

Reconstructing and resampling: a guide to utilising posterior samples from gravitational wave observations

Gregory Ashton^{1,2,*}

¹Physics Department, Royal Holloway, University of London, Egham Hill, Egham, TW20 0EX, United Kingdom

²Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, United Kingdom

11 February 2026

ABSTRACT

The LIGO, Virgo, and KAGRA (LVK) gravitational-wave observatories have opened new scientific research in astrophysics, fundamental physics, and cosmology. The collaborations that build and operate these observatories release the interferometric strain data as well as a catalogue of observed signals with accompanying Bayesian posterior distributions. These posteriors, in the form of equally-weighted samples, form a dataset that is used by a multitude of further analyses seeking to constrain the population of merging black holes, identify lensed pairs of signals, and much more. However, many of these analyses rely, often implicitly, on the ability to reconstruct the likelihood and prior from the inputs to the analysis and apply resampling (a statistical technique to generate new samples varying the underlying analysis assumptions). In this work, we first provide a guide on how to reconstruct and modify the posterior density accurately from the inputs for analyses performed with the `Billby` inference library. We then demonstrate and compare resampling techniques to produce new posterior sample sets and discuss Pareto-smoothing to improve the efficiency. Finally, we provide examples of how to use resampling to study observed gravitational-wave signals. We hope this guide provides a useful resource for those wishing to use open data products from the LVK for gravitational-wave astronomy.

Key words: gravitational waves, methods: data analysis

1 INTRODUCTION

Gravitational wave observations of merging black holes and neutron stars are now routinely made by interferometric gravitational wave detectors. There are currently four operational kilometre-scale interferometers: Laser Interferometer Gravitational-Wave Observatory (LIGO: Aasi et al. 2015), Virgo (Acernese et al. 2015), and KAGRA (Akutsu et al. 2021). Their raw output is a time series of calibrated strain, representing the relative difference in arm lengths for each instrument. The LIGO–Virgo–KAGRA (LVK) Collaborations publish this data (Abac et al. 2025a) in addition to the Gravitational-Wave Transient Catalog (GWTC: Abac et al. 2025b), containing a list of candidate signals and measurements of their Bayesian posterior distributions (Abac et al. 2025c).

The posterior distribution (see Thrane & Talbot 2019, for a discussion on why a Bayesian framework is used) is defined as the probability density of a set of source parameters θ conditional on the data d and model M (which implicitly defines the parameters) and can be calculated up to a normalisation factor from

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

where $\mathcal{L}(d|\theta, M)$ is the likelihood¹ and $\pi(\theta|M)$ is the prior proba-

bility distribution of the parameters conditioned only on the model. The normalisation factor that ensures $\int p(\theta|d, M) d\theta = 1$ can also be calculated and is referred to as the evidence:

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

The LVK analysis methodology for approximating Eq. (1) for observed signals is described in detail in Section 5 of Abac et al. (2025b). In brief, for transient gravitational-wave data analysis, analysis is performed on a time series of consecutive strain data observations, which we denote by d and define its duration to be T and sampling frequency f_s . Inference is performed on the frequency series \tilde{d} generated by applying a fast Fourier transform and dividing by f_s (a normalisation choice; see Thrane & Talbot (2019)). For signals lasting up to a few hundred seconds, we assume the non-astrophysical noise to be generated by a stationary additive coloured Gaussian noise process with a power spectral density (PSD) P such that the likelihood for the j th frequency bin is given by

$$\mathcal{L}(d_j|\theta, M) = \frac{1}{2\pi P_j} \exp\left(-\frac{2}{T} \frac{|\tilde{d}_j - \tilde{\mu}_j(\theta)|^2}{P_j}\right), \quad (3)$$

where $\tilde{\mu}_j(\theta)$ is the frequency-domain astrophysical signal model projected onto the response of the detector. The full likelihood is then calculated from the product over the j frequency bins, i.e.

$$\mathcal{L}(d|\theta, M) = \prod_j \mathcal{L}(d_j|\theta, M). \quad (4)$$

* E-mail: gregory.ashton@ligo.org

¹ Formally, the likelihood is the conditional probability of the data conditional on the parameters, but in parameter estimation, where the observed data is fixed, we treat it as a function of the parameters.

In practise, this is computed as the sum in log-space for numerical stability.

Typically, $\tilde{\mu}_j(\theta)$ is a waveform approximant that models the inspiral–merger–ringdown of the compact binary coalescence (CBC) signal utilising either phenomenological fits (Ajith et al. 2007), the effective one-body approach (Buonanno & Damour 1999, 2000), or numerical-relativity surrogates (Blackman et al. 2017). To date, `Bilby` uses the `LALSuite` (LIGO Scientific Collaboration et al. 2018; Wette 2020) library to access these waveforms.

Meanwhile, the prior distribution $\pi(\theta|M)$ is either generally weakly astrophysically motivated or informed by the observed population of CBC sources (Abac et al. 2025b).

Eq. (1) is not solvable in closed form for the CBC likelihood and prior, and the distribution is typically non-Gaussian, multimodal, and contains strong curved degeneracies. Therefore, computational Bayesian inference methods such as Markov chain Monte Carlo (MCMC; Hastings 1970) and nested sampling (Skilling 2006) are used with post-processing to produce a set of equally-weighted posterior samples $\theta_i \sim p(\theta|d, M)$ and an estimate of the evidence $\mathcal{Z}(d|M)$. To date, the LVK has used `LALInference` (Veitch et al. 2015), `RIFT` (Pankow et al. 2015; Lange et al. 2017; Wysocki et al. 2019), and `Bilby` (Ashton et al. 2019; Romero-Shaw et al. 2020), with an adaptation of the `Dynesty` (Speagle 2020) nested sampling implementation, to produce posterior samples. In this work, we will consider the LVK posteriors produced by `Bilby` and packaged using `PESummary` (Hoy & Raymond 2021). This is motivated by the ease of use in reconstructing the posteriors using `Bilby` (part of its design philosophy) and the wide applicability: `Bilby` posteriors have been estimated for all signals above threshold for parameter estimation from the first observing run until the most recent release, see GWTC-2.1 (Abbott et al. 2024), GWTC-3.0 (Abbott et al. 2023), and GWTC-4.0 (Abac et al. 2025c).

The posterior samples from a given observed signal form a data set that can be used for further analyses. For example, taking samples from a set of highly probable signals, the samples form an input data set that can be used to constrain the astrophysical distribution of merging black holes (see, e.g. Talbot & Thrane 2018; Abac et al. 2025d) or measure the cosmological properties of our universe (see, e.g. Schutz 1986; Abac et al. 2025e). These methods typically perform inference on a hierarchical population model using the GWTC posteriors as input data (Thrane & Talbot 2019). On the other hand, considering individual events, studies can be conducted using the posterior samples to examine the inferences under alternative assumptions about the waveform model, prior distribution, or data. Underlying all these methods is the technique known as *resampling*, which utilises a set of weights, constructed from the ratio of the new to the old posterior density, to generate a new set of samples under some changed assumptions. Resampling is an alternative to a re-analysis (i.e. applying stochastic sampling to estimate the posterior afresh), which can be computationally prohibitive.

Two types of resampling are in common use: rejection sampling (RS) and importance sampling (IS). For RS (discussed further in Section 3.1), samples are accepted into the new set in proportion to their weight. Meanwhile, IS (discussed further in Section 3.2) is a broad term, often without a consensus definition, but generally refers to the process of using the weights to determine the importance of the samples. In some use cases, this involves using the weights and samples together (e.g. to create a weighted histogram). Alternatively, some use cases apply multinomial sampling to produce a new sample set according to the probabilities provided by the weights (either with or without replacement). In this work, we refer to the former as IS and the latter as multinomial-IS.

Resampling techniques have long been used in the field. For example, both MCMC and nested sampling utilise resampling to produce sample sets; Rover et al. (2006) used multinomial-IS to initialise a MCMC sampler; Payne et al. (2019) demonstrated the use of IS to utilise higher-order gravitational-wave models; Payne et al. (2020) extended this to marginalise over the calibration model (and this was implemented in Abbott et al. (2024)); Baka et al. (2025) explored RS to correct for calibration misspecification; Tiwari (2018) and Talbot et al. (2019) used IS to estimate the binary black hole (BBH) population sensitivity; RS is used to apply a cosmological prior to LVK analyses (Abac et al. 2025b), and an astrophysically-motivated prior (Chattopadhyay et al. 2024; Abac et al. 2025c); Dax et al. (2023) uses IS with a proposal distribution provided by neural posterior estimation (but generally the technique can be applied to any pre-computed sample set with a tractable posterior); Williams et al. (2025) investigated sequential-Monte Carlo approaches, an extension of IS; and finally, Hourihane & Chatziioannou (2025), use IS changing the data in the likelihood to study the impact of transient non-Gaussian noise. There are many more examples besides this in the literature. However, we caution that the terms are often confused and there are additional terms such as “reweighting” that we choose to avoid since this is used to describe both RS and IS.

The goal of this work is to provide a simple guide to the application of resampling, along with a discussion of reconstructing the likelihood and prior distribution. We begin in Section 2 by discussing how the likelihood and prior can be calculated from the released posterior samples, including a comparison with the original calculations stored in data files. We also provide estimates of the accuracy achievable given practical hardware and software considerations. Then, in Section 3, we go on to introduce and compare the methodology of resampling, including RS and IS and introduce Pareto-smoothing as a means to improve the efficiency. In Section 4, we provide examples using publicly available data sets and applying resampling to illustrate the possibilities. Finally, we conclude with a discussion in Section 5. Readers may find it useful to refer to the data release associated with this work (Ashton 2026), which contains worked examples.

2 RECONSTRUCTING THE LIKELIHOOD AND PRIOR

In this Section, we will detail how to reconstruct the likelihood and prior for samples drawn from the posterior distribution presented in the digital GWTC. Specifically, we consider existing analyses of gravitational-wave signals using the `Bilby` inference library and then packaged using `PESummary`. Examples of such files can be found as part of the data released with GWTC-4.0 (LIGO Scientific Collaboration et al. 2025a).

Each file stores a set of posterior samples $\{\theta_i\}$. Additionally, users will find columns corresponding to the log-likelihood and log-prior (in all cases, these are natural logarithms). We refer to these as the *stored* values and denote them with a sub-script *S* (e.g. $\ln \mathcal{L}_S$ and $\ln \pi_S$) to differentiate them from the *reconstructed* values that we denote with a sub-script *R* (e.g. $\ln \mathcal{L}_R$ and $\ln \pi_R$). For accurate reconstruction of the log-likelihood and log-prior for a sample, it is important to match the original computation. Any differences can manifest as either systematic shifts (which could entirely bias any subsequent resampling) or random shifts (which, if sufficiently small, may be negligible). In the following, we address each point in turn.

Analysis	conda environment	Python	Bilby	BilbyPipe	Dynesty	LALSimulation
GWTC-2.1	igwn-py38-20210107	3.8	1.0.4	1.0.2	1.0.1	2.4.0
GWTC-3.0	igwn-py38-20210107	3.8	1.0.4	1.0.2	1.0.1	2.4.0
GWTC-4.0	igwn-py310-20241106	3.10	2.2.2.1	1.4.0	2.1.4	6.0.0

Table 1. A table detailing the conda environment and specific top-level packages used in recent GWTC updates. Full details of the environments can be found in the data release associated with this work.

2.1 Computing environment

There are two crucial aspects to consider regarding the computing environment used to reconstruct the likelihood and prior.

First, the versions of the software should match those used in the analysis: this is because changes to the software may reflect changes to the underlying assumptions. For example, if the default assumptions about data processing change between versions of the software, this will induce systematic shifts in the likelihood. Similarly, if the waveform model is not identical (including specific options hard-coded into the environment), this can also induce systematic and random shifts in the likelihood. Information about the software packages used in the analysis is packaged as part of the PESummary results files (see the data release for details on how to access the information).

However, for all LVK analyses, the conda (docs.anaconda.com/) package manager has been used to fully specify the environment, and typically it is easier to use a conda environment rather than attempt to match packages individually. Therefore, in Table 1, we list the conda environments used for pertinent updates to the GWTC alongside the versions of the packages that have the most significant impact (namely the sampling and simulation libraries); full details of the environments can be found in the data release of this work.

Second, for exact reconstruction, the hardware should also be identical. This is because the underlying hardware determines the implementation of floating-point arithmetic: differences in this implementation will change the byte-level results. There may also be platform-specific optimisation of linear algebra libraries that can differ even if the hardware is identical. Comparing an analysis run on an Intel Xeon Processor (Skylake) and then reconstructed on an Apple M3 Pro, we find the distributions of log-likelihoods are centred on zero but with a standard deviation of $\approx 3 \times 10^{-6}$. Meanwhile, when they are reconstructed on identical hardware, we find the log-likelihoods can be reconstructed up to floating-point precision. However, typically, it is not possible to match the hardware used in the analysis. But, as we will show later, differences of less than 1 part in a million are sufficiently small that they have a negligible impact on the capacity to resample the posterior.

2.2 Likelihood

To reconstruct the likelihood, Eq. (3), for a given sample, we must match the data, PSD, waveform model, and likelihood configuration. In the data release, we demonstrate how this can be achieved, taking the settings from the configuration file packaged as part of the data release. In the following, we provide a discussion of specific elements.

The likelihood ratio – Eq. (3) is the likelihood of the data under the model comprised of an additive signal and noise. However, we can see that if $\mu_j = 0 \forall j$, i.e. there is no signal, just noise, then we can also introduce $\mathcal{L}(d|N)$, the noise likelihood. Since the noise likelihood is θ -independent, it is common during sampling to use the likelihood ratio

$$\Lambda(d|\theta, M) = \frac{\mathcal{L}(d|\theta, M)}{\mathcal{L}(d|N)}, \quad (5)$$

rather than the likelihood itself. While this choice does not change the parameter estimates, it is often used because the evidence estimate obtained (e.g. from nested sampling) is the signal vs noise Bayes factor, i.e. $\mathcal{B} = \mathcal{Z}(d|M)/\mathcal{Z}(d|N)$, which is easier to interpret as a diagnostic during sampling rather than $\mathcal{Z}(d|M)$ itself.

When reconstructing the likelihood, it is important to distinguish between the likelihood and the likelihood ratio. For resampling, one only needs to be consistent. However, evidence estimates computed from the likelihood will yield measurements of the signal evidence. In contrast, those computed from the likelihood ratio will yield measurements of the signal vs noise Bayes factor. Bilby can compute either as demonstrated in the data release.

Data – The raw strain data used in LVK analyses is available from the Gravitational Wave Open Science Center (GWOSC) at <https://gwosc.org/data/> and available either at the original sampling frequency of 16384 Hz or downsampled to 4096 Hz. While there is not expected to be any signal content above a few thousand Hertz, for accurate reconstruction, it is important to start from the data sampled at 16382 Hz since the details of the downsampling algorithm do impact the entire data set: using the pre-downsampled data can result in significant deviations between the reconstructed and original likelihoods. For some events, the raw data is affected by transient non-Gaussian noise artefacts known as *glitches* (Nuttall 2018; Glanzer et al. 2023; Soni et al. 2025). In many cases, the glitches can be modelled and removed from the data using either a modelled Bayesian approach (BayesWave: Pankow et al. 2018; Cornish et al. 2021; Chatzioannou et al. 2021; Hourihane et al. 2022), or a linear noise subtraction based on auxiliary witness channels (Davis et al. 2022). Therefore, to recreate the likelihood, the glitch-subtracted data should be used (see, e.g. LIGO Scientific Collaboration et al. 2025b, for glitch-subtracted data for events from the first part of the fourth observing run).

Downsampling – For LVK analyses, Bilby uses the ResampleTimeSeries downsampling routines implemented in LALSuite which implements a time-domain Butterworth filter (Butterworth et al. 1930). This filter was selected to match the downsampling performed by the tools used to generate the PSD (discussed shortly). The specific choice of sampling frequency is stored in the configuration file for the Bilby analysis.

Windowing – The strain data used for inference is a subset of a longer time series of strain taken during the observation run. Typically, a short section of data is cut out, containing the observable signal and sufficient padding on either side. As a result, this time series is not periodic, which is required to approximately diagonalise the noise covariance matrix (see the discussion in Talbot et al. 2025). Therefore, a window factor is applied, gradually tapering the time series to zero at either end. Following standard methods (Abac et al. 2025b), a Tukey window (Harris 2005) is typically applied with a roll-off determined by a factor determining the fraction of the window that is tapered. Before passing the data to likelihood, this window should be applied to the time-domain strain.

PSD – The GWTC analyses use BayesWave (Cornish & Littenberg 2015; Littenberg & Cornish 2015; Cornish et al. 2021; Gupta & Cornish 2024) to estimate the on-source PSD for each detector. This

is packaged within the data release and can be extracted and provided to the `Bilby` likelihood. Note that the data pre-processing should be identical for the analysis data and PSD construction.

Waveform – The `Bilby` likelihood uses a waveform approximant provided by `LALSimulation` to simulate the frequency-domain signal $\mu(\theta)$ in Eq. (3). This is implemented within `Bilby` in a `WaveformGenerator` object that needs to be configured with the name of the waveform approximant to use, in addition to any configuration settings.

Explicit marginalization – The likelihood used for inference, at base, is the Whittle likelihood as given in Eq. (3). However, `Bilby` implements several forms of explicit marginalisation, and these are routinely used to improve the convergence properties of the samplers (Abac et al. 2025b). As further described in Thrane & Talbot (2019); Romero-Shaw et al. (2020), after sampling `Bilby` then reconstructs the full posterior distribution in post-processing. Therefore, it is important to differentiate between the marginalised likelihood, which is used for sampling and the non-marginalised likelihood, which corresponds to the full posterior distribution. In the data release, we show how both of these can be calculated. In addition, we show that the log-likelihood that is stored in the data release corresponds to the marginalised likelihood. The not-marginalised likelihood is not stored within the packaged data release, but can be calculated from the matched-filter and optimal signal-to-noise ratios (SNRs) that are stored (as shown in the data release).

Other configuration – While formally the likelihood for inference is a sum over the frequency bins of the Fourier transform, it is standard practise to omit data below a minimum frequency f_{\min} and above a maximum frequency f_{\max} . Typically, f_{\min} is set to 20 Hz, except in cases where the data is affected by low-frequency non-Gaussian noise that cannot be subtracted, in which case a higher minimum frequency is chosen. Meanwhile, the maximum frequency is by default set at a nominal choice above the maximum frequency of the signal. In addition, `Bilby` also has options to modify the choice of parameterisation. For example, while the analyst may set a prior distribution on the geocentric arrival time (Abac et al. 2025f), it is often advantageous for sampling to instead sample in the merger time relative to a fiducial detector.

2.3 Prior distribution

`Bilby` provides an interface for defining the prior in terms of a Python dictionary. The *analysis* prior is packaged as part of the data release and can be easily reconstructed; this prior corresponds to the prior for the full posterior (e.g. after reconstruction of any marginalised parameters). It is also helpful to define the *run* prior that is used during stochastic sampling by `Bilby`. The run prior is created by modifying the *analysis* prior during the instantiation of the likelihood. For example, if a parameter is explicitly marginalised, the run prior will replace the prior on this parameter with a delta-function prior at a fiducial value. As with the likelihood, the log-prior stored in the result is from the run prior used during sampling.

In addition, some data releases apply a resampling of the posterior to utilise a cosmological prior on the luminosity distance. Care must therefore be taken to ensure the prior is appropriately constructed to match the posterior samples.

2.4 Calibration

The strain has a known calibration uncertainty, which is represented in the frequency domain by a distribution on the phase and amplitude (see, e.g. Sun et al. 2020). For standard LVK analyses, `Bilby`

marginalises over the calibration uncertainty by taking as input a calibration envelope from which it constructs a spline approximation with an associated set of calibration parameters. These are then sampled during inference to marginalise over the calibration uncertainty (Farr et al. 2014; Abbott et al. 2016b). The calibration envelope is packaged as part of the data release and can be read in and used to construct a prior on the spline nodes. A calibration model must also be provided to the interferometer instances that contain the strain data. A demonstration of how this is done is provided in the data release.

2.5 Demonstrations

In the data release, we provide notebooks demonstrating how to implement the setting described within this section for two analyses and calculate the reconstructed likelihood (both marginalised and non-marginalised) and prior. We also show how to extract the stored values of the likelihood and prior. We choose GW150914, the first observed binary black hole merger (Abbott et al. 2016a), and GW230627_015337, a high-SNR binary black hole observed during the first part of the fourth observing run (Abac et al. 2025c). For GW150914, we use the GWTC-2.1 analysis (Abbott et al. 2024; LIGO Scientific Collaboration and Virgo Collaboration 2022), which, amongst other things, utilised upgraded waveform models relative to the initial studies. In Fig. 1, we show histograms of the residual between the reconstructed and stored log-likelihood and log-prior; in Table 2, we summarise fits to the residuals. In both instances, the residuals are centred on zero, but while for the prior reconstruction we can achieve floating-point precision, for the likelihood, the residual values are scattered with random offsets below the level of one in a thousand. For the log-likelihood, we fit both a normal distribution and a Student's T distribution, showing that the typical variations better fit the Student's T distribution with heavy tails.

While we do not extend this analysis to all observed events, we believe the notebooks in the data release should allow users to reconstruct the likelihood and prior for any observed event where `Bilby` analyses have been produced by the LVK.

3 RESAMPLING THE POSTERIOR

In this Section, we will introduce the formalism of resampling in the context of applications to a set of released posterior samples $\{\theta_i\}$. We note that this presentation is intended to be specific to the use-case of gravitational-wave astronomy; for a more general guide, see, e.g. MacKay (2003).

Given a sample θ_i , we can calculate the unnormalised posterior density

$$p(\theta_i) = \mathcal{L}(d|\theta_i, M)\pi(\theta_i|M), \quad (6)$$

where we use the term density to distinguish it from the distribution indicated in Eq. (1) and drop the conditional statements to simplify the notation. In the language of resampling, we refer to $p(\theta)$ as the *proposal* distribution. It is assumed that the set of samples $\{\theta_i\}$ is properly drawn from the distribution; if this is not the case, for example, if the stochastic sampler has not converged, then resampling will inherit the bias.

Next, we wish to resample to a *target* distribution $p'(\theta)$ with an associated density:

$$p'(\theta_i) = \mathcal{L}'(d'|\theta_i, M')\pi'(\theta_i|M') \quad (7)$$

where the prime could mean a variation in the choice of likelihood

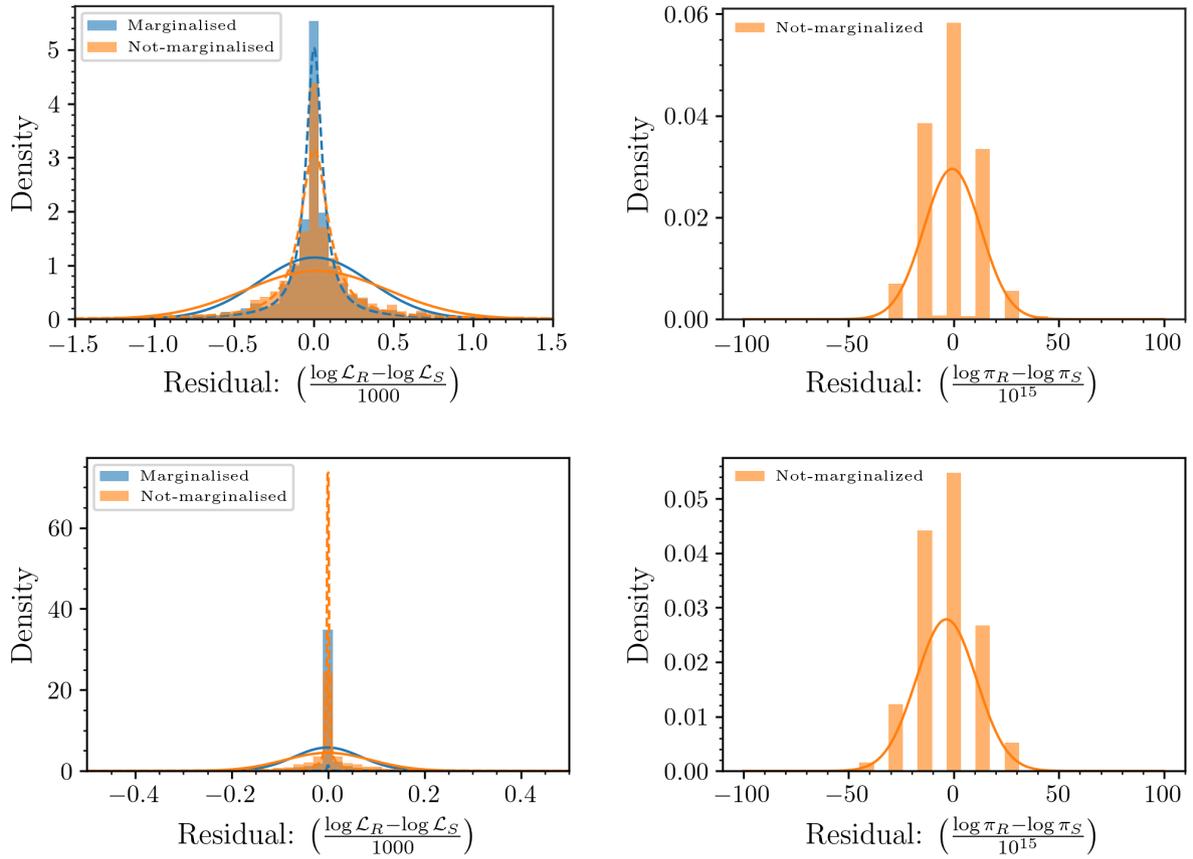


Figure 1. Histograms showing the residual difference between the reconstructed and stored log likelihood (left-hand column) and log-prior (right-hand column) for the GWTC-2.1 analysis of GW150914 (top row) and GWTC-4.0 analysis of GW230627_015337 (bottom row). We fit a normal (solid curve) and Student's T distribution (dashed curve) to the likelihood residuals and present summaries of these fits in Table 2.

Event	Dist.	Range	Normal	Student's T
GW150914	Marginalised	(-4, 4)	(0.005, 0.3)	(0.002, 0.9, 0.06)
	Not-marginalised	(-4, 5)	(0.02, 0.4)	(0.003, 1, 0.1)
	Prior	(-40, 40)	(-0.7, 10)	—
GW230627_015337	Marginalised	(-0.6, 0.4)	(-0.0008, 0.07)	(2×10^{-12} , 0.3, 9×10^{-11})
	Not-marginalised	(-0.5, 0.4)	(-0.0004, 0.09)	(2×10^{-6} , 0.3, 0.0005)
	Prior	(-40, 40)	(-4, 10)	—

Table 2. Summary statistics for the distribution of the residuals presented in Fig. 1. We give the minimum and maximum of the range, the fitted mean and standard deviation of a normal distribution, and the fitted mean, standard deviation, and degree of freedom for the Student's T distribution (but this is not applied to the prior residual). For the likelihood residuals, all values are scaled by 1000, while for the prior residuals, values are scaled by 10^{15} .

$\mathcal{L} \rightarrow \mathcal{L}'$ (which could include, for example, changes to the data processing or assumptions about the noise), a change in the generative model $M \rightarrow M'$, or a change in the choice of prior $\pi \rightarrow \pi'$ (under some interpretations, the prior is part of the model, but here we distinguish it as a separate choice to make it clear if instead the generative model alone is changing). Note that while it is possible to change more than one of these at a time, it would be advisable to change one at a time, at least initially, to understand their individual impact.

Given the proposal distribution and target distribution, we define a generalised weight function

$$w(\theta) = \frac{p'(\theta)}{p(\theta)}. \quad (8)$$

Then, given the set of samples and the computed proposal and target values, we construct a set of weights calculated as the ratio of the densities:

$$w_i = w(\theta_i) = \frac{p'(\theta_i)}{p(\theta_i)}. \quad (9)$$

The goal of the following subsections is to outline different strategies to resample the posterior using these weights as an alternative to reanalysis, applying stochastic sampling directly to estimate $p'(\theta|M, d)$. We note that in some applications, the calculation of the weights greatly simplifies. For example, if only the prior changes, then the likelihood terms cancel. However, our examples are given with the fully general case to enable the reader to apply it in arbitrary contexts.

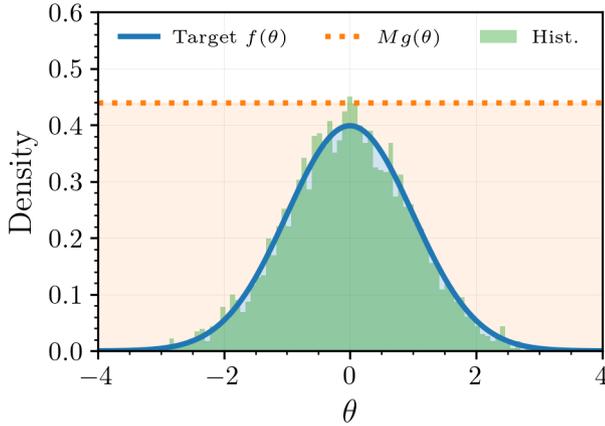


Figure 2. A simple illustration of rejection sampling for a standard normal target density $f(x)$ and uniform proposal density $g(x)$. We plot the target density (blue solid curve), scaled proposal density (orange dashed curve), and a histogram of the accepted samples from 10 000 iterations of the algorithm. Proposal samples, θ_i drawn from $p(\theta)$ are accepted with a probability proportional to $p'(\theta)/Mp(\theta)$. Visually, in the figure above, this is the ratio of the solid blue curve to the orange dashed line.

3.1 Rejection sampling

RS is a classical Monte Carlo technique which draws samples from a proposal distribution $p'(\theta)$ using a generating distribution $p(\theta)$ and a constant M which bounds the ratio:

$$M \geq \sup_{\theta} \frac{p'(\theta)}{p(\theta)}, \quad (10)$$

where $\sup_{\theta} f(\theta)$ is the *supremum* of $f(\theta)$ for all possible values of θ ; in the context of continuous functions with a maximum, the supremum is the maximum. RS starts by drawing a sample $\theta_i \sim p(\theta)$ and then defining an acceptance probability for the sample $p'(\theta_i)/Mp(\theta_i)$ which is proportional to how well the proposed sample matches the target distribution $p'(\theta)$ normalised by the envelope $Mp(\theta)$. Algorithmically, this is implemented by first drawing $u \sim \text{Unif}(0, 1)$ and then accepting the sample if

$$u < \frac{p'(\theta_i)}{Mp(\theta_i)}, \quad (11)$$

and rejecting the sample otherwise. After applying this test to all samples in $\{\theta_i\}$, we end up with a resampled posterior distribution $\{\theta'_i\}$ under the assumptions of the primed distribution. To illustrate the basic idea, in Fig. 2 we provide a plot from a simple example applying rejection sampling for a uniform proposal distribution and standard normal target distribution.

The bound M ensures that $p'(\theta) \leq Mp(\theta)$ for all possible values of θ and is ideally chosen to be as small as possible while respecting Eq. (10). Visually, in Fig. 2, the closer the $Mg(x)$ curve is to the peak of $f(\theta)$, the more efficient the sampler, as the smaller the rejection region is. More formally, if we define RS efficiency as

$$\eta_{\text{RS}} = \frac{N'}{N}, \quad (12)$$

where N' is the number of accepted samples. Then the efficiency can be calculated from averaging the acceptance ratio:

$$\eta_{\text{RS}} = \left\langle \frac{p'(\theta)}{Mp(\theta)} \right\rangle_{p(\theta)} = \frac{1}{M} \int_{\theta} p'(\theta) d\theta, \quad (13)$$

where we introduce $\langle g(x) \rangle_{f(x)}$ as the expectation of x under $f(x)$:

$$\langle g(x) \rangle_{f(x)} = \int g(x)f(x) dx. \quad (14)$$

Hence, the smallest value of M that satisfies Eq. (10) maximises the efficiency. In practice, the optimal value of M is unknown in advance and must be chosen instead. Typically, this is done by selecting

$$M = \max(w_i) \quad (15)$$

which ensures the condition is met over the set of samples. However, it does not ensure the condition is met in general. (From Fig. 2, it can be seen that if Eq. (10) is not met, the posterior will be clipped for all θ where $p'(\theta) > Mp(\theta)$).

If M is approximated by Eq. (15), the acceptance criteria is

$$u < \frac{w_i}{\max(w_i)}, \quad (16)$$

which is the standard implementation often used in the field.

With this choice of M , the efficiency can be computed directly from the weights as

$$\eta_{\text{RS}} = \frac{\langle \{w_i\} \rangle}{\max(\{w_i\})}, \quad (17)$$

where $\langle \{w_i\} \rangle$ is the mean of the weights. This demonstrates that efficient rejection sampling requires the distribution of weights to be compact, with a maximum not much larger than the mean. If, on the other hand, the maximum weight is much larger than the mean, sampling will be inefficient.

We can further explore the efficiency by considering the case when the weights are log-normally distributed (i.e. that the distribution of $\ln w_i$ is approximately normal) with a mean of μ and variance σ^2 . Then, the mean of the weights is given by $\exp(\mu + \sigma^2/2)$ while we can approximate the maximum of a set N weights by $\max(\{\theta_i\}) \approx \mu + \sigma\sqrt{2\log N}$ yielding

$$\eta_{\text{RS}} \approx \exp\left(-\sigma\sqrt{2\log N} + \frac{\sigma^2}{2}\right). \quad (18)$$

This approximation breaks down for $\sigma \gtrsim 1$, but provides a useful estimate of the expected efficiency for $\sigma < 1$. Furthermore, for $\sigma \ll 1$ we obtain

$$\eta_{\text{RS}} \approx 1 - \sigma\sqrt{2\log N}. \quad (19)$$

We can use this approximation to provide a bound on the required accuracy of likelihood reconstruction in Section 2, assuming they are log-normally distributed (in practise we find a Student's T distribution is a better fit to the residual of the log-likelihood weights, but it nevertheless provides an approximate bound). To ensure the sampling efficiency is impacted at the 1% level or less (for $N = 10000$ samples), we require $\sigma < 0.0035$. Comparing this to the distributions in Fig. 1, we can conclude that the reconstructions are sufficiently accurate to have a negligible impact on the resampling efficiency. Of course, in practise the distribution will depend on the changed assumptions and will likely not be log-normally distributed. Nevertheless, the approximations can provide a useful heuristic to predict the performance.

In the straightforward approach described above, the number of samples produced by RS, N' , is determined by the acceptance ratio (see Eq. (12)). However, we note that it is also possible to repeatedly apply the RS algorithm until a pre-defined number of samples is produced, N_t . If $N_t > N'$, the resulting sample set will no longer be independent and identically distributed (IID), but this technique is nevertheless useful for smoothing histograms (however, see the discussion in the next section for alternatives).

3.2 Importance sampling

IS is an alternative Monte Carlo approach which, in its textbook definition, rather than creating a new set of equally-weighted samples (as done by RS), instead utilises the samples and associated weights directly. For example, to visualise the re-weighted posterior distribution, a histogram could be produced using the weights to modify each sample's contribution to the histogram density. Or alternatively, for summary statistics, the weights can be used to modify the contribution from each sample. E.g., the weighted mean of a distribution can be computed as

$$\langle \theta \rangle_{p(\theta)} = \int p(\theta) \theta d\theta \approx \sum_i \bar{w}_i \theta_i, \quad (20)$$

where \bar{w}_i are the normalised weights

$$\bar{w}_i = \frac{w_i}{\sum_i w_i}. \quad (21)$$

There is no direct way to compare the efficiency of RS and IS because they are fundamentally different approaches. However, it can nevertheless be useful to consider a different measure of efficiency for IS, the Kish effective sample size (ESS):

$$N_{\text{ess}} = N \frac{\langle w \rangle^2}{\langle w^2 \rangle} = \frac{(\sum_i w_i)^2}{\sum_i w_i^2} = \frac{1}{\sum_i \bar{w}_i^2}, \quad (22)$$

where we give three common and equivalent definitions: in terms of the mean of the weights and squared weights, in terms of the weights, and in terms of the normalised weights.

The ESS was introduced (Kish 1965) as a measure of the loss of precision due to unequal weights. Specifically, it can be interpreted as the number of samples that would need to be drawn from the target distribution to match the variance of the estimate taken from the weighted distribution. However, the definitions given in Eq. (22) are an approximation that can be computed from the samples themselves (see Kong (1992) for the derivation and Elvira et al. (2022) for a detailed discussion).

We then define the IS efficiency as

$$\eta_{\text{IS}} = \frac{N_{\text{ess}}}{N}, \quad (23)$$

and hence, considering Eq. (22) and assuming the weights are log-normally distributed, the efficiency of IS can then be estimated from

$$\eta_{\text{IS}} = \frac{\langle w \rangle^2}{\langle w^2 \rangle} = \exp(-\sigma^2). \quad (24)$$

In Fig. 3, we compare the predicted efficiency of the RS and IS approaches assuming the weights are log-normally distributed. We also add a comparison with a numerical example, validating the predictions. This suggests that IS is more efficient for moderate values of σ , though both of course become highly inefficient as σ approaches unity. However, we note that these efficiencies should not be taken at face value, as they present fundamentally different statistical concepts: while η_{RS} is the acceptance rate of the RS algorithm, η_{IS} measures the ratio of the effective sample size of the weighted sample set. Nevertheless, they help us to identify a key difference between RS and IS, namely a trade-off between smoothness and noise: while RS produces pure equally-weighted samples, the level of noise (given by the square root of the number of accepted samples) is larger compared to IS, which keeps the entire set of samples (maximising the smoothness of the resulting histograms).

A common issue with IS is the requirement to package the weights alongside the samples. Since many downstream methods assume a

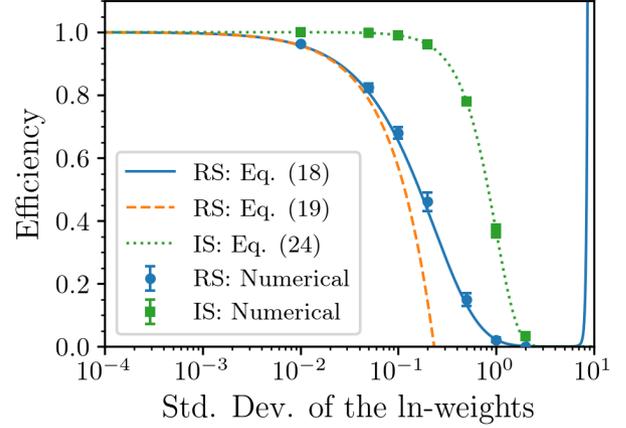


Figure 3. The efficiency of the RS and IS approaches as predicted by Eq. (18), Eq. (19) and Eq. (24). To numerically validate the theoretical predictions, we add numerical results from a toy model where the target distribution is a standard normal $x \sim N(0, 1)$, and the proposal is a shifted normal distribution $x' \sim N(\mu, 1)$. For this setup, the log-weight is $\mu x - \mu^2/2$, which has standard deviation $\sigma = |\mu|$ when sampling $x \sim N(0, 1)$. By varying μ , we can control σ and compare empirical efficiency measurements against the theoretical predictions, demonstrating agreement. For further details, please see the software behind the figure in the data release (Ashton 2026).

set of equally weighted samples, this can be problematic. An alternative is therefore to instead apply multinomial-IS in which we draw N_t samples (with replacement) from the original samples $\{\theta_i\}$ with probabilities given by the normalised weights in Eq. (21). N_t could either be taken to be the size of the original sample set, or a more conservative choice is to take N_{ess} . In either case, the resulting resampled posterior $\{\theta'_i\}$ is no longer IID (because it will contain repeated samples), but it can present smoother histograms compared to RS. We note that multinomial-IS is identical in behaviour to repeated application of RS (as discussed at the end of the last Section). However, multinomial-IS is computationally more efficient.

3.3 Pareto-smoothing

When the distribution of weights has a heavy right tail, both RS and IS can be highly inefficient. This typically occurs because the proposal and target distribution differ. A technique commonly applied in the field of statistics (but not yet common in gravitational-wave astronomy, though see Mould et al. (2025)) is the use of Pareto-smoothing, which modifies the right tail of the weight distribution to stabilise resampling and reduce the error of IS estimates (Vehtari et al. 2024). The approach was developed as an extension of IS and named Pareto-smoothed IS (PSIS). A generalised Pareto distribution is fitted to the logarithm of the largest weights (typically assigned by the user, with a typical value of the largest 20%). These weights are then replaced with the expected order statistics of the fitted distribution, and then the entire set of weights is re-normalised. This process smoothes the tails of the distribution, and enables the entire set of samples to be used in IS with improved efficiency. Pareto smoothing does induce a small bias, but this is often acceptable since only the tail of the distribution is affected. However, the algorithm yields a diagnostic \hat{k} , the tail index of the fitted distribution, that can be used to assess the impact. For $\hat{k} < 0.5$, the bias is expected to be negligible, and results can be trusted. For $0.5 \leq \hat{k} < 0.7$, care should be taken. And for $\hat{k} > 0.7$, the estimates are unstable or have a high

degree of bias. \hat{k} can also be negative, indicating that the tail of the weight distribution decays faster than a Pareto distribution would, and smoothing will have a negligible impact.

In Fig. 4, we present a demonstration of Pareto smoothing for some simulated weights with a heavy right tail using the *arviz* implementation (Kumar et al. 2019). We vary N , the number of samples, to show that smoothing is most effective when N is small, such that there is significant variance in the tail.

While PSIS is predominantly used in the context of IS, it can also be used to smooth the weights for RS. We refer to this as Pareto-smoothed RS (PSRS). We note that, like PSIS, the samples produced from PSRS will be biased as the distribution is an approximation to the target distribution. However, often this bias is worthwhile for the improvement in efficiency.

3.4 Evidence estimation

Typically, alongside the initial sample set $\{\theta_i\}$, the Bayesian evidence, Eq. (2), is also estimated during the stochastic sampling process. When resampling, we can use the weights, Eq. (9), to obtain a Bayes factor between the primed and unprimed evidence values as follows. First, we write the definition of the primed evidence

$$\mathcal{Z}'(d'|M') = \int \mathcal{L}'(d'|\theta, M') \pi'(\theta) d\theta. \quad (25)$$

Then multiply the integrand by the proposal likelihood and prior

$$\mathcal{Z}'(d'|M') = \int \frac{\mathcal{L}'(d'|\theta, M') \pi'(\theta|M)}{\mathcal{L}(d|\theta, M) \pi(\theta|M)} \mathcal{L}(d|\theta, M) \pi(\theta|M) d\theta. \quad (26)$$

$$= \int w(\theta) \mathcal{L}(d|\theta, M) \pi(\theta|M) d\theta. \quad (27)$$

where we have used the definition of the generalised weight function, Eq. (8). Now, we recognise that $\mathcal{L}(d|\theta, M) \pi(\theta|M) = \mathcal{Z}(d|M) p(\theta|d, M)$ and therefore

$$\mathcal{Z}'(d'|M') = \mathcal{Z}(d|M) \int w(\theta) p(\theta|d, M) d\theta, \quad (28)$$

and so finally, if we have N samples $\{\theta_i\}$ drawn from $p(\theta|d, M)$, then we can replace the integral with a Monte Carlo approximation and hence form a Bayes factor

$$\mathcal{B} = \frac{\mathcal{Z}'(d'|M')}{\mathcal{Z}(d|M)} = \frac{1}{N} \sum_i w_i = \langle w \rangle. \quad (29)$$

(Note: a similar derivation can be found in Payne et al. (2019) when only the likelihood is changing). For numerical stability, it is usually better to estimate the logarithm of the Bayes factor from

$$\ln \mathcal{B} = \text{LSE}(\{\ln w_i\}) - \ln(N), \quad (30)$$

where $\text{LSE}(\{x_i\}) \equiv \ln(\sum_i \exp(x_i))$ is the logarithm of the sum of the exponentials, a function for which computational libraries such as SciPy (Virtanen et al. 2020) provide convenient methods that preserve numerical stability.

Given a set of weights, Eq. (29) can be used to perform a comparison of the evidence under the two differing assumptions. If an estimate of $\mathcal{Z}(d|M)$ is available, then multiplying by the Monte Carlo average provides a new estimate of $\mathcal{Z}'(d'|M')$.

To estimate the uncertainty in the Bayes factor introduced by the Monte Carlo sum, we first note that the standard error on the Bayes factor is given by

$$\sigma_{\langle w \rangle} = \frac{\sigma_w}{\sqrt{N}} \quad (31)$$

where σ_w^2 is the true variance of the weights.

For a given set of weights, computing σ_w and inserting this into Eq. (31) can be used to estimate the uncertainty on the Bayes factor calculated in Eq. (29). If instead, the log-Bayes factor is calculated, Eq. (30), then propagating the uncertainty:

$$\sigma_{\ln \mathcal{B}} = \frac{\sigma_w}{\mathcal{B} \sqrt{N}}. \quad (32)$$

Starting from Eq. (31), we can understand the performance in more detail by estimating σ_w (the true variance of the weights) with the sample variance:

$$\sigma_w^2 \approx \frac{1}{(N-1)} \sum_i (w_i - \langle w \rangle)^2 \quad (33)$$

$$= \frac{1}{N-1} \left(\sum_i w_i^2 - N \langle w \rangle^2 \right), \quad (34)$$

where we have rearranged the expression in the second line. Next, recalling the definition of the ESS, we can rewrite Eq. (22) as

$$N_{\text{ess}} = \frac{N^2 \langle w \rangle^2}{\sum_i w_i^2}. \quad (35)$$

Now combining with Eq. (34), we can identify that

$$\sigma_w^2 = \langle w \rangle^2 \left(\frac{N}{N_{\text{ess}}} - 1 \right) \quad (36)$$

and hence the relative error in the Bayes factor, Eq. (29), is

$$\frac{\sigma_{\langle w \rangle}}{\langle w \rangle} = \frac{1}{\sqrt{N_{\text{ess}}}} \sqrt{1 - \frac{N_{\text{ess}}}{N}} = \frac{1}{\sqrt{N_{\text{ess}}}} \sqrt{1 - \eta_{\text{IS}}}. \quad (37)$$

This expression shows that when the efficiency is high, the variance in the estimator of the Bayes factor is small, and we recover the usual variance scaling $\propto 1/\sqrt{N}$. Meanwhile, when the efficiency is small, such that $N_{\text{ess}} \ll N$, the variance instead scales as $\propto 1/\sqrt{N_{\text{ess}}}$. In other words, not only is the N_{ess} useful in determining the efficiency of IS, it also determines the accuracy of the Monte Carlo integration approach to estimating the Bayes factor, Eq. (29).

4 EXAMPLES

In this Section, we will provide several simple examples of the use of resampling. We begin with a simple toy model that investigates the sampling approaches discussed in Section 3, then move on to demonstrations from gravitational-wave astronomy. Details of how to reproduce these examples can be found in the data release (Ashton 2026).

4.1 Validating and exploring resampling methods

We now demonstrate the application of the resampling methods described in Section 3 to a toy model. Of course, the methods are already well-validated within the statistics literature, but our goal is to use tools familiar to gravitational-wave astronomers to demonstrate the qualities of the methods.

To construct a toy problem, we consider a one-dimensional inference problem in which a variable x is measured to have a posterior distribution $p(x)$ that follows a truncated normal distribution bounded on the unit interval. We simulate an ensemble of measurements of x , drawing true values μ_x from a population prior distribution $\mu_x = \text{Unif}(0, 1)$. To each true value, we then simulate measurement error by adding a random variable $\Delta\mu_x \sim \text{Norm}(\delta\mu_e, \sigma_e)$,

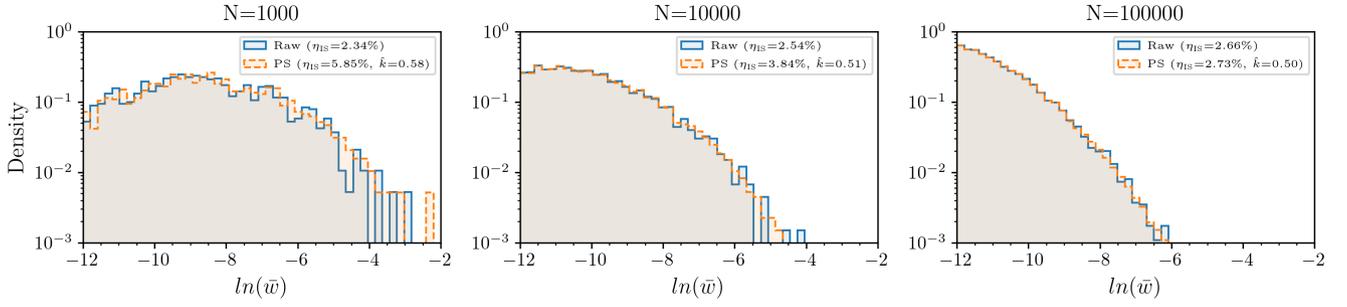


Figure 4. Histograms of simulated weights (Raw: blue solid curve) and Pareto-smoothed versions (PS: orange dashed curves). The weights are simulated from a distribution $w \sim e^n$, where n is a random variable drawn from a normal distribution with zero mean and standard deviation of 2. We plot histograms of the natural logarithm of the normalised weights and truncate the histogram at a lower edge to focus on the right-hand tail. In the legend, we provide efficiency metrics and the Pareto-smoothing diagnostic \hat{k} .

where $\delta\mu_e$ and σ_e are the bias and standard deviation in the measurement error. We then draw samples from the simulated posterior distribution, $x \sim \text{Norm}(\mu_x + \delta\mu_e, \sigma_p)$ where we introduce σ_p as the standard deviation of the Gaussian posterior. For unbiased posteriors, the mean of the measurement error is zero $\delta\mu_e = 0$ and the standard deviations of the measurement error σ_e and simulated posterior distributions σ_p are identical, $\sigma_e = \sigma_p$. However, if $\delta\mu_e \neq 0$, this simulates bias. Meanwhile, if $\sigma_p \neq \sigma_e$, this simulates a scale-bias (over-constrained or under-constrained).

We concoct a scenario in which the posteriors are under-constrained $\sigma_e/\sigma_p = 2$, but unbiased, with $\delta\mu_e = 0$, and then generate many realisations of the posterior distribution. In Fig. 5, we calculate a probability probability (PP) test (Siderly et al. 2014; Veitch et al. 2015; Romero-Shaw et al. 2020) showing that the concocted posterior distributions are scale-biased (do not follow the expected 1-to-1 relation) and fail a p -value test. We discuss the specifics of the PP test further in Appendix A and extend the investigation to also consider $\delta\mu_e \neq 0$, showing how bias and scale-bias can be diagnosed from the failure of the PP test.

Taking the posterior samples, we then calculate the weights as the ratio of the likelihood from the correct normal likelihood with standard deviation σ_p to the erroneous distribution σ_e . In Fig. 5, we then perform PP tests for RS and multinomial-IS with and without Pareto smoothing. For all four cases, the PP test illustrates that the resampled posteriors are unbiased.

We then repeat the simulation study, varying σ_p while keeping σ_e fixed. In Fig. 6, we plot the efficiencies as calculated directly from Eq. (12) and Eq. (23). This allows us to study the performance of the different methods in practice (though, as discussed in Section 3.2, the efficiency measures of RS and IS are not an apples-to-apples comparison).

For $\sigma_e/\sigma_p < 1$, the generating distribution is broader than the target distribution (i.e. the concocted posterior is under-constrained). This is the ideal case for resampling, and we find good efficiency with RS behaving linearly. Multinomial-IS does yield greater efficiency (and hence will produce smoother histograms), but at the cost of repeated samples. We do not find that Pareto smoothing has any impact in this regime. This is further confirmed by the measurements of \hat{k} (right-hand axes), which are less than 0 for all $\sigma_s/\sigma_p < 1$.

For $\sigma_e/\sigma_p > 1$, the generating distribution is narrower than the target distribution (i.e. the concocted posterior is over-constrained). In this case, the efficiency of RS rapidly decays while IS exhibits a more gradual decrease. Pareto smoothing improves the performance of IS by a few tens of per cent (and RS to a lesser extent). However,

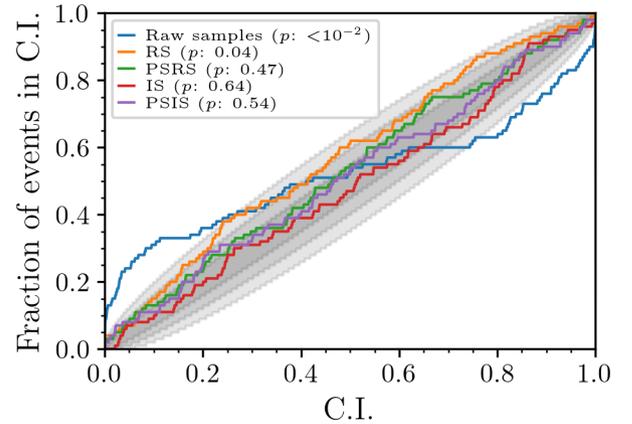


Figure 5. A PP test for the concocted under-constrained posterior distributions with $\sigma_e/\sigma_p = 2$ as described in Section 4.1. In the legend, we give the p -value of the PP test applied to the raw samples (which fail at a threshold of 0.01) while all four resampled posterior sample sets pass at this threshold. In Appendix A, we describe the details of how the PP test is applied and how the p -values provided in the legend are calculated.

as shown on the right-hand axis, above $\sigma_e/\sigma_p > 2$, the median of the \hat{k} distribution exceeds 0.7, a threshold above which the results can be biased (Vehtari et al. 2024). For cases like this (and more extreme), where the generating distribution does not cover the target, it is generally advisable to use an alternative approach such as a re-analysis of the data, or sequential Monte Carlo as explored in Williams et al. (2025).

4.2 Changing the waveform model for GW150914

To illustrate the application of resampling to gravitational-wave observations, we take the GWTC-2.1 data released for the LVK analysis of GW150914 Abbott et al. (2024); LIGO Scientific Collaboration and Virgo Collaboration (2022) with the IMRPhenomXPHM waveform model (Pratten et al. 2021). We then reconstruct the likelihood, using the open data from GWOSC (Abac et al. 2025a), and then calculate a new likelihood replacing the waveform with IMRPhenomXP (a special case of IMRPhenomXPHM which does not contain higher-order modes). This is in contrast to how resampling would usually be used (see, e.g. Payne et al. 2019), but we provide it as a simple illustra-

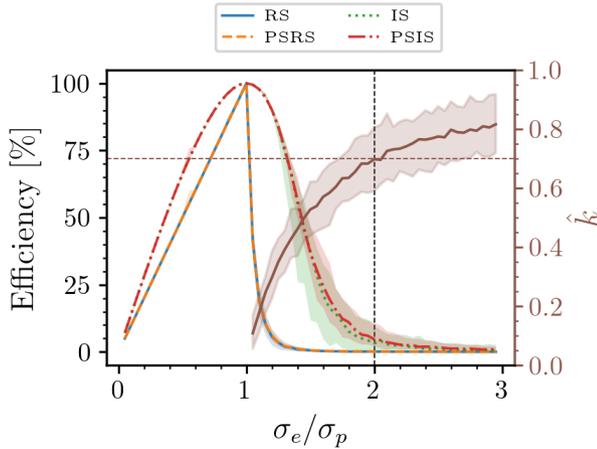


Figure 6. Measures of the efficiency of four resampling algorithms applied to concocted posterior distributions, described in Section 4.1. For RS and PSRS, we calculate the efficiency from the acceptance ratio of the resulting distribution as defined in Eq. (12). For IS and PSIS, we calculate the efficiency as the ratio of the ESS to the number of original samples, as defined in Eq. (23). On the right-hand axes (brown solid curve), we plot the Pareto-smoothing diagnostic \hat{k} . For all quantities, we repeat the simulation study multiple times and then plot the median and 90% interval across the simulated realisations. We provide a vertical dashed line at $\sigma_e/\sigma_p = 2$, corresponding to the simulation studied in Fig. 5. We also provide a horizontal dashed line at $\hat{k} = 0.7$ on the right-hand axis, a threshold above which the Pareto smoothing diagnostics suggest the estimates may be unstable or have a high degree of bias.

tion of the techniques rather than a scientific result. In particular, the IMRPhenomXP primary pass posterior distribution is wider than the posterior found using the IMRPhenomXPHM model. This is expected, since the higher-order modes contain additional information about the source parameters that isn't captured by the dominant mode alone.

In Fig. 7, we plot the posterior distributions, comparing the original analysis (IMRPhenomXPHM) with two sets of resampled posteriors using PSRS and multinomial-PSIS changing the waveform model. To validate the resampling procedure, we also include a new re-analysis of GW150914, using an identical setup to the GWTC-2.1 analysis but changing only the waveform model. The samples from this analysis agree within the resampling uncertainty with both PSRS and PSIS, demonstrating their capacity to correctly resample the posterior. However, we do find that the Pareto smoothing diagnostic is greater than the threshold of 0.7. Finally, by showing the uncertainty in the inferred density (calculated by repeated resampling), we illustrate the key difference between IS and RS: the PSIS interval is narrower than the PSRS interval, indicating a reduction in the variance and a smoother resulting histogram. However, it should of course be recalled that the cost of using multinomial-PSIS is repeated samples.

Finally, we also compute the Bayes factor between the IMRPhenomXP and IMRPhenomXPHM models (using the weights before Pareto smoothing) to be

$$\ln \mathcal{B} = 0.16 \pm 0.03 \quad (38)$$

which provides mild support in favour of the IMRPhenomXP waveform model over IMRPhenomXPHM. This finding is consistent with Payne et al. (2019): the log-Bayes factor is within 2 standard deviations of their value $\ln \mathcal{B} = 0.20$ (here we have changed the sign of

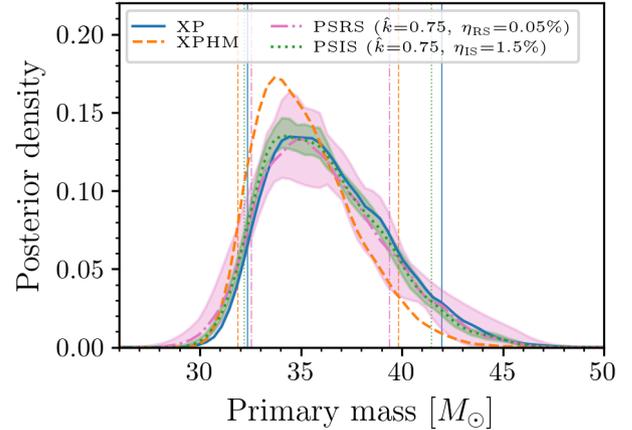


Figure 7. The posterior distribution (presented using a Gaussian KDE) on the primary mass of GW150914 as presented in the GWTC-2.1 analysis using IMRPhenomXPHM (orange dashed curve). We compare this with the posterior distributions for the IMRPhenomXP waveform calculated from resampling the GWTC-2.1 results with PSRS (pink dash-dotted curve) and PSIS (green dotted curve). We also include a new re-analysis using identical settings to GWTC-2.1, but with the IMRPhenomXP model (blue solid curve). For the resampled posteriors, we repeat the resampling multiple times and plot the median and 90% intervals to illustrate the variance in the result.

the Bayes factor found in Table 1 of that work so that the definition of the Bayes factors are consistent).

4.3 Changing the PSD for GW150914

As another example of the application of resampling, we again consider the GWTC-2.1 re-analysis of GW150914. For that analysis, BayesWave (Littenberg & Cornish 2015) was used to calculate a PSD, using data surrounding the signal and modelling the noise properties with a parameterised model including cubic splines and Lorentzian lines. However, the version of BayesWave used to produce LVK parameter estimates has since been upgraded (Gupta & Cornish 2024), for example, using Akima splines rather than cubic splines. We use this upgraded version to calculate a new PSD with settings similar to those used in the GWTC-4.0 analysis (Abac et al. 2025b). In Fig. 8, we compare the ASD (amplitude spectral density: the square root of the PSD) for the two LIGO detectors that observed the event, illustrating that while the overall shape is consistent, there are small differences between the two.

Resampling can provide an easy way to check the impact these differences may have on the scientific results. However, we have to be careful in our choice of likelihood since the PSD enters Eq. (3) in a normalisation term that is often ignored (cf. Thrane & Talbot 2019). This correction is not included in the Bilby likelihood computation by default, and so we add it explicitly (see the data release for details). We calculate the likelihood and likelihood ratio under the old and new PSD then using the difference in log-likelihoods, resample from the initial sample set. In Fig. 9, we plot the distribution of the detector-frame (i.e. redshifted) chirp mass. We select this parameter for visualisation as it is the best measured value. We find that the change in PSD results in a few-per-cent shift in the posterior, consistent with the findings of Biscoveanu et al. (2020). We also find that the resampling is efficient, with a good Pareto-smoothing diagnostic and efficiency for both PSRS and PSRS.

Now turning to the Bayes factor, from Eq. (5), we see that three

Bayes factor	Estimate
$\ln \left(\frac{\mathcal{Z}(d M, P_{\text{new}})}{\mathcal{Z}(d M, P_{\text{old}})} \right)$	-1684.86 ± 0.02
$\ln \left(\frac{\mathcal{Z}(d N, P_{\text{new}})}{\mathcal{Z}(d N, P_{\text{old}})} \right)$	-1726.90
$\ln \left(\frac{\mathcal{Z}(d M, P_{\text{new}})}{\mathcal{Z}(d N, P_{\text{new}})} \frac{\mathcal{Z}(d N, P_{\text{old}})}{\mathcal{Z}(d M, P_{\text{old}})} \right)$	42.04 ± 0.02

Table 3. Estimates of the three log Bayes factors that can be computed from the study changing the PSD in the analysis of GW150914. For the first and last lines, the evidences are computed from Eq. (29) using different choices of the likelihood or likelihood ratio. Meanwhile, for the second row, the ratio of noise evidence, there is no uncertainty, as this is computed directly from the data and PSD. Note that all three are computed directly from the likelihoods themselves, but the final row can also be computed as the difference in the first two rows.

different Bayes factors can be computed when changing from the “old” (GWTC-2.1) to our “new” PSD (using the GWTC-4.0 settings). First, the ratio of the signal evidence, second, the ratio of the noise evidence, and third, the ratio of the signal vs noise evidence. We tabulate these in Table 3. The first row, the Bayes factor of signal evidence, indicates a strong preference for the old PSD over the new PSD. However, this comparison includes information about both how well the signal model explains the data and how well the PSD explains the noise properties of the data. We can separate these by considering the second row, the ratio of noise evidence, which indicates an even stronger preference for the old PSD. Finally, when we compute the ratio of the signal vs noise Bayes factors, we effectively ask “Which PSD makes the signal stand out more strongly?” and now we find evidence in favour of the new PSD. This suggests the new PSD, while assigning a lower overall probability to the data, provides a noise model against which the signal is more distinguishable. We do not investigate this point further, but note that Bayes factors are not typically used as a means to determine the choice of PSD: for details of the development of the BayesWave see Gupta & Cornish (2024). Moreover, we highlight that the differences could be related to the settings we chose for the PSD generation.

5 DISCUSSION

In this work, we have provided a comprehensive guide to reconstructing likelihoods and priors from gravitational-wave posterior samples and applying resampling techniques to generate new posterior distributions under modified assumptions. This capability is fundamental to many analyses in gravitational-wave astronomy, from population studies to tests of general relativity, yet the practical implementation details have not been systematically documented until now. Our analysis demonstrates that accurate reconstruction of the likelihood and prior is achievable when the computing environment and analysis configuration are properly matched. The residual differences we observe (Fig. 1) are typically at the level of one part in a thousand or better for the likelihood, which we have shown is sufficiently accurate to have a negligible impact on resampling efficiency. This level of precision requires careful attention to software versions, data processing details, and numerical implementation choices. The hardware-dependent variations we identified highlight an important but often overlooked aspect of reproducible gravitational-wave analysis. While exact byte-level reproduction requires identical hardware, the small random shifts introduced by different floating-point implementations are typically negligible for scientific applications, which should reassure researchers that resampling can be performed reliably across different computing environments.

A key limitation of our approach is the requirement that the orig-

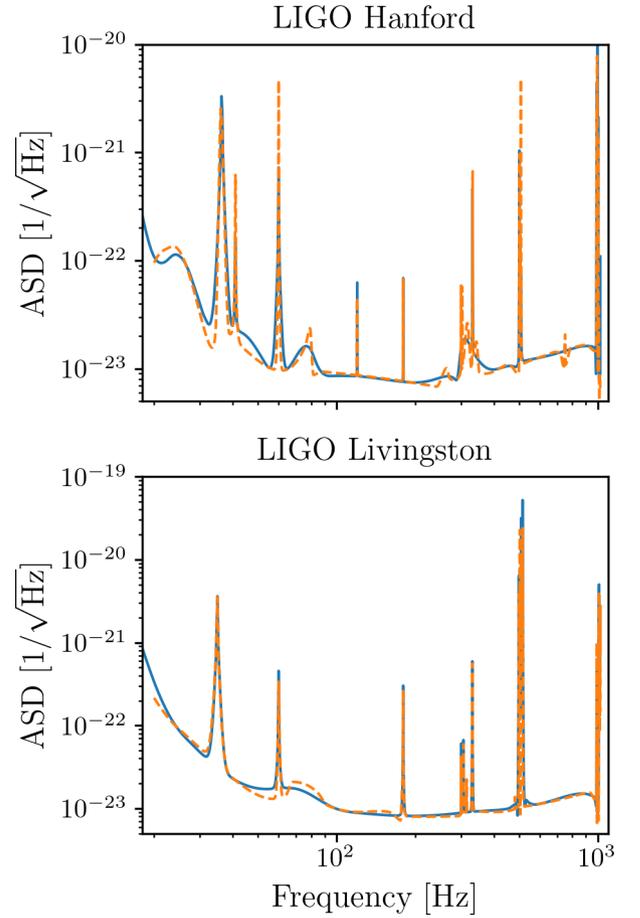


Figure 8. A comparison of the BayesWave ASD estimates from the original GWTC-2.1 analysis (solid blue curves) with a new ASD created using the upgraded version of BayesWave (orange dashed curves).

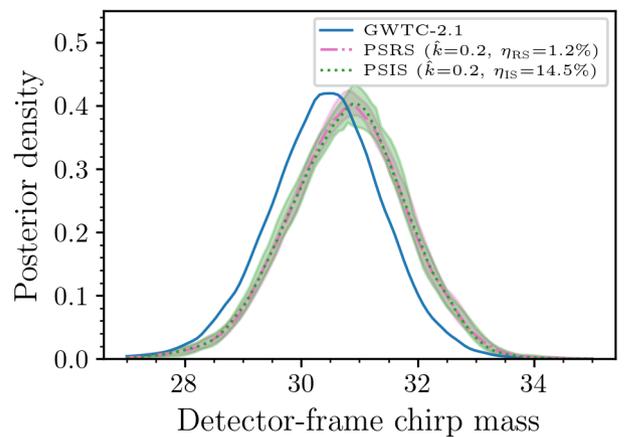


Figure 9. The redshifted (detector-frame) chirp mass posterior distribution (presented as a Gaussian KDE) of GW150914 from the original GWTC-2.1 analysis (solid blue curve) and resampling using an updated BayesWave PSD (see Fig. 8) using PSRS (pink dash-dotted curve) and PSIS (green dotted curve). For the resampled distributions, we repeat the process multiple times and plot the median of the distribution and 90% interval to illustrate the variance in the procedure.

inal analysis was performed using Bilby with PESummary packaging. While this covers a significant fraction of LVK analyses since GWTC-2.1, other analyses using LALInference or RIFT would require adapted approaches. Additionally, our reconstruction methodology assumes that all necessary configuration information has been preserved in the data release, which may not always be the case for older analyses.

Our comparison of resampling techniques reveals important trade-offs that practitioners should consider. RS produces clean, equally-weighted samples but can be inefficient when the proposal and target distributions differ significantly. Multinomial-IS can yield smoother posterior estimates, but at the cost of introducing sample correlations through repeated samples in multinomial resampling. For modest changes to analysis assumptions, both approaches perform well. For more dramatic changes, IS may be preferable to avoid significant variance in the resulting posterior samples, though a re-analysis would be better if computationally feasible.

We also introduce Pareto-smoothing, demonstrating that it can provide moderate improvements in efficiency and comes with an automated diagnostic which provides a useful indicator of when Pareto-smoothing can be trusted.

Our real-data examples demonstrate that some typical analysis modifications (waveform model changes, PSD updates) can fall within the regime where resampling is efficient and reliable. However, we caution that more extreme modifications—such as changing the waveform family, changing from a precessing to non-precessing waveform or dramatically altering the frequency band—may fail altogether or require careful validation of the resampling assumptions. The GW150914 examples illustrate that even seemingly minor changes, such as waveform model updates or improved PSD estimation algorithms, can have measurable impacts on inferred parameters, making resampling a valuable tool for systematic uncertainty assessment.

The methodology presented here enables several important classes of scientific analysis that would otherwise require computationally expensive re-analysis. Therefore, resampling can allow researchers without access to high-performance computing resources to explore “what if” questions that would otherwise require full re-analysis.

Based on our analysis, we recommend several best practices for gravitational-wave posterior resampling. When possible, researchers should match the conda environment used in the original analysis, using the environment specifications provided in Table 1 and our accompanying data release. For method selection, rejection sampling is preferred for modest analysis changes where clean, equally-weighted samples are needed, while importance sampling with Pareto-smoothing should be used for larger changes or when smooth posterior estimates are critical, carefully monitoring the \hat{k} diagnostic. When possible, we recommend validation against a full re-analysis for at least one representative event. Calculating and reporting resampling efficiency helps readers assess the reliability of results, as low efficiency may indicate that the analysis modification is too extreme for reliable resampling.

We hope this guide provides a useful resource for the gravitational-wave community and enables new scientific investigations that leverage the wealth of information contained in the LVK data releases. The combination of theoretical understanding, practical implementation details, and working code examples should lower the barrier for researchers to incorporate resampling techniques into their analyses, ultimately leading to more robust and comprehensive studies of the gravitational-wave universe.

ACKNOWLEDGEMENTS

We are thankful to Charlie Hoy for help with the software packaging provided by PESummary, to Colm Talbot and Tamasz Baka for useful discussions that refined the description of the resampling methods, and to Michael Williams for useful comments in reviewing this work.

This work is supported by the Science and Technology Facilities Council (STFC) grant UKRI2488. The authors are grateful for the computational resources provided by the LIGO Laboratory and supported by the National Science Foundation Grants PHY-0757058 and PHY-0823459. This material is based upon work supported by NSF’s LIGO Laboratory, which is a major facility fully funded by the National Science Foundation.

We utilise the Numpy (Harris et al. 2020) and Scipy library (Virtanen et al. 2020) for data processing and analysis, and we also use the Matplotlib (Hunter 2007) library for visualisation.

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

We provide a data release associated with this paper (Ashton 2026), which contains the Python programs to recreate the results. This includes programs to reconstruct the likelihood and priors from LVK public data releases, and programs to resample the distribution for the given examples. We include a lightweight Python module which implements the various resampling algorithms discussed within that can be easily ported or modified to users’ projects. If using this program (with modification), we ask authors to reference this work.

REFERENCES

- Aasi J., et al., 2015, *Class. Quant. Grav.*, 32, 074001
 Abac A. G., et al., 2025a, *arXiv e-prints*, p. arXiv:2508.18079
 Abac A. G., et al., 2025b, *arXiv e-prints*, p. arXiv:2508.18081
 Abac A. G., et al., 2025c, *arXiv e-prints*, p. arXiv:2508.18082
 Abac A. G., et al., 2025d, *arXiv e-prints*, p. arXiv:2508.18083
 Abac A. G., et al., 2025e, *arXiv e-prints*, p. arXiv:2509.04348
 Abac A. G., et al., 2025f, *ApJ*, 995, L18
 Abbott B. P., et al., 2016a, *Phys. Rev. Lett.*, 116, 061102
 Abbott B. P., et al., 2016b, *Phys. Rev. Lett.*, 116, 241102
 Abbott R., et al., 2023, *Phys. Rev. X*, 13, 041039
 Abbott R., et al., 2024, *Phys. Rev. D*, 109, 022001
 Acernese F., et al., 2015, *Class. Quant. Grav.*, 32, 024001
 Ajith P., et al., 2007, *Class. Quant. Grav.*, 24, S689
 Akutsu T., et al., 2021, *PTEP*, 2021, 05A101
 Ashton G., 2026, Data behind: “Reconstructing and resampling: a guide to utilising posterior samples from gravitational wave observations”, doi:10.5281/zenodo.18337354, <https://doi.org/10.5281/zenodo.18337354>
 Ashton G., et al., 2019, *Astrophys. J. Suppl.*, 241, 27
 Baka T., et al., 2025, Correcting misspecification of calibration uncertainties in gravitational-wave data analysis with efficient reweighting, <https://dcc.ligo.org/LIGO-T2500295/public>
 Biscoveanu S., Haster C.-J., Vitale S., Davies J., 2020, *Phys. Rev. D*, 102, 023008
 Blackman J., Field S. E., Scheel M. A., Galley C. R., Hemberger D. A., Schmidt P., Smith R., 2017, *Phys. Rev. D*, 95, 104023
 Buonanno A., Damour T., 1999, *Phys. Rev. D*, 59, 084006
 Buonanno A., Damour T., 2000, *Phys. Rev. D*, 62, 064015
 Butterworth S., et al., 1930, *Wireless Engineer*, 7, 536
 Chattopadhyay D., Al-Shammari S., Antonini F., Fairhurst S., Miles B., Raymond V., 2024, *Mon. Not. Roy. Astron. Soc.*, 536, L19
 Chatziioannou K., Cornish N., Wijngaarden M., Littenberg T. B., 2021, *Phys. Rev. D*, 103, 044013

- Cornish N. J., Littenberg T. B., 2015, *Class. Quant. Grav.*, 32, 135012
- Cornish N. J., Littenberg T. B., Bécsy B., Chatziioannou K., Clark J. A., Ghonge S., Millhouse M., 2021, *Phys. Rev. D*, 103, 044006
- Davis D., Littenberg T. B., Romero-Shaw I. M., Millhouse M., McIver J., Di Renzo F., Ashton G., 2022, *Class. Quant. Grav.*, 39, 245013
- Dax M., Green S. R., Gair J., Pürrer M., Wildberger J., Macke J. H., Buonanno A., Schölkopf B., 2023, *Phys. Rev. Lett.*, 130, 171403
- Elvira V., Martino L., Robert C. P., 2022, *International Statistical Review*, 90, 525
- Farr W., Farr B., Littenberg T., 2014, Technical Report DCC-T1400682, Modelling calibration errors in CBC waveforms, <https://dcc.ligo.org/LIGO-T1400682/public>. LIGO, <https://dcc.ligo.org/LIGO-T1400682/public>
- Glanzer J., et al., 2023, *Class. Quant. Grav.*, 40, 065004
- Gupta T., Cornish N. J., 2024, *Phys. Rev. D*, 109, 064040
- Harris F. J., 2005, *Proceedings of the IEEE*, 66, 51
- Harris C. R., et al., 2020, *Nature*, 585, 357
- Hastings W. K., 1970, *Biometrika*, 57, 97
- Hourihane S., Chatziioannou K., 2025, *arXiv e-prints*, p. arXiv:2506.21869
- Hourihane S., Chatziioannou K., Wijngaarden M., Davis D., Littenberg T., Cornish N., 2022, *Phys. Rev. D*, 106, 042006
- Hoy C., Raymond V., 2021, *SoftwareX*, 15, 100765
- Hunter J. D., 2007, *Computing in Science & Engineering*, 9, 90
- Kish L., 1965, *Am Polit Sci Rev*, 59, 1025
- Kong A., 1992, University of Chicago, Dept. of Statistics, Tech. Rep. 348, 14
- Kumar R., Carroll C., Hartikainen A., Martin O., 2019, *Journal of Open Source Software*, 4, 1143
- LIGO Scientific Collaboration and Virgo Collaboration 2022, GWTC-2.1: Deep Extended Catalog of Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run - Parameter Estimation Data Release, doi:10.5281/zenodo.6513631, <https://doi.org/10.5281/zenodo.6513631>
- LIGO Scientific Collaboration Virgo Collaboration KAGRA Collaboration 2018, LVK Algorithm Library - LALSuite, Free software (GPL), doi:10.7935/GT1W-FZ16
- LIGO Scientific Collaboration Virgo Collaboration KAGRA Collaboration 2025b, GWTC-4.0: Glitch modelling for events, doi:10.5281/zenodo.16857060, <https://doi.org/10.5281/zenodo.16857060>
- LIGO Scientific Collaboration Virgo Collaboration KAGRA Collaboration 2025a, GWTC-4.0: Parameter estimation data release, doi:10.5281/zenodo.16053484, <https://doi.org/10.5281/zenodo.16053484>
- Lange J., et al., 2017, *Phys. Rev. D*, 96, 104041
- Littenberg T. B., Cornish N. J., 2015, *Phys. Rev. D*, 91, 084034
- MacKay D. J., 2003, *Information theory, inference and learning algorithms*. Cambridge university press
- Mould M., Wolfe N. E., Vitale S., 2025, *Phys. Rev. D*, 111, 123049
- Nuttall L. K., 2018, *Phil. Trans. Roy. Soc. Lond. A*, 376, 20170286
- Pankow C., Brady P., Ochsner E., O'Shaughnessy R., 2015, *Phys. Rev. D*, 92, 023002
- Pankow C., et al., 2018, *Phys. Rev. D*, 98, 084016
- Payne E., Talbot C., Thrane E., 2019, *Phys. Rev. D*, 100, 123017
- Payne E., Talbot C., Lasky P. D., Thrane E., Kissel J. S., 2020, *Phys. Rev. D*, 102, 122004
- Pratten G., et al., 2021, *Phys. Rev. D*, 103, 104056
- Romero-Shaw I. M., et al., 2020, *Mon. Not. Roy. Astron. Soc.*, 499, 3295
- Rover C., Meyer R., Christensen N., 2006, *Class. Quant. Grav.*, 23, 4895
- Schutz B. F., 1986, *Nature*, 323, 310
- Sidery T., et al., 2014, *Phys. Rev. D*, 89, 084060
- Skilling J., 2006, *Bayesian Analysis*, 1, 833
- Soni S., et al., 2025, *Class. Quant. Grav.*, 42, 085016
- Speagle J. S., 2020, *Mon. Not. Roy. Astron. Soc.*, 493, 3132
- Sun L., et al., 2020, *Class. Quant. Grav.*, 37, 225008
- Talbot C., Thrane E., 2018, *Astrophys. J.*, 856, 173
- Talbot C., Smith R., Thrane E., Poole G. B., 2019, *Phys. Rev. D*, 100, 043030
- Talbot C., et al., 2025, *Classical and Quantum Gravity*, 42, 235023
- Thrane E., Talbot C., 2019, *Publ. Astron. Soc. Austral.*, 36, e010
- Tiwari V., 2018, *Class. Quant. Grav.*, 35, 145009
- Vehtari A., Simpson D., Gelman A., Yao Y., Gabry J., 2024, *Journal of Machine Learning Research*, 25, 1
- Veitch J., et al., 2015, *Phys. Rev. D*, 91, 042003
- Virtanen P., et al., 2020, *Nature Methods*, 17, 261
- Wette K., 2020, *SoftwareX*, 12, 100634
- Williams M. J., Karamanis M., Luo Y., Seljak U., 2025, *MNRAS*,
- Wysocki D., O'Shaughnessy R., Lange J., Fang Y.-L. L., 2019, *Phys. Rev. D*, 99, 084026

APPENDIX A: PROBABILITY-PROBABILITY TESTS

The PP test is a standard diagnostic tool used to assess the reliability of Bayesian posterior distributions in gravitational-wave parameter estimation (Veitch et al. 2015; Romero-Shaw et al. 2020). The test compares the cumulative distribution of confidence intervals from a set of simulated injections against the uniform distribution expected for unbiased posteriors. We follow the Bilby implementation of the PP test and provide a simplified version in the data release notebooks. Specifically, we take the result from injection studies and compute the fraction of injected parameter values that fall below various quantile levels of their corresponding posterior distributions. For each parameter of interest, the PP test calculates the cumulative probability $P(\theta < \theta_{\text{sim}})$ where θ_{sim} is the simulated value and the probability is computed using the recovered posterior samples. If the posteriors are unbiased and properly calibrated, these p -values should be uniformly distributed between 0 and 1. The PP plot displays the empirical cumulative distribution function of these p -values against the theoretical uniform distribution, with deviations from the diagonal indicating systematic biases in the parameter estimation. We include a Kolmogorov-Smirnov test to compare the empirical distribution of p -values against the uniform distribution, providing a p -value that indicates the probability of observing such deviations by chance. In Fig. A1, we also include confidence bands around the diagonal using the beta distribution to account for finite sample size effects. The width of these bands depends on the number of injections, with larger injection campaigns providing tighter constraints on systematic biases.

In the main text, Section 4, we use the PP test to demonstrate that the resampling methods can unbiased a posterior given a corrected posterior density. The specific case given considers over- and under-constrained posteriors generated by varying the ratio of σ_s (the standard deviation of the simulated measurement error) to σ_p (the standard deviation of the simulated posteriors). Specifically, in Fig. 5, we consider the case when the posteriors are too narrow $\sigma_s/\sigma_p = 2$. This results in a characteristic ‘S’ shape. This is one possible characteristic failure mode of the PP plot. In Fig. A1, we extend this to study the other three failure modes: a case where the posterior is under-constrained (producing an inversion of the ‘S’); biased positively (producing a bow-like deviation above the diagonal), and negatively (producing a bow-like deviation below the diagonal). We include these figures to help users diagnose bias from resampling (or from direct sampling) approaches.

We note that this behaviour is specific to PP tests, which use quantile levels to calculate the fraction of injected parameters that fall outside the posterior at a given level. If instead, the highest-posterior density is used, the behaviour will differ.

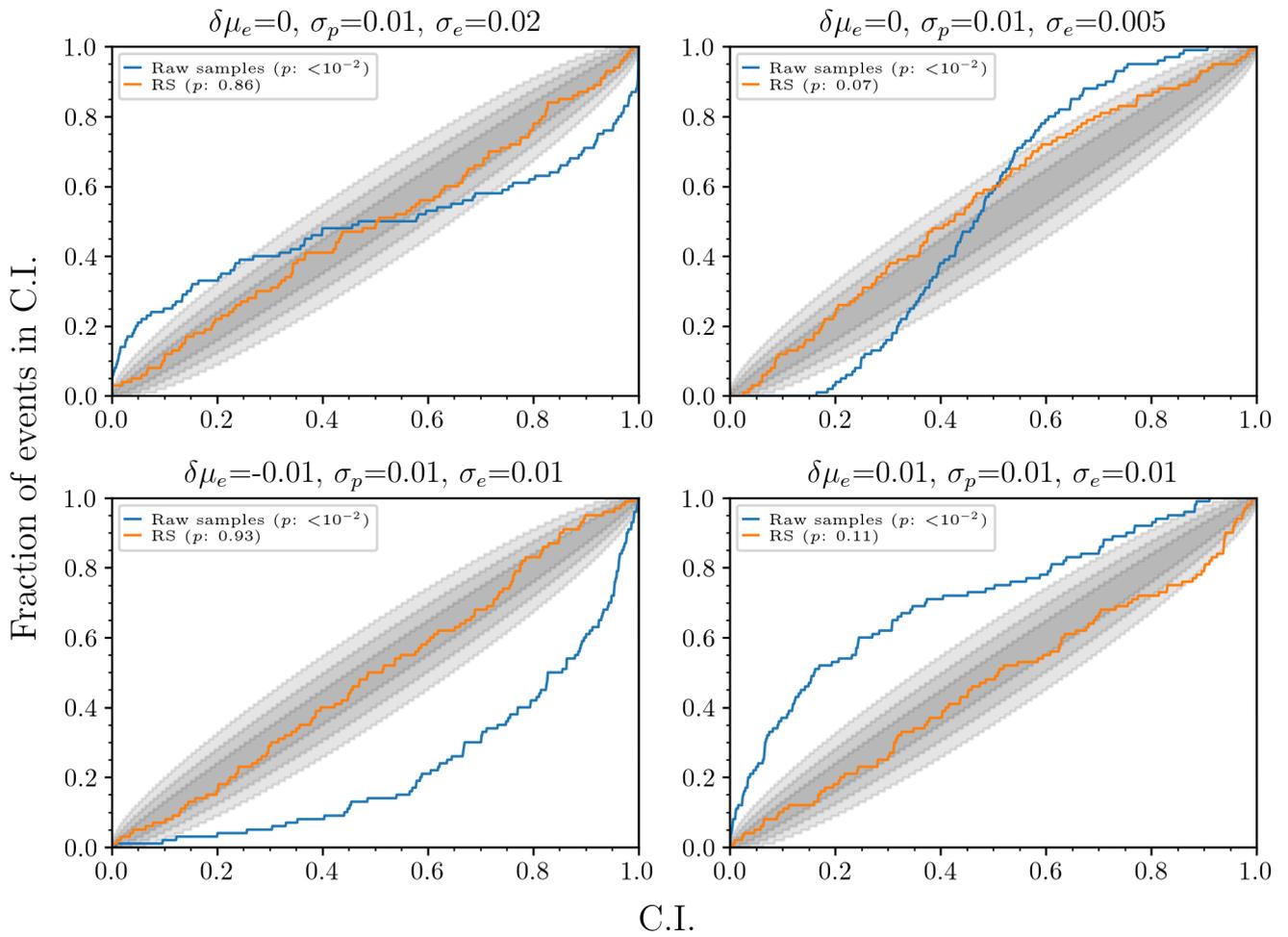


Figure A1. A set of four PP tests showing the four failure modes of the test. In blue, we plot the biased tests, and in orange, we plot the corrected posteriors. For each plot, we provide the parameters of the simulation in the title. Further details of the plots can be found in the notebooks as part of the data release.

ORIGINAL UNEDITED