typical examples of this category. As seen in the top panels of Fig. 13, these events also correspond to unusually long durations compared to other glitches in this period, and nearly 200 *Koi fish* glitches occur within the selected time window.

B. Analyzing a BBH injection into a highly non-Gaussian data segment

We extract 3600 s of strain data around each glitch and process these segments following the procedure described in Sec. VI. This yields one simulation per second of data, providing a dataset from which we train an RNLE likelihood for each glitch independently. We then inject a GW signal, using the same source parameters as in Sec. V, into the data segment containing the glitch. The trigger time of the injection is shown as a red dashed vertical line in the lower panels of Fig. 14. Throughout this discussion, we

define the offset as the relative temporal separation between the glitch and the injected GW signal. A positive offset therefore, indicates that the glitch occurs before the signal, whereas a negative offset means that it occurs after the signal. We take the time of the glitch to be the event time provided by GravitySpy. The simulated signal is always placed at the trigger time, i.e., 0.8 seconds after the start of the analyzed segment and 0.2 seconds before its end.

The posterior distributions shown in the bottom panels of Fig. 14 correspond to the case of zero offset, in which the signal and glitch are temporally coincident. We display the chirp-mass posterior recovered with the Whittle likelihood in blue and the posterior obtained using the RNLE likelihood in orange, for Glitch A, left panel, and Glitch B, right panel. In this analysis, the chirp mass is the only free parameter. For both glitches, the RNLE likelihood yields a posterior distribution that more accurately reflects the injected chirp mass compared to the Whittle likelihood. For Glitch A, the injected value, vertical black line, lies within the 90% credible interval, dotted lines. For Glitch B, although the 90% interval does not fully contain the injected value, the full posterior contains it, while the Whittle posterior rails against the upper prior bound. These results provide encouraging proof-of-principle evidence that RNLE can improve the robustness of parameter recovery in the presence of strongly non-Gaussian noise. However, the performance depends on several factors, including the chosen dataset, the relative timing between the glitch and the signal, the signal-to-noise ratios of both, and the number of simulations used during training.

VIII. ANALYZING BBH INJECTIONS AROUND A BLIP GLITCH

We now investigate the performance of the RNLE algorithm in the presence of glitches, focusing in

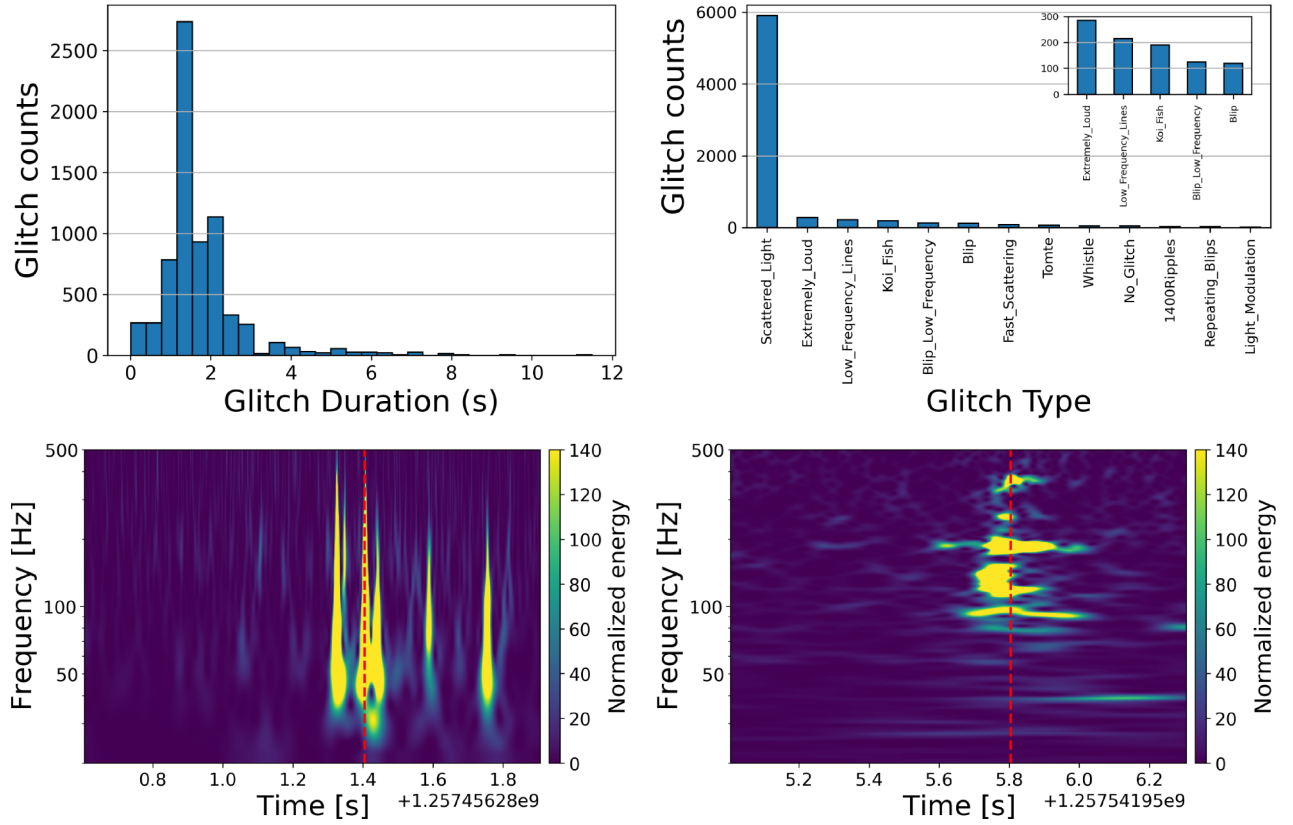


FIG. 13. Top panels: Analysis of the properties of glitches in a 30-hour window in LHO from O3b around the two selected glitches. In the left panel the glitch duration versus the glitch count. In the right panel, a histogram of the glitch labels assigned by GravitySpy. The inset highlights the counts for the *Koi fish* glitches, which correspond to the glitches analyzed in this study. Bottom panels: Time-frequency plot of 1 second of data from the Hanford detector to visualize the glitches selected for the study. These glitches stand out for their long duration, high SNR, and frequency content in the injected GW signal's band. The red dotted vertical line indicates the trigger time of the injection, corresponding to the coalescence of the BBH system and the moment when the (2,2) mode of the waveform attains its maximum amplitude.

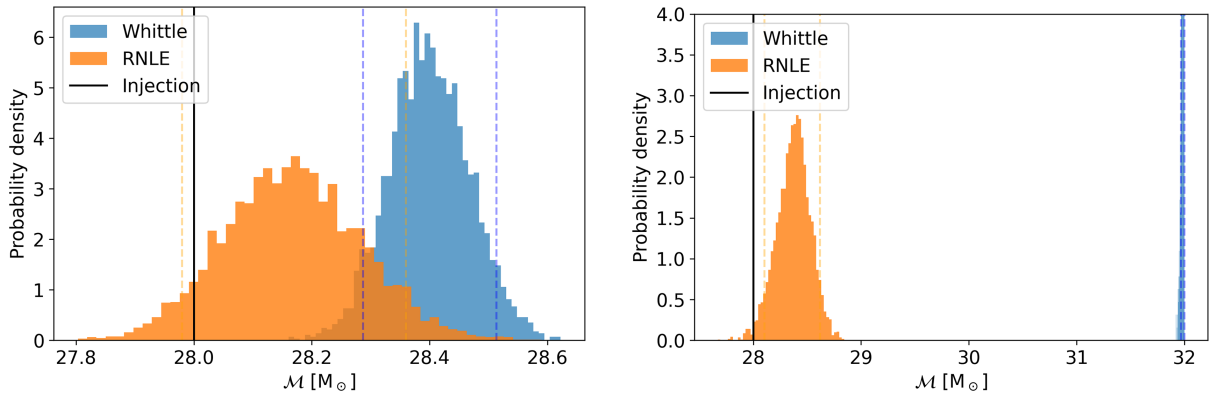


FIG. 14. Posterior probability distribution of the chirp mass obtained running Bilby with the DYNesty sampler using the Whittle likelihood, in blue, and the RNLE likelihood, in orange. The dotted lines indicate the 90% contours and the black vertical line shows the injection value. The left panel shows the results for Glitch A and the right panel shows the same for Glitch B.

particular on how the results depend on the construction of the training dataset and on the relative timing between the glitch and the signal. The goal of this section is to study RNLE in a setup that is closely related to the

scenarios considered in recent deglitching analyses, investigating how the performance of RNLE depends on the construction of the training dataset in this regime.

We follow an approach similar to that of Ghonge *et al.* [69], who injected gravitational-wave signals on and around various types of glitches and performed parameter-estimation analyses before and after deglitching the data using the BayesWave algorithm, finding that parameter estimation results are biased for injections within < 0.1 s of the glitch. This provides a useful reference point for constructing controlled injection studies in strongly non-Gaussian data.

Ghonge *et al.* [69] analyzed injections at offsets between -0.075 and 0.075 s with a timestep of 0.025 s. We use the same offsets for our injections. We analyze one of the blip glitches from their study, for which BayesWave was not always able to unbiased the posterior distributions. Blip glitches are characterized by their short duration, approximately 10 ms, and broad frequency bandwidth, around 100 Hz [151]. They are observed in both LIGO detectors, with an average occurrence rate of about two per hour. They are of particular interest because their morphology closely resembles the gravitational-wave signal from high-mass compact binary mergers, while their physical origin remains largely unknown.

A. Identification of RNLE training datasets for the analysis of a blip glitch

The analysis focuses on a blip glitch observed in the LHO detector during the second observing run (O2)

observing run at GPS time 1165578732.45, corresponding to 11:51:55 UTC on December 12, 2016. The glitch exhibits a signal-to-noise ratio of 15.34, a duration of 0.25 s, and a central frequency of 549 Hz. The top panels of Fig. 15 show, respectively, the distributions of glitch durations and glitch types obtained from GravitySpy within a two-hour window centered on the selected event. The duration of our chosen glitch is representative of the local glitch population, and we note that only 14 blip glitches are present in this interval. The bottom-left panel of Figure 15 displays the time-frequency representation of one second of data surrounding the glitch. The bottom-right panel shows the same data segment with an injected GW signal, highlighted by the overlaid red dotted trajectory. Ghonge *et al.* [69] found that this blip glitch can substantially bias parameter-estimation results and were unable to recover unbiased posteriors for all the tested offsets after glitch subtraction.

To study this behavior with the RNLE likelihood, we analyze injections around the blip glitch at the same set of temporal offsets, with the goal of recovering accurate posterior distributions in all cases. For this purpose, we construct several datasets to train the conditional density estimator. These datasets, summarized in Table I, differ in the number of simulations and in their defining characteristics. In contrast to previous analyses—where we restricted

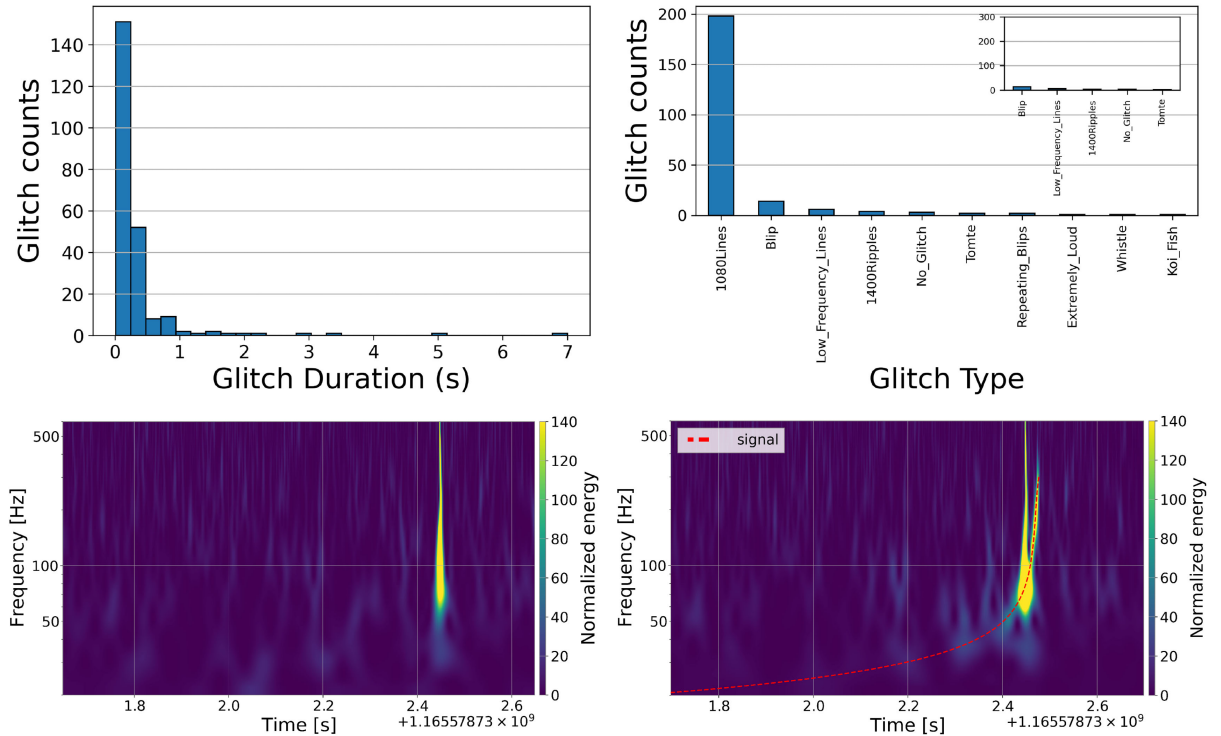


FIG. 15. Top panels: Histograms of the glitch duration, left, and glitch type, right, for all glitches identified by GravitySpy in a 2-hour window centered on the selected blip glitch. Bottom panels: Time-frequency plots of the selected blip glitch, left, and the glitch and an injected GW signal, right. In the latter, the red dotted line indicates the evolution of the signal. The offset for the glitch is 0.05 seconds and we double the signal's SNR with respect to the analysis so that it is clearly visible.

TABLE I. Characteristics of the different datasets employed to train the conditional density estimator for RNLE. Selected blip glitches satisfy the criteria: duration > 0.2 s, SNR > 10 , and $f_{\text{peak}} > 60$ Hz. Unless otherwise specified, glitches are taken from O2.

Dataset label	Number of simulations	Characteristics
GD1.0	9240	Selected blip glitches with offset 0.0 s
GD1.1	9240	Selected blip glitches with offset -0.025 s
GD1.2	9240	Selected blip glitches with offset -0.05 s
GD1.3	9240	Selected blip glitches with offset -0.075 s
GD1.4	9240	Selected blip glitches with offset 0.025 s
GD1.5	9240	Selected blip glitches with offset 0.05 s
GD1.6	9240	Selected blip glitches with offset 0.075 s
GD2.0	184000	Selected blip glitches with offset $\mathcal{U}[-0.1, 0.1]$, each taken 100 times
GD2.1	394990	Selected blip glitches from O2 and first half of the third observing run (O3a) with offset $\mathcal{U}[-0.1, 0.1]$, each taken 100 times
GD3.0	36751	Selected blip glitches with discrete offsets at $[-0.075, -0.05, -0.025, 0.025, 0.05, 0.075]$
T1	14300	Seconds of data around the trigger time

ourselves to simulations based solely on data adjacent to the analyzed segment—here we develop two classes of training sets: (i) datasets containing data drawn from the vicinity of the observed glitch time, and (ii) datasets in which each simulated example includes a blip glitch taken from the full GravitySpy repository for LHO during O2. This strategy is motivated by the fact that the small number of local blip glitches might be insufficient for the network to learn their variability, whereas O2 contains thousands of labeled blip glitches that can be leveraged to provide a more robust and representative training distribution. To construct the glitch datasets we selected glitches with SNR > 10 , $f_{\text{peak}} > 60$, and duration > 0.2 s to maximize the similarity to the analyzed glitch.

B. Analyzing a BBH injection around a blip glitch

Following the approach of Ghonge *et al.* [69], we conduct a standard set of seven analyses, varying the offset between the glitch and the injection. The SNR of the injected BBH signal is fixed at 12.88. Parameter estimation is carried out using both the Whittle likelihood and RNLE trained with different datasets. We show the results using violin plots in Fig. 16. For each violin, the left side shows posterior distributions obtained with the Whittle likelihood, and the right side shows RNLE posteriors. The black dashed lines inside the violins represent the mean and $1\text{-}\sigma$ credibility intervals for each posterior. The red dashed horizontal line represents the injection value. We again fix all parameters apart from the chirp mass.

For panel 16A in the top left, we use dataset T1, composed of 14300 s around the glitch, to train the

conditional density estimator. We note that the most bias for the Whittle likelihood results is recovered for an offset of 0.0 and 0.05 s. Unfortunately, the RNLE likelihood is similarly biased for those offset values, while accurate results are recovered for every other offset. This is not unexpected, as the data around the blip glitch contains only a few other blip glitches, and we would not expect the estimator to capture their properties. For panel 16B in the top right, we use dataset GD2.0 to train the conditional density estimator. This is composed of blip glitches, each taken 100 times and placed at a randomly sampled offset from the uniform distribution $\mathcal{U}[-0.1, 0.1]$. While in this case the RNLE likelihood recovers accurate results for the 0.05 s offset, i.e., the posterior distribution includes the true value, it recovers biased results for offsets for which the Whittle likelihood does not display any bias. For panel 16C in the bottom left, we use dataset GD3.0, composed of blip glitches each taken 20 times and placed at discrete offsets coinciding with the offsets used for the observed data, to train the conditional density estimator. The resulting RNLE likelihood recovers accurate posteriors for both the offsets 0.0 and 0.05 s. However, also in this case it recovers biased results for offset values for which the Whittle likelihood is not biased. We attribute the oscillating behavior of these RNLE likelihoods to the limited amount of training data and the nonconvergence of the conditional density estimator. To test this hypothesis, we create a new set of datasets, labeled as GD1 in Table I. We create one dataset for each of the offset values used in our analysis. Each of them is composed of the same set of glitches from LHO during O2 placed at the same offset, i.e., for GD1.0 all glitches are placed at a null offset. Using those datasets, we train seven different RNLE likelihoods and use them to analyze the corresponding observed data realization, i.e., we use the likelihood obtained from GD1.0 to analyze the data with null offset. Panel 16D shows the results obtained when each RNLE likelihood trained on a given GD1 dataset is applied to the observed data realization with the matching offset. We recover accurate posteriors for all values of the offset, demonstrating that the approach is capable of learning an accurate likelihood even in the presence of non-Gaussian noise transients when the training and analysis conditions are closely matched. While these results validate the feasibility of the method, they do not yet provide a practical strategy for realistic analyses. Training a separate likelihood for each specific glitch position would implicitly assume prior knowledge of the transient location, require a large number of independent trainings, and reduce the comparability and robustness of the inferred posteriors. More broadly, these results show that the construction of the training dataset is a central ingredient of RNLE in the strongly non-Gaussian regime. Event-specific or offset-specific datasets are useful for controlled proof-of-principle studies, but a deployable implementation would require a broader detector- and epoch-specific training distribution,

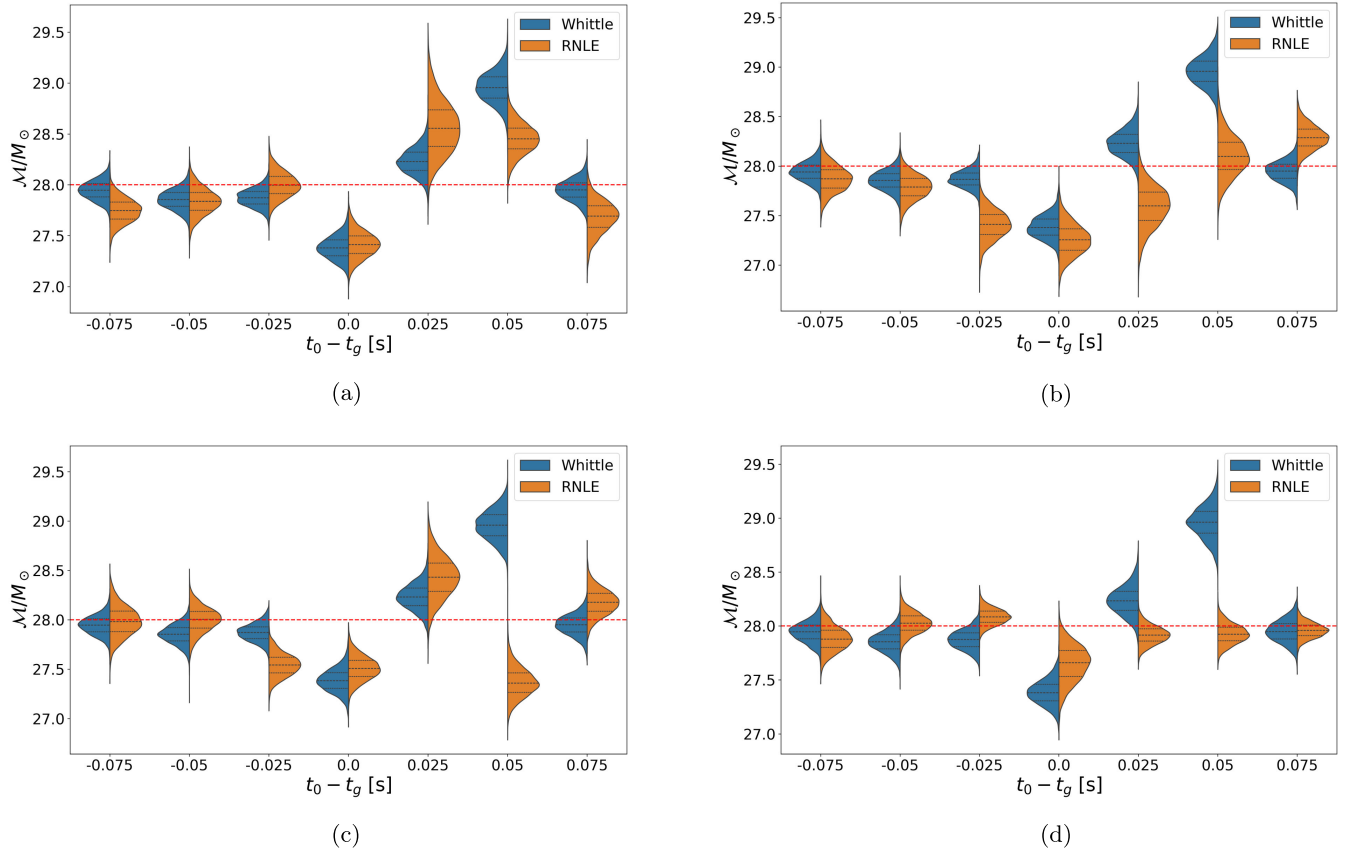


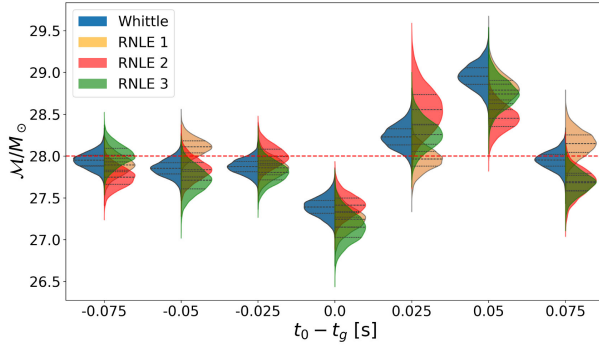
FIG. 16. Violin plots comparing Whittle (blue) and RNLE (orange) results for various offsets between a blip glitch observed on December 12, 2016 and an injected GW signal. Panels A–D correspond to the descriptions in the subcaptions. (a) Results obtained using an RNLE likelihood trained on the T1 dataset containing 14 300 seconds of data around the blip glitch. (b) Results obtained with an RNLE likelihood trained on the GD2.0 dataset containing all LHO glitches from O2, each taken one hundred times and repositioned using offsets sampled from $\mathcal{U}[-0.1, 0.1]$. (c) Results obtained with an RNLE likelihood trained on the GD3.0 dataset containing all LHO glitches from O2, each taken ten times and positioned at discrete offsets. (d) Results obtained with RNLE likelihoods trained on the GD1 datasets (Table I), which include all LHO blip glitches from O2 always positioned at the offset time.

including a representative population of transient artifacts with random relative timing. Future work should therefore focus on optimizing the balance between local background information and broad glitch coverage in order to construct a single likelihood that can be applied consistently across different offsets and data realizations. In this context, adaptive strategies for constructing or refining the training dataset, for example by combining information from data surrounding the event with a broader library of glitch realizations, may provide a useful route toward a deployable RNLE analysis pipeline [152]. As a robustness test for such a study, we suggest the use of calibration plots similar to the one shown in Fig. 11.

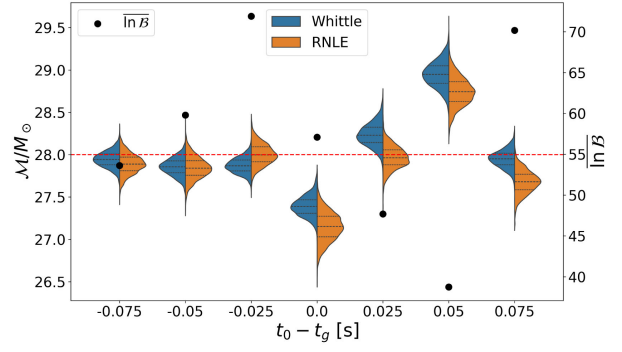
IX. INVESTIGATING TRAINING NOISE

As an initial step toward understanding how different sources of variability affect the performance of the RNLE likelihood, we investigate the impact of training noise in the presence of blip glitches. Building on the approach

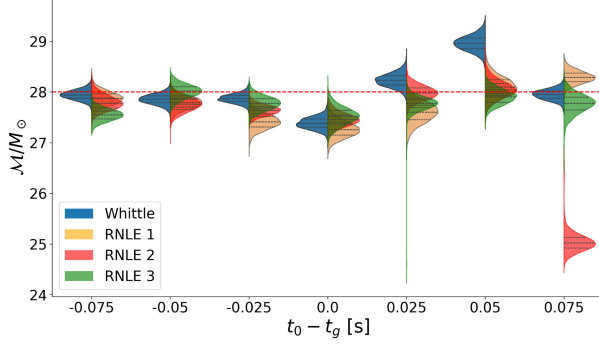
described in Sec. IV C, where the stability of the conditional density estimator was assessed using ensembles of independently trained likelihoods, we observed that convergence of the learned likelihood corresponds to stabilization of the Bayesian evidence values across realizations (cf. Fig. 6). Motivated by this insight, we now examine how variability in the RNLE posteriors and evidence manifests when analyzing the blip glitch contaminated data. To this end, we retrain the RNLE conditional density estimator multiple times on an identical dataset, while changing the optimization seed of the conditional density estimator and the order of the training data realizations. After performing stochastic sampling, we quantify the variability among the resulting posterior distributions. This procedure isolates the intrinsic stochasticity introduced by the training process itself, analogous to the Gaussian-noise case discussed in Sec. IV C. The left panels of Fig. 17 show violin plots analogous to those in Fig. 16, but now including three independent realizations of the RNLE posteriors produced by retraining the likelihood. For this study, the GD1



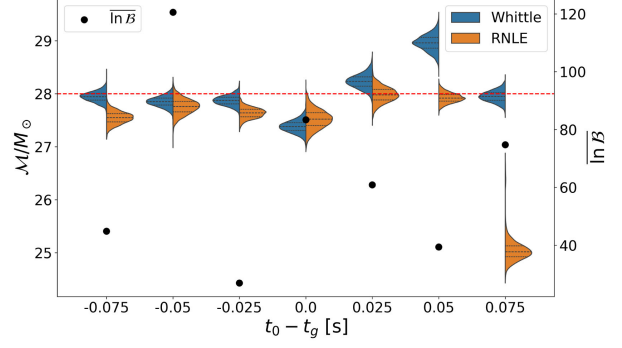
(a1) RNLE posteriors — T1 dataset



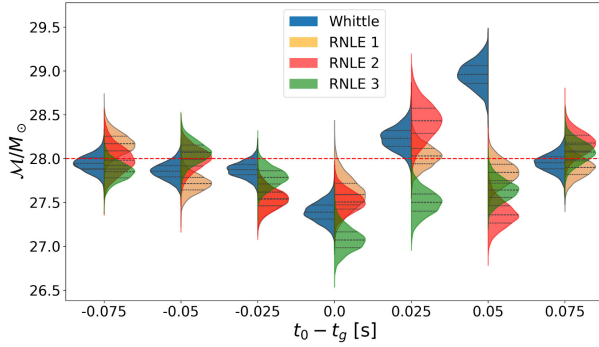
(a2) Evidence-weighted posterior — T1 dataset



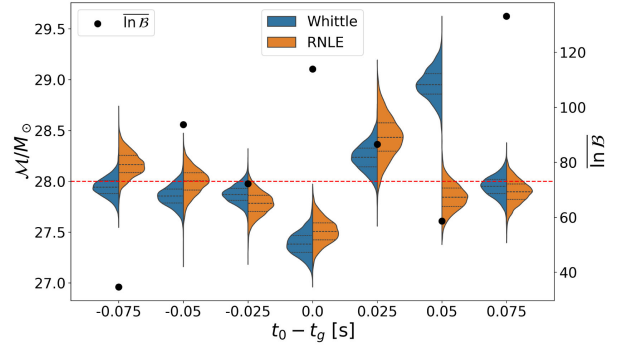
(b1) RNLE posteriors — GD2.0



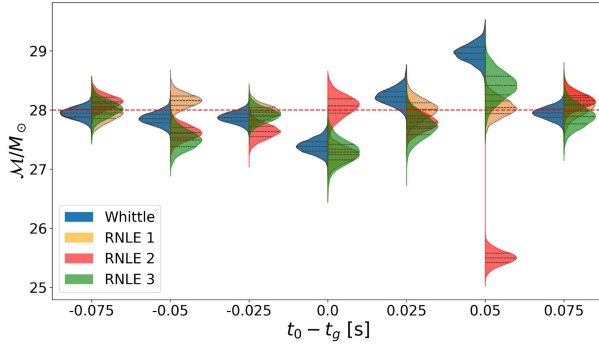
(b2) Evidence-weighted posterior — GD2.0



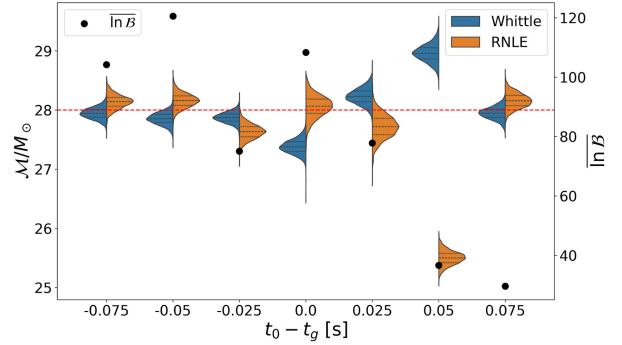
(c1) RNLE posteriors — GD3.0



(c2) Evidence-weighted posterior — GD3.0



(d1) RNLE posteriors — GD2.1



(d2) Evidence-weighted posterior — GD2.1

FIG. 17. Comparison of posterior distributions obtained with the Whittle likelihood and multiple RNLE realizations across four different training datasets. Each row corresponds to a dataset: (a) T1, (b) GD2.0, (c) GD3.0, (d) GD2.1. The left columns show three RNLE realizations. The right columns show the evidence-weighted combination of the posteriors. The black dots represent the mean of the log-Bayes factors.

datasets are excluded as we want to focus on obtaining a single likelihood to analyze all offset realizations. From top to bottom, the panels correspond to likelihoods trained with the T1, GD3.0, GD2.0, and GD2.1 datasets. The GD2.1 dataset extends GD2.0 by additionally incorporating blip glitches from Hanford during O3a, selected using the same criteria as GD2.0, with each glitch used 100 times and offsets drawn from $\mathcal{U}[-0.1, 0.1]$. Across all datasets, the posteriors obtained from the different RNLE realizations exhibit substantial variability. In particular, for GD3.0 and GD2.1, at every offset value, there exists at least one realization whose posterior samples include the injected parameter value at the 98% credibility level, even for offsets where the Whittle likelihood yields biased estimates. This behavior is different from the results presented in Fig. 6 for the simulated colored Gaussian noise case, where independently trained likelihoods yielded consistent evidences and highly similar posterior distributions, exhibiting both accuracy and small uncertainties. The high-variability seen in the non-Gaussian scenario indicates that an ensemble of independently trained realizations enables a more reliable assessment of the variability induced by the training procedure itself, and that the mean log-Bayes can be an indicator of the robustness of the trained likelihood. Taken together, these findings suggest that the increased inter-realization variability observed here is driven by the complexity of the non-Gaussian noise features present in real data. The long-term objective is that increasing the amount of training data and improving model capacity will reduce this variability, approaching the level of stability observed in the Gaussian-noise case and yielding both more precise and more accurate parameter estimates. To combine the different RNLE realizations into a single posterior distribution, each posterior is weighted by its corresponding Bayesian evidence, using

$$w_i = \frac{\mathcal{Z}_i}{\sum_j \mathcal{Z}_j}, \quad (16)$$

where $\mathcal{Z}_i$ is the evidence associated with the i th posterior, and the sum runs over all posteriors being combined. Because all analyses employ identical priors, this procedure is equivalent to model averaging with Bayesian odds, as discussed by Thrane and Talbot [43]. Naïvely, one might hope that this approach will favor the most accurate posterior at each offset. However, the right panels of Fig. 17 show the evidence-weighted posteriors corresponding to the left panels. Interestingly, for each offset, the combined posterior is almost entirely dominated by a single realization, with little to no mixing of posteriors. Moreover, the selected realization is not always the most accurate: in rows C and D, clearly biased posteriors receive the highest weight. This behavior, while seemingly counterintuitive, is consistent with the understanding from Sec. IV C that evidence integrates the likelihood over the full parameter

space and does not depend solely on the local posterior for a specific data realization. Consequently, a realization yielding an accurate posterior at a particular offset may not correspond to a higher evidence value. The black dots in the right-hand panels further illustrate this behavior, representing the mean log-Bayes factor computed across the three posterior realizations at each offset. These results differ markedly from those shown in Fig. 6 for the analogous study using simulated colored Gaussian noise. In that case, the mean evidence was observed to approach zero as the number of simulations increased, indicating that effective mixing would occur if the weighting scheme introduced in this section was applied. By contrast, in Fig. 17 the mean evidence remains consistently large, exceeding 40 across all offsets. Such values correspond to substantial variability among the inferred posterior distributions. Overall, the pronounced variance in the evidence estimates across all datasets suggests that the number of simulations employed is insufficient to ensure full convergence of the conditional density estimator. A systematic investigation of convergence properties and alternative training strategies is therefore deferred to future work. In this context, the ensemble weighting approach based on the evidence, introduced here, provides a promising diagnostic tool to assess convergence and to mitigate the stochasticity inherent in the training procedure of the conditional density estimator.

X. CONCLUSION

We have presented the development and application of a novel SBI algorithm for gravitational-wave parameter estimation based on a modified neural likelihood estimation scheme, termed RNLE. The method exploits both the SBI implementation of NLE and the additive structure of gravitational-wave signals and detector noise to directly learn the likelihood of the noise contribution. A key feature of RNLE is that the conditional density estimator can be trained exclusively on background noise, without requiring gravitational-wave signal injections. This substantially reduces the dimensionality of the parameter space explored during training and enables the modeling of non-Gaussian likelihoods associated with glitch-contaminated data. The RNLE algorithm is implemented within the Bilby framework, allowing seamless integration with standard stochastic samplers.

The performance of the algorithm was validated through a sequence of controlled studies using a sine-Gaussian toy model, simulated gravitational-wave detector noise, and real interferometer data. These investigations included injection campaigns benchmarking RNLE against the Whittle likelihood, with a systematic exploration of its dependence on the analyzed data duration and on the size and composition of the training dataset. Under approximately Gaussian noise conditions, RNLE accurately learns the likelihood function and recovers posterior distributions consistent with those

obtained using the Whittle likelihood. When applied to real detector data, this agreement persists whenever the assumptions underlying the Whittle likelihood are satisfied. Crucially, in the presence of loud non-Gaussian noise transients, RNLE outperforms the Whittle likelihood, recovering posterior distributions consistent with the injected true values, while the Whittle likelihood leads to biased posteriors that do not include the true parameters. Analyses of loud glitches from the O3b run, and a blip glitch from O2 demonstrate that glitch morphology, duration, and temporal overlap with the gravitational-wave signal play a critical role in determining likelihood accuracy. These results provide clear practical guidance: parameter estimation in the presence of glitches benefits from training datasets that closely resemble the data being analyzed, i.e., same GravitySpy glitch class, same frequency range, and similar relative timing between the signal and noise transients.

Beyond accuracy, we investigated sources of variability intrinsic to the RNLE likelihood estimation procedure. By retraining the conditional density estimator multiple times on identical datasets, we quantified the stochastic variability induced by training noise. We observed substantial variability among posterior realizations in the presence of non-Gaussian noise. In contrast, for Gaussian noise, increasing the number of simulations led to consistent likelihood estimates across retraining runs. We propose an evidence-weighting scheme to combine multiple realizations into a single posterior; however, the Bayesian evidence was found to often strongly favor a single realization, sometimes even a biased one, reflecting its sensitivity to global likelihood structure rather than local accuracy near the true parameters. Comparing multiple likelihood realizations estimated from the same dataset provides a diagnostic for assessing the reliability of the RNLE results. In the Gaussian noise case, the evidence of the RNLE realizations converged when increasing the size of the training dataset. Convergence was not observed in the non-Gaussian noise analysis, hinting at an insufficient training dataset size to capture the relevant non-Gaussian features of the noise. This behavior underscores the importance of careful training dataset construction and highlights the need for stability-aware inference strategies when deploying neural likelihood estimators.

Overall, RNLE provides a promising pathway toward more robust gravitational-wave parameter estimation in the presence of non-Gaussian noise transients, emphasizing likelihood-level modeling rather than *post hoc* glitch mitigation. The present results demonstrate proof-of-principle performance in real detector noise, while also showing that reliability in the strongly non-Gaussian regime depends sensitively on training-dataset construction and likelihood stability. Key next steps are large-scale calibration studies based on injections into representative detector noise, together with the development of training strategies that remain effective across realistic distributions of non-

Gaussian data realizations. At the same time, the presented analyses do not allow for a deployment-ready prescription for real-event inference in non-Gaussian data. A practical implementation will likely require carefully designed and possibly adaptive training datasets that combine information from data surrounding the event with a broader distribution of glitch realizations. This need for a calibrated and broadly applicable treatment of non-Gaussian noise is further motivated by recent work showing that glitch subtraction approaches generally leave residuals that can bias inference, suggesting that joint modeling of the signal and noise contributions offers a more robust route to inference [71]. In this context, RNLE offers a flexible framework for incorporating complex noise behavior directly into the inference process.

Future efforts will focus on systematic studies of training-dataset design, likelihood stability, and computational performance [152,153], with the goal of more fully characterizing the applicability and limitations of RNLE across a broad range of gravitational-wave analyses. Future work should also investigate the convergence of the estimated likelihood more systematically, with the aim of reducing retraining variability and establishing practical recommendations for the training settings required to obtain stable results.

ACKNOWLEDGMENTS

We would like to thank Alessio Spurio Mancini for the useful discussions during the development of the project, Julianna Ostrovska for the work done on NLE in the early stages of the project and the anonymous referee for the useful comments on this manuscript. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants No. PHY-0757058 and No. PHY-0823459. This work is supported by the Science and Technology Facilities Council (STFC) Grant No. UKRI2488. The authors are also grateful for computational resources provided by Cardiff University and supported by STFC Grants No. ST/I006285/1 and No. ST/V005618/1. This research has made use of data or software obtained from the Gravitational Wave Open Science Center [154], a service of the LIGO Scientific Collaboration, the Virgo Collaboration, and KAGRA. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation, as well as the Science and Technology Facilities Council (STFC) of the United Kingdom, the Max-Planck-Society (MPS), and the State of Niedersachsen/Germany for support of the construction of Advanced LIGO and construction and operation of the GEO600 detector. Additional support for Advanced LIGO was provided by the Australian Research Council. Virgo is funded, through the European Gravitational Observatory (EGO), by the French Centre National de Recherche

Scientifique (CNRS), the Italian Istituto Nazionale di Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by institutions from Belgium, Germany, Greece, Hungary, Ireland, Japan, Monaco, Poland, Portugal, Spain. K. A. G. R. A. is supported by Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan Society for the Promotion of Science (JSPS) in Japan; National Research Foundation (NRF) and Ministry of Science and ICT (MSIT) in Korea; Academia Sinica

(AS) and National Science and Technology Council (NSTC) in Taiwan.

DATA AVAILABILITY

The data that support the findings of this article are openly available [\[138\]](#).

APPENDIX A: TABLES OF PARAMETER VALUES AND PRIORS

TABLE II. Priors on the parameters of the sine-Gaussian toy model.

Parameter	α	f	A	σ
Prior	$\mathcal{U}[1, 3]$	$\mathcal{U}[1, 3]$	$\mathcal{U}[9, 11]$	$\mathcal{U}[0, 2]$

TABLE III. Parameter values for simulated gravitational wave signals corresponding to precessing and aligned-spins BBH mergers. Values are given for the chirp mass ($\mathcal{M}$), mass ratio (q), luminosity distance (D_L), dimensionless spins (a_1 and a_2), tilt angles (θ_1 and θ_2), phase angles (ϕ_{12} and ϕ_{jl}), geocentric time (t_{geocent}), angular parameter (θ_{jn}), polarization angle (ψ), and overall phase (ϕ). These values serve as the ground truth for the simulated signals. Right to each parameter value we report the JS divergence value between the posterior distribution obtained with the Whittle and RNLE likelihood.

Parameter	10D precessing BBH	JS values	15D precessing BBH	JS values	11D aligned-spin BBH	JS values
$\mathcal{M}$	28	0.002	28	0.008	28	0.005
q	0.82	0.003	0.82	0.010	0.94	0.001
D_L	1400	0.001	1400	0.003	1400	0.002
χ_1	...	...	...	...	0.1	0.008
χ_2	...	...	...	...	0.1	0.006
a_1	0.32	0.001	0.32	0.024	...	...
a_2	0.44	0.001	0.44	0.027	...	...
θ_1	0.5	0.001	0.5	0.004	0.0	...
θ_2	1.0	0.001	1.0	0.025	0.0	...
ϕ_{12}	0.0	...	1.7	0.002	0.0	...
ϕ_{jl}	0.0	...	0.3	0.001	0.0	...
DEC	-1.21	...	0.7219	0.007	0.7219	0.060
RA	1.375	...	1.375	0.012	0.1904	0.040
t_{geocent}	1126259642.0	0.002	1126259642.0	0.001	1126259642.0	0.010
θ_{jn}	0.4	0.001	0.4	0.008	0.4	0.060
ψ	2.66	0.001	2.66	0.003	2.66	0.001
ϕ	1.3	0.001	1.3	0.002	1.3	0.001

TABLE IV. Prior distributions employed in the parameter estimation of simulated gravitational wave signals arising from precessing and spin-aligned BBH mergers. The symbol $\mathcal{U}$ denotes a uniform prior, and λ^2 represents a power law prior within the specified ranges. The table details the priors for chirp mass ($\mathcal{M}$), mass ratio (q), dimensionless spins (χ_1 and χ_2), tilt angles (θ_1 and θ_2), declination (DEC), luminosity distance (D_L), right ascension (RA), the geocentric time, and various angle parameters.

Parameter	Precessing BBH	Aligned-spin BBH
$\mathcal{M}$	$\mathcal{U}$ [24, 32]	$\mathcal{U}$ [24, 32]
q	$\mathcal{U}$ [0.125, 1]	$\mathcal{U}$ [0.125, 1]
χ_1	...	$\mathcal{U}$ [0, 0.99]
χ_2	...	$\mathcal{U}$ [0, 0.99]
a_1	$\mathcal{U}$ [0, 0.99]	...
a_2	$\mathcal{U}$ [0, 0.99]	...
θ_1	Sin	...
θ_2	Sin	...
ϕ_{12}	$\mathcal{U}$ [0, 2π]	...
ϕ_{jl}	$\mathcal{U}$ [0, 2π]	...
D_L	λ^2 [100,10000]	λ^2 [100,10000]
DEC	Cos	Cos
RA	$\mathcal{U}$ [0, 2π]	$\mathcal{U}$ [0, 2π]
t_{geocent}	$\mathcal{U}$ [$t_{\text{trigger}} - 0.1$, $t_{\text{trigger}} + 0.1$]	$\mathcal{U}$ [$t_{\text{trigger}} - 0.1$, $t_{\text{trigger}} + 0.1$]
θ_{jn}	Sin	Sin
ψ	$\mathcal{U}$ [0, π]	$\mathcal{U}$ [0, π]
ϕ	$\mathcal{U}$ [0, 2π]	$\mathcal{U}$ [0, 2π]

APPENDIX B: SAMPLING SEED POSTERIOR VARIABILITY

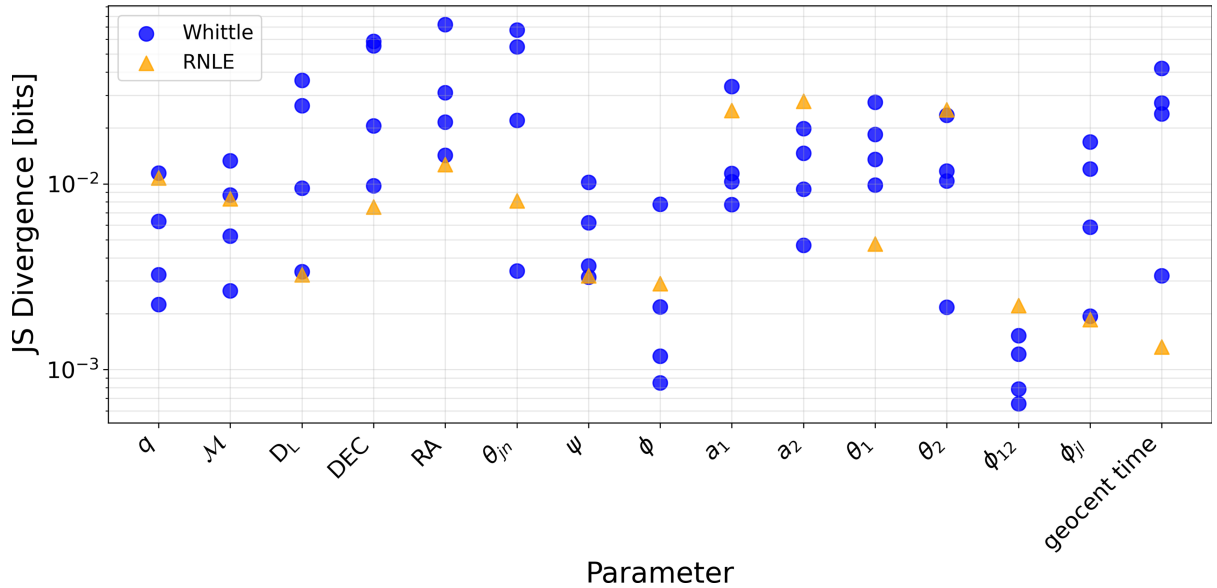


FIG. 18. Comparison of the JS divergence values for all parameters between an RNLE run and Whittle likelihood run, in orange, and between Whittle likelihood runs using different sampling seeds for the full 15-dimensional runs.

APPENDIX C: PSD WINDOWING COMPARISON

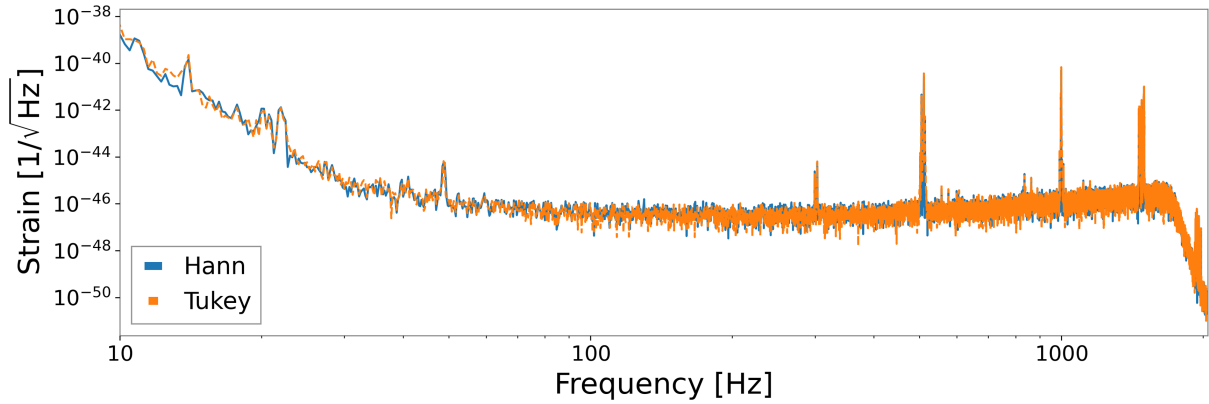


FIG. 19. Comparison of a PSD computed using the same data and method, using a Hann window, in blue, and a Tukey window, in orange.

-
- [1] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Phys. Rev. Lett.* **116**, 061102 (2016).
 - [2] J. Aasi *et al.* (LIGO Scientific Collaboration), *Classical Quantum Gravity* **32**, 074001 (2015).
 - [3] F. Acernese *et al.* (Virgo Collaboration), *Classical Quantum Gravity* **32**, 024001 (2015).
 - [4] T. Akutsu *et al.* (KAGRA Collaboration), *Prog. Theor. Exp. Phys.* **2021**, 05A101 (2021).
 - [5] A. G. Abac *et al.* (LIGO Scientific Collaboration, VIRGO Collaboration, and KAGRA Collaboration), *arXiv*: 2508.18082.
 - [6] B. P. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), *Phys. Rev. Lett.* **116**, 221101 (2016); **121**, 129902(E) (2018).
 - [7] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Phys. Rev. D* **100**, 104036 (2019).
 - [8] R. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Phys. Rev. D* **103**, 122002 (2021).
 - [9] R. Abbott *et al.* (LIGO Scientific Collaboration, VIRGO Collaboration, and KAGRA Collaboration), *Phys. Rev. D* **112**, 084080 (2025).
 - [10] B. P. Abbott *et al.* (LIGO Scientific, Virgo, 1M2H, Dark Energy Camera GW-E, DES, DLT40, Las Cumbres Observatory, VINROUGE, MASTER Collaborations), *Nature (London)* **551**, 85 (2017).
 - [11] A. G. Abac *et al.* (LIGO Scientific Collaboration, VIRGO Collaboration, and KAGRA Collaboration), *arXiv*: 2509.04348.
 - [12] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Phys. Rev. X* **9**, 011001 (2019).
 - [13] B. Margalit and B. D. Metzger, *Astrophys. J. Lett.* **850**, L19 (2017).
 - [14] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Phys. Rev. Lett.* **121**, 161101 (2018).
 - [15] R. Nair, S. Perkins, H. O. Silva, and N. Yunes, *Phys. Rev. Lett.* **123**, 191101 (2019).
 - [16] A. G. Abac *et al.* (LIGO Scientific Collaboration, VIRGO Collaboration, and KAGRA Collaboration), *arXiv*: 2508.18083.
 - [17] N. Christensen and R. Meyer, *Rev. Mod. Phys.* **94**, 025001 (2022).
 - [18] F. Echeverria, *Phys. Rev. D* **40**, 3194 (1989).
 - [19] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *J. Chem. Phys.* **21**, 1087 (1953).
 - [20] W. K. Hastings, *Biometrika* **57**, 97 (1970).
 - [21] N. Christensen and R. Meyer, *Phys. Rev. D* **58**, 082001 (1998).
 - [22] J. Veitch *et al.*, *Phys. Rev. D* **91**, 042003 (2015).
 - [23] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Phys. Rev. Lett.* **116**, 241102 (2016).
 - [24] R. Meyer, M. C. Edwards, P. Maturana-Russel, and N. Christensen, *WIREs Comput. Stat.* **14**, e1532 (2022).
 - [25] J. Skilling, *AIP Conf. Proc.* **735**, 395 (2004).
 - [26] J. Skilling, *Bayesian Analysis* **1**, 833 (2006).
 - [27] G. Ashton *et al.*, *Astrophys. J. Suppl. Ser.* **241**, 27 (2019).
 - [28] I. M. Romero-Shaw *et al.*, *Mon. Not. R. Astron. Soc.* **499**, 3295 (2020).
 - [29] A. G. Abac *et al.* (LIGO Scientific Collaboration, VIRGO Collaboration, and KAGRA Collaboration), *arXiv*: 2508.18081.
 - [30] J. Lange, R. O'Shaughnessy, and M. Rizzo, *arXiv*: 1805.10457.
 - [31] D. Wysocki, R. O'Shaughnessy, J. Lange, and Y.-L. L. Fang, *Phys. Rev. D* **99**, 084026 (2019).

- [32] J. Wofford *et al.*, [arXiv:2210.07912](#).
- [33] C. Pankow, P. Brady, E. Ochsner, and R. O’Shaughnessy, *Phys. Rev. D* **92**, 023002 (2015).
- [34] R. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Phys. Rev. X* **11**, 021053 (2021).
- [35] R. Abbott *et al.* (KAGRA Collaboration, VIRGO Collaboration, and LIGO Scientific Collaboration), *Phys. Rev. X* **13**, 041039 (2023).
- [36] M. J. Williams, J. Veitch, and C. Messenger, *Mach. Learn. Sci. Tech.* **4**, 035011 (2023).
- [37] M. Prathaban, H. Bevins, and W. Handley, [arXiv:2411.17663](#).
- [38] L. Pathak, A meshfree approach for rapid reconstruction of compact binary sources detected in the network of international gravitational wave detectors, Ph.D. thesis, Indian Institute of Technology Gandhinagar, 2025, [arXiv:2509.20104](#).
- [39] D. Yallup, M. Prathaban, J. Alvey, and W. Handley, in *2nd European AI for Fundamental Physics Conference* (SciPost, Amsterdam, 2025).
- [40] M. J. Williams, M. Karamanis, Y. Luo, and U. Seljak, *Mon. Not. R. Astron. Soc.* **1479**, 1493 (2025).
- [41] P. Whittle, Hypothesis Testing in Time Series Analysis, Ph.D. thesis, Uppsala University, 1951.
- [42] A. Contreras-Cristan, E. Gutierrez-Pena, and S. G. Walker, *Commun. Stat. Simul. Comput.* **35**, 857 (2006).
- [43] E. Thrane and C. Talbot, *Pub. Astron. Soc. Aust.* **36**, e010 (2019); **37**, e036(E) (2020).
- [44] S. S. Rao and J. Yang, [arXiv:2001.06966](#).
- [45] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Classical Quantum Gravity* **37**, 055002 (2020).
- [46] D. Davis *et al.* (LIGO Collaboration), *Classical Quantum Gravity* **38**, 135014 (2021).
- [47] S. Soni *et al.* (LIGO Collaboration), *Classical Quantum Gravity* **42**, 085016 (2025).
- [48] J. C. Driggers *et al.* (LIGO Scientific Collaboration), *Phys. Rev. D* **99**, 042001 (2019).
- [49] J. Powell, *Classical Quantum Gravity* **35**, 155017 (2018).
- [50] O. Edy, A. Lundgren, and L. K. Nuttall, *Phys. Rev. D* **103**, 124061 (2021).
- [51] D. Davis, T. B. Littenberg, I. M. Romero-Shaw, M. Millhouse, J. McIver, F. Di Renzo, and G. Ashton, [arXiv:2207.03429](#).
- [52] S. Kumar, A. H. Nitz, and X. J. Forteza, *Astrophys. J.* **982**, 67 (2025).
- [53] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Phys. Rev. Lett.* **119**, 161101 (2017).
- [54] B. P. Abbott *et al.* (LIGO Scientific, Virgo, Fermi GBM, INTEGRAL, IceCube, AstroSat Cadmium Zinc Telluride Imager Team, IPN, Insight-Hxmt, ANTARES, Swift, AGILE Team, 1M2H Team, Dark Energy Camera GW-EM, DES, DLT40, GRAWITA, Fermi-LAT, ATCA, ASKAP, Las Cumbres Observatory Group, OzGrav, DWF (Deeper Wider Faster Program), AST3, CAASTRO, VINROUGE, MASTER, J-GEM, GROWTH, JAGWAR, CaltechNRAO, TTU-NRAO, NuSTAR, Pan-STARRS, MAXI Team, TZAC Consortium, KU, Nordic Optical Telescope, ePESSTO, GROND, Texas Tech University, SALT Group, TOROS, BOOTES, MWA, CALET, IKI-GW Follow-up, H.E.S.S., LOFAR, LWA, HAWC, Pierre Auger, ALMA, Euro VLBI Team, Pi of Sky, Chandra Team at McGill University, DFN, ATLAS Telescopes, High Time Resolution Universe Survey, RIMAS, RATIR, SKA South Africa/MeerKAT Collaborations), *Astrophys. J. Lett.* **848**, L12 (2017).
- [55] B. P. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration, Fermi-GBM Collaboration, and INTEGRAL Collaboration), *Astrophys. J. Lett.* **848**, L13 (2017).
- [56] E. Troja *et al.*, *Nature (London)* **551**, 71 (2017).
- [57] C. Pankow *et al.*, *Phys. Rev. D* **98**, 084016 (2018).
- [58] D. Davis, T. J. Massinger, A. P. Lundgren, J. C. Driggers, A. L. Urban, and L. K. Nuttall, *Classical Quantum Gravity* **36**, 055011 (2019).
- [59] G. Vajente, Y. Huang, M. Isi, J. C. Driggers, J. S. Kissel, M. J. Szczepanczyk, and S. Vitale, *Phys. Rev. D* **101**, 042003 (2020).
- [60] G. Vajente, *Phys. Rev. D* **105**, 102005 (2022).
- [61] J. Merritt, B. Farr, R. Hur, B. Edelman, and Z. Doctor, *Phys. Rev. D* **104**, 102004 (2021).
- [62] M. A. T. Chowdhury, [arXiv:2406.01318](#).
- [63] C. Plunkett, S. Hourihane, and K. Chatziioannou, *Phys. Rev. D* **106**, 104021 (2022).
- [64] G. Ashton, *Mon. Not. R. Astron. Soc.* **520**, 2983 (2023).
- [65] A.-K. Malz and J. Veitch, *Phys. Rev. D* **112**, 024071 (2025).
- [66] N. J. Cornish and T. B. Littenberg, *Classical Quantum Gravity* **32**, 135012 (2015).
- [67] K. Chatziioannou, N. Cornish, M. Wijngaarden, and T. B. Littenberg, *Phys. Rev. D* **103**, 044013 (2021).
- [68] S. Hourihane, K. Chatziioannou, M. Wijngaarden, D. Davis, T. Littenberg, and N. Cornish, *Phys. Rev. D* **106**, 042006 (2022).
- [69] S. Ghonge, J. Brandt, J. M. Sullivan, M. Millhouse, K. Chatziioannou, J. A. Clark, T. Littenberg, N. Cornish, S. Hourihane, and L. Cadonati, *Phys. Rev. D* **110**, 122002 (2024).
- [70] N. J. Cornish, T. B. Littenberg, B. Bécsy, K. Chatziioannou, J. A. Clark, S. Ghonge, and M. Millhouse, *Phys. Rev. D* **103**, 044006 (2021).
- [71] R. Udall, S. Bini, K. Chatziioannou, D. Davis, S. Hourihane, Y. Lecoëuche, J. McIver, and S. Miller, [arXiv:2510.05029](#).
- [72] E. Cuoco *et al.*, *Mach. Learn. Sci. Tech.* **2**, 011002 (2021).
- [73] J. Powell, D. Trifirò, E. Cuoco, I. S. Heng, and M. Cavaglià, *Classical Quantum Gravity* **32**, 215012 (2015).
- [74] M. Razzano and E. Cuoco, *Classical Quantum Gravity* **35**, 095016 (2018).
- [75] S. Soni *et al.*, *Classical Quantum Gravity* **38**, 195016 (2021).
- [76] R. Biswas *et al.*, *Phys. Rev. D* **88**, 062003 (2013).
- [77] M. Cavaglia, K. Staats, and T. Gill, *Commun. Comput. Phys.* **25**, 963 (2025).
- [78] M. Cavaglia, S. Gaudio, T. Hansen, K. Staats, M. Szczepanczyk, and M. Zanolin, *Mach. Learn. Sci. Tech.* **1**, 015005 (2020).
- [79] E. Payne, S. Hourihane, J. Golomb, R. Udall, R. Udall, D. Davis, and K. Chatziioannou, *Phys. Rev. D* **106**, 104017 (2022).
- [80] R. Macas, A. Lundgren, and G. Ashton, *Phys. Rev. D* **109**, 062006 (2024).

- [81] R. Udall, S. Hourihane, S. Miller, D. Davis, K. Chatziioannou, M. Isi, and H. Deshong, *Phys. Rev. D* **111**, 024046 (2025).
- [82] A. Ray, S. Banagiri, E. Thrane, and P.D. Lasky, [arXiv:2510.07228](#).
- [83] C. Chatterjee and K. Jani, *Astrophys. J.* **982**, 102 (2025).
- [84] R. Legin, M. Isi, K.W.K. Wong, Y. Hezaveh, and L. Perreault-Levasseur, *Astrophys. J. Lett.* **985**, L46 (2025).
- [85] S. Tavaré, D.J. Balding, R.C. Griffiths, and P. Donnelly, *Genetics* **145**, 505 (1997).
- [86] J.K. Pritchard, M.T. Seielstad, A. Perez-Lezaun, and M.W. Feldman, *Mol. Biol. Evol.* **16**, 1791 (1999).
- [87] Y. Chen, M. U. Gutmann, and A. Weller, in *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)* (JMLR.org, Honolulu, Hawaii, USA, 2023), p. 178.
- [88] M. Deistler, J. Boelts, P. Steinbach, G. Moss, T. Moreau, M. Gloeckler, P.L.C. Rodrigues, J. Linhart, J.K. Lappalainen, B.K. Miller, P.J. Gonçalves, J.-M. Lueckmann, C. Schröder, and J.H. Macke, [arXiv:2508.12939](#).
- [89] G. Papamakarios, D.C. Sterratt, and I. Murray, [arXiv:1805.07226](#).
- [90] C. Durkan, G. Papamakarios, and I. Murray, [arXiv:1811.08723](#).
- [91] G. Papamakarios, D. Sterratt, and I. Murray, in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (AISTATS 2019)*, Proceedings of Machine Learning Research, edited by K. Chaudhuri and M. Sugiyama (PMLR, Naha, Okinawa, Japan, 2019), Vol. 89, pp. 837–848.
- [92] S. Dirmeier, C. Albert, and F. Perez-Cruz, [arXiv:2308.01054](#).
- [93] G. Papamakarios and I. Murray, [arXiv:1605.06376](#).
- [94] D.S. Greenberg, M. Nonnenmacher, and J.H. Macke, [arXiv:1905.07488](#).
- [95] B.K. Miller, A. Cole, P. Forré, G. Louppe, and C. Weniger, in *35th Conference on Neural Information Processing Systems* (2021).
- [96] B.K. Miller, A. Cole, C. Weniger, F. Nattino, O. Ku, and M.W. Grootes, *J. Open Source Software* **7**, 4205 (2022).
- [97] K. Cranmer, J. Brehmer, and G. Louppe, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 30055 (2020).
- [98] H. Gabbard, C. Messenger, I. S. Heng, F. Tonolini, and R. Murray-Smith, *Nat. Phys.* **18**, 112 (2022).
- [99] M. Crisostomi, K. Dey, E. Barausse, and R. Trotta, *Phys. Rev. D* **108**, 044029 (2023).
- [100] B. Wang, J. Leja, V. A. Villar, and J. S. Speagle, *Astrophys. J. Lett.* **952**, L10 (2023).
- [101] N. Anau Montel, Simulation-based inference for astrophysical data, Ph.D. thesis, Amsterdam U., 2024.
- [102] B. Liang and H. Wang, [arXiv:2507.11192](#).
- [103] E. Marx, D. Chatterjee, M. Desai, R. Kumar, W. Benoit, A. Sasli, L. Singer, M.W. Coughlin, P. Harris, and E. Katsavounidis, [arXiv:2509.22561](#).
- [104] P. Graff, F. Feroz, M. P. Hobson, and A. Lasenby, *Mon. Not. R. Astron. Soc.* **421**, 169 (2012).
- [105] L. Negri and A. Samajdar, [arXiv:2509.17606](#).
- [106] U. Bhardwaj, J. Alvey, B.K. Miller, S. Nissanke, and C. Weniger, *Phys. Rev. D* **108**, 042004 (2023).
- [107] J. Alvey, U. Bhardwaj, S. Nissanke, and C. Weniger, [arXiv:2308.06318](#).
- [108] I.M. Vilchez and C.F. Sopena, *J. Cosmol. Astropart. Phys.* **04** (2025) 022.
- [109] M. Dax, S.R. Green, J. Gair, J.H. Macke, A. Buonanno, and B. Schölkopf, *Phys. Rev. Lett.* **127**, 241103 (2021).
- [110] M. Dax, S.R. Green, J. Gair, M. Pürrer, J. Wildberger, J.H. Macke, A. Buonanno, and B. Schölkopf, *Phys. Rev. Lett.* **130**, 171403 (2023).
- [111] M. Dax, S.R. Green, J. Gair, N. Gupte, M. Pürrer, V. Raymond, J. Wildberger, J.H. Macke, A. Buonanno, and B. Schölkopf, *Nature (London)* **639**, 49 (2025).
- [112] G. Ashton, [arXiv:2510.11197](#).
- [113] A. Spurio Mancini, M.M. Docherty, M.A. Price, and J.D. McEwen, *RAS Tech. Instrum.* **2**, 710 (2023).
- [114] A. Tejero-Cantero, J. Boelts, M. Deistler, J.-M. Lueckmann, C. Durkan, P.J. Gonçalves, D.S. Greenberg, and J.H. Macke, [arXiv:2007.09114](#).
- [115] J.S. Speagle, *Mon. Not. R. Astron. Soc.* **493**, 3132 (2020).
- [116] M. Emma and G. Ashton, SBILBY: Residual neural likelihood evaluation for gravitational-wave inference (2026), Python package. Installable via pip: `pip install sbilby==0.0.1` <https://pypi.org/project/sbilby/>.
- [117] H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory*, Studies in Managerial Economics (Division of Research, Graduate School of Business Administration, Harvard University, Boston, 1961), pp. xxviii+356.
- [118] G. Cowan, *Statistical Data Analysis* (Clarendon Press, Oxford, 1998).
- [119] J. Alsing, T. Charnock, S. Feeney, and B. Wandelt, *Mon. Not. R. Astron. Soc.* **488**, 4440 (2019).
- [120] Y. Jo, S. Genel, B. Wandelt, R.S. Somerville, F. Villaescusa-Navarro, G.L. Bryan, D. Angles-Alcazar, D. Foreman-Mackey, D. Nelson, and J.-h. Kim, *Astrophys. J.* **944**, 67 (2023).
- [121] K. Lehman, S. Krippendorff, J. Weller, and K. Dolag, [arXiv:2411.08957](#).
- [122] G. Papamakarios, T. Pavlakou, and I. Murray, [arXiv:1705.07057](#).
- [123] B. Uria, M.-A. Côté, K. Gregor, I. Murray, and H. Larochelle, [arXiv:1605.02226](#).
- [124] D.J. Rezende and S. Mohamed, [arXiv:1505.05770](#).
- [125] M. Germain, K. Gregor, I. Murray, and H. Larochelle, [arXiv:1502.03509](#).
- [126] X. Glorot, A. Bordes, and Y. Bengio, in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (Fort Lauderdale, USA, 2011), Vol. 15, pp. 315–323.
- [127] J. Lin, *IEEE Trans. Inf. Theory* **37**, 145 (1991).
- [128] P. Virtanen *et al.*, *Nat. Methods* **17**, 261 (2020).
- [129] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, *Phys. Rev. Lett.* **113**, 151101 (2014).
- [130] L. Santamaria *et al.*, *Phys. Rev. D* **82**, 064016 (2010).
- [131] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. Jiménez Forteza, and A. Bohé, *Phys. Rev. D* **93**, 044006 (2016).
- [132] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez Forteza, and A. Bohé, *Phys. Rev. D* **93**, 044007 (2016).

- [133] B. P. Abbott *et al.* (KAGRA Collaboration, LIGO Scientific Collaboration, and Virgo Collaboration), *Living Rev. Relativity* **21**, 69 (2018).
- [134] C. Talbot *et al.*, [arXiv:2508.11091](#).
- [135] J. W. Tukey, in *Analysis of Time Series*, edited by B. Harris (Wiley, New York, USA, 1967), pp. 25–46.
- [136] A. G. Abac *et al.* (LIGO Scientific Collaboration, VIRGO Collaboration, and KAGRA Collaboration), [arXiv:2508.18080](#).
- [137] M. Maggiore, *Gravitational Waves. Vol. 1: Theory and Experiments* (Oxford University Press, New York, 2007).
- [138] M. Emma and G. Ashton, Scripts for: “RNLE: Residual neural likelihood estimation and its application to gravitational-wave astronomy” (2026) [10.5281/zenodo.19353018](#).
- [139] E. Baird, S. Fairhurst, M. Hannam, and P. Murphy, *Phys. Rev. D* **87**, 024035 (2013).
- [140] S. A. Usman, J. C. Mills, and S. Fairhurst, *Astrophys. J.* **877**, 82 (2019).
- [141] A. Trovato (LIGO Scientific Collaboration and VIRGO Collaboration), *Proc. Sci. Asterics2019* (2020) 082.
- [142] A. G. Abac *et al.* (LIGO Scientific Collaboration, VIRGO Collaboration, and KAGRA Collaboration), [arXiv:2508.18079](#).
- [143] D. M. Macleod, J. S. Areeda, S. B. Coughlin, T. J. Massinger, and A. L. Urban, *SoftwareX* **13**, 100657 (2021).
- [144] B. F. Schutz, *Classical Quantum Gravity* **28**, 125023 (2011).
- [145] H.-Y. Chen, D. E. Holz, J. Miller, M. Evans, S. Vitale, and J. Creighton, *Classical Quantum Gravity* **38**, 055010 (2021).
- [146] F. Robinet, Omicron: An algorithm to detect and characterize transient events in gravitational-wave detectors, Tech. Report No. LIGO-T1500073 (LIGO Scientific Collaboration, 2015).
- [147] F. Robinet, N. Arnaud, N. Leroy, A. Lundgren, D. Macleod, and J. McIver, *SoftwareX* **12**, 100620 (2020).
- [148] M. Zevin *et al.*, *Classical Quantum Gravity* **34**, 064003 (2017).
- [149] J. Glanzer *et al.*, *Classical Quantum Gravity* **40**, 065004 (2023).
- [150] M. Zevin *et al.*, *Eur. Phys. J. Plus* **139**, 100 (2024).
- [151] M. Cabero *et al.*, *Classical Quantum Gravity* **36**, 15 (2019).
- [152] H. Lyu, J. Alvey, N. Anau Montel, M. Pieroni, and C. Weniger, *Mach. Learn. Sci. Tech.* **7**, 025036 (2026).
- [153] S. Dirmeier, S. Ulzega, A. Mira, and C. Albert, [arXiv:2409.19435](#).
- [154] [gwosc.org](#).