Case studies with GPBilby of glitch-contaminated transient gravitational waves

Mattia Emma,[1,*] Ann-Kristin Malz,[1] Adriana Dias,[1] and Gregory Ashton[1,2]
[1]*Royal Holloway University of London, Egham Hill TW20, United Kingdom*
[2]*Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, United Kingdom*

(Received 3 April 2026; accepted 11 September 2026; published 7 October 2026)

In their fourth observing run, the LIGO-Virgo-KAGRA gravitational-wave observatories have found hundreds of new signals, but many are contaminated by non-Gaussian transient noise artifacts known as glitches. Left unaddressed, glitches can bias parameter inference and lead to misleading astrophysical conclusions. We present a series of case studies using GPBilby, a parameter estimation tool that employs a time-domain likelihood jointly modeling the astrophysical signal with a physical waveform and non-Gaussian noise with a Gaussian process. We first show that when the detector noise is Gaussian, GPBilby produces results consistent with those obtained with the standard Gaussian-noise likelihood, and then consider events affected by non-Gaussian features. For GW231123, the highest-mass binary black hole candidate observed to date, analyses using `IMRPhenomXPHM` reveal coherent residual structure that leads to measurable shifts in inferred source parameters. In contrast, analyses employing `NRSur7dq4` show no significant excess residual power and remain consistent across likelihood choices. This demonstrates that waveform systematics and flexible noise modeling are intrinsically coupled, as the Gaussian process terms can partially absorb coherent waveform mismatches. For GW191109, we find that evidence for spin misalignment remains robust despite glitches in both LIGO detectors. For GW230630_070659, excluded from GWTC-4.0 owing to poor data quality, we find the data to be consistent with a binary black hole waveform model, with no additional residual power identified by the Gaussian process component. Overall, these results highlight how GPBilby can be used to perform glitch-robust inference and as a tool to understand waveform modeling systematics.

DOI: 10.1103/pnmb-78p3

I. INTRODUCTION

The LIGO-Virgo-KAGRA (LVK) Collaborations have recently presented GWTC-4.0, the fourth iteration of the Gravitational Wave Transient Catalogue [1–3]. Analyzing data from the Laser Interferometer Gravitational-Wave Observatory (LIGO) [4], Virgo [5], and KAGRA [6] interferometric gravitational-wave detectors, LVK has identified 218 signals with a probability of astrophysical origin greater than 0.5. This rich treasure trove of astrophysical observations is bringing new insights into the dark Universe, such as studies of the mass and spin spectrum [7] from which we can learn about the deaths of massive stars, measurements of the expansion rate of the Universe [8], and tests of the nature of gravity [9–11].

*Contact author: mattia.emma@ligo.org

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

However, astrophysical inferences rest on the assumptions made during the individual event analyses: that the background noise is stationary, colored Gaussian noise [12,13]. If this is not the case, the inferred parameters from individual events can be biased [14–18] and this may have a potential systematic effect when considering the population as a whole. To mitigate this, the LVK performs detailed studies of all events that undergo parameter estimation [2]. Checks are performed using the strain data and auxiliary channels to understand the data quality and identify any non-Gaussian transient artifacts known as glitches [13]. Of the new signals discovered in GWTC-4.0, 86 met the criteria for parameter estimation, and of these, 44 were found to have non-Gaussian transient artifacts in the data surrounding the event. We note, however, that for many of these, the observed glitches fall outside the region in which bias is known to occur [19].

To mitigate the potential impact glitches may have on parameter estimation, a glitch subtraction approach is typically used (see, e.g., Refs. [13,20,21]), in which the glitch is modeled and subtracted from the data. However, as described in Udall *et al.* [18], this approach is suboptimal: glitch subtraction always leaves behind residual signal-to-noise

ratio (SNR), and subtraction of low-SNR glitches can leave residuals that further bias inferences. Therefore, it is better to perform joint inference of the glitch alongside the astrophysical signal, enabling measurements of the astrophysical source properties marginalized over the uncertainty introduced by the glitch. Several approaches to joint inference have been developed, such as those using a deterministic glitch model [22–24] and those using a data-informed glitch model trained on past observations [25]. Alternative approaches aim to address these biases at the likelihood level, for example through modified noise models such as the hyperbolic likelihood proposed by Sasli *et al.* [26]. In this work, we continue to develop and validate GPBilby [27], a Gaussian process (GP)-enhanced [28] approach for parameter estimation of transient gravitational wave signals.

GPBilby performs inference on the time series strain data $\mathbf{d}$ after whitening using a given power spectral density (PSD). The core idea is to model the astrophysical signal using a deterministic waveform approximant model with astrophysical source parameters θ, and the glitch using a GP with an associated set of hyperparameters α. Given a set of source parameters, we define the time-domain residual vector $\mathbf{r}_\theta \equiv \mathbf{d}_w - \boldsymbol{\mu}_w(\theta)$ between the whitened data $\mathbf{d}_w$ and the whitened model $\boldsymbol{\mu}_w(\theta)$. The GP log-likelihood is then

$$\ln \mathcal{L}(\mathbf{d}|\theta,\alpha) = -\frac{1}{2}\mathbf{r}_\theta^T \Sigma_\alpha^{-1} \mathbf{r}_\theta - \frac{1}{2}\ln\left((2\pi)^N |\Sigma_\alpha|\right), \quad (1)$$

where Σ_α is the covariance matrix predicted by the GP model with hyperparameters α, and $|\Sigma_\alpha|$ is its determinant. In practice, GPBilby uses CELERITE [29], which approximates Eq. (1), enabling rapid evaluation of the likelihood within an order of magnitude of the standard likelihood evaluation time.

The CELERITE software provides a rich interface for building the GP kernel used to predict the covariance matrix in Eq. (1) by combining different *terms*. We use this interface in the following way. To model the whitened strain data in the absence of an astrophysical signal or glitch, we find that a single white noise *jitter* term is sufficient; its associated parameter σ corresponds to the standard deviation of the whitened strain. To model white noise alongside a glitch with a single frequency component, we find that the simple harmonic oscillator (SHO) kernel term works well. The SHO terms in CELERITE have three parameters: an angular frequency ω_0, a quality factor Q, and an amplitude parameter S_0 (see Section 4 of [29]). All of these terms are parameterized in terms of their natural logarithm. We also find this term is effective when the data contains multiple temporally isolated artifacts at the same frequency. For glitches with multiple frequency components, we use multiple numbered SHO terms with order statistics to circumvent the label-switching degeneracy.

We begin this paper in Sec. II by discussing updates to the GPBilby software; then in Sec. III we provide a set of case studies on interesting gravitational-wave observations. Finally, in Sec. IV, we discuss the results and future prospects.

II. UPDATES TO THE SOFTWARE

Since the initial version of GPBilby [27], we have improved the software, introducing new CELERITE terms for GP kernel building, developing a new approach to incorporate calibration uncertainty, extending the usability of GPBilby to a full range of waveform models, as well as making numerous minor improvements. Alongside this work, we release Version 0.0.1 of the software [30]. Below, we describe the first two of these updates in detail with some background on GPBilby.

A. Kernel building parameterization

The frequency dependence of the SHO kernel term provided by CELERITE is parameterized by the natural logarithm of the angular frequency ω_0. However, since the frequency of the SHO term corresponds directly to the frequency of the glitch we are trying to model, we have reparameterized the SHO term in GPBilby in terms of the frequency $f_0 = \omega_0/2\pi$. Beyond the simple parameterization change, we also use uniform priors on f_0 rather than on $\ln \omega_0$; this will provide more prior weight to larger frequencies, but we are typically interested in glitch frequencies that overlap with the signal in the 20 to 1000 Hz band. Therefore, we consider a uniform-in-frequency prior appropriate. Throughout the remainder of this work, we use this new frequency parametrization of the SHO term. We note that this reparameterization and the associated change of prior do not alter the underlying SHO kernel or its covariance structure. Accordingly, we expect the qualitative behavior of GPBilby to remain consistent with that reported in Ref. [27], including the improved modeling of longer-duration transients such as scattering glitches.

B. Calibration

In the initial implementation of GPBilby [27], the strain errors δh provided to CELERITE were fixed as a user-defined input parameter: for all results in that work, we used a fixed absolute error of 0.01. In the inference that followed, the GP model was flexible enough to model additional error in the kernel by tuning the hyperparameters of the jitter term. Therefore, in effect, we provided the GP with an initial estimate, and then the jitter and SHO terms could model the uncertainty.

However, CELERITE can take as input a vector of heteroscedastic symmetric measurement errors: a 1-standard-deviation error for each data point in the time series. Therefore, we implemented a new approach to estimate the measurement errors on the whitened strain data from the frequency-domain calibration model. In standard

gravitational-wave parameter estimation, the calibration uncertainty model is constructed as a frequency-domain spline on the amplitude and phase with an associated set of calibration parameters. Estimates of the uncertainty are obtained from the calibration process as a calibration envelope [31] which provides the 16%, 50%, and 84% quantiles as a function of frequency. The envelope is then interpolated with a spline and provided as a prior on the calibration parameters, which are marginalized over during inference [32,33].

To approximate the measurement uncertainty on the whitened strain data in GPBilby, we take as input the calibration envelope, interpolate over frequency, and fit the quantiles with a skew-normal distribution. From the fitted skew-normal distribution, we then sample new calibration parameters, use these to perturb the frequency-domain strain, whiten, and Fourier transform, producing many realizations of the whitened time-domain strain, including the effects of calibration uncertainty. Finally, we take the residuals between the time-domain whitened strain and the perturbed realizations and calculate the standard deviation (as a function of time) across the realizations. This standard-deviation time series is then provided to CELERITE as the vector of measurement errors on the whitened strain data.

In Fig. 1, we show this procedure for the LIGO hanford observatory (LHO) data for the event GW150914 [34]. The upper panel shows the whitened time-domain strain, while the lower panel compares the one-standard-deviation measurement error with the nonsymmetric 68% interval computed across the distribution of residuals. The PSD used for whitening the nominal strain and the calibration-perturbed realizations is taken from the GWTC-2.1 reanalysis of the event [35]. The agreement shown in the lower panel confirms that our calibration uncertainties are well approximated by a symmetric interval.

III. CASE STUDIES

To test and validate GPBilby, we perform a series of case studies on selected events from the O1–O4a observing runs. We chose the events GW150914 [34], GW170814 [36], and GW230814 [37] as examples where the data is clean, i.e., free from glitches. We then study GW191109 [38], GW231123 [39], and GW231113_122623 to study the performance of our algorithm on events known to be contaminated by glitches. Finally, we analyze a trigger from the first part of the fourth observing run (O4a), which was deemed to be a glitch rather than a gravitational wave event, GW230630_070659.

A. GW150914

GW150914 was the first observed binary black hole (BBH) merger [40] and, with a network SNR of 24, remains one of the highest-SNR events to date. Analyses have not identified any data quality issues surrounding the event [32,35,40]. Therefore, as an initial case study, analyzing GW150914 with GPBilby provides a test of the capacity to reproduce standard analysis results when the Gaussian noise assumption holds.

We take as a base analysis the GWTC-2.1 analysis of GW150914 [35] using the `IMRPhenomXPHM` [41] waveform model. From this, we define six new reanalyses of GW150914, each using 8 s data around the event. In terms of the likelihood, we use three different configurations: first, an analysis using the standard WL, then a GPBilby analysis with a single white noise jitter term (GP-J), and an analysis with a white noise jitter term and a single SHO term (GP-JS). For each likelihood configuration, we then repeat the analysis using the original `BayesWave` PSD developed in GWTC-2.1 and an analysis using a new PSD created using the version of `BayesWave` applied in GWTC-4.0 [42–45]. The new PSD is available from Ashton [46], where it was shown

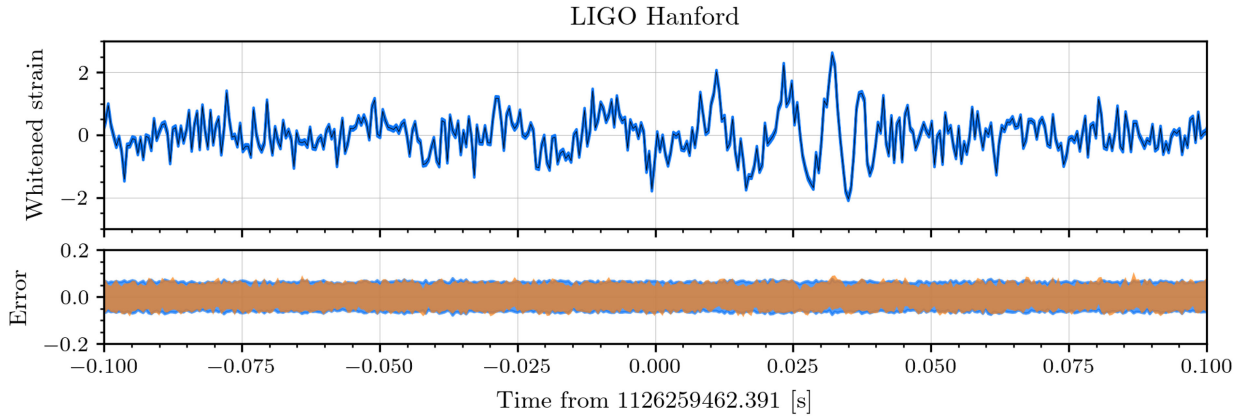


FIG. 1. Whitened time-domain strain from the LHO data surrounding GW150914. In the upper panel, we show the raw whitened strain as a black curve and the measurement errors (estimated from the calibration envelope) as a one-standard-deviation blue band around the whitened strain. In the lower panel, we show the measurement error as a blue band, and we show the 68% nonsymmetric interval estimated from the residual directly to validate that the errors are approximately symmetric. Differences in the whitened strain relative to Fig. 6 of Abbott *et al.* [32] arise from the use of a different PSD and data-conditioning procedures.

that the differing PSD realization produced small changes in the inferred parameters (consistent with the level expected from PSD uncertainty, see Biscoveanu *et al.* [47]). Therefore, the mirrored analysis using different PSDs provides a point of comparison for comparing the level of difference from the Whittle likelihood and the GPBilby analyses to the typical variation due to PSD uncertainty.

In Fig. 2, we show violin plots comparing the one-dimensional posterior distributions across the six analyses for the detector-frame source mass $\mathcal{M}^d$, mass ratio q,

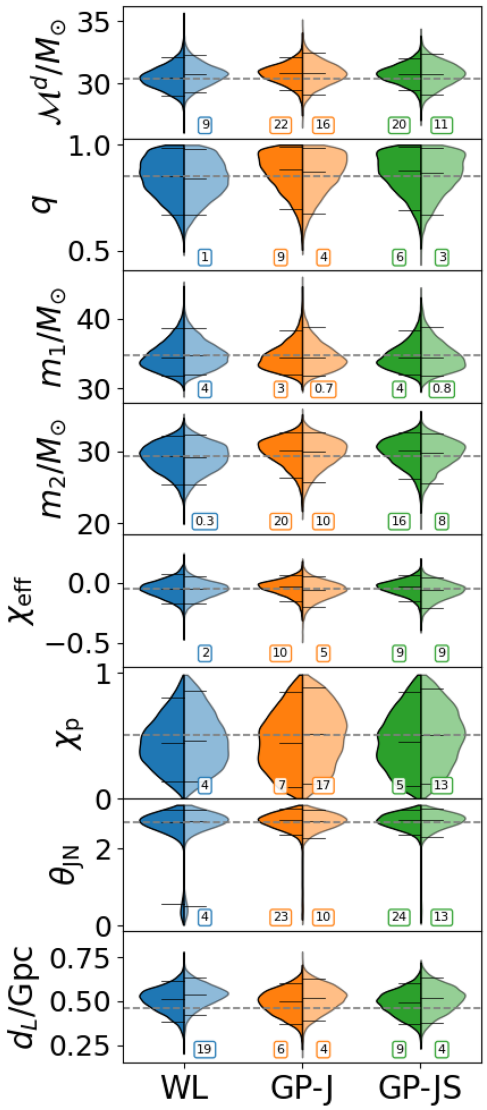


FIG. 2. Violin plots of selected GW150914 source parameter posteriors for the Whittle likelihood (WL), GPBilby with a single jitter term (GP-J), and GPBilby with a jitter and SHO term (GP-JS). The left side of each violin shows the posterior obtained using the GWTC-2.1 PSD, while the right side shows the posterior obtained using the newly computed PSD and the GWTC-4.0 settings. The boxed numbers give the JS divergence in millibits between each posterior and that of the reference analysis, taken here to be the Whittle likelihood with the GWTC-2.1 PSD; the reference analysis itself therefore carries no value.

primary mass m_1, secondary mass m_2, effective inspiral spin $\chi_{\rm eff}$, effective precession $\chi_{\rm p}$, source inclination angle θ_{JN}, and luminosity distance $D_{\rm L}$ (see Table 3 of Abac *et al.* [48] for definitions and further discussion). The detector-frame qualifier refers to the mass as measured from the data without including the multiplicative factor $1+z$ required to correct for the redshift z [49]. Following the GWTC-2.1 methodology, all other masses are given in the source frame and utilize the standard reference cosmology [see Ref. [48]], namely column TT + lowP + lensing + ext of Table 4 in Planck Collaboration [50].

To quantify the level of agreement between analyses, we compute the JS divergence between the one-dimensional marginal posteriors of each parameter [51]. For each comparison we draw an equal number of samples $N = 3000$ from the two posteriors, fit a kernel density estimator to each, and evaluate the resulting probability densities on a common grid. The divergence is obtained using the SciPy implementation [52], which returns the JS distance in bits, i.e., using a base-2 logarithm such that $\max(\mathrm{JS}) = 1$; we square this quantity to recover the JS divergence. The procedure is repeated one hundred times with independent resamplings, and we quote the median across these trials, which also provides an estimate of the scatter introduced by the finite number of posterior samples. Throughout this work we express JS divergences in millibits (mbit), i.e., in units of 10^{-3} bits. Two posteriors are conventionally regarded as indistinguishable when their divergence falls below $10/N$ [53], which for $N = 3000$ corresponds to ∼3 mbit. As a rough guide for interpreting the values quoted in the figures, divergences below ∼3 mbit indicate posteriors that are indistinguishable at the resolution of our analyses; values of order ten to a few tens of mbit correspond to modest but discernible shifts in the inferred distributions; and values above ∼100 mbit indicate substantial disagreement, readily visible by eye in the marginal distributions. This procedure is applied identically for all the events considered in the following sections, with only the choice of reference analysis differing between cases.

Broadly, Fig. 2 shows agreement between the posterior distributions across all configurations of the likelihood. There are mostly percent-level differences in the inferred quantiles for most parameters, which are at a similar level to the differences between estimates of the PSD. For this event we take the Whittle likelihood analysis using the GWTC-2.1 PSD as the reference, so that the quoted divergences measure both the effect of changing the likelihood and, for the mirrored analyses, the effect of the differing PSD realization. The majority of the divergences lie at or below the level identified above as consistent with sampling noise, confirming that the GPBilby analyses reproduce the standard result for this event. The exception is the inclination angle θ_{JN}, where the Whittle likelihood analyses show more weight in a secondary mode than is measured from any of the GPBilby analyses. This is also the parameter with the

largest JS divergences in Fig. 2. We discuss this feature and its possible interpretation in Sec. IV.

To investigate the inferences from the SHO term in the two GP-JS analyses, in Fig. 3, we provide the posterior distribution for SHO parameters. Interestingly, we find a spike in the LHO [LIGO Livingston observatory (LLO)] analysis at ≈60 Hz (≈320 Hz) for the GWTC-2.1 PSD (new PSD). That these frequencies lie close to known lines in the strain data [12], caused by the 60 Hz power line and its harmonics, suggests that the PSD has not perfectly whitened the data, and GPBilby can find and model these features.That the PSD is unable to model the power lines is not surprising, since it is a coherent deterministic signal with a narrow band: the colored Gaussian noise assumption is weak under these conditions. However, standard intuition within the field argues that, since the PSD is large close to the lines, these regions have a minimal impact on the likelihood due to the noise weighting within the Whittle likelihood. To validate that the power line does not have any impact on the inference of the astrophysical source parameters, we perform a *notching* study. We take the H1 and L1 PSDs, identify frequency-domain windows ranging from 4 to 600 Hz widths around the 60 Hz line, and replace the PSD within these windows to a fixed value of 1. We then repeat the analysis under the Whittle likelihood, varying the size of the windows and estimating the posterior distribution. In Fig. 4,

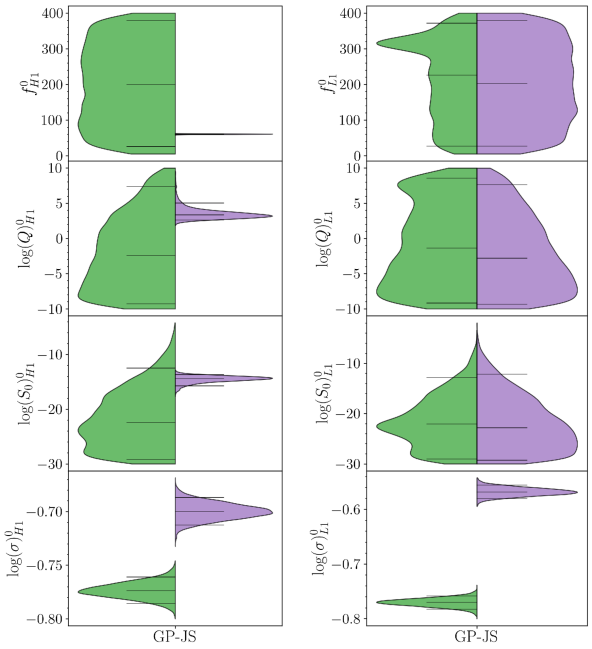


FIG. 3. Posterior distributions for the SHO term parameters of the GPBilby GP-JS analyses of GW150914 (left panel: LHO, right panel: LLO). The analyses using the GWTC-2.1 PSD are shown in green, while those using the newly computed PSD are shown in purple.

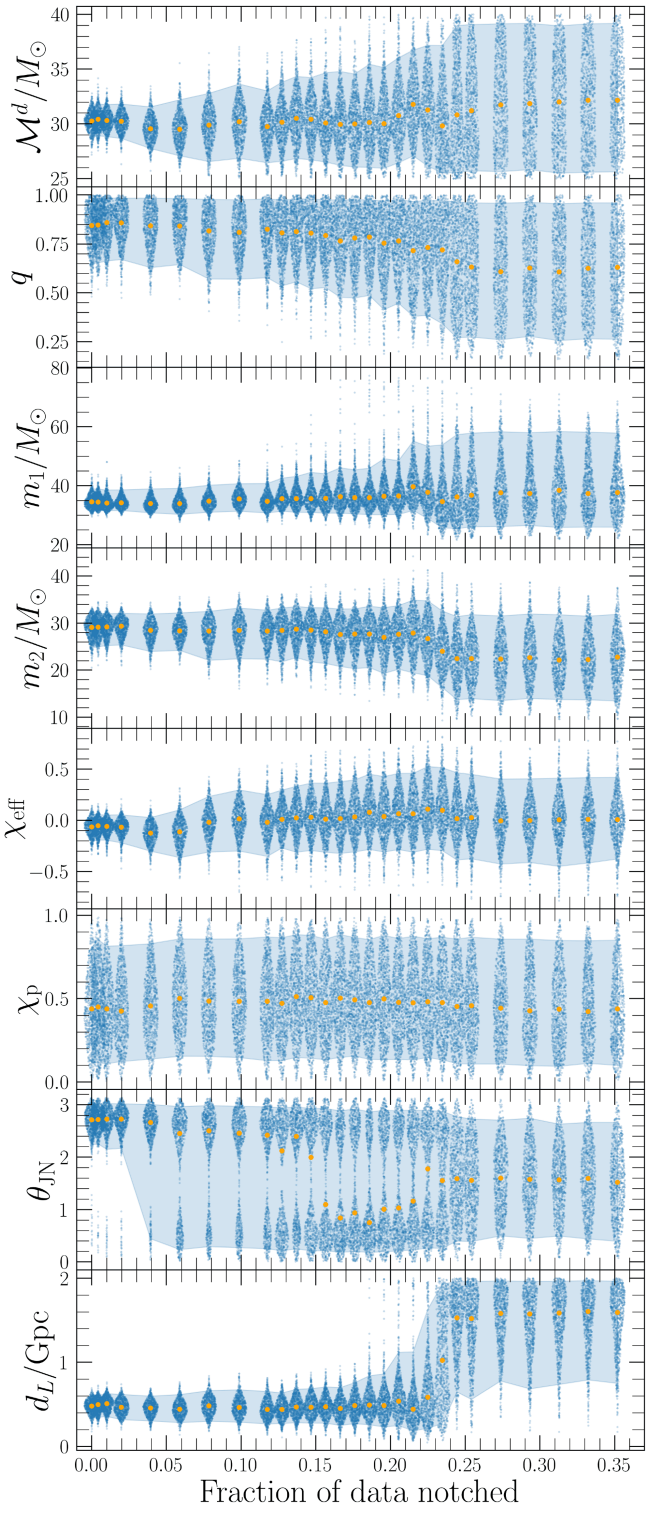


FIG. 4. Plots of selected GW150914 source parameter posteriors, with respect to the fraction of data notched. The blue filled regions show how the 90% interval changes as the proportion of notched data increases. The orange dots represent the median values and the blue shaded scatter show the full distribution of posterior samples.

we plot the posterior distribution for a set of selected source parameters as a function of the fraction of data removed. In the extremes, we find that when a negligible amount of data is removed, the posterior inference is robust, while when a significant amount is removed, the posteriors widen and become uninformative as expected. Between these extremes, we do not find any region in which the posterior is inconsistent with the inferences without notching. This validates that, while the PSD is unable to model the contributions from the power lines, their impact on the inferred posteriors is negligible, agreeing with the standard intuition within the field.

In summary, the case study of GW150914 further validates the absence of any transient noise artifacts (at least that GPBilby can find and model) impacting the measured source parameters.

B. GW170814

GW170814 is the first gravitational wave event detected by three detectors, namely the two LIGO and the Virgo observatories [36]. It has a network SNR of 18, which, combined with the triple detector observation, enabled a tight sky-localization area of 60 deg^2 (90% credible region). Similarly to the approach taken for GW150914, we take as a reference point the GWTC-2.1 [35] analysis and perform three reanalyses using 4 s of data around the event. We perform a WL analysis, a GP-J analysis, and a GP-JS analysis, keeping all the other settings fixed. In Fig. 5, we compare the one-dimensional posterior distributions for the three analyses. The results show a broad agreement between the posteriors in the different configurations, although there is a general tightening of the 90% confidence levels from the WL to the GP-JS analyses. For this event we take the WL analysis as the reference for the JS divergences, so that the quoted values isolate the effect of introducing the GP terms at otherwise fixed settings. As for GW150914, the posterior for the inclination angle θ_{JN} shows more weight for a secondary mode in the WL configuration than in either of the GP analyses. This support is partially there in the GP-J configuration, but vanishes in the GP-JS analysis. Consistently with this, θ_{JN} shows the largest JS divergence of the parameters considered. Considering the GP terms in the GP-JS analysis, we find that the posterior on the frequency parameter for both detectors is consistent with the prior, suggesting no excess power is found. Similarly, both the quality and amplitude parameters rail against the lower prior boundary, confirming that there are no significant non-Gaussian features in those data segments. Finally, the variance of the white noise term peaks at $\simeq 0.78$ for both detectors, as for the GWTC-2.1 PSD analysis of GW150914. Since these posteriors are largely uninformative, the corresponding plot is not included here, but is provided as part of the accompanying data release [54].

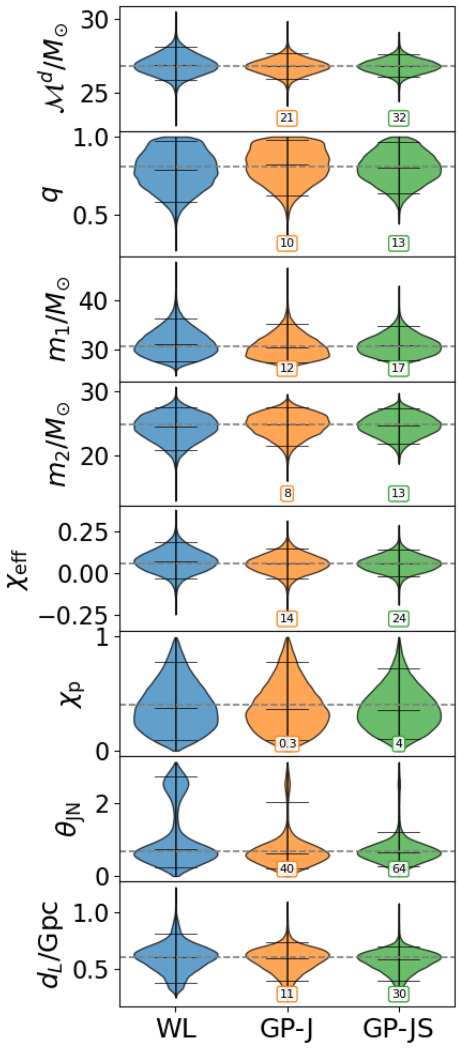


FIG. 5. Violin plots of selected GW170814 source parameter posteriors for the WL, GPBilby with a single jitter term (GP-J), and GPBilby with a jitter and SHO term (GP-JS). The left and right side of each panel are identical. The boxed numbers give the JS divergence in millibits with respect to the WL analysis, computed as described in Sec. III A; since each violin is symmetric, a single value is quoted per analysis, and the reference analysis itself carries no value.

C. GW230814

To further test our algorithm, we analyze the single-detector event GW230814_230901 (hereafter GW230814) from O4a [37]. This event was observed only in LIGO Livingston, where it reached a single-detector signal-to-noise ratio of 42.4, making it one of the loudest gravitational-wave candidates reported to date [55]. The LVK analysis found no data-quality issues associated with the event [37]. It therefore provides a useful case study for testing the performance of GPBilby in the high-SNR regime for a clean signal; the corresponding time-frequency representation is shown in Fig. 16.

We perform a WL, a GP-J and a GP-JS analysis using the settings employed in the `IMRPhenomXPHM_SpinTaylor` analysis from Abac *et al.* [39]. For each analysis, we repeat it with a 4 and 8 s data duration, and use in both cases an end time of 2 s after the trigger time.

The inference results for selected parameters are shown in Fig. 6. The Bilby analyses using 4 and 8 s of data yield consistent results, and these in turn are consistent with the jitter-only GPBilby analysis (GP-J), although the credible intervals are narrower. For this event we take the 4 s WL analysis as the reference for the JS divergences, computed as described in Sec. III A, so that the quoted values measure jointly the effect of changing the likelihood and of extending the analysis segment.

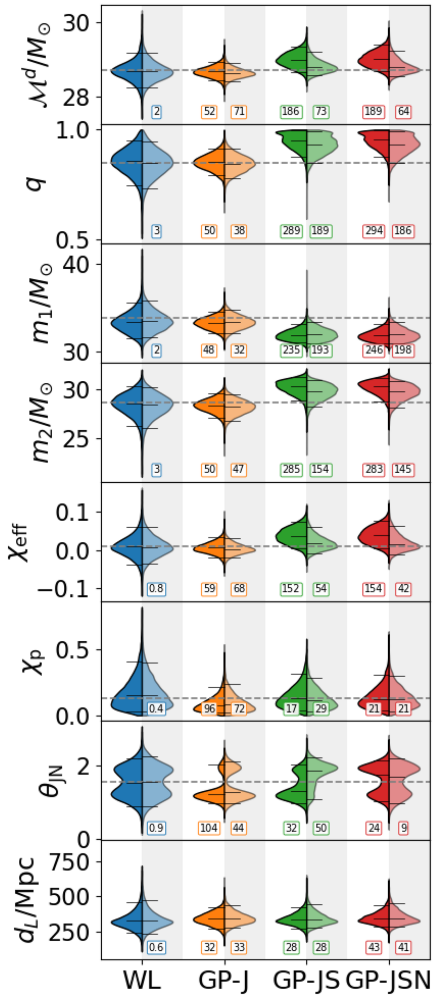


FIG. 6. Violin plots showing selected source-parameter posteriors for GW230814 obtained with different analyses. For each violin, the left half corresponds to the posterior derived using 4 s of data, while the right half shows the posterior obtained using 8 s of data; the 8 s halves are indicated by the grey shaded background. From left to right, the violins correspond to analyses performed with Bilby using the Whittle likelihood, GPBilby including only a jitter term, GPBilby including both jitter and SHO terms, and GPBilby including both jitter and SHO terms applied to data in which a 30 Hz band around 345 Hz has been notched, as discussed in the text. The boxed numbers give the JS divergence in millibits with respect to the 4 s WL analysis, computed as described in Sec. III A; the reference analysis itself carries no value.

However, the GPBilby analyses that include a SHO term (GP-JS) recover markedly different posterior distributions for several intrinsic parameters. Specifically, we observe: (i) a slightly higher chirp mass, driven by a lower primary mass and higher secondary mass, (ii) a preference for more equal-mass systems (higher mass ratio), and (iii) a shift in the effective spin parameter $\chi_{\rm eff}$ toward positive values, away from zero. Notably, while the chirp mass median from the SHO analysis remains within the 90% bounds of the Whittle likelihood results, the posteriors show narrower credible intervals. These shifts are reflected in the JS divergences quoted in Fig. 6, which for the GP-JS analyses generally exceed the level attributable to sampling noise for the affected parameters, in contrast to the GP-J analyses. Furthermore, the 4 and 8 s SHO analyses show some differences: the 8 s analysis yields chirp mass and effective spin posteriors more similar to the Bilby results than the 4 s SHO analysis.

Examining the inferred values of the SHO terms in the GP-JS analysis, Fig. 7 demonstrates that the SHO term identifies power at ∼345 Hz and ∼340 Hz for the 4 s and 8 s analyses, respectively. Examination of the time-frequency spectrogram, included in the data release [54] and shown in Collaboration *et al.* [37], reveals no clearly visible power at these frequencies, neither do spectral lines appear in the PSD in this frequency range. Interestingly, the jitter term is unconstrained for this event, signaling difficulties in modeling the distribution of the Gaussian and stationary part of the noise.

To investigate whether the inference changes are driven by the modeling of this ∼345 Hz power as a glitch, we perform additional GP-JS runs using a PSD notched in the frequency range 330 to 360 Hz, following the procedure described in Sec. III A (we designate these notched analyses as GP-JSN). The posterior distributions for these notched runs, shown in the rightmost violins of Fig. 6, agree with those obtained using unnotched data (i.e., GP-JS) for both the 4 and 8 s datasets. This suggests that the ∼345 Hz power is not responsible for the differences between the GP-JS analysis and the GP-J and WL analyses.

In Fig. 7, we also include the SHO posterior distributions from the GP-JSN analysis. Interestingly, the SHO term still identifies power at 340 to 345 Hz. This persistence can be attributed to the fact that notching is performed in the frequency domain, while the data analysis is conducted in the time domain. The inverse Fourier transform necessarily reintroduces some power to the notched frequency bins, allowing the SHO term to continue modeling features in this frequency range.

To further investigate the discrepancies between the GP-JS and GP-J analyses, we compare the signal and noise

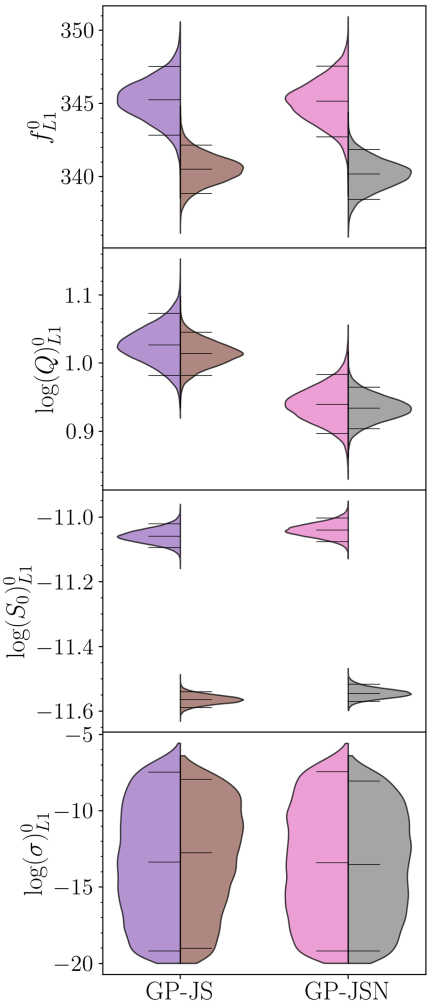


FIG. 7. Posterior distributions for the SHO term parameters of the GPBilby GP-JS analyses of GW230814 (LLO detector). For each violin, the left half corresponds to the posterior derived using 4 s of data, while the right half shows the posterior obtained using 8 s of data.

predictions from the GP model across different runs. Figure 8 summarizes the comparison between the different GPBilby analyses. The left panel, comparing the 8 s SHO runs with notched and unnotched data, shows that the main differences arise from the notching procedure, appearing as high-frequency beating patterns that are most evident during and after the merger phase. Additional diagnostics based on a time-frequency spectrogram of the residuals show that their dominant frequency content coincides with the high-frequency part of the signal, where the largest residual amplitudes are observed. The comparisons between the SHO and jitter-only runs, shown for both the 4 s (right panel) and 8 s analyses (middle panel), reveal that the largest discrepancies occur in the premerger and merger stages, where the two models predict different signal morphologies. These differences are more pronounced in the 4 s analysis, which exhibits larger-amplitude residuals, consistent with the stronger shifts observed in the inferred source parameters.

Overall, these results indicate that the inference differences between the SHO and jitter-only analyses are primarily driven by the modeling of the premerger and merger portions of the signal. In the high-SNR regime, subtle structures or fluctuations in these phases can be absorbed either by the signal model or by the noise model, depending on the chosen GP kernel. The SHO term, owing to its ability to capture narrow-band features, absorbs power not accounted for by the jitter-only model, leading to compensatory changes in the inferred source parameters to maintain agreement with the data.

D. GW191109

After validating GPBilby's ability to accurately infer source parameters for gravitational-wave signals unaffected by nearby glitches, we now turn to events with data quality issues, starting with GW191109_010717 (hereafter GW191109). This event is of particular astrophysical interest, as it exhibits strong support for an effective inspiral spin that is antialigned with the orbital angular momentum, favoring a dynamical formation scenario for the binary system [38]. GW191109 is also interesting from the perspective of tests of fundamental physics: consistency tests performed as part of the GWTC-3 analysis identified anomalous behaviour for this event [56], which was subsequently attributed to the presence of glitches in both detectors. These features make GW191109 a compelling case study for assessing the robustness of parameter inference and tests of general relativity in the presence of non-Gaussian noise.

1. Data description: Raw and deglitched frames

Glitches were identified in both the LHO (H1) and LLO (L1) detectors around the time of GW191109. In LHO, a glitch is present approximately 1.2 s to 2 s before merger at a frequency of ∼36 Hz. This transient lies outside the signal track and is therefore not expected to significantly affect source parameter inference.

In LLO, the noise is more complex. The data contain a series of slow-scattering arches [57] that repeat in time and frequency. The strongest of these occurs approximately 4 s before the trigger time and is followed by additional arches near 24 Hz and 36 Hz, both of which overlap in time with the gravitational-wave signal track. The LHO glitch and the LLO feature at 24 Hz were subtracted in the deglitched data products used for the GWTC-3 parameter estimation [58]. In contrast, the excess power at 36 Hz in LLO was not removed, as it could not be unambiguously separated from the signal.

Figure 17, in the Appendix, shows time–frequency representations of the raw data, the reconstructed glitches, and the glitch-subtracted frames for both detectors. In the following, we analyze both the raw and deglitched data in

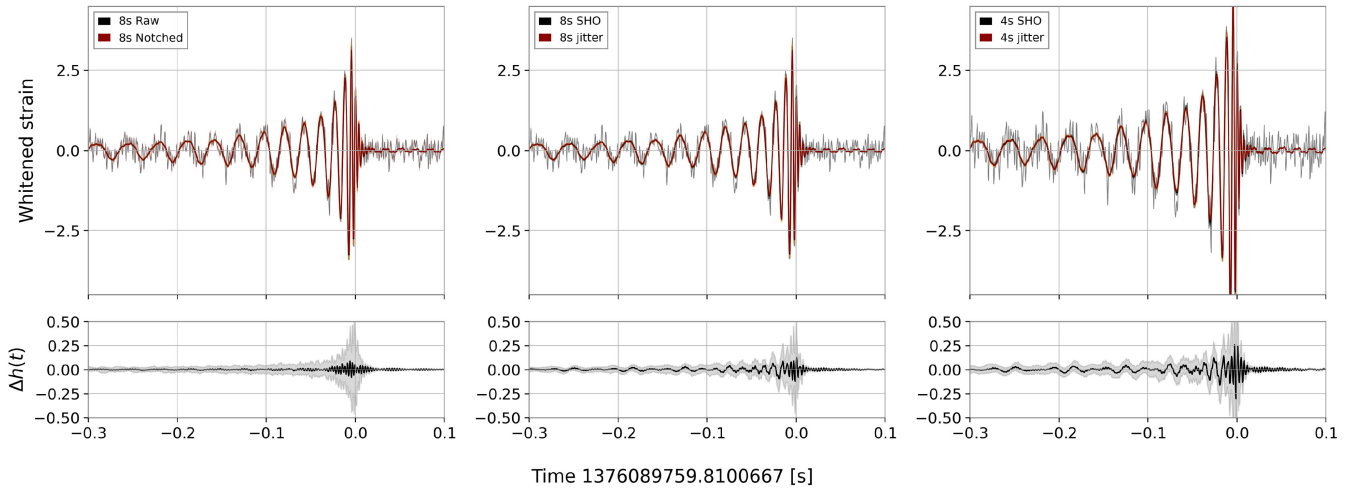


FIG. 8. Posterior predictive plots for the event GW230814 for the LLO detector. In each panel we compare the strain prediction given by the maximum likelihood waveform for two analyses using the `IMRPhenomXPHM_SpinTaylor` waveform. The upper plots show the whitened strain data and the predicted strains with 90% confidence levels, while the bottom plots show the residuals between the predicted strains and their 90% confidence levels. In the left panel we compare two GP-JS analyses, one using 8 s of the full raw data, and one using 8 s of data notched from 330 to 360 Hz. The middle panel compares the 8 s GP-JS and GP-J analyses. The right panel compares the 4 s GP-JS and GP-J analyses. Differences in the whitened strain relative to Fig. 1 of The LIGO Scientific Collaboration *et al.* [37] arise from the use of different data-conditioning procedures.

order to assess the impact of glitch modeling on parameter inference.

2. Joint signal-glitch inference with GPBilby

We perform a systematic analysis of GW191109 using GPBilby, with the goal of jointly modeling the gravitational-wave signal and non-Gaussian noise features. For both raw and deglitched data, we consider data segments of 4 and 8 s duration, with the end time fixed to 2 s after merger. For each dataset we perform a WL analysis, a GP-J and a GP-JS analysis using the settings employed in the `IMRPhenomXPHM_SpinTaylor` analysis from Abbott *et al.* [58]. The inference results for selected parameters are shown in the left panel of Fig. 9. Since in the preliminary results, we found support for multiple modes in the GP parameters, when using the raw data frames, we performed additional analyses employing two and three SHO terms with those data frames. The inference results for those runs are shown in the right panel of Fig. 9, and the posteriors of the GP parameters for all the runs employing at least one SHO term and raw data are shown in Fig. 10. For this event we take the 4 s WL analysis of the raw data as the reference for the JS divergences, computed as described in Sec. III A. The same reference is used for both panels of Fig. 9, so that the values quoted for the multiple-SHO analyses are directly comparable with those of the single-SHO configurations.

Including an SHO term in the analysis of the raw data reveals evidence for excess power in both detectors. The median frequency values found in our analyses are overplotted as red dashed horizontal lines in Fig. 17 in the Appendix. For the 8 s analysis, the characteristic frequencies inferred for the dominant SHO components are $36.23^{0.65}_{0.63}$ Hz in H1 and $22.35^{+3.63}_{-1.88}$ Hz in L1. Comparison with the time-frequency spectograms in Fig. 17 indicates that the LLO feature can be associated with the slow-scattering glitch occurring approximately ∼3.2 s before the trigger. The interpretation of the LHO feature is evident, as the inferred frequencies correspond to the glitch occurring ∼4.5 s before merger, and the same excess power is only partially found in the 4 s analysis. The power found in the 4 s analysis for LHO can be related to the residuals of the glitch subtraction.

Overall, the inferred source-parameter posteriors for the BBH are largely consistent across configurations, indicating that the parameter estimation is relatively robust to the details of the glitch modeling. This is borne out by the JS divergences quoted in Fig. 9, which for most parameters and configurations remain close to the level attributable to sampling noise, apart from the analyses using multiple SHO terms and 4 s of data, for which the results remain qualitatively comparable. The inferred posterior on the effective inspiral spin $\chi_{\rm eff}$ remains qualitatively similar and consistently favors negative values for all analyses. Interestingly, the nonzero support for non-negative spins seen in the analysis with Bilby using the deglitched frames for 4 and 8 s data lengths, vanishes in all the GPBilby analyses, especially those including at least one SHO term. The primary differences arise in the behaviour of the SHO terms themselves. In particular, the SHO components recover stronger and more sharply constrained excess power in the LHO detector than in LLO, consistent with the higher-amplitude glitch observed in H1. In LLO, excess

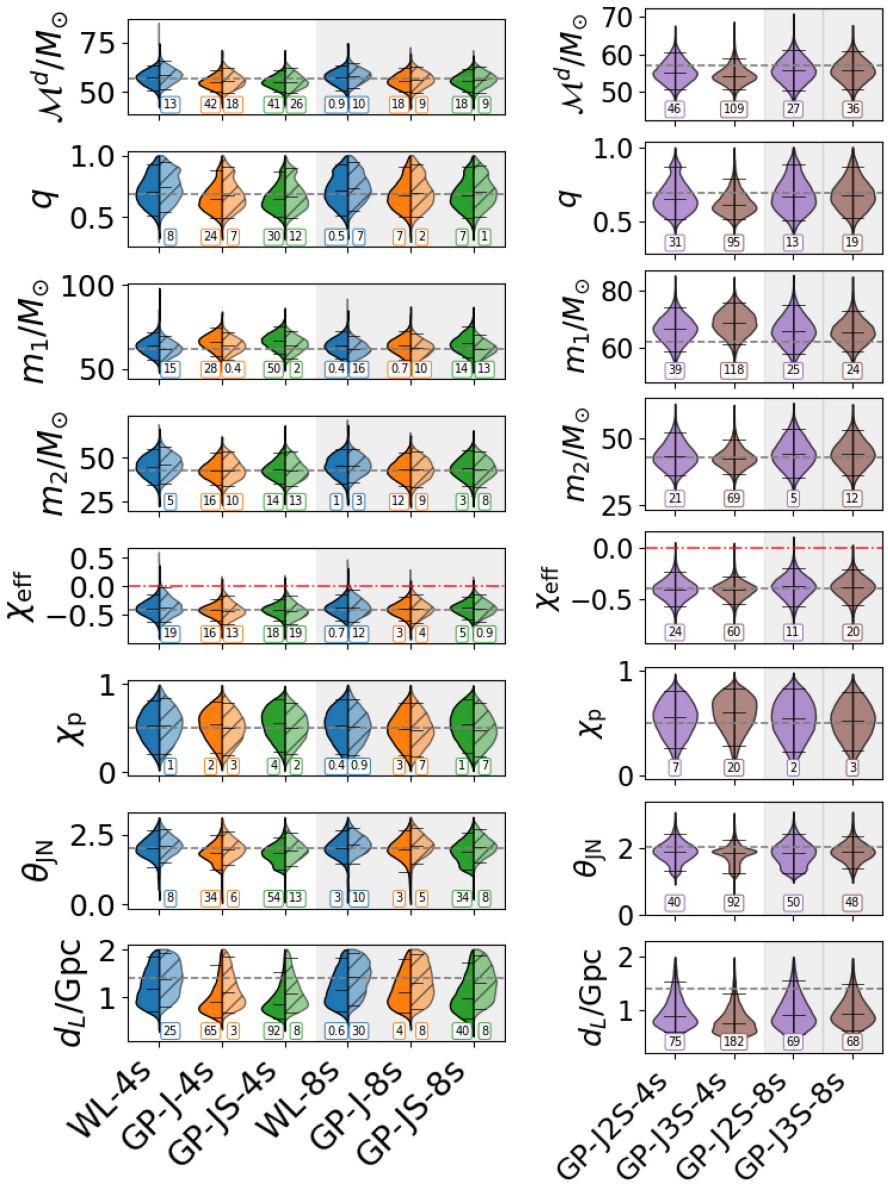


FIG. 9. Posterior distributions for the source parameters of GW191109 obtained with the WL and the Gaussian-process likelihoods with jitter only (GP-J) and jitter plus one SHO term (GP-JS). Left: each model is shown for both the 4 and 8 s analysis segments; for every model the left (solid) half violin corresponds to the raw data and the right (hatched, translucent) half violin to the deglitched data. The 8 s analyses are indicated by the gray shaded background, and colors denote the likelihood model, shared between the 4 and 8 s analyses. Right: analyses using additional SHO terms, i.e., GP-J with two (GP-J2S) and three (GP-J3S) terms, for the 4 and 8 s segments, using the raw data only; no deglitched comparison is shown in this panel, so the two halves of each violin are identical and a single value is quoted per analysis. In both panels the dashed gray line marks the reference value and, for $\chi_{\rm eff}$, the red dash-dotted line marks zero. The boxed numbers give the JS divergence in millibits with respect to the 4 s WL analysis of the raw data, computed as described in Sec. III A; the reference analysis itself carries no value.

power is also identified, but with smaller amplitudes and broader 90% confidence levels.

When allowing multiple SHO terms, we find that three SHO components are required to adequately model the glitch structure in the LHO data, whereas two SHO terms are sufficient in LLO. Notably, none of the configurations shows evidence for excess power at ∼36 Hz in LLO being captured by the Gaussian-process noise model, consistent with the difficulty of disentangling this feature from the gravitational-wave signal.

While the inclusion of explicit glitch modeling does not lead to substantial shifts in the source-parameter posteriors relative to the production analysis, it provides a coherent framework in which the observed non-Gaussian noise features—particularly the stronger glitches in LHO—are explicitly identified and marginalized over.

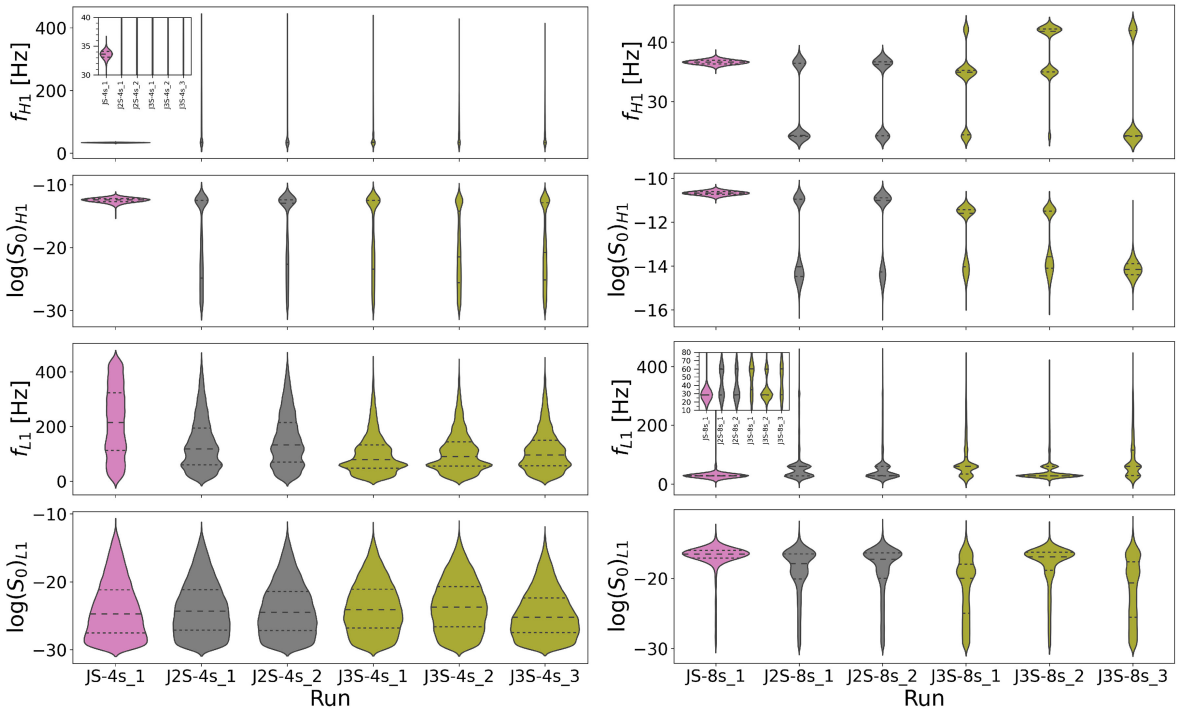


FIG. 10. Left: posterior distributions of the SHO-term parameters for GW191109 analysed on 4 s of data with the GP-JS likelihood, for models using one, two, and three SHO terms. From top to bottom, the panels show the central frequency f and log-amplitude $\log(S_0)$ of the terms in LHO (H1) and LLO (L1). Each violin corresponds to one term of one model. Runs are labelled JnS for a model with n SHO terms. The horizontal position groups the terms by run, and the colour encodes the number of SHO terms in the model. pink for one SHO term, gray for two, and green for three. The interior lines mark the quartiles of each posterior. The inset in the top panel enlarges into the 30–40 Hz range, where the dominant scattered-light feature is recovered. Right: as in the left panel, but for the 8 s analysis of GW191109. The panels show the central frequency f and log-amplitude $\log(S_0)$ of the SHO terms in LHO (H1) and LLO (L1), with the color encoding the number of SHO terms in the model and the interior lines marking the quartiles. The inset in the third panel zooms into the 10–80 Hz range.

3. Comparison with previous glitch-modeling studies

A detailed investigation of GW191109 was recently presented by Udall *et al.* [16], who analyzed this event using a range of waveform approximants and glitch-modeling strategies. In particular, they focused on the excess power at 35 to 40 Hz in the LLO detector and performed joint signal-glitch inference using both a physically motivated slow-scattering model and a flexible wavelet-based approach implemented in `BayesWave` [42,44]. Considering several analysis configurations, they found no decisive evidence favouring an interpretation of the excess power as either a glitch or part of the astrophysical signal.

Our results are complementary to those of Udall *et al.* [16]. Rather than attempting to assign the excess power to a specific origin, the GPBilby framework marginalizes over non-Gaussian noise contributions through a stochastic process model. This allows us to recover inference results consistent with the production analysis while simultaneously accounting for excess power in the raw data. However, we do not interpret the GP component as a diagnostic of the physical origin of the 35 Hz feature; if this power is associated with glitchlike noise, our analysis marginalizes over its contribution rather than definitively attributing its provenance. The agreement between the GPBilby analyses and the results from [16] therefore provides additional evidence that GW191109 is characterized by a negative effective spin, but can resolve the tension in the interpretation of the 35 Hz power.

4. Inspiral-merger-ringdown consistency test

Finally, we perform an inspiral-merger-ringdown (IMR) consistency test by analyzing different portions of the data separately using the same GPBilby configuration. This enables a direct comparison of the posterior distributions for the final mass and spin inferred from the inspiral and postinspiral regimes, while consistently modeling non-Gaussian noise.

This analysis is particularly noteworthy because GW191109 was excluded from the IMR tests performed in the third observing run tests of general relativity due to concerns about data quality [56]. By explicitly modeling glitches alongside the signal, GPBilby enables an IMR test for this event despite the presence of strong non-Gaussianities. The results of this test are shown in

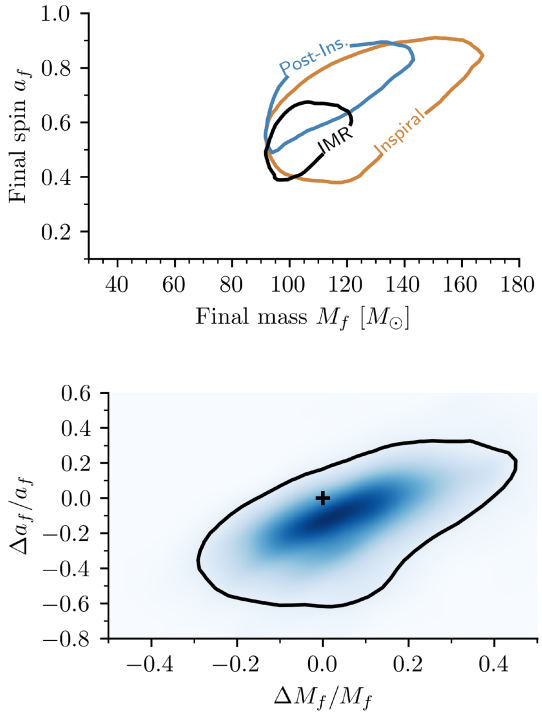


FIG. 11. Inspiral-merger-ringdown consistency test for GW191109. The top panel shows the remnant mass and spin inferred from the full signal (black), the inspiral and merger portion (orange), and the postmerger portion of the signal (blue). The bottom panel shows the posterior on the fractional differences $\Delta M_f/M_f$ and $\Delta a_f/a_f$ between the inspiral and postinspiral estimates. The black contour denotes the 90% credible region, and the point (0,0) corresponds to the value expected for consistency with general relativity.

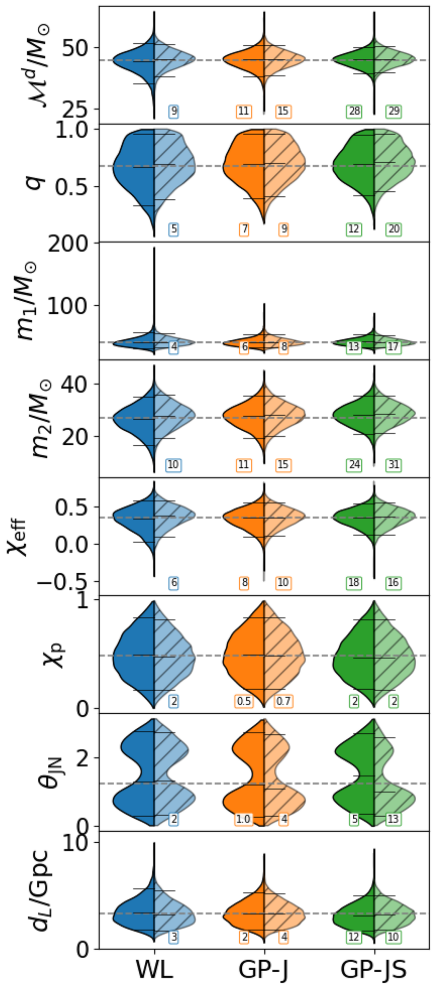


FIG. 12. Violin plots of selected GW231113_122623 source parameter posteriors for the WL, GPBilby with a single jitter term (GP-J), and GPBilby with a jitter and SHO term (GP-JS). The left side of each violin shows the posterior obtained from the raw data, while the right side shows the posterior obtained from the deglitched data, indicated by the hatching. The dashed grey line indicates the median value for the `IMRPhenomXPHM_SpinTaylor` posteriors reported in the discovery paper [3]. The boxed numbers give the JS divergence in millibits with respect to the WL analysis of the raw data, computed as described in Sec. III A; the reference analysis itself carries no value.

Fig. 11. The remnant mass and spin inferred from the inspiral and postinspiral portions of the signal, in the top plot, show broad agreement, with overlapping credible regions in the M_f–a_f plane. This consistency is also visible in the posterior on the fractional differences in the bottom plot, where the GR-consistent value $(\Delta M_f/M_f, \Delta a_f/a_f) = (0, 0)$ lies within the 90% credible region. The constraints are broad, particularly for the final spin, but we find no evidence for a statistically significant inconsistency between the inspiral and postinspiral estimates. These results demonstrate the potential of Gaussian-process-based methods to extend precision tests of gravity to events previously deemed unsuitable for such analyses.

E. GW231113_122623

We analyze the event GW231113_122623, first presented in the GWTC-4.0 catalog [55] and observed by the LHO and LLO detectors. The signal is relatively weak, with a network SNR of just 8.6. Additionally, the data from LLO was found to have data quality issues. Therefore, deglitched frames were created using `BayesWave` and targeting a region from 70 to 120 Hz in the 0.2 s window after the merger time. We therefore use this case study to investigate how well GPBilby performs for a typical analysis where we have a single detector impacted by data quality issues and the glitch and signal tracks are reasonably well separated (see Fig. 18 in the Appendix). We perform a WL analysis, a GP-J and a GP-JS analysis using the settings employed in the `IMRPhenomXPHM_SpinTaylor` analysis from Abac *et al.* [39]. We perform each of these analyses twice, once using the raw data and once using the deglitched data. In Fig. 12, we show the posteriors of selected source parameters using a violin plot comparing the inference from the raw and deglitched frames. For this event we take

the WL analysis of the raw data as the reference for the JS divergences, computed as described in Sec. III A, so that the quoted values measure jointly the effect of changing the likelihood and of using the deglitched frames. We find overall strong consistency among all the analyses, with the only marked difference being that the WL analysis finds a long tail to high mass in the primary (mirroring the results of the initial analysis [55]). Meanwhile, the two GPBilby analyses are more constraining and do not find support for this long tail. Accordingly, the largest JS divergences in Fig. 12 are found for the mass parameters, while the remaining parameters agree at the level attributable to sampling noise. Examining the posteriors for the GP parameters from the GP-JS analysis (plot included in the data release [54]), we find evidence of glitch mitigation behavior. For LHO, the posteriors on the frequency are broad and uninformative, while the amplitude and quality factor parameters are peaked at their arbitrary lower bound, indicating no significant glitch power (in agreement with the original analysis). For LLO, the frequency posteriors peak at ∼220 Hz for both the raw and deglitched frames, suggesting GPBilby is fitting glitch power at a different frequency than that mitigated in the deglitched frames. However, the agreement in source parameter estimates (see Fig. 12) indicates that this additional power has no meaningful impact on the astrophysical inferences.

F. GW231123

GW231123_135430 (hereafter GW231123) is probably the most massive BBH reported to date [39], with a total mass between $188M_\odot$ to $265M_\odot$ (though see Mandel [59]). The primary component lies within or above the theorized mass gap, where black holes in the range $60M_\odot$ to $130M_\odot$ are expected to be rare due to pair-instability processes, while the secondary lies within this gap. The detection of GW231123 was therefore interpreted as evidence that black holes can form through channels other than standard stellar collapse, and that intermediate-mass black holes with masses of order $\sim 200M_\odot$ may be assembled through gravitational-wave-driven hierarchical mergers. This discovery has prompted renewed discussion of currently accepted theories of stellar evolution and binary black hole formation pathways [60].

The signal was observed by both LIGO detectors, and both data sets are affected by glitches. In LHO, a glitch occurred between 1.7 and 1.1 s before the event in the frequency range 15 to 30 Hz; this glitch was modeled and subtracted using `BayesWave`, and the LVK parameter estimates [39,55] use the resulting deglitched data. Abac *et al.* [39] also report additional broadband non-stationary noise in LHO around the time of the event. In the LLO detector, a glitch was reported in the frequency range 10 to 20 Hz within 3.0 to 2.0 s before the event. However, its time–frequency morphology indicated that it would have no measurable impact on the analysis, and therefore no glitch subtraction was performed. In Fig. 19, we show time-frequency spectrograms of the raw data for the two LIGO detectors together with the deglitched data for LHO. Because of the short duration of the signal and the presence of non-Gaussian noise transients in the surrounding data, several studies have challenged the current interpretation of the event and proposed alternative explanations of the signal [see, e.g., Refs. [17,61–66]]. Furthermore, Jan *et al.* [67] have investigated the possibile measurement of eccentricity.

We analyze GW231123 with GPBilby to test whether the astrophysical signal can be cleanly separated from terrestrial disturbances. We use the same settings as the `IMRPhenomXPHM_SpinTaylor` analysis of Abac *et al.* [39] and perform analyses on both the raw and deglitched data. For each data product, we perform three runs: a WL analysis, a GP-J analysis, and a GP-JS analysis.

The results are presented in Fig. 13 using violin plots to compare the inferred source parameters for both deglitched and raw data. For this event the JS divergences, computed as described in Sec. III A, are quoted with respect to the Whittle analysis of the raw data using the same waveform family: the `IMRPhenomXPHM_SpinTaylor` runs are referenced to the WL analysis and the `NRSur7dq4` runs to the WL-NR analysis. The divergences therefore quantify the effect of the likelihood and data product within each waveform family, and are not directly comparable between the two families. The WL analysis shows good agreement with the original LVK results. By contrast, the GP-J and GP-JS analyses favor a smaller secondary mass, while the primary mass estimates remain consistent. Even with this reduction, the inference still identifies GW231123 as one of the most massive events observed: for the GP-JS analysis on deglitched data, we infer a total mass of $227^{+22}_{-23}M_\odot$. In addition, the GPBilby analyses show support for a negative effective spin parameter. This suggests that at least one component has significant spin misalignment with the orbital angular momentum, which is more consistent with dynamical assembly than with isolated binary evolution.

To further investigate these results, we perform an additional analysis using the `NRSur7dq4` waveform model with the settings employed in Abac *et al.* [39]. We analyze both the raw and deglitched data using the Whittle likelihood and GPBilby with a jtter term and a single SHO term. The results, shown in the two rightmost violins in Fig. 13, indicate that in this case no significant difference is recovered between the Bilby and GPBilby analyses. This contrast is quantified by the JS divergences within each family: the GP analyses of the `IMRPhenomXPHM_SpinTaylor` runs depart from their Whittle reference for the parameters discussed above, whereas the `NRSur7dq4` GP analysis remains closer to its own reference across all parameters.

A clearer picture emerges when directly comparing the waveform reconstructions obtained with the two

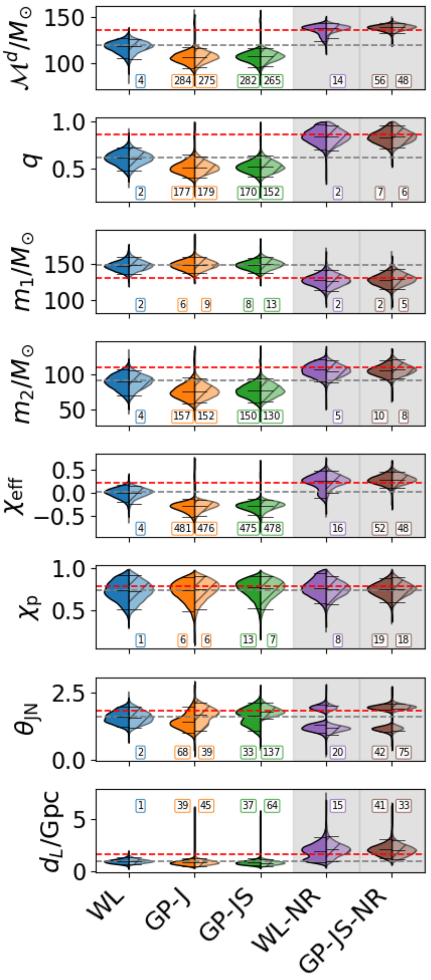


FIG. 13. Violin plots of selected GW231123 source parameter posteriors for the WL, GPBilby with a single jitter term (GP-J), and GPBilby with a jitter and SHO term (GP-JS) with `IMRPhenomXPHM_SpinTaylor`, and for the Whittle likelihood (WL-NR) and GPBilby with a jitter and SHO term (GP-JS-NR) with `NRSur7dq4`, the latter two indicated by the gray background. The left side of each violin shows the posterior from the raw data, the right side (hatched) that from the deglitched data. The dashed grey and red lines indicate the median values for the `IMRPhenomXPHM_SpinTaylor` and `NRSur7dq4` posteriors reported in the discovery paper [37], respectively. Boxed numbers give the JS divergence in mbit with respect to the Whittle analysis of the raw data of the same waveform family, and are therefore not comparable across families.

waveform families, shown in Fig. 14. Although both models provide acceptable fits to the data at the level of the overall signal morphology, the reconstructed waveforms exhibit systematic differences. In particular, the NRSur-based reconstruction predicts noticeably different amplitudes and phases around the merger peaks, while smaller but coherent deviations are also visible during the late inspiral. These differences are most apparent in the LLO detector owing to its higher SNR ratio; however, when considered relative to the signal amplitude, the residual structures are comparable in both detectors. The bottom panels, showing the residuals between the two reconstructions, indicate that these deviations are coherent features concentrated primarily in the highest-frequency portion of the signal rather than random fluctuations.

The presence of such coherent residuals provides a natural explanation for the behavior observed in the parameter-estimation results. In particular, the inclusion of Gaussian-process terms leads to shifts in the inferred parameters for the `IMRPhenomXPHM_SpinTaylor` analyses, while the corresponding `NRSur7dq4` results remain stable. This suggests that, for `IMRPhenomXPHM_SpinTaylor`, part of the mismatch between waveform and data is absorbed by the GP model, whereas for `NRSur7dq4` the waveform reconstruction leaves comparatively little coherent structure to be captured by the GP. The improved agreement between waveform and data in `NRSur7dq4` allows for the inference to remain consistent between the WL and GP analyses.

This interpretation highlights that the behavior of the GP-enhanced likelihood can be intrinsically coupled to waveform accuracy. If the waveform model captures the signal morphology well, the GP terms remain subdominant and primarily account for genuine noise features. If instead the waveform leaves coherent residuals, the GP can absorb part of this structure, effectively redistributing power between the signal and noise models. The resulting posterior shifts therefore provide indirect evidence for modeling systematics.

These findings should be interpreted in the context of the analysis of Bini *et al.* [68], who argued that waveform systematics and Gaussian-noise effects can account for the observed behavior of GW231123 and that the `NRSur7dq4`-based interpretation remains robust. While our results are broadly consistent with the overall astrophysical picture, they highlight an aspect that becomes apparent when using a flexible likelihood model. Although Bini *et al.* note that the waveform models are highly similar (as quantified by differences in the whitened signal), our waveform reconstructions reveal coherent and visibly significant discrepancies in key parts of the signal, both in the merger cycles and in the preceding portion of the waveform. These differences are large compared to the local signal structure and lead to systematic shifts in the inferred source parameters between the two waveform families, even though the qualitative astrophysical interpretation remains unchanged.

Such localized discrepancies leave structured residuals after subtraction that can be captured with a GP-enhanced likelihood. From this perspective, the GP analysis acts as a diagnostic of waveform mismodeling: waveform differences that may appear subdominant in global comparisons can become clearly visible through their residual structure and their effect on inference.

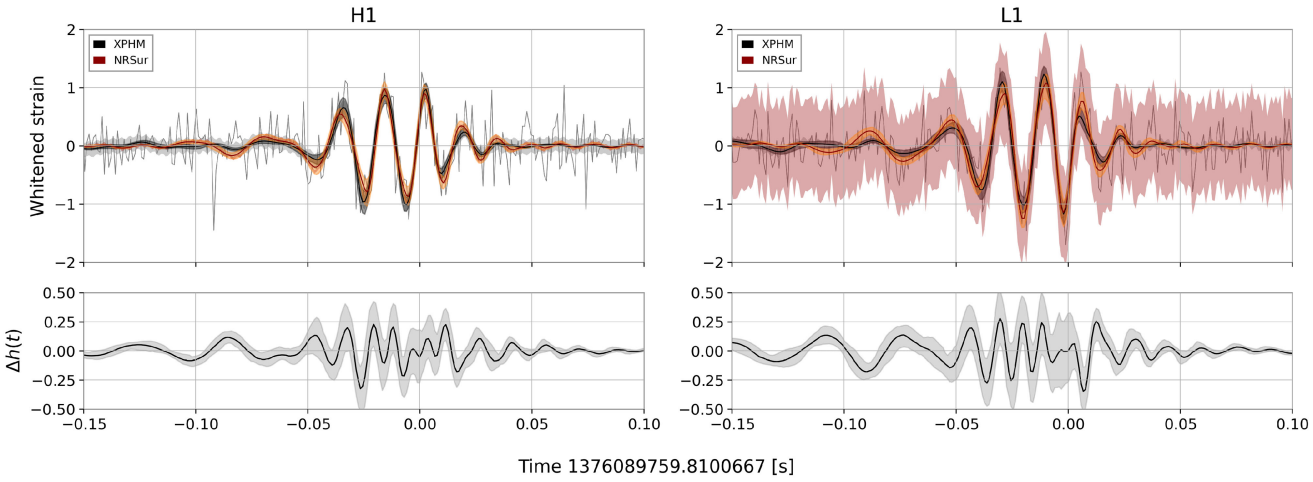


FIG. 14. Posterior predictive plots for the event GW231123 for the LHO and LLO detectors, shown in the left and right columns, respectively. The upper panels display the reconstructed strain obtained from a GPBilby analysis using one SHO term with the `IMRPhenomXPHM_SpinTaylor` waveform and the `NRSur7dq4` waveform, for the Hanford and Livingston detectors on the left and right, respectively. Three shaded bands are shown in each upper panel, all giving 90% confidence levels. The grey band is the `IMRPhenomXPHM_SpinTaylor` reconstruction and the orange band the `NRSur7dq4` reconstruction; both are narrow and follow the respective median waveforms closely. The broader brown band is the full posterior predictive distribution, i.e. the combined signal-plus-Gaussian-process description of the data. Its width is therefore set predominantly by the inferred amplitude of the GP noise component: it remains wide in the pre- and postmerger regions, where the signal amplitude is negligible and where uncertainty in the waveform parameters alone would produce a vanishing band, and its high-frequency structure is characteristic of the correlated-noise component. The comparatively large width of this band in the LLO panel reflects a larger uncertainty in the GP contribution in that analysis. The residuals between the two waveform reconstructions are shown in the bottom panels. Differences in the whitened strain relative to Fig. 1 of Abac *et al.* [39] arise from the use of different data-conditioning procedures.

GW231123 therefore provides a useful case study illustrating how waveform systematics, noise modeling, and parameter inference are tightly coupled within the analysis. Within a GP-enhanced likelihood framework, coherent residuals arising from waveform differences can be partially absorbed by the noise model, translating modeling discrepancies into shifts in the inferred astrophysical parameters. The comparison between `IMRPhenomXPHM_SpinTaylor` and `NRSur7dq4` shows that this behavior depends sensitively on waveform accuracy. When coherent waveform residuals are present, GP terms can absorb part of the signal structure and modify the balance between signal and noise contributions. In contrast, when the waveform provides a closer representation of the data, the GP contribution remains subdominant and the inferred parameters are stable across likelihood choices. In this sense, the GP analysis acts as a probe of residual structure that may not be evident from global waveform comparisons alone. These results motivate treating waveform modeling and likelihood flexibility as jointly interacting components when interpreting individual high-mass events.

G. GW230630_070659

GW230630_070659 is a candidate gravitational-wave signal identified in O4a by the GstLAL pipeline [69–73], with a false alarm rate of 0.47 per year and a p_{astro} of 0.88 [55]. However, offline follow-up (as described in Ref. [74]) identified this event as likely of instrumental origin, due to excess power in both detectors, consistent with scattered light [75]. In Fig. 20, we reproduce the time-frequency spectrograms first reported in Abac *et al.* [55], but with an extended time span.

To further investigate GW230630_070659, we perform WL and GP-JS analyses of the event using the `IMRPhenomXPHM` waveform. We know that the signal must have a reasonably good match with the BBH model, since it was picked up by a modeled search pipeline. However, our motivation is to test if, after fitting the BBH signal, there is any residual power modeled by the GP. This would be indicative of a non-BBH source, e.g., an unfortunate coincidence between two otherwise independent glitches.

Since parameter estimation was not performed in the original analyses, we set up new analyses, but followed the same approach as applied to other GWTC-4.0 candidates [13].

In Fig. 15, we present the inferred source parameters. The comparison demonstrates good agreement between the two analyses, with shifts in the median below the uncertainty in the posterior distribution itself. Taking the WL analysis as the reference, the JS divergences, computed as described in Sec. III A, remain at or close to the level attributable to sampling noise for all parameters considered,

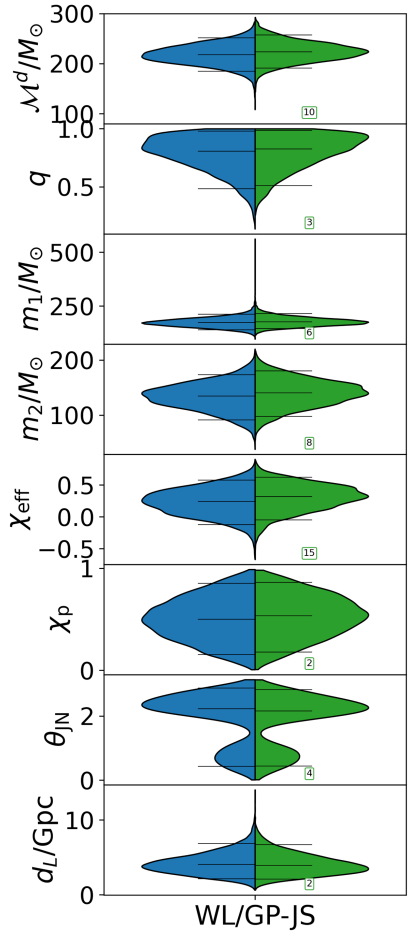


FIG. 15. Violin plots of selected GW230630_070659 source parameter posteriors for the WL and GPBilby with a jitter and SHO term (GP-JS). The left and right side of each violin are identical. Boxed numbers give the JS divergence in mbit with respect to the WL analysis.

confirming the good agreement. We infer that, if real, GW230630_070659 is an exceptionally high-mass event: the primary source mass is inferred to be $173^{+39}_{-33}M_\odot$ by the WL analysis. This is significantly larger than GW231123, which has a primary mass of 137^{+23}_{-18} [39]. Therefore, this would make the primary of GW230630_070659 one of the largest black holes observed to date. Moreover, the secondary of GW230630_070659 is also massive, at $135^{+39}_{-44}M_\odot$. Combined, the total mass would be 306^{+64}_{-53}, again exceeding that of GW231123, the current record holder.

Meanwhile, for the inferred GP parameters (the plot is provided in the data release [54]), the frequency posteriors are uninformative and the amplitude posterior peaks at the arbitrary lower bound. This indicates that, within the sensitivity of our model, we do not identify additional residual power beyond that captured by the BBH waveform. We emphasize, however, that this result should be interpreted with caution. In particular, the absence of significant residual power in the GPBilby analysis does not imply that the data are consistent with purely Gaussian noise or that the event is astrophysical in origin. Our method assumes a BBH signal model and tests for additional structured deviations, but it is not designed to distinguish between astrophysical signals and coincident or incoherent noise artifacts. This is especially relevant in light of the GWTC-4.0 analysis, which identified excess power in both detectors inconsistent with a compact binary coalescence signal and attributed the candidate to instrumental noise. The apparent consistency of the data with a BBH model in our analysis may therefore reflect the limited duration and low signal-to-noise ratio of the candidate, which make it difficult to robustly distinguish between competing interpretations. We therefore conclude that, while GW230630_070659 can be fit by a BBH waveform without requiring additional glitch modeling in this framework, our analysis does not provide evidence in favour of an astrophysical origin.

IV. DISCUSSION AND CONCLUSION

In this work, we have presented a set of case studies using GPBilby to model observed gravitational-wave signals in both Gaussian and non-Gaussian noise environments. By jointly modeling the astrophysical signal and transient noise artifacts with a time-domain Gaussian process, we demonstrate a flexible alternative to traditional glitch-subtraction approaches.

For events consistent with Gaussian noise, such as GW150914 and GW170814, GPBilby yields posterior distributions in good agreement with the standard Whittle likelihood. In GW170814, the GP analysis produces tighter credible intervals and reduces the weight of secondary modes in the inclination-angle posterior. In GW150914, GPBilby also identifies narrowband instrumental features, with the SHO terms recovering the 60 Hz power line, while leaving the astrophysical inference effectively unchanged. The main visible difference for GW150914 is the reduced support for the secondary mode in the inclination angle, $\theta_{\rm JN}$, relative to the Whittle-likelihood analyses. This behavior may indicate that explicitly modeling narrowband residual structure allows the likelihood to use the phase and amplitude information in the data more effectively, thereby reducing degeneracies in parameters such as the inclination. However, because the inclination posterior is sensitive to waveform systematics, calibration, and details of the noise model, we do not interpret this difference as evidence of a bias in either analysis [76]. A more systematic study using controlled injections will be required to determine what drives the different relative weights of the inclination modes.

We then investigated glitch-contaminated events. For GW191109, the evidence for a negative effective inspiral spin remains robust when analysing raw data with GPBilby, showing that this conclusion is not driven by transient

noise. This event additionally enabled an inspiral-merger-ringdown consistency test that was previously excluded owing to data-quality concerns. Similarly, for GW231113, GPBilby proved more constraining than standard analyses by identifying glitch power without introducing significant bias in the inferred astrophysical parameters.

The case study of GW231123 provides a particularly informative example of how GP-enhanced likelihoods can probe waveform performance in high-mass events. When using `IMRPhenomXPHM`, GPBilby identifies coherent residual structure that leads to measurable shifts in the inferred mass and spin parameters, indicating that non-negligible signal power remains after waveform subtraction. By contrast, analyses based on the `NRSur7dq4` waveform remain stable across likelihood choices and show little excess residual power, suggesting that this model captures the signal morphology more completely. This comparison highlights the diagnostic role of the GP framework: rather than introducing additional bias, the GP terms reveal residual coherent structure when waveform mismatches are present and remain subdominant when the waveform accurately describes the data. Our results, therefore, reinforce the robustness of the NRSur-based interpretation of GW231123, while showing that flexible likelihood models can make subtle waveform differences directly visible through their impact on inference. Consistent with the analysis of Bini *et al.* [68], we find that the overall astrophysical interpretation remains unchanged, but that GP-based analyses provide a complementary way to identify and quantify waveform-related systematics.

A related behavior is observed in our analysis of GW230814, another high-SNR BBH, where the inclusion of an SHO term leads to different parameter inference results and to nonzero residual structure and a small amount of power assigned to the GP component, particularly near the peaks of the time-domain signal. In this case, the effect is likely driven by the very high SNR, which renders subtle waveform inaccuracies directly visible in the residuals, rather than by challenges associated with the underlying parameter space. Together with GW231123, this illustrates that GP-enhanced analyses can act as sensitive probes of small waveform mismatches, both in very high-SNR events and in regions of parameter space in which waveform systematics are expected to be significant.

Finally, we analyzed GW230630_070659, an event excluded from GWTC-4.0 due to suspected instrumental contamination. We find that the data can be fit by a BBH waveform without requiring additional structured residuals in the GPBilby analysis. However, we emphasize that this does not provide evidence for an astrophysical origin, and is not in tension with the GWTC-4.0 classification. The limited duration and low signal-to-noise ratio of the candidate make it difficult to distinguish between a true signal and noise artefacts within this framework. If astrophysical, it would correspond to one of the highest-mass binaries observed to date, motivating further investigation by the broader community.

Overall, our results demonstrate that GPBilby is a powerful diagnostic tool for studying exceptional gravitational-wave events and for separating non-Gaussian noise features from astrophysical signals. At the same time, the GW231123 and GW230814 analyses show that GP-based likelihoods implicitly assume waveform accuracy: when this assumption fails, residual signal power can be absorbed by the noise model. In addition, subthreshold or weakly coherent glitches may not be fully captured by the current GP parametrization, and could therefore remain partially hidden within the inferred noise model. This highlights the importance of continued development of GP kernel choices and more systematic studies of how kernel flexibility, waveform systematics, and noise transients interact in realistic detector data. Waveform modeling and flexible noise descriptions must therefore be considered jointly when interpreting high-mass or otherwise challenging gravitational-wave events.

We utilize the NumPy [77] and SciPy libraries [78] for data processing and analysis, and we also use the Matplotlib [79] library for visualization.

ACKNOWLEDGMENTS

We would like to thank Ling (Lilli) Sun for the discussion that inspired our treatment of calibration errors in GPBilby. We are also grateful to Michael Williams and Sylvia Biscoveanu for their useful comments and suggestions on the analysis of GW231123. This work is supported by the Science and Technology Facilities Council (STFC) Grant No. UKRI2488. 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. The authors are also grateful for computational resources provided by Cardiff University and supported by STFC Grants No. ST/I006285/1 and No. vST/V005618/1. This research has made use of data or software obtained from the Gravitational Wave Open Science Center [80], a service of the LIGO Scientific, Virgo, and KAGRA Collaborations. 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, and Spain. KAGRA is supported by the Ministry of Education, Culture, Sports, Science, and Technology (MEXT), The Japan Society for the Promotion of Science (JSPS) in Japan; the National Research Foundation (NRF) and Ministry of Science and ICT (MSIT) in Korea; and Academia Sinica (AS) and the National Science and Technology Council (NSTC) in Taiwan.

DATA AVAILABILITY

The notebooks and scripts to reproduce the results, as well as the result files to reproduce the plots of this study, are publicly available on the Zenodo repository [54] and can be used with the GPBilby package available on PyPI [30].

APPENDIX: SPECTROGRAMS OF GRAVITATIONAL-WAVE EVENTS

We present the time-frequency representations used to motivate and contextualize the case studies discussed in the main text. These figures are intended as diagnostic visualizations of the data quality and of the relative morphology of the astrophysical signal and non-Gaussian noise features. They provide useful qualitative information about which parts of the data motivate the use of flexible noise modeling. For each relevant event, we show spectrograms of the raw strain data and, where applicable, the reconstructed glitch contribution and

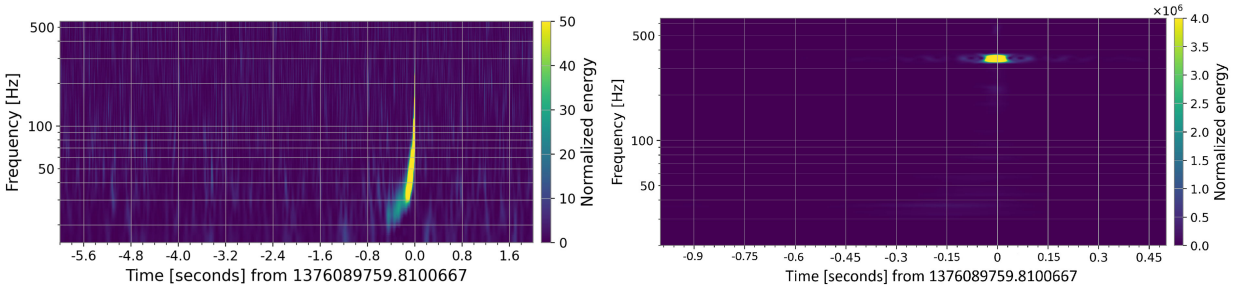


FIG. 16. Left: the time-frequency spectrogram showing LLO data surrounding GW230814. Right: the time-frequency spectrogram showing the residuals obtained from the maximum likelihood waveform for the GPBilby analysis on the unnotched and notched data.

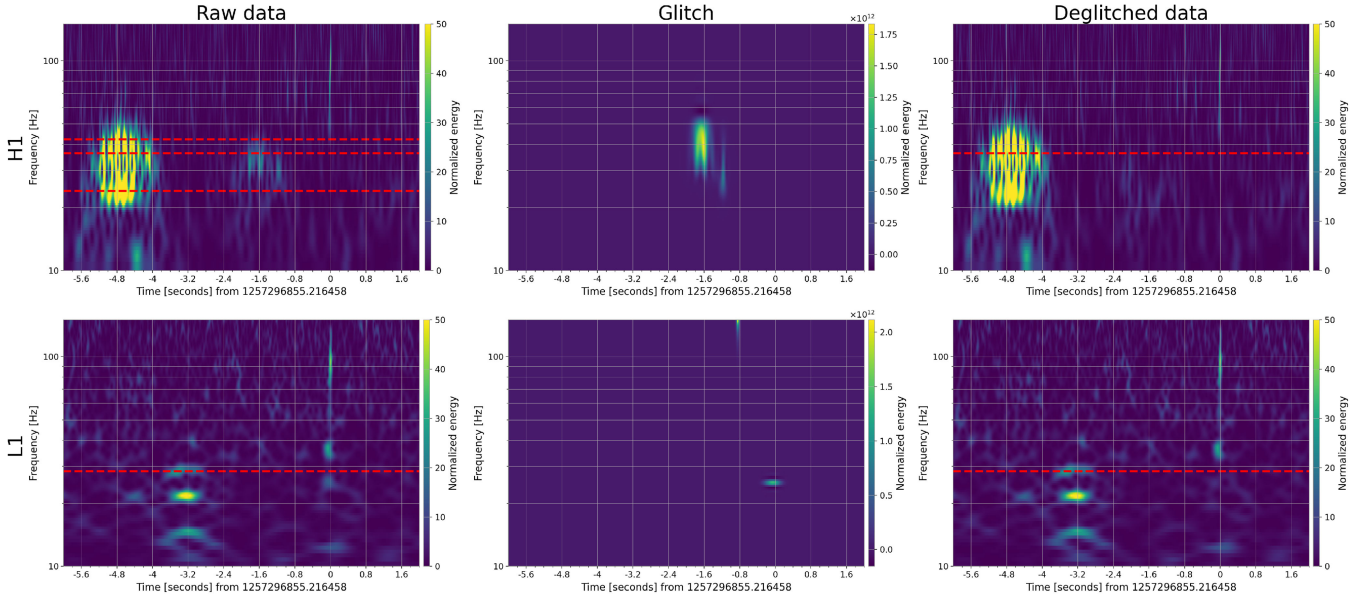


FIG. 17. Time-frequency spectrograms of 8 s of data surrounding GW191109_010717 for the LHO detector (top) and the LLO detector (bottom). For each detector, the panels show from left to right the raw data, the reconstructed glitch, and the glitch-subtracted data.

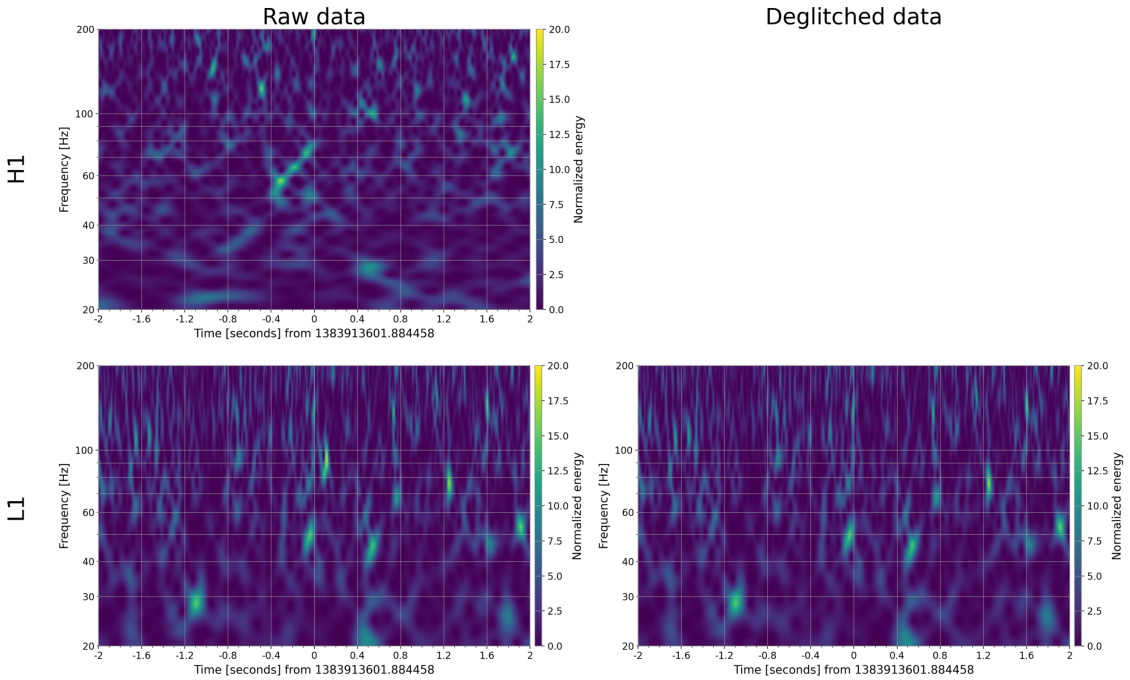


FIG. 18. Time-frequency spectrograms of the data surrounding GW231113_122623. The left panels show the raw data for the LHO and LLO detectors, upper and lower panels, respectively. The right panels show the corresponding deglitched data; no deglitched frames were produced for LHO, so the raw data are shown for both columns. The glitch subtracted from the LLO data had a frequency range of ∼80–100 Hz and occurred ∼0.1 s after the trigger time. We include two seconds of data before and two seconds of data after the trigger time.

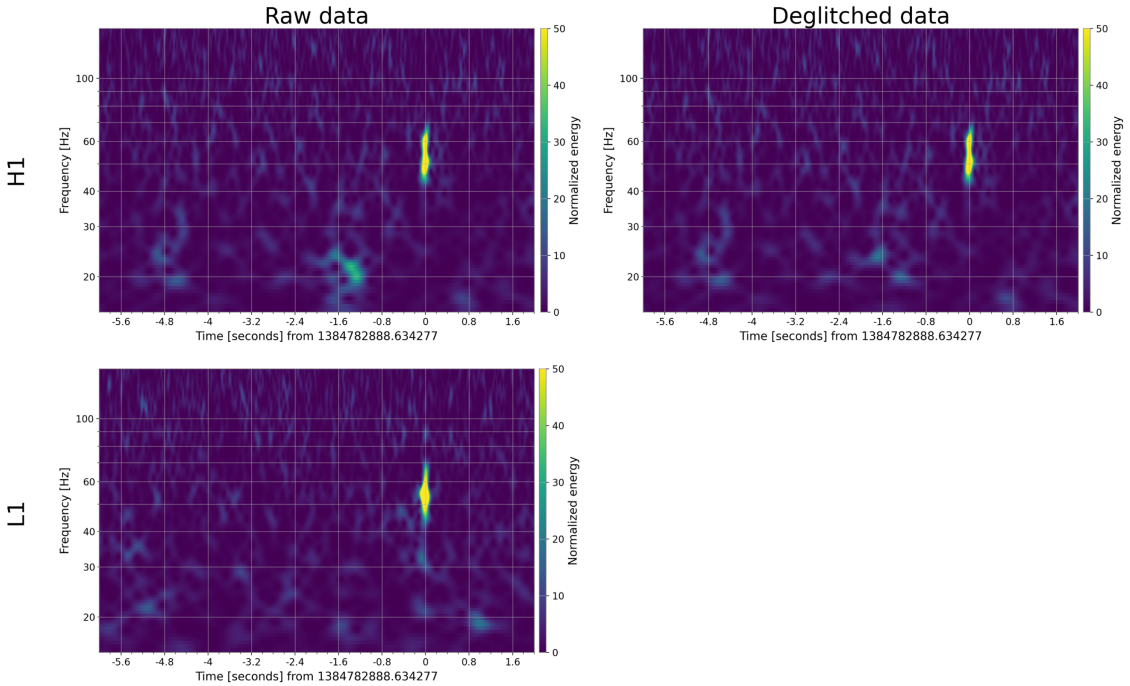


FIG. 19. Time-frequency spectrograms of the data surrounding GW231123. The left panels show the raw data for the LHO and LLO detectors, upper and lower panels, respectively. The right panel shows the deglitched LHO data; no deglitched frames were produced for LLO. The glitch subtracted from the LHO data had a frequency range of ∼15–25 Hz and occurred ∼1.5 s before the trigger time. We include six seconds of data before and two seconds of data after the trigger time.

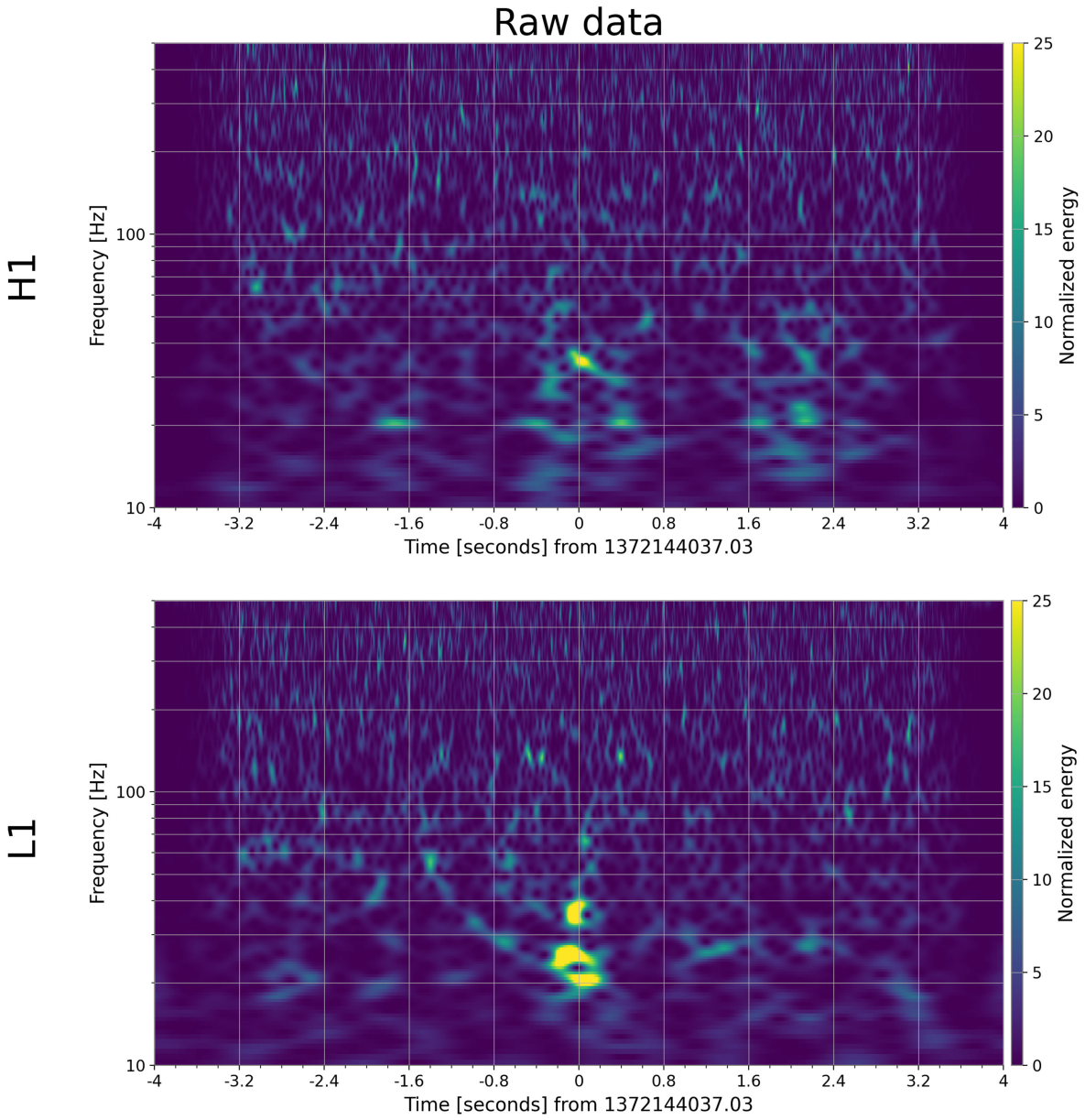


FIG. 20. Time-frequency spectrogram of 8 s of data surrounding GW230630_070659 for the LHO and LLO detectors.

deglitched data. These comparisons illustrate whether excess power is temporally or morphologically separated from the compact-binary signal, whether it appears coherently across detectors, and whether subtraction or marginalization over non-Gaussian noise is expected to affect the astrophysical inference. The spectrograms therefore complement the posterior-comparison figures in the main text: the latter quantify the impact on parameter estimation, while the figures in this appendix show the corresponding time-frequency structure in the data.

For GW230814, the left panel of Fig. 16 shows the time-frequency structure of a high-SNR signal in data that are otherwise sufficiently clean for standard analyses. This event is included to illustrate the behaviour of GPBilby in the high-SNR regime, where small residuals or narrow band features can become visible even when an obvious transient glitch does not dominate the data. The right panel presents the spectrogram of the residuals between the maximum likelihood waveform of the GPBilby runs including a SHO term using the raw unnotched and notched data. As expected, the residuals are significant in the notched data portion.

For GW191109, Fig. 17 shows the raw data, reconstructed glitches, and glitch-subtracted data for both detectors. The red-dotted lines indicate the mean frequency values recovered by the GP SHO terms in the raw and deglitched data analyses. The values are $f_1 = 36.23$ Hz, $f_2 = 23.98$ Hz, and $f_3 = 42.24$ Hz for H1 and $f_1 = 28.42$ Hz for L1 with raw data. In the deglitched data

analyses, we only recover the f_1 frequency for each detector.

For GW231113, Fig. 18 shows a case in which one detector is affected by data-quality issues while the signal and glitch tracks are reasonably well separated. This morphology makes the event a useful example of a typical single-detector contamination problem, where flexible noise modeling should absorb the transient feature without significantly perturbing the coherent astrophysical signal.

For GW231123, Fig. 19 shows the raw data in the two LIGO detectors, together with the deglitched data for LHO. The spectrograms provide context for the waveform-systematics study in the main text: they show the data-quality features that motivate modeling the residual power.

Finally, for GW230630_070659, Fig. 20 reproduces the time–frequency view of the candidate over an extended time span. This broader view helps place the candidate in its surrounding data-quality environment and motivates our cautious interpretation: although the data can be fit with a compact-binary waveform, the limited duration and low signal-to-noise ratio make it difficult to distinguish an astrophysical signal from non-Gaussian noise using this information alone.

[1] A. G. Abac, I. Abouelfettouh, F. Acernese, K. Ackley, S. Adhicary, D. Adhikari, N. Adhikari, R. X. Adhikari, V. K. Adkins, S. Afroz *et al.*, GWTC-4.0: An introduction to version 4.0 of the gravitational-wave transient catalog, Astrophys. J. **995**, L18 (2025).

[2] A. G. Abac, I. Abouelfettouh, F. Acernese, K. Ackley, S. Adhicary, D. Adhikari, N. Adhikari *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), GWTC-4.0: Methods for identifying and characterizing gravitational-wave transients, arXiv:2508.18081.

[3] A. G. Abac *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), GWTC-4.0: Updating the gravitational-wave transient catalog with observations from the first part of the fourth LIGO-Virgo-KAGRA observing run, Astrophys. J. Lett. **1004**, L22 (2026).

[4] J. Aasi, B. P. Abbott, R. Abbott, T. Abbott, M. R. Abernathy, K. Ackley, C. Adams, T. Adams, P. Addesso *et al.* (LIGO Scientific Collaboration), Advanced LIGO, Classical Quantum Gravity **32**, 074001 (2015).

[5] F. Acernese, M. Agathos, K. Agatsuma, D. Aisa, N. Allemandou, A. Allocca, J. Amarni, P. Astone, G. Balestri, G. Ballardin *et al.*, Advanced Virgo: A second-generation interferometric gravitational wave detector, Classical Quantum Gravity **32**, 024001 (2015).

[6] T. Akutsu, M. Ando, K. Arai, Y. Arai, S. Araki, A. Araya, N. Aritomi, Y. Aso, S. Bae, Y. Bae *et al.*, Overview of KAGRA: Detector design and construction history, Prog. Theor. Exp. Phys. **2021**, 05A101 (2021).

[7] A. G. Abac *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), GWTC-4.0: Population properties of merging compact binaries, Astrophys. J. Lett. **1005**, L51 (2026).

[8] A. G. Abac *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), GWTC-4.0: Constraints on the cosmic expansion rate and modified gravitational-wave propagation, Astrophys. J. Lett. **1007**, L17 (2026).

[9] A. G. Abac *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), GWTC-4.0: Tests of general relativity. I. Overview and general tests, arXiv:2603.19019.

[10] A. G. Abac *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), GWTC-4.0: Tests of general relativity. II. Parameterized tests, arXiv:2603.19020.

[11] A. G. Abac *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), GWTC-4.0: Tests of general relativity. III. Tests of the remnants, arXiv:2603.19021.

[12] B. P. Abbott, R. Abbott, T. D. Abbott, S. Abraham, F. Acernese, K. Ackley, C. Adams, V. B. Adya, C. Affeldt, M. Agathos *et al.*, A guide to LIGO-Virgo detector noise and extraction of transient gravitational-wave signals, Classical Quantum Gravity **37**, 055002 (2020).

[13] A. G. Abac *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), GWTC-4.0: Methods for identifying and characterizing gravitational-wave transients, Astrophys. J. Lett. **1004**, L21 (2026).

[14] J. Powell, Parameter estimation and model selection of gravitational wave signals contaminated by transient detector noise glitches, Classical Quantum Gravity **35**, 155017 (2018).

[15] R. Macas, J. Pooley, L. K. Nuttall, D. Davis, M. J. Dyer, Y. Lecoeuche, J. D. Lyman, J. McIver, and K. Rink, Impact of noise transients on low latency gravitational-wave event localization, Phys. Rev. D **105**, 103021 (2022).

[16] R. Udall, S. Hourihane, S. Miller, D. Davis, K. Chatziioannou, M. Isi, and H. Deshong, Antialigned spin of GW191109: Glitch mitigation and its implications, Phys. Rev. D **111**, 024046 (2025).

[17] A. Ray, S. Banagiri, E. Thrane, and P. D. Lasky, GW231123: Extreme spins or microglitches?, arXiv:2510.07228.

[18] R. Udall, S. Bini, K. Chatziioannou, D. Davis, S. Hourihane, Y. Lecoeuche, J. McIver, and S. Miller, Inferring the spins of merging black holes in the presence of data-quality issues, Phys. Rev. D **113**, 042005 (2026).

[19] S. Hourihane and K. Chatziioannou, Glitches far from transient gravitational-wave events do not bias inference, Phys. Rev. D **112**, 084006 (2025).

[20] D. Davis, T. B. Littenberg, I. M. Romero-Shaw, M. Millhouse, J. McIver, F. Di Renzo, and G. Ashton, Subtracting glitches from gravitational-wave detector data during the third LIGO-Virgo observing run, Classical Quantum Gravity **39**, 245013 (2022).

[21] R. Macas, A. Lundgren, and G. Ashton, Revisiting the evidence for precession in GW200129 with machine learning noise mitigation, Phys. Rev. D **109**, 062006 (2024).

[22] R. P. Udall and D. Davis, Bayesian modeling of scattered light in the LIGO interferometers, Appl. Phys. Lett. **122**, 094103 (2023).
[23] S. Hourihane, K. Chatziioannou, M. Wijngaarden, D. Davis, T. Littenberg, and N. Cornish, Accurate modeling and mitigation of overlapping signals and glitches in gravitational-wave data, Phys. Rev. D **106**, 042006 (2022).
[24] S. Ghonge, J. Brandt, J. M. Sullivan, M. Millhouse, K. Chatziioannou, J. A. Clark, T. Littenberg, N. Cornish, S. Hourihane, and L. Cadonati, Assessing and mitigating the impact of glitches on gravitational-wave parameter estimation: A model agnostic approach, Phys. Rev. D **110**, 122002 (2024).
[25] A.-K. Malz and J. Veitch, Joint inference for gravitational wave signals and glitches using a data-informed glitch model, Phys. Rev. D **112**, 024071 (2025).
[26] A. Sasli, M. Karamanis, N. Karnesis, M. W. Coughlin, V. Mandic, U. Seljak, and N. Stergioulas, Beyond Gaussian assumptions: A new robust statistical framework for gravitational-wave data analysis, Phys. Rev. D **114**, 024004 (2026).
[27] G. Ashton, Gaussian processes for glitch-robust gravitational-wave astronomy, Mon. Not. R. Astron. Soc. **520**, 2983 (2023).
[28] C. E. Rasmussen, Gaussian processes in machine learning, in *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, 2003, Tübingen, Germany, 2003, Revised Lectures*, edited by O. Bousquet, U. von Luxburg, and G. Rätsch (Springer, Berlin, Heidelberg, 2004), pp. 63–71.
[29] D. Foreman-Mackey, E. Agol, S. Ambikasaran, and R. Angus, Fast and scalable Gaussian process modeling with applications to astronomical time series, Astron. J. **154**, 220 (2017).
[30] M. Emma, A.-K. Malz, and G. Ashton, GPBilby: A Python package for gravitational-wave data analysis, Version 0.0.1, Python package, https://pypi.org/project/gpbilby/ (2026).
[31] L. Sun, E. Goetz, J. S. Kissel, J. Betzwieser, S. Karki, A. Viets, M. Wade, D. Bhattacharjee, V. Bossilkov, P. B. Covas *et al.*, Characterization of systematic error in Advanced LIGO calibration, Classical Quantum Gravity **37**, 225008 (2020).
[32] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari *et al.*, Properties of the binary black hole merger GW150914, Phys. Rev. Lett. **116**, 241102 (2016).
[33] W. Farr, B. Farr, and T. Littenberg, Modelling calibration errors in CBC waveforms, Technical Report No. DCC-T1400682, LIGO, 2014.
[34] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari *et al.*, GW150914: First results from the search for binary black hole coalescence with Advanced LIGO, Phys. Rev. D **93**, 122003 (2016).
[35] R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, N. Adhikari, R. X. Adhikari, V. B. Adya, C. Affeldt, D. Agarwal *et al.*, GWTC-2.1: Deep extended catalog of compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run, Phys. Rev. D **109**, 022001 (2024).
[36] B. P. Abbott, R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, V. B. Adya *et al.*, GW170814: A three-detector observation of gravitational waves from a binary black hole coalescence, Phys. Rev. Lett. **119**, 141101 (2017).
[37] LIGO Scientific, Virgo, and KAGRA Collaborations, GW230814: Investigation of a loud gravitational-wave signal observed with a single detector, Astrophys. J. Lett. **1004**, L23 (2026).
[38] R. C. Zhang, G. Fragione, C. Kimball, and V. Kalogera, On the likely dynamical origin of GW191109 and binary black hole mergers with negative effective spin, Astrophys. J. **954**, 23 (2023).
[39] A. G. Abac, I. Abouelfettouh, F. Acernese, K. Ackley, C. Adamcewicz, S. Adhicary, D. Adhikari, N. Adhikari, R. X. Adhikari, V. K. Adkins *et al.*, GW231123: A binary black hole merger with total mass 190–265$M_\odot$, Astrophys. J. Lett. **993**, L25 (2025).
[40] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari *et al.*, Observation of gravitational waves from a binary black hole merger, Phys. Rev. Lett. **116**, 061102 (2016).
[41] G. Pratten, C. García-Quirós, M. Colleoni, A. Ramos-Buades, H. Estellés, M. Mateu-Lucena, R. Jaume, M. Haney, D. Keitel, J. E. Thompson *et al.*, Computationally efficient models for the dominant and subdominant harmonic modes of precessing binary black holes, Phys. Rev. D **103**, 104056 (2021).
[42] N. J. Cornish and T. B. Littenberg, BayesWave: Bayesian inference for gravitational wave bursts and instrument glitches, Classical Quantum Gravity **32**, 135012 (2015).
[43] T. B. Littenberg and N. J. Cornish, Bayesian inference for spectral estimation of gravitational wave detector noise, Phys. Rev. D **91**, 084034 (2015).
[44] N. J. Cornish, T. B. Littenberg, B. Bécsy, K. Chatziioannou, J. A. Clark, S. Ghonge, and M. Millhouse, BayesWave analysis pipeline in the era of gravitational wave observations, Phys. Rev. D **103**, 044006 (2021).
[45] T. Gupta and N. J. Cornish, Bayesian power spectral estimation of gravitational wave detector noise revisited, Phys. Rev. D **109**, 064040 (2024).
[46] G. Ashton, Reconstructing and resampling: A guide to utilising posterior samples from gravitational wave observations, RAS Tech. Instrum. **5**, rzag012 (2026).
[47] S. Biscoveanu, C.-J. Haster, S. Vitale, and J. Davies, Quantifying the effect of power spectral density uncertainty on gravitational-wave parameter estimation for compact binary sources, Phys. Rev. D **102**, 023008 (2020).
[48] A. G. Abac *et al.*, GWTC-4.0: An introduction to version 4.0 of the gravitational-wave transient catalog, Astrophys. J. Lett. **995**, L18 (2025).
[49] A. Krolak and B. F. Schutz, Coalescing binaries—Probe of the universe, Gen. Relativ. Gravit. **19**, 1163 (1987).
[50] P. A. R. Ade, N. Aghanim, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. J. Banday, R. B. Barreiro, J. G. Bartlett *et al.* (Planck Collaboration), Planck 2015 results. XIII. Cosmological parameters, Astron. Astrophys. **594**, A13 (2016).

[51] J. Lin, Divergence measures based on the Shannon entropy, IEEE Trans. Inf. Theory **37**, 145 (1991).
[52] P. Virtanen *et al.*, SciPy 1.0–Fundamental algorithms for scientific computing in Python, Nat. Methods **17**, 261 (2020).
[53] I. M. Romero-Shaw, C. Talbot, S. Biscoveanu, V. D'Emilio, G. Ashton, C. P. L. Berry, S. Coughlin, S. Galaudage, C. Hoy, M. Hübner *et al.*, Bayesian inference for compact binary coalescences with Bilby: Validation and application to the first LIGO-Virgo gravitational-wave transient catalogue, Mon. Not. R. Astron. Soc. **499**, 3295 (2020).
[54] E. Mattia, A.-K. Malz, A. Dias, and G. Ashton, Scripts for: "Case studies with GPBilby of glitch-contaminated transient gravitational waves", Zenodo, 10.5281/zenodo.19386108 (2026).
[55] A. G. Abac *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), GWTC-4.0: Updating the gravitational-wave transient catalog with observations from the first part of the fourth LIGO-Virgo-KAGRA observing run, Astrophys. J. Lett. **1004**, L22 (2026).
[56] R. Abbott, H. Abe, F. Acernese, K. Ackley, N. Adhikari, R. X. Adhikari, V. K. Adkins, V. B. Adya, C. Affeldt, D. Agarwal *et al.*, Tests of general relativity with GWTC-3, Phys. Rev. D **112**, 084080 (2025).
[57] S. Soni, C. Austin, A. Effler, R. M. S. Schofield, G. González, V. V. Frolov, J. C. Driggers, A. Pele, A. L. Urban, G. Valdes *et al.*, Reducing scattered light in LIGO's third observing run, Classical Quantum Gravity **38**, 025016 (2021).
[58] R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, N. Adhikari, R. X. Adhikari, V. B. Adya, C. Affeldt, D. Agarwal *et al.*, GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run, Phys. Rev. X **13**, 041039 (2023).
[59] I. Mandel, What is the most massive gravitational-wave source?, Astrophys. J. **996**, L4 (2026).
[60] D. Croon, D. Gerosa, and J. Sakstein, Can GW231123 have a stellar origin?, Mon. Not. R. Astron. Soc. **546**, stag073 (2026).
[61] L. Passenger, S. Banagiri, E. Thrane, P. D. Lasky, A. Borchers, M. Fishbach, and C. S. Ye, Is GW231123 a hierarchical merger?, Astrophys. J. **999**, 236 (2026).
[62] C. Yuan, Z.-C. Chen, and L. Liu, GW231123 mass gap event and the primordial black hole scenario, Phys. Rev. D **112**, L081306 (2025).
[63] S. Goyal, H. Villarrubia-Rojo, and M. Zumalacarregui, Across the universe: GW231123 as a magnified and diffracted black hole merger, Astrophys. J. Lett. **1008**, L12 (2026).
[64] Q. Hu, H. Narola, J. Heynen, M. Wright, J. Veitch, J. Janquart, and C. Van Den Broeck, GW231123: Overlapping gravitational wave signals?, Astrophys. J. **1003**, 187 (2026).
[65] I. Cuceu, M. A. Bizouard, N. Christensen, and M. Sakellariadou, GW231123: Binary black hole merger or cosmic string?, Phys. Rev. D **113**, L021302 (2026).
[66] V. De Luca, G. Franciolini, and A. Riotto, GW231123: A possible primordial black hole origin, Phys. Rev. Lett. **136**, 201401 (2026).
[67] A. Jan, S. Nicolella, D. Shoemaker, and R. O'Shaughnessy, Measuring eccentricity and addressing waveform systematics in GW231123, Phys. Rev. D **113**, 084052 (2026).
[68] S. Bini, K. Król, K. Chatziioannou, and M. Isi, The impact of waveform systematics and Gaussian noise on the interpretation of GW231123, Phys. Rev. D **113**, 083036 (2026).
[69] C. Messick, K. Blackburn, P. Brady, P. Brockill, K. Cannon, R. Cariou, S. Caudill, S. J. Chamberlin, J. D. E. Creighton, R. Everett *et al.*, Analysis framework for the prompt discovery of compact binary mergers in gravitational-wave data, Phys. Rev. D **95**, 042001 (2017).
[70] S. Sachdev, S. Caudill, H. Fong, R. K. L. Lo, C. Messick, D. Mukherjee, R. Magee, L. Tsukada, K. Blackburn, P. Brady *et al.*, The GstLAL search analysis methods for compact binary mergers in Advanced LIGO's second and Advanced Virgo's first observing runs, arXiv:1901.08580.
[71] L. Tsukada, P. Joshi, S. Adhicary, R. George, A. Guimaraes, C. Hanna, R. Magee, A. Zimmerman, P. Baral, A. Baylor *et al.*, Improved ranking statistics of the GstLAL inspiral search for compact binary coalescences, Phys. Rev. D **108**, 043004 (2023).
[72] S. Sakon, L. Tsukada, H. Fong, J. Kennington, W. Niu, C. Hanna, S. Adhicary, P. Baral, A. Baylor, K. Cannon *et al.*, Template bank for compact binary mergers in the fourth observing run of Advanced LIGO, Advanced Virgo, and KAGRA, Phys. Rev. D **109**, 044066 (2024).
[73] P. Joshi, L. Tsukada, C. Hanna, S. Adhicary, D. Mukherjee, W. Niu, S. Sakon, D. Singh, P. Baral, A. Baylor *et al.*, New methods for offline GstLAL analyses, arXiv:2506.06497.
[74] S. Soni, B. K. Berger, D. Davis, F. Di Renzo, A. Effler, T. A. Ferreira, J. Glanzer, E. Goetz, G. González, A. Helmling-Cornell *et al.*, LIGO detector characterization in the first half of the fourth observing run, Classical Quantum Gravity **42**, 085016 (2025).
[75] D. J. Ottaway, P. Fritschel, and S. J. Waldman, Impact of upconverted scattered light on advanced interferometric gravitational wave detectors, Opt. Express **20**, 8329 (2012).
[76] S. A. Usman, J. C. Mills, and S. Fairhurst, Constraining the inclinations of binary mergers from gravitational-wave observations, Astrophys. J. **877**, 82 (2019).
[77] C. R. Harris *et al.*, Array programming with NumPy, Nature (London) **585**, 357 (2020).
[78] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright *et al.*, SciPy 1.0: Fundamental algorithms for scientific computing in Python, Nat. Methods **17**, 261 (2020).
[79] J. D. Hunter, Matplotlib: A 2D graphics environment, Comput. Sci. Eng. **9**, 90 (2007).
[80] https://gwosc.org.