Investigation of the effects of non-Gaussian noise transients and their mitigation in parameterized gravitational-wave tests of general relativity

Jack Y. L. Kwok[1,*] Rico K. L. Lo,[2] Alan J. Weinstein,[2] and Tjonnie G. F. Li[1,3,4]

[1]*Department of Physics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong*
[2]*LIGO, California Institute of Technology, Pasadena, California 91125, USA*
[3]*Institute for Theoretical Physics, KU Leuven, Celestijnenlaan 200D, B-3001 Leuven, Belgium*
[4]*Department of Electrical Engineering (ESAT), KU Leuven,*
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

The detection of gravitational waves from compact binary coalescence by Advanced LIGO and Advanced Virgo provides an opportunity to study the strong-field, highly relativistic regime of gravity. Gravitational-wave tests of general relativity (GR) typically assume Gaussian and stationary detector noise and, thus, do not account for non-Gaussian, transient noise features (glitches). We present the results obtained by performing parametrized gravitational-wave tests on simulated signals from binary-black-hole coalescence overlapped with three classes of frequently occurring instrumental glitches with distinctly different morphologies. We then review and apply three glitch mitigation methods and evaluate their effects on reducing false deviations from GR. By considering nine cases of glitches overlapping with simulated signals, we show that the short-duration, broadband blip and tomte glitches under consideration introduce false violations of GR, and using an inpainting filter and glitch model subtraction can consistently eliminate such false violations without introducing additional effects.

I. INTRODUCTION

Over a century after its formulation in 1915, Einstein's general relativity (GR) remains as the accepted theory of gravity, passing all precision tests to date [1]. In the weak-field, slow-motion regime, where the effects of metric theories of gravity can be approximated as higher-order *post-Newtonian* (PN) corrections to the Newtonian theory [2], GR lies within the stringent bounds set by solar-system tests and pulsar tests [3,4]. Recent attention has turned to testing GR in the strong-field, highly relativistic regime [3], which potentially suggests high-energy corrections to the Einstein-Hilbert action [5], making GR compatible with standard quantum field theory [1]. One approach of probing the strong-field regime is through the detection of gravitational waves (GWs), which carry information about their astrophysical origins [6].

Of all strong-field astrophysical events that could be probed using GWs, the *coalescence* of stellar-mass binary black holes (BBHs), which can be schematically divided into *inspiral*, *merger*, and *ringdown* (IMR) stages, plays a crucial role in testing GR [1]. Since the orbital separation of BBHs can reach far below the last stable orbit before merging, the generated gravitational field can be many orders of magnitudes stronger than other astrophysical

events observed so far [7–13]. Moreover, GWs emitted by coalescing BBHs offer one of the cleanest tests of GR, as matter and electromagnetic fields are negligible for most sources [8,14], and the emitted GWs essentially propagate through matter unimpeded [8], enabling precision tests of the strong-field dynamics of GR. Since 2015, Advanced LIGO [15] and Advanced Virgo [16] have jointly announced over 40 confident detections of GWs from coalescing BBHs [17–19].

Several GW tests of GR using coalescing BBHs are developed to test for *generic* deviations from GR without the need for signal models from competing theories of gravity [8]. For example, consistency tests search for excess power in the residual noise after subtracting a best-fit GR waveform [20] or compare the source parameters inferred using only high-frequency data to that inferred using only low-frequency data [20]; parametrized tests introduce parametrized deformations to waveform approximations to GR and infer the extent of deviation using Bayesian parameter estimation [9]. To date, no evidence for violations of GR has been identified using GWs emitted by coalescing BBHs [21,22].

Aside from GWs, output from GW detectors is attributed to many independent sources of random noise [23]. Assuming that noise characteristics remain stationary over observation timescales, detector noise is typically modeled as stationary and Gaussian in GW data analysis in light of
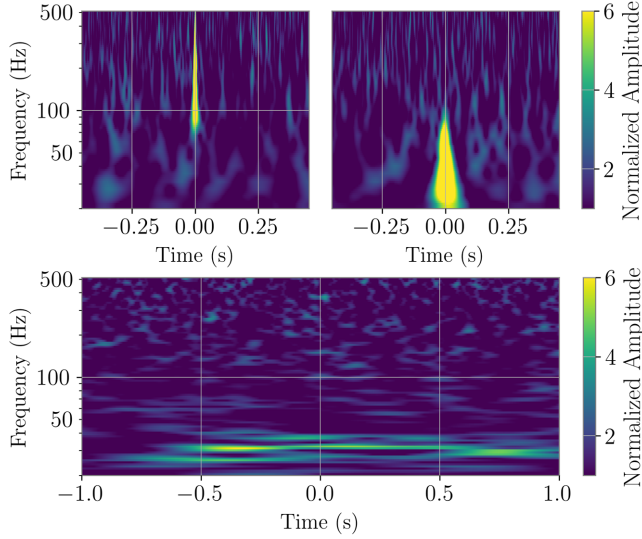
---

*jackkwok@link.cuhk.edu.hk

FIG. 1. Glitches with similar morphology can be algorithmically categorized into different classes [27]. A time-frequency representation, called a *Q scan* (or *Omega scan*) [29], where the duration of each time-frequency bins varies inversely with frequency and linearly with a parameter Q, is commonly used to visualize glitches [27,30]. Q scans of three frequently occurring glitches (top left, blip; top right, tomte; bottom, scattered light) during the O3 observing run are shown. The value of Q used is 8, 8, and 40, respectively. The color represents the normalized amplitude (square root of the normalized power) in each time-frequency bin.

the central limit theorem [24,25]. However, these assumptions cannot account for transient, non-Gaussian noise features, commonly referred to as *glitches* [26–28]. Glitches pose significant problems to GW searches [27] and may bias GW data analysis by violating the noise model. Three glitches from commonly seen glitch classes during the O3 observing run are shown in Fig. 1.

Many efforts are made to identify and classify glitches [27,30–34]. Once a glitch is identified, the data containing the glitch can be removed using various mitigation methods [35–39]. The effects of glitches and their mitigation on the inference of source parameters have been studied in the context of glitches similar to that affecting GW170817 [40]. It is of interest to extend the study to parametrized tests of GR, as the additional degree(s) of freedom introduced by parametrized deformations of the signal model may enhance such effects.

This article is structured as follows: Section II describes the typical data model used in GW data analyses [24,25], which comprises of a GW signal in additive stationary and Gaussian noise. Section III introduces a parametrized test of GR involving the phase parametrization of an IMR waveform model [41]. Section IV introduces three glitch mitigation methods to be applied in our investigation, namely, bandpass filtering, inpainting, and glitch model subtraction. Section V presents the results obtained by performing

the parametrized test of GR to glitch-overlapped BBH-coalescence GW signals before and after glitch mitigation.

II. DATA MODEL

A GW detector is designed to respond linearly to the fractional change in arm length, or *strain* [23]. The time series of detector output data $\boldsymbol{d}$, sampled at time t_k at constant sampling interval Δt, can thus be expressed as a linear superposition of a time series of the GW strain signal $\boldsymbol{h}$ and a time series of detector noise $\boldsymbol{n}$:

$$\boldsymbol{d}(t_k) = \boldsymbol{h}(t_k) + \boldsymbol{n}(t_k). \tag{1}$$

In Eq. (1) and in subsequent discussion, boldface denotes the matrix representation of specified quantities.

A. Stationary Gaussian noise model

Assuming that a *large* number of independent noise sources contribute linearly to the detector noise $\boldsymbol{n}$, the central limit theorem states that the probability density distribution of the noise $\boldsymbol{n}$ tends to follow a multivariate *Gaussian* distribution [42]:

$$P(\boldsymbol{n}) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{n}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{n}-\boldsymbol{\mu})}, \tag{2}$$

which is uniquely defined by the *covariance matrix* $\Sigma_{ij} = E[(n(t_i) - \mu(t_i))(n(t_j) - \mu(t_j))]$ and the mean vector $\mu_i = E[n(t_i)]$, where $E[\cdot]$ and $|\cdot|$ denote the expectation and determinant operation, respectively. The diagonal (off-diagonal) terms of the covariance matrix are the variances at each instance of time (correlations between data from different instances of time).

If the number of samples N is large, it is undesirable to invert the $N \times N$ covariance matrix in Eq. (2). Instead, we consider the joint probability density in the frequency domain, which is also a multivariate Gaussian distribution [42]. With the assumption of stationarity—i.e., the joint probability density distribution is time invariant—the covariance matrix in the frequency domain is diagonalized in the infinite-duration limit [43]. This relation can be approximated for the finite-duration discretely sampled time series, giving the following approximation to the joint probability density in the frequency domain [43] (for even N), also known as the *Whittle likelihood* [44] in the context of statistical inference:

$$P(\boldsymbol{n}) \simeq \prod_{j=0}^{N/2-1} \frac{2\Delta f}{\pi S_n(f_j)} \exp\left(-\Delta f \frac{2|\tilde{n}_j|^2}{S_n(f_j)}\right), \tag{3}$$

where $f_j \equiv j/N\Delta t$. The quantity $S_n(f_j) \equiv 2|\tilde{n}(f_j)|^2/T$ is scaled from the diagonal terms of the covariance matrix in the frequency domain, $\Delta f \equiv 1/T$ is the *frequency*

resolution, and the tilde denotes a discrete Fourier transformed (DFT) quantity:

$$\tilde{n}_j \equiv \Delta t \, \text{DFT}[n(t_k)] = \Delta t \sum_{k=0}^{N-1} n(t_k) e^{-2\pi i j k / N}. \quad (4)$$

To motivate the quantity $S_n(f_j)$, called the *one-sided power spectral density* (PSD), we invoke Parseval's theorem [43]:

$$\sum_{j=0}^{N/2-1} S_n(f_j) \Delta f \equiv \frac{2}{T} \sum_{j=0}^{N/2-1} |\tilde{n}(f_j)|^2 \Delta f = \frac{1}{T} \sum_{k=0}^{N-1} |n(t_k)|^2 \Delta t, \quad (5)$$

and note that the rightmost side of Eq. (5) returns the *power* of the time series. Since the time series is real, we have $\tilde{n}(f_j) = \tilde{n}^*(-f_j)$. Consequently, we can sample only the frequency bins from 0 Hz to up to the *Nyquist frequency* $1/2\Delta t$ and introduce the factor of 2 in Eqs. (3) and (5).

B. Signal model

Since the two-body self-gravitating problem cannot be solved analytically in GR, we generate simulated GW strain signals from coalescing BBHs using the frequency-domain precessing IMR waveform model IMRPhenomPv2 [41] in virtue of its good match with numerical relativity (NR) waveforms [45] and low computational costs.

IMRPhenomPv2 is a phenomenological waveform model constructed by combining PN-like inspiral waveforms with NR-calibrated merger-ringdown ansatz [46]. In natural units, the *inspiral* stage of IMRPhenomPv2 is modeled up to $f \sim 0.018/M$, where M is the total mass of the system. The region with $Mf \geq 0.018$ is subdivided into an *intermediate* stage with $0.018 \geq Mf \geq 0.5 f_{\text{RD}}$, which bridges the inspiral stage to the *merger-ringdown* stage modeled above half the ringdown frequency f_{RD} [46]. Figure 2 illustrates the stages of coalescence of an example IMRPhenomPv2 GW strain and its frequency evolution over time.

The phase of IMRPhenomPv2 consists of terms with known frequency dependence. The coefficients of these terms, denoted as the *phase coefficients* p_i, are the subjects of parametrized tests of GR to be discussed in Sec. III. The phase coefficients p_i and the orbital evolution of the BBH depend only on the masses and spin angular momentum vectors of the component black holes [45], denoted as the *intrinsic* parameters. The phase coefficients p_i can be categorized into three groups, depending on the stages of coalescence in which they predominantly assert their effect on [9,46] (i) the *inspiral* PN coefficients $\{\varphi_0, ..., \varphi_5, \varphi_{5l}, \varphi_6, \varphi_{6l}, \varphi_7\}$ and phenomenological coefficients $\{\sigma_0, ..., \sigma_4\}$; (ii) the *intermediate* phenomenological coefficients $\{\beta_0, ..., \beta_3\}$; and (iii) the
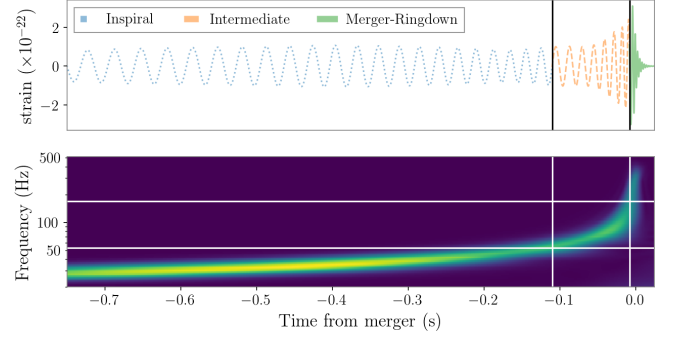

FIG. 2. An example IMRPhenomPv2 time-domain GW waveform (upper figure) and the corresponding instantaneous frequency (lower figure) plotted against time. In natural units, the two horizontal lines in the lower figure correspond to the frequencies $0.018/M$ (lower line) and $f_{\text{RD}}/2$ (upper line), which defines the boundaries of the inspiral, intermediate, and merger-ringdown stages in the frequency domain. The corresponding time-domain boundaries (vertical lines) are determined as the times when the instantaneous frequency of the signal intersects with the frequency-domain boundaries.

merger-ringdown phenomenological and black hole perturbation theory coefficients $\{\alpha_0, ..., \alpha_5\}$.

Seven additional *extrinsic* parameters, including the sky location, luminosity distance, polarization angle of the source, and the spatial orientation and orbital phase of the system at a reference frequency and time, respectively, are needed to determine the response of the GW detectors.

III. PARAMETRIZED TESTS OF GR

We will focus on a test of GR which tests for parametrized deviations from GR. It assumes the stationary Gaussian noise model. As such, this test provides quantitative indicators of whether glitches can result in false deviations of GR and whether glitch mitigation will reduce or amplify them.

In the test, *fractional* deviations δp_i, also known as *dephasing coefficients*, are introduced to the IMRPhenomPv2 phase coefficients p_i [9]:

$$p_i \mapsto p_i[1 + \delta p_i]. \quad (6)$$

For the exceptional case where $p_i = 0$, such as φ_1, an *absolute* deviation is instead introduced [9]. In practice, we do not allow some of the IMRPhenomPv2 phase coefficients to deviate from their prescribed values, as they have large uncertainties or are degenerate with other coefficients or physical parameters [9]. We therefore perform tests with the remaining 14 dephasing coefficients, henceforth denoted as the *testing* dephasing coefficients [9]:

$$\{\delta p_i\} = \{\delta\varphi_0, ..., \delta\varphi_4, \delta\varphi_{5l}, \delta\varphi_6, \delta\varphi_{6l}, \delta\varphi_7,$$
$$\delta\beta_2, \delta\beta_3, \delta\alpha_2, \delta\alpha_3, \delta\alpha_4\}. \quad (7)$$

TABLE I. The frequency dependence of the IMRPhenomPv2 testing parameters used in parametrized tests of GR. The table is reproduced from Table 1 of Ref. [20]. The coefficients a and b in the f dependence of $\delta\alpha_4$ are functions of the component masses and spins [46].

Stage of coalescence	δp_i	f dependence
Inspiral	$\delta\varphi_0$	$f^{-5/3}$
	$\delta\varphi_1$	$f^{-4/3}$
	$\delta\varphi_2$	f^{-1}
	$\delta\varphi_3$	$f^{-2/3}$
	$\delta\varphi_4$	$f^{-1/3}$
	$\delta\varphi_{5l}$	$\log f$
	$\delta\varphi_6$	$f^{1/3}$
	$\delta\varphi_{6l}$	$f^{1/3}\log f$
	$\delta\varphi_7$	$f^{2/3}$
Intermediate	$\delta\beta_2$	$\log f$
	$\delta\beta_3$	f^{-3}
Merger-ringdown	$\delta\alpha_2$	f^{-1}
	$\delta\alpha_3$	$f^{3/4}$
	$\delta\alpha_4$	$\tan^{-1}(af+b)$

The frequency dependence of the testing parameters δp_i is shown in Table I [20,47].

To quantify a deviation from GR, we can infer the most probable values of δp_i through Bayesian parameter estimation, as discussed in the following subsection.

A. Parameter estimation

Recall our data model $d = h + n$. We denote $\theta(\theta, \delta p_i)$ as the parameter vector generating the signal h. It consists of the parameters θ generating the IMRPhenomPv2 waveform and the testing parameters δp_i generating the phase deviations from the IMRPhenomPv2 waveform. In practice, the testing parameters are introduced *one at a time*, which is expected to capture a deviation from GR present in multiple phase coefficients while returning narrower credible intervals than introducing multiple coefficients at a time [20].

Given the detector output d and prior information I, we wish to infer the conditional probability density of θ, referred to as the *posterior*, by invoking Bayes' theorem

$$P(\theta|d, I) = \frac{P(d|\theta, I) \times P(\theta|I)}{P(d|I)}, \qquad (8)$$

which relates the posterior to three probability densities: the *likelihood* $P(d|\theta, I)$, the *prior* $P(\theta|I)$, and the *evidence* $P(d|I)$. During parameter estimation, the evidence, which does not depend explicitly on θ, can be seen as a proportionality constant, since d and I are kept fixed. The likelihood and prior are separately discussed below.

Given $h(\theta)$, the time series of the output data d uniquely defines a time series of the residual noise $d - h$, which is modeled as Gaussian and stationary. As such, the

likelihood is approximated by the Whittle likelihood in Eq. (3):

$$P(d|\theta, I) \propto \exp\left[-\frac{1}{2}(d - h|d - h)\right], \qquad (9)$$

where $(\cdot|\cdot)$ is the *noise-weighted inner product* [48]:

$$(a|b) \equiv \sum_{j=0}^{N/2-1} 4\mathrm{Re}\left(\frac{\tilde{a}_j^* \tilde{b}_j}{S_n(f_j)}\right)\Delta f. \qquad (10)$$

Assuming that noise from multiple detectors, indexed l, is uncorrelated, the joint likelihood takes the form

$$P(d|\theta, I) \propto \exp\left[-\frac{1}{2}\sum_l (d_l - h_l|d_l - h_l)\right]. \qquad (11)$$

The prior $P(\theta|I)$ incorporates our beliefs about θ prior to the observation. We follow the default choice of prior in LALInference [25], which includes uniform priors for the component masses m_1 and m_2, with $m_2 \leq m_1$, a log-uniform prior for the luminosity distance, an isotropic prior for the sky location of the source and the spin angular momentum vectors of the component black holes, and uniform priors for the remaining parameters. We note that, in LALInference, the uniform priors specified for component masses are transformed to nonuniform, correlated priors for the chirp mass $\mathcal{M} \equiv (m_1 m_2)^{3/5}(m_1 + m_2)^{-1/5}$ and the mass ratio $q \equiv m_2/m_1$ for more efficient sampling [25].

In parametrized tests of GR, parameters of primary interest are the testing parameters δp_i, while the posterior distribution spans the full 16-dimensional parameter space. We therefore compute the *marginalized* posterior distribution for introduced the testing parameter δp_i:

$$P(\delta p_i|d, I) = \int P(\theta|d, I)d\theta, \qquad (12)$$

where θ denotes the parameters generating the underlying IMRPhenomPv2 waveform.

IV. GLITCH MITIGATION METHODS

In this section, we review four methods that can be applied to mitigate data containing glitches, in which three are used in our investigation, including a frequency-domain filtering method of bandpass filtering, a time-domain filtering method of inpainting, and a glitch model subtraction method using the BayesWave algorithm.

A. Bandpass filtering in frequency domain

Assuming stationary and Gaussian noise, components of the noise-weighted inner product from different frequency bins of equal bandwidth and from different detectors contribute linearly to the log likelihood, as seen from

Eq. (11). A direct way of removing the glitch in the frequency domain is by excluding the frequency bins containing the glitch from the likelihood calculation. In LALInference, this can be done by specifying the high-pass and low-pass cutoff frequency for the affected detector such that data containing the glitch are filtered out. Only the passed frequency bins are considered in the likelihood calculation. By default, data are high passed at 20 Hz in LALInference [25].

B. Gating and inpainting in time domain

A similar procedure can be done in the time domain, commonly known as *gating*, in which data containing the glitch are zeroed out by multiplying an inverse window function. The inverse window function reduces the spectral leakage in the frequency domain due to discontinuity of data at the boundary of the region to be zeroed out [49].

Gating was adopted in the mitigation of the glitch-overlapped GW170817 signal in LIGO-Livingston during the rapid localization of the source [50], which successfully led to follow-up electromagnetic observations [51]. However, gating was not used for parameter estimation purposes for the first half of the O3 observing run (O3a) [52]. There are concerns over mitigating glitches by gating. For example, as remarked in Ref. [40], gating can introduce errors to parametrized tests of GR, as it affects the signal power in frequency bins that count toward the noise-weighted inner product.

A new method, called *inpainting* or *hole filling* [39], is developed to address the noise artifacts and statistical bias that may result from gating. After specifying the time interval to be mitigated, new values are assigned for data within the interval, or *hole*, according to an *inpainting filter*, while data outside the hole are unaffected. The inpainting filter depends on the PSD of the stationary Gaussian noise. For inpainted data d_{inp}, the quantity $(d_{\mathrm{inp}}|h)$ is, by design of the filter, independent of the template waveform h inside the interval, and d_{inp} within the hole is identically zero upon twice whitening by the same PSD [39]. Since the hole can be made arbitrarily narrow, inpainting affects the minimal amount of data if the glitch is localized in time.

Reexpressing the noise-weighted inner product in the likelihood calculation,

$$P(d_{\mathrm{inp}}|h) \propto \exp\left[-\frac{1}{2}(d_{\mathrm{inp}} - h|d_{\mathrm{inp}} - h)\right]$$
$$= \exp\left[-\frac{1}{2}(d_{\mathrm{inp}}|d_{\mathrm{inp}}) + (d_{\mathrm{inp}}|h) - \frac{1}{2}(h|h)\right]. \quad (13)$$

Given inpainted data d_{inp}, only the terms $(d_{\mathrm{inp}}|h)$ and $(h|h)$ differ across waveform templates h; between these two terms, only $(d_{\mathrm{inp}}|h)$ explicitly depends on the inpainted data. As $(d_{\mathrm{inp}}|h)$ is independent of the template waveform

inside the hole by design of the filter, inpainted data inside the hole are not expected to contribute to the outcome of parametrized tests.

C. Glitch model subtraction

The BayesWave [37,38] algorithm models the GW signal and glitches in each detector using a variable number of wavelets, such as sine-Gaussian wavelets. Mitigated, or *deglitched*, data are generated by subtracting the glitch model from the original data. Using Bayesian inference, the output data in each detector are modeled as a superposition of a GW signal h, stationary Gaussian noise n_G, and glitches g:

$$d = h + n_G + g. \quad (14)$$

While both the GW signal and glitches are nonstationary and non-Gaussian, coherent features across data from multiple detectors are modeled by the signal model and independent features are modeled by the glitch model [53]. A transdimensional reversible jump Markov chain Monte Carlo algorithm is used to sample models with different number of wavelets or with wavelets of different parameter values. The most probable model is inferred through Bayesian inference by comparing the evidence $P(d|M_i, I)$ for different models M_i: Given data d and prior information I, we define the *odds* O_2^1 between two competing models M_1 and M_2 as

$$O_2^1 \equiv \frac{P(M_1|d, I)}{P(M_2|d, I)} = \frac{P(M_1|I)}{P(M_2|I)} \times \frac{P(d|M_1, I)}{P(d|M_2, I)}, \quad (15)$$

where the equality on the right is obtained by invoking Bayes' theorem. The model M_1 will be more probable than model M_2 if the odds O_2^1 are larger than 1. To express our ignorance toward the probability of models prior to observation, we can set the first term on the rightmost in Eq. (15), called the *prior odds*, to unity. The odds can then be obtained by comparing the evidences of the two models. In BayesWave, the evidences are calculated through thermodynamic integration [37]. Once the most probable glitch + signal model is inferred, the glitch model is subtracted from the data.

The BayesWave algorithm was first used to remove the glitch which overlapped with the GW170817 signal during parameter estimation [50] and was regularly used to mitigate glitch-overlapped signals during O3a [18]. Reference [40] concluded that parameter recovery results using data reconstructed by BayesWave are unbiased. In the context of tests of GR, which are designed to detect small deviations from GR waveforms, the subtraction of sine-Gaussian wavelets by BayesWave may alter the GW signal to an extent which may be reported as a false violation of GR. This is not observed in our results.
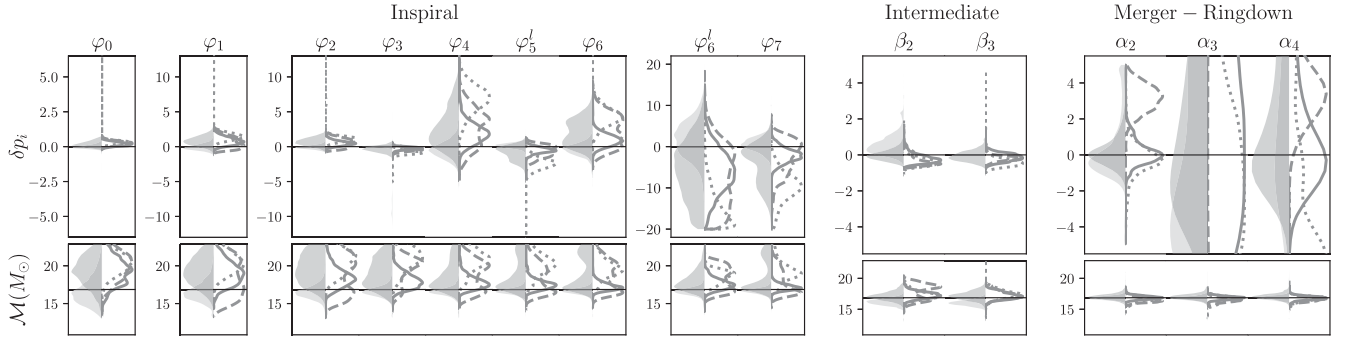
FIG. 3.   Posterior distributions of testing parameters (top) and recovered chirp mass (bottom) obtained by performing parametrized tests of GR on five data realizations of a simulated GW190828_065509-like signal in stationary Gaussian noise. Each shading and line style (left, light-gray shade and dark-gray shade; right, solid, dotted, and dashed line) represents one data realization. The simulated noise is colored by the representative best LIGO-Hanford, LIGO-Livingston, and Virgo detector PSD estimates during O3a.

V. RESULTS OF GLITCHES OVERLAPPING A GW190828_065509-LIKE SIGNAL

We are motivated to consider a signal similar to that of the high-mass-ratio BBH-merger event GW190828_065509 [18], in which the mitigation of potential glitches overlapping the event in L1 through bandpass filtering resulted in pathological features in parametrized tests of GR [54]. Values of some selected generating parameters of the GW190828_065509-like signal are tabulated in Table III.

We first present the expected results of parametrized tests of GR in the absence of glitches by coherently injecting the simulated GW190828_065509-like signal, generated with an IMRPhenomPv2 waveform model, into five realizations of simulated stationary, Gaussian noise colored with the representative best (cleaned) PSD of the LIGO-Hanford (H1), LIGO-Livingston (L1), and Virgo (V1) detectors during O3a. The posterior distributions of the testing parameters and the recovered chirp mass are plotted in Fig. 3.

Beneath the posteriors of testing parameters, the posteriors of recovered chirp mass are also plotted as an indicator of the sampling performance. With an extra degree of freedom introduced by inspiral PN testing parameters, broad and occasional multimodal distributions can be observed due to strong correlations between the inspiral PN testing parameters and chirp mass, as demonstrated in Figs. 4(a) and 4(b). Such degeneracies could also bring the one-dimensional marginalized distributions of the inspiral PN testing parameters away from zero, e.g., for the dotted distributions in Figs. 4(a) and 4(b), where the GR value of 0 is excluded at 90% credibility. These exclusions of 0 should not be counted toward a violation of GR, as the bias introduced by sampling can be clearly identified.

Correlations between inspiral PN testing parameters and mass parameters can be expected from theoretical grounds, as $\delta\varphi_j$ contain the term proportional to $\eta^{-1}M^{(j-5)/3}$, where η and M are the symmetric mass ratio and total mass, respectively, with $\delta\varphi_0$ being completely degenerate with chirp mass. The degeneracy of inspiral PN testing
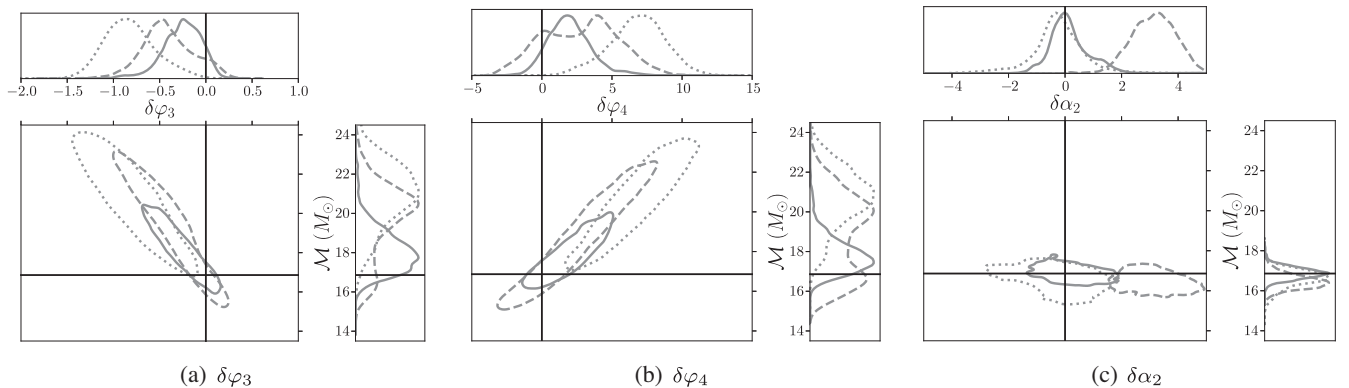

FIG. 4.   Marginalized distributions of testing parameters (left, $\delta\varphi_3$; middle, $\delta\varphi_4$; right, $\delta\alpha_2$) and chirp mass. The contours of the two-dimensional distributions show 90% credible regions. Each line style represents the data realization of the simulated GW190828_065509-like signal in stationary Gaussian noise with the corresponding line style in Fig. 3. The vertical and horizontal black lines denote the GR value of the testing parameters and the injected value of chirp mass, respectively.

TABLE II.   Key specifications of the three mitigation methods.

Mitigation method	Specification	Blip	Tomte	Scattered light
Bandpass	High-pass cutoff (Hz)	20	105	40
	Low-pass cutoff (Hz)	60	511.875	511.875
Inpainting	Hole duration (s)	0.005	0.040	$\cdots$*
	Sampling rate (Hz)	4096	4096	$\cdots$*
Glitch model subtraction	Segment length (s)	4	4	8
	High-pass cutoff (Hz)	20	20	8
	Sampling rate (Hz)	2048	2048	2048
	Q_{max}	40	40	200
	D_{max}	100	100	200

*Inpainting is replaced by discarding data from the detector in which the scattered-light glitch is present.

parameters with the chirp mass used during sampling may be amplified by the weakness of the signal, as we are injecting a weak signal [55].

We, however, note that the exclusion of 0 in simulated Gaussian noise of the phenomenological merger-ringdown parameter $\delta\alpha_2$ for one considered data realization cannot be explained in a similar way, as a degeneracy with chirp mass cannot be observed in Fig. 4(c). The source of this bias is not identified in the present study, but it may be due to the weakness of the injected signal, as similar anomalies arise when substantial signal power is discarded in Fig. 9.

The same signal is then injected into real data from three detectors (H1, L1, and V1) at times when all three detectors are operating in the science mode and with glitches present in either H1 or L1 [56]. Glitches from the *blip*, *tomte*, and *scattered-light* classes are chosen, as these classes of glitches have the highest occurrence rates in O3a [58]. The glitches used in our study are further chosen so that their duration and peak frequency are representative of their corresponding glitch classes. The chosen blip, tomte, and scattered-light glitch are present at GPS time around 1253103382, 1252901859, and 1253416025 at H1, L1, and H1, respectively.

The GW190828_065509-like signal is injected coherently into the three detectors such that each glitch overlaps with the signal at the inspiral, intermediate, and merger-ringdown stage in the time domain for different data samples. The three stages in the time domain are defined as the time intervals when the instantaneous frequencies of the signal are in the corresponding three stages in frequency domain discussed in Sec. II B, respectively. The boundaries of the three stages of the signal in the time and frequency domain are marked in the Q scans by vertical and horizontal white lines, respectively.

After preparing the data samples, we applied the glitch mitigation methods of bandpass filtering, inpainting, and BayesWave glitch subtraction as described in Sec. IV on detector data in which glitches are present. We then performed parametrized tests of GR on the unmitigated and mitigated samples. The specifications of the three

glitch mitigation methods are tabulated in Table II. We referred to Ref. [59] for BayesWave specifications, whereas bandpass cutoff frequencies and inpainting hole duration are chosen to exclude time-frequency bins affected by the glitch.

A. Blip glitch

Blip glitches are short-duration, broadband glitches characterized by their teardrop shape as seen in time-frequency representations. A Q scan of a blip glitch is shown on the top left in Fig. 1 [27]. The sources and coupling of blip glitches are not well understood [18].

The simulated GW190828_065509-like signal is coherently injected into H1, L1, and V1 in a way that a blip glitch at H1 overlaps with the signal at the inspiral, intermediate, and merger-ringdown stages in the time domain. Mitigation methods are applied to H1 data, and parametrized tests of GR are performed on the unmitigated and mitigated data. The posteriors of the testing parameters are plotted on the left and right side of each violin plot in Fig. 5, respectively,

TABLE III.   Injected values of some selected generating parameters of a GW190828_065509-like signal using the IMRPhenomPv2 waveform model. The GW190828_065509-like signal is taken to be the *maximum likelihood* waveform inferred for real GW190828_065509 data using the IMRPhenomPv2 template waveform model. Despite the large injected values for the component spins, the inferred posterior distributions of the component spins are flat throughout the prior range for the GW190828_065509 and simulated GW190828_065509-like signals.

Waveform parameter	Value
Chirp mass $\mathcal{M}$ ($M_\odot$)	16.86
Mass ratio q	0.14
Dimensionless primary spin magnitude a_1	0.92
Dimensionless secondary spin magnitude a_2	0.75
Right ascension α (rad)	2.54
Declination δ (rad)	−0.84
Luminosity distance D_L (Mpc)	1021

(a) Simulated GW190828_065509-like signal overlapped with a H1 blip glitch at inspiral stage in the time domain.

(b) Simulated GW190828_065509-like signal overlapped with a H1 blip glitch at intermediate stage in the time domain.

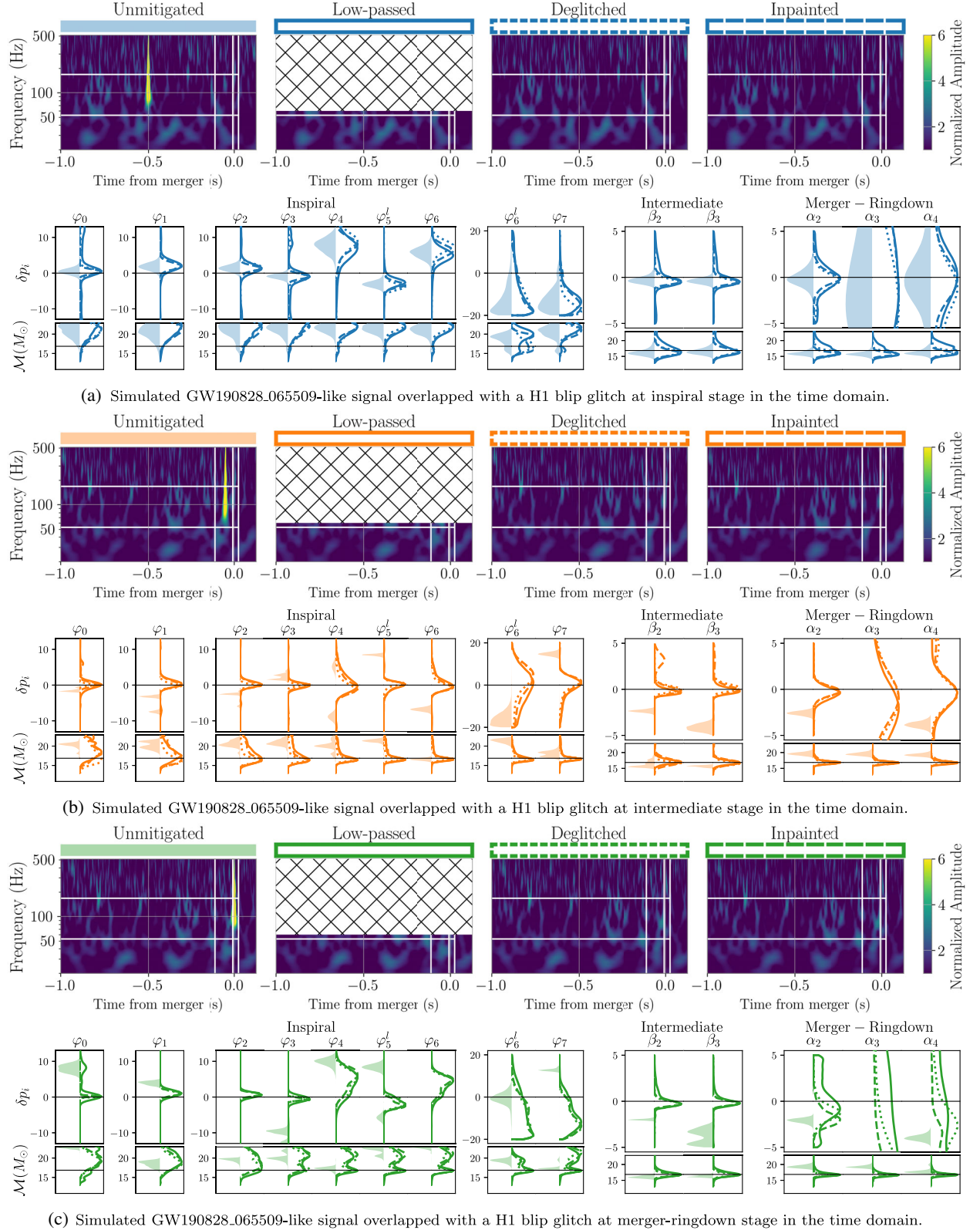(c) Simulated GW190828_065509-like signal overlapped with a H1 blip glitch at merger-ringdown stage in the time domain.

FIG. 5. Top subfigures: Q scans of the unmitigated and mitigated data samples. Vertical and horizontal white lines denote the boundaries of different stages of coalescence in the time and frequency domain, respectively. Bottom subfigures: Posterior distributions of testing parameters (top) and the recovered chirp mass (bottom) obtained by performing parametrized tests on unmitigated (left of violin plot) blip-glitch-overlapped signals during a three-detector observation. The corresponding mitigated cases (right of violin plot) with bandpass filtering (solid line), BayesWave glitch model subtraction, also called *deglitching* (dotted line), and inpainting (dashed line) are also plotted. The GR value of the testing parameters and the injected value of chirp mass are indicated by vertical black lines.
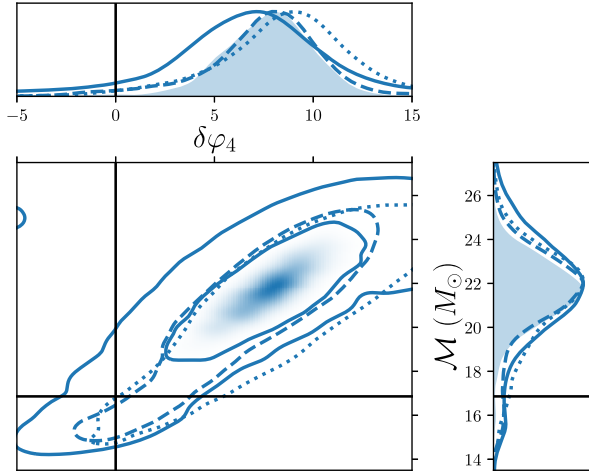
FIG. 6. Marginalized distributions of $\delta\varphi_4$ and chirp mass. The contours of the two-dimensional distributions show 90% credible regions. The distributions are obtained by performing parametrized tests on unmitigated (shaded), bandpassed (solid line), glitch-subtracted (dotted line), and inpainted (dashed line) blip-glitch-overlapped signal as shown in Fig. 5(a). The vertical and horizontal black lines denote the GR value of the testing parameter and the injected value of chirp mass, respectively.

while the unmitigated and mitigated data from H1 are represented by Q scans.

The stages of coalescence where violations of GR are observed show no correlation with those overlapped by the glitch in the time or frequency domain: Violations of GR can be observed for testing parameters from all stages of coalescence when the blip glitch overlaps with the signal in the intermediate or merger-ringdown stage in the time domain, even though the blip glitch contributes excess power only to intermediate and merger-ringdown frequency bands.

No observable effects on parametrized tests of GR are observed when the blip glitch temporally overlaps the signal at the inspiral stage, suggested by the matching posterior distributions without and with the glitch removed through independent methods of low passing to 60 Hz, deglitching, and inpainting, though the GR value of 0 is located only at the far tails of distributions for the inspiral PN testing parameters due to the degeneracy with chirp mass, as demonstrated in Fig. 6.

Comparing the unmitigated and mitigated results of the glitch overlapping the intermediate and merger-ringdown stage in the time domain, all three mitigation methods of low passing, inpainting, and deglitching can reduce false violations of GR by bringing posteriors of testing parameters from distributions that exclude the GR value of 0 at 90% credibility to one that peaks close to 0 [e.g., $\delta\varphi_3$, $\delta\varphi_4$, $\delta\beta_2$, $\delta\beta_3$, and $\delta\alpha_2$ in Fig. 5(c)]. The posterior distributions of the testing parameters for mitigated samples match each other closely, indicating that the mitigation methods did not contribute extra effects on parametrized tests of GR in these

three cases. Significant improvements in parametrized tests of GR after removal of the blip glitch suggest that false violations are attributed to the presence of the glitch.

Upon introduction of an extra degree of freedom by the parametrized deviations, the blip glitch leads to bias in other intrinsic variables such as the chirp mass. This is reflected by the sharp peaks in the chirp mass posteriors which exclude the injected value at 90% credibility [e.g., $\delta\alpha_2$, $\delta\alpha_3$, and $\delta\alpha_4$ in Fig. 5(b) and $\delta\varphi_2$, $\delta\varphi_{5l}$, and $\delta\varphi_6$ in Fig. 5(c)]. The independent glitch mitigation methods bring the peak values close to the injected value, although multimodal and broad distributions for chirp mass can still be noted after mitigation due to its degeneracy with inspiral PN testing parameters.

We quote representative figures of the network matched-filter signal-to-noise ratio (henceforth SNR) for unmitigated and mitigated samples, recovered with IMRPhenomPv2 (without introducing parametrized deviations) using data samples in Fig. 5(c), where the simulated signal overlaps with the glitch at the merger-ringdown stage in the time domain. The SNR are $20.76^{+0.18}_{-0.28}$ before mitigation, $8.26^{+0.41}_{-1.07}$ after low passing, $9.67^{+0.27}_{-0.53}$ after glitch subtraction, and $9.93^{+0.26}_{-0.47}$ after inpainting. Symmetric 90% credible intervals are denoted by the subscripts and superscripts.

B. Tomte glitch

Tomte glitches are short-duration, broadband glitches characterized by their triangular shape as seen in time-frequency representations. A Q scan of a tomte glitch is shown on the top right in Fig. 1. The sources and coupling of tomte glitches are not well understood.

The simulated GW190828_065509-like signal is coherently injected into H1, L1, and V1 in a way that a tomte
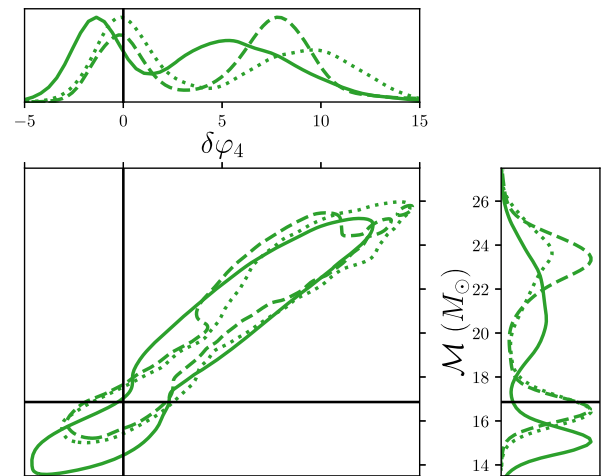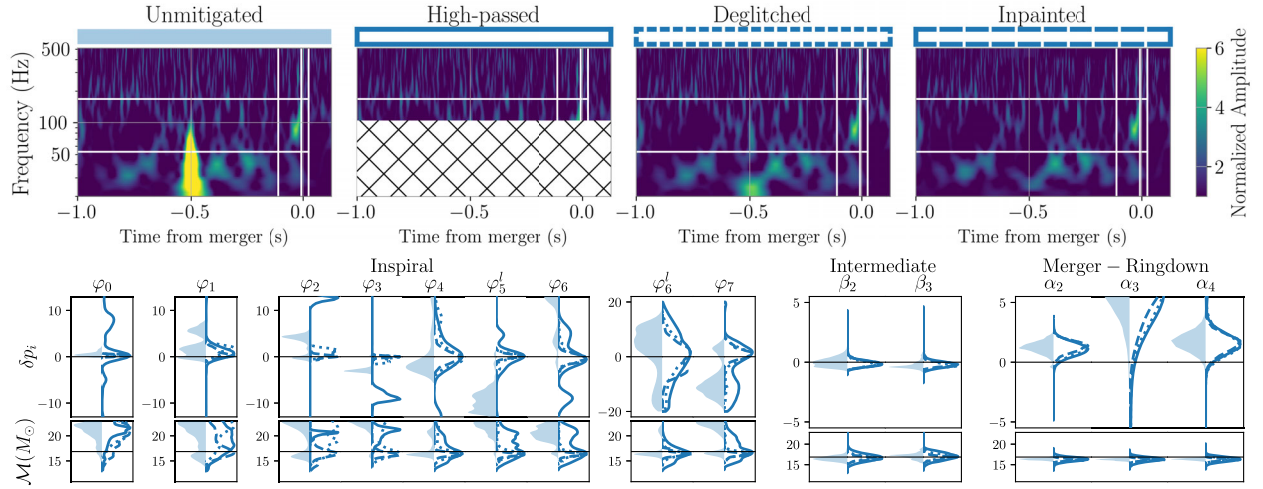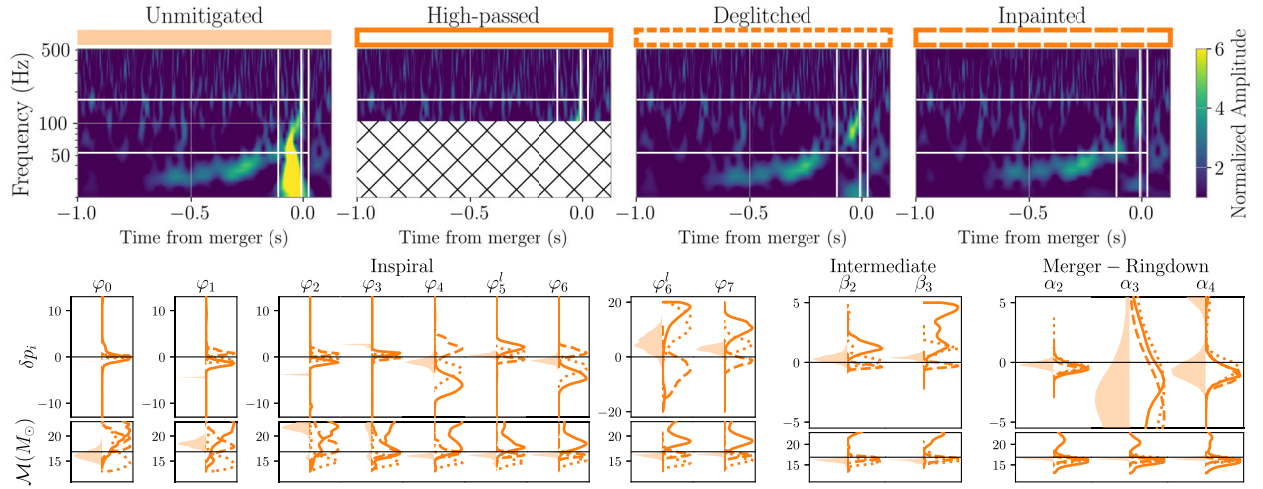

FIG. 7. The same as Fig. 6, but for the tomte-glitch-overlapped signal as shown in Fig. 8(c). The unmitigated distributions are out of the boundaries of the figure.

(a) Simulated GW190828_065509-like signal overlapped with a L1 tomte glitch at inspiral stage in the time domain.

(b) Simulated GW190828_065509-like signal overlapped with a L1 tomte glitch at intermediate stage in the time domain.
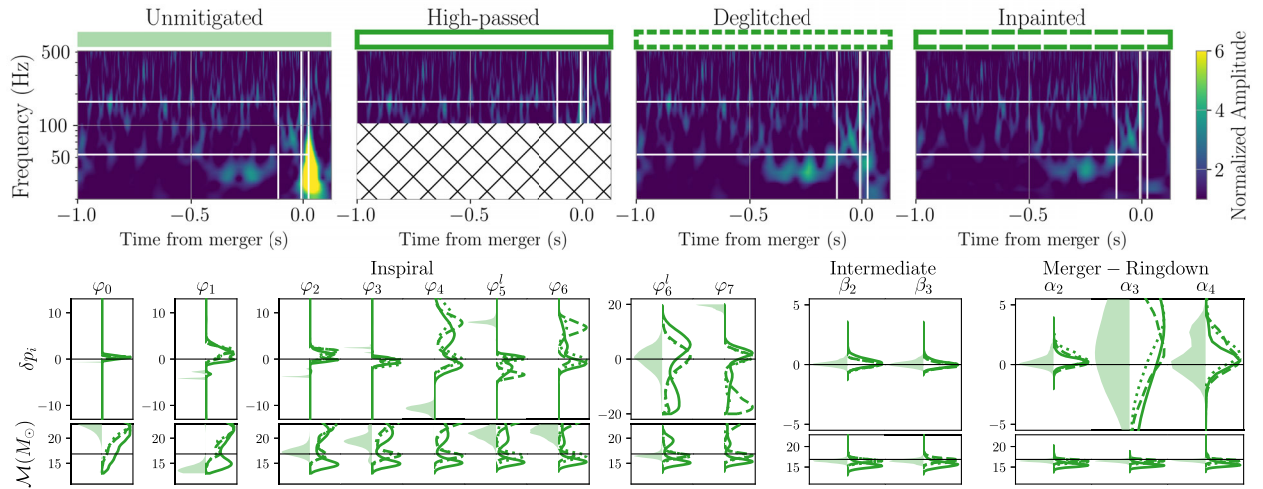
(c) Simulated GW190828_065509-like signal overlapped with a L1 tomte glitch at merger-ringdown stage in the time domain.

FIG. 8.   Similar to Fig. 5, but for tomte-glitch-overlapped signals.

glitch at L1 overlaps with the signal at the inspiral, intermediate, and merger-ringdown stages in the time domain. Mitigation methods are applied to L1 data, and

parametrized tests of GR are performed on the unmitigated and mitigated data. The posteriors of the testing parameters are plotted on the left and right side of each violin plot in

Fig. 8, respectively, while the unmitigated and mitigated data from L1 are represented by Q scans.

The stages of coalescence where violations of GR are observed show no correlation with those overlapped by the glitch in the time or frequency domain: Despite shifting the time of overlap of the signal with the tomte glitch, which contributes considerable excess power also to the intermediate frequency bands, false violations of GR are observed only in inspiral testing parameters for the unmitigated data samples (left of violins). For example, exclusions of the GR value of 0 at 90% credibility are observed in lower PN orders such as $\delta\varphi_2$, $\delta\varphi_3$, and $\delta\varphi_4$.

Again, we can observe bias in the inference of chirp mass when the tomte glitch overlaps the simulated signal, such that the posterior distributions peak far away and exclude the injected value at 90% credibility [e.g., $\delta\varphi_3$ in Fig. 8(a), $\delta\varphi_2$ in Fig. 8(b), and $\delta\varphi_4$ in Fig. 8(c)]. The independent glitch mitigation methods bring the peak values of the chirp mass close to the injected value, yet multimodal features can still be observed for both chirp mass and inspiral PN parameters due to the degeneracies between them, as demonstrated in Fig. 7.

Comparing the unmitigated and mitigated results, both inpainting and glitch model subtraction can reduce the false violations in the lower PN order testing parameters, resulting in strong support for the GR value of 0 in most testing parameters. Improvements in parametrized tests of GR after removal of the glitch by inpainting and glitch model subtraction suggest that the false violations in the inspiral parameters are attributed to the presence of the tomte glitch, which contributes significant excess power in inspiral frequency bands.

Meanwhile, high passing up to 105 Hz is not a robust glitch mitigation method, as false deviations of GR can be amplified [e.g., $\delta\varphi_2$ and $\delta\varphi_3$ in Fig. 8(a)] or introduced by the mitigation [e.g., $\delta\beta_2$ and $\delta\beta_3$ in Fig. 8(b)]. Figure 9 shows an example of increasing the high-pass cutoff from the default specifications of 20 Hz for unmitigated samples by increments of 25 or 35 Hz up to 130 Hz and compares the distributions with that obtained by only using data from H1 and V1. It is found that, as more and more signal (glitch) power is discarded by high passing, the distributions converge to a peak away from the injected value (and away from the biased distributions due to the glitch). It is then found that using only H1 and V1 data in the absence of glitches would recover this bias. This indicates that further reduction of signal power from the weak signal can lead to bias. While the source of this bias, or its relationship with the bias observed in Fig. 4(c) for $\delta\alpha_2$ in simulated Gaussian noise, are unidentified, this anomalous effect can be easily decoupled from the effects due to glitches by performing independent glitch mitigation methods as illustrated in Figs. 8 and 9.

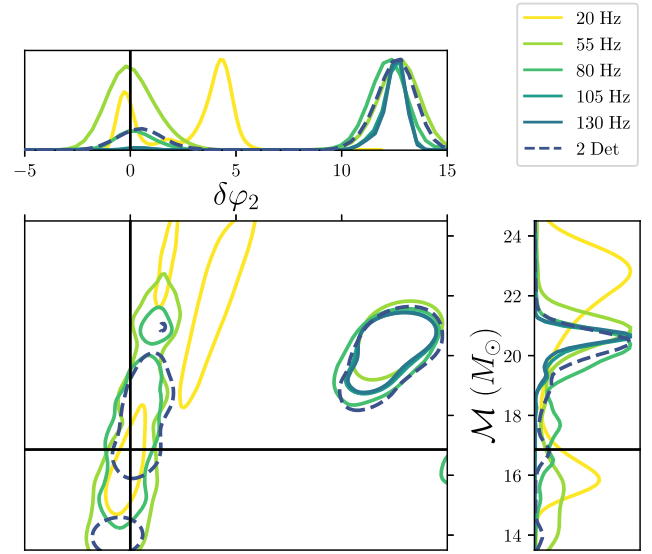Representative figures of the SNR for unmitigated and mitigated samples are obtained in the same way as in

FIG. 9. Marginalized distributions of $\delta\varphi_2$ and chirp mass. The contours of the two-dimensional distributions show 90% credible regions. Solid curves with different colors represent different distributions obtained by high passing the tomte-glitch-overlapped data in L1 shown in Fig. 5(a) to different frequencies. The dotted curves represent distributions obtained by discarding data from L1 and performing two-detector observation with H1 and V1 data only. The vertical and horizontal black lines denote the GR value of the testing parameter and the injected value of chirp mass, respectively.

Sec. V A and are $13.93^{+0.22}_{-0.36}$ before mitigation, $12.91^{+0.25}_{-0.43}$ after high passing, $13.51^{+0.27}_{-0.36}$ after glitch subtraction, and $13.99^{+0.23}_{-0.38}$ after inpainting.

C. Scattered-light glitch

Scattered-light glitches are produced by laser light scattering out and reentering the main laser beam, and their correlation with seismic motion is well understood [18]. Scattered-light glitches are characterized by their arch shape as seen in a time-frequency representation such as the bottom Q scan in Fig. 1 [27].

The simulated GW190828_065509-like signal is coherently injected into H1, L1, and V1 in a way that a scattered-light glitch at H1 overlaps with the signal at the inspiral, inspiral-intermediate-merger-ringdown, and merger-ringdown stages in the time domain. Mitigation methods are applied to H1 data, and parametrized tests of GR are performed on the unmitigated and mitigated data. The posteriors of the testing parameters are plotted on the left and right side of each violin plot in Fig. 10, respectively, while the unmitigated and mitigated data from H1 are represented by Q scans.

Since the typical timescale of scattered-light glitches (>1 s) is large compared to that of BBH-coalescence GW signals (in the LIGO band), it is likely for an overlapping scattered-light glitch to overlap with the entire GW signal.
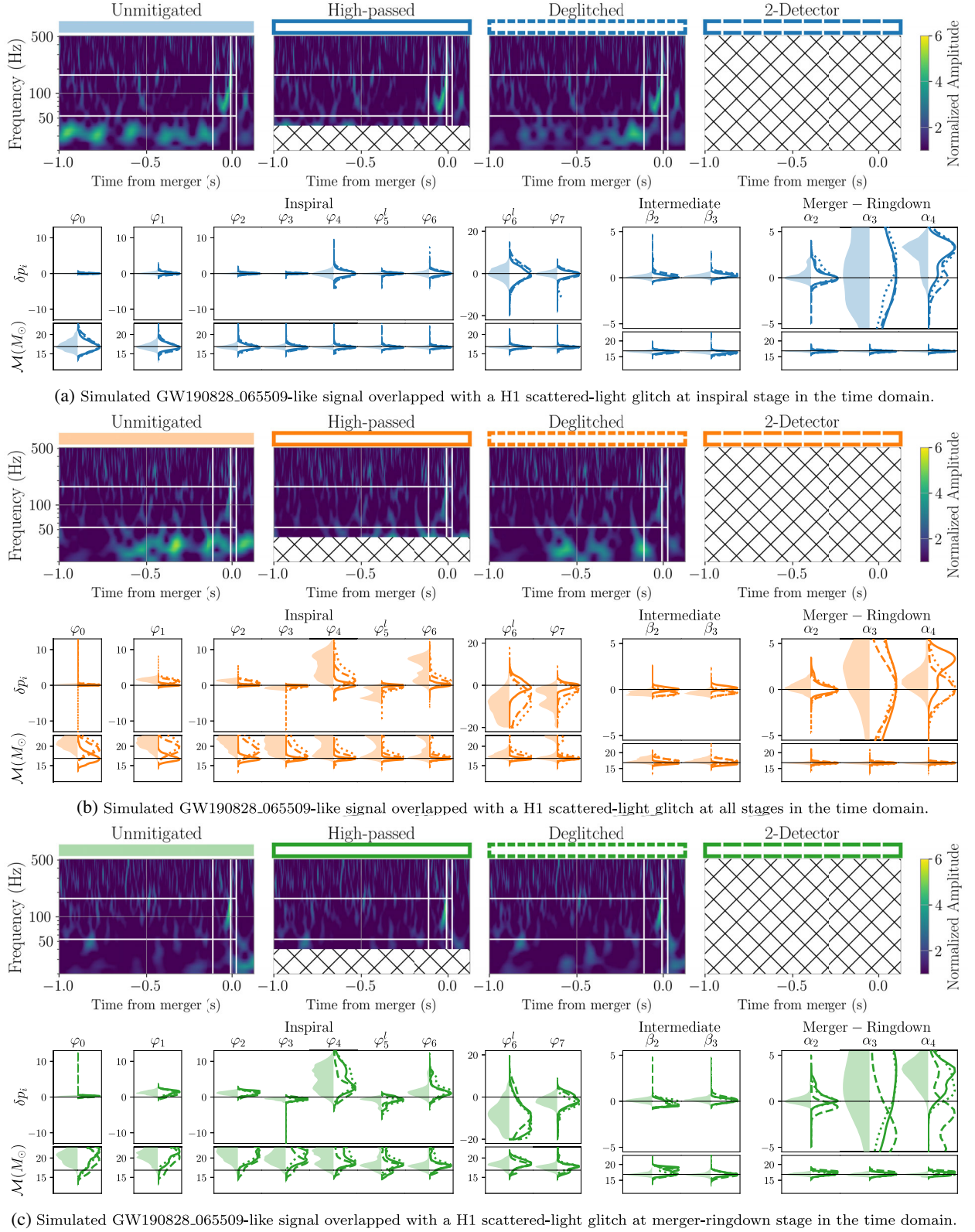
(a) Simulated GW190828_065509-like signal overlapped with a H1 scattered-light glitch at inspiral stage in the time domain.

(b) Simulated GW190828_065509-like signal overlapped with a H1 scattered-light glitch at all stages in the time domain.

(c) Simulated GW190828_065509-like signal overlapped with a H1 scattered-light glitch at merger-ringdown stage in the time domain.

FIG. 10. Similar to Fig. 5, but for scattered-light-glitch-overlapped signals and with inpainting replaced with using only data from detectors without glitches.

As such, we choose not to apply time-domain filtering methods, such as inpainting, to the study of scattered-light-glitch-overlapped signals. To compare the results of the other mitigation methods, we replace inpainting with using data only from the remaining two detectors in which glitches are not present. However, such a method is not preferred, in general, as much useful information is lost, and is omitted from later discussions of suitable mitigation methods; the example illustrated in Fig. 9 also suggests that such a method may as well lead to bias.

For all three cases of glitch overlapping, the value of 0 is not excluded at 90% credibility from the posterior distributions of testing parameters for the unmitigated samples (left of violin plots). Furthermore, the glitch has no observable effect on parametrized tests of GR when it temporally overlaps with the inspiral stage of the signal [Fig. 10(a)], as the posterior distributions of all testing parameters for the unmitigated case match with that with the glitch removed in three independent methods.

In most cases, the posterior distributions of testing parameters obtained by the three independent glitch mitigation methods match closely with each other. This is a good indication that the mitigation methods do not introduce additional effects to the results. In particular, in the case where the scattered-light glitch overlaps all stages of coalescence in the time domain [Fig. 10(b)], the glitch removal through high passing to 40 Hz improves the inference of inspiral testing parameters most significantly (see $\delta\varphi_4$, $\delta\varphi_{5l}$, $\delta\varphi_6$, $\delta\varphi_{6l}$, and $\delta\varphi_7$.) However, from the investigation in Fig. 9, we would warn against discarding data from the affected detector entirely, as substantial reduction of signal power may lead to bias.

Representative figures of the SNR for unmitigated and mitigated samples are obtained in the same way as in Sec. V A and are $15.06^{+0.23}_{-0.37}$ before mitigation, $15.04^{+0.24}_{-0.37}$ after high passing, $14.21^{+0.25}_{-0.39}$ after glitch subtraction, and $12.19^{+0.27}_{-0.43}$ after discarding data from H1 where the scattered-light glitch is present.

VI. CONCLUSION AND OUTLOOK

We overlapped a simulated high-mass-ratio coalescing BBH signal with three glitches from most frequently occurring glitch classes in O3a. We then investigated the effects on parametrized tests of GR of the glitches and their mitigation through bandpass filtering, inpainting, and BayesWave glitch model subtraction. Although the number of glitches considered in this investigation is not sufficient for us to give quantitative statements about the effects of certain glitch classes or mitigation methods on tests of GR, our analysis covered all stages of BBH coalescence in the time and frequency domain, and we are able to identify the effects case by case by comparing the unmitigated results with that mitigated by independent methods and expected GR results when the noise model is not violated.

No false violations of GR are identified for data samples overlapped with the scattered-light glitch, while false violations are observed for that overlapped with the tomte and blip glitches. For the latter cases, we found no clear correlation between the stages of coalescence in which false violations occurred and those overlapped by the glitch in the time or frequency domain.

Out of the three mitigation methods, we find that inpainting and BayesWave glitch model subtraction consistently reduce false violations of GR, and the results match closely with each other. This indicates that the two methods did not introduce additional effects to parametrized tests and suggests successful glitch removals. Bandpass filtering, on the other hand, can also reduce false violations in most cases. However, false violations are amplified or new violations are introduced in more than one case after substantial removal of signal power through high passing or discarding data from the glitch-affected detector. We suggest the application of inpainting or BayesWave glitch model subtraction for glitch mitigation, as they are found to be effective even when an extra degree of freedom is involved with the introduction of parametrized deviation to the signal model.

A major improvement on the LIGO detectors is expected to be completed in a few years. The increased sensitivity, in turn, suggests more frequent occurrence of glitches overlapping signals. As mitigating signals overlapped with glitches may become a regularity in the future, a systematic study on the effects of glitches and their mitigation to tests of GR will be crucial to the next generation of GW astronomy. Such a study would likely involve similar methodologies to that presented by this work, applied repeatedly to study different conditions, making this work an important first study on this subject.

2

[1] E. Berti, E. Barausse, V. Cardoso, L. Gualtieri, P. Pani, U. Sperhake, L. C. Stein, N. Wex, K. Yagi, T. Baker *et al.*, Testing general relativity with present and future astrophysical observations, Classical Quantum Gravity **32,** 243001 (2015).

[2] C. W. Misner, K. S. Thorne, J. A. Wheeler *et al.*, *Gravitation* (MacMillan, London, 1973).

[3] C. M. Will, The confrontation between general relativity and experiment, Living Rev. Relativity **9,** 3 (2006).

[4] C. M. Will, The confrontation between general relativity and experiment, Living Rev. Relativity **17,** 4 (2014).

[5] K. Stelle, Renormalization of higher-derivative quantum gravity, Phys. Rev. D **16,** 953 (1977).

[6] B. S. Sathyaprakash and B. F. Schutz, Physics, astrophysics and cosmology with gravitational waves, Living Rev. Relativity **12,** 2 (2009).

[7] N. Yunes and X. Siemens, Gravitational-wave tests of general relativity with ground-based detectors and pulsar-timing arrays, Living Rev. Relativity **16,** 9 (2013).

[8] N. Yunes, K. Yagi, and F. Pretorius, Theoretical physics implications of the binary black-hole mergers GW150914 and GW151226, Phys. Rev. D **94,** 084002 (2016).

[9] J. Meidam, K. W. Tsang, J. Goldstein, M. Agathos, A. Ghosh, C.-J. Haster, V. Raymond, A. Samajdar, P. Schmidt, R. Smith, K. Blackburn, W. Del Pozzo, S. E. Field, T. Li, M. Pürrer, C. Van Den Broeck, J. Veitch, and S. Vitale, Parametrized tests of the strong-field dynamics of general relativity using gravitational wave signals from coalescing binary black holes: Fast likelihood calculations and sensitivity of the method, Phys. Rev. D **97,** 044033 (2018).

[10] K. G. Arun, B. R. Iyer, M. S. S. Qusailah, and B. S. Sathyaprakash, Probing the nonlinear structure of general relativity with black hole binaries, Phys. Rev. D **74,** 024006 (2006).

[11] K. G. Arun, B. R. Iyer, M. S. S. Qusailah, and B. S. Sathyaprakash, LETTER TO THE EDITOR: Testing post-Newtonian theory with gravitational wave observations, Classical Quantum Gravity **23,** L37 (2006).

[12] T. G. F. Li, W. Del Pozzo, S. Vitale, C. Van Den Broeck, M. Agathos, J. Veitch, K. Grover, T. Sidery, R. Sturani, and A. Vecchio, Towards a generic test of the strong field dynamics of general relativity using compact binary coalescence, Phys. Rev. D **85,** 082003 (2012).

[13] M. Agathos, W. Del Pozzo, T. G. F. Li, C. Van Den Broeck, J. Veitch, and S. Vitale, Tiger: A data analysis pipeline for testing the strong-field dynamics of general relativity with gravitational wave signals from coalescing compact binaries, Phys. Rev. D **89,** 082001 (2014).

[14] E. Barausse, V. Cardoso, and P. Pani, Can environmental effects spoil precision gravitational-wave astrophysics?, Phys. Rev. D **89,** 104059 (2014).

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

[16] 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).

[17] B. Abbott, R. Abbott, T. Abbott, S. Abraham, F. Acernese, K. Ackley, C. Adams, R. Adhikari, V. Adya, C. Affeldt *et al.*, Gwtc-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo During the First and Second Observing Runs, Phys. Rev. X **9,** 031040 (2019).

[18] R. Abbott, T. Abbott, S. Abraham, F. Acernese, K. Ackley, A. Adams, C. Adams, R. Adhikari, V. Adya, C. Affeldt *et al.*, GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run, Phys. Rev. X **11,** 021053 (2021).

[19] 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.* (The LIGO Scientific Collaboration, the Virgo Collaboration), GWTC-2.1: Deep extended catalog of compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run, arXiv:2108.01045.

[20] 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.*, Tests of General Relativity with GW150914, Phys. Rev. Lett. **116,** 221101 (2016).

[21] B. Abbott, R. Abbott, T. Abbott, S. Abraham, F. Acernese, K. Ackley, C. Adams, R. Adhikari, V. Adya, C. Affeldt *et al.*, Tests of general relativity with the binary black hole signals from the LIGO-Virgo catalog GWTC-1, Phys. Rev. D **100,** 104036 (2019).

[22] R. Abbott, T. Abbott, S. Abraham, F. Acernese, K. Ackley, A. Adams, C. Adams, R. Adhikari, V. Adya, C. Affeldt *et al.*, Tests of general relativity with binary black holes from the second LIGO-Virgo gravitational-wave transient catalog, Phys. Rev. D **103,** 122002 (2021).

[23] P. R. Saulson, *Fundamentals of Interferometric Gravitational Wave Detectors* (World Scientific, Singapore, 2017).

[24] LIGO Scientific Collaboration, LIGO Algorithm Library— LALSuite, free software (GPL), 2018.

[25] J. Veitch *et al.*, Parameter estimation for compact binaries with ground-based gravitational-wave observations using the lalinference software library, Phys. Rev. D **91,** 042003 (2015).

[26] L. K. Nuttall, T. Massinger, J. Areeda, J. Betzwieser, S. Dwyer, A. Effler, R. Fisher, P. Fritschel, J. Kissel, A. Lundgren *et al.*, Improving the data quality of advanced LIGO based on early engineering run results, Classical Quantum Gravity **32,** 245005 (2015).

[27] M. Zevin, S. Coughlin, S. Bahaadini, E. Besler, N. Rohani, S. Allen, M. Cabero, K. Crowston, A. K. Katsaggelos, S. L. Larson *et al.*, Gravity Spy: Integrating advanced LIGO detector characterization, machine learning, and citizen science, Classical Quantum Gravity **34,** 064003 (2017).

[28] B. P. Abbott, R. Abbott, T. Abbott, M. Abernathy, F. Acernese, K. Ackley, M. Adamo, C. Adams, T. Adams, P. Addesso *et al.*, Characterization of transient noise in advanced LIGO relevant to gravitational wave signal GW150914, Classical Quantum Gravity **33,** 134001 (2016).

[29] S. Chatterji, L. Blackburn, G. Martin, and E. Katsavounidis, Multiresolution techniques for the detection of gravitational-wave bursts, Classical Quantum Gravity **21,** S1809 (2004).

[30] F. Robinet, N. Arnaud, N. Leroy, A. Lundgren, D. Macleod, and J. McIver, Omicron: A tool to characterize transient noise in gravitational-wave detectors, SoftwareX **12,** 100620 (2020).

[31] J. R. Smith, T. Abbott, E. Hirose, N. Leroy, D. MacLeod, J. McIver, P. Saulson, and P. Shawhan, A hierarchical method for vetoing noise transients in gravitational-wave detectors, Classical Quantum Gravity **28,** 235005 (2011).

[32] T. Isogai (L. S. Collaboration, V. Collaboration) *et al.*, Used percentage veto for LIGO and Virgo binary inspiral searches, J. Phys. Conf. Ser. **243,** 012005 (2010).

[33] R. Essick, L. Blackburn, and E. Katsavounidis, Optimizing vetoes for gravitational-wave transient searches, Classical Quantum Gravity **30,** 155010 (2013).

[34] R. Biswas, L. Blackburn, J. Cao, R. Essick, K. A. Hodge, E. Katsavounidis, K. Kim, Y.-M. Kim, E.-O. Le Bigot, C.-H. Lee, J. J. Oh, S. H. Oh, E. J. Son, Y. Tao, R. Vaulin, and X. Wang, Application of machine learning algorithms to the study of noise artifacts in gravitational-wave data, Phys. Rev. D **88,** 062003 (2013).

[35] C. Messick *et al.*, Analysis framework for the prompt discovery of compact binary mergers in gravitational-wave data, Phys. Rev. D **95,** 042001 (2017).

[36] S. A. Usman, A. H. Nitz, I. W. Harry, C. M. Biwer, D. A. Brown, M. Cabero, C. D. Capano, T. Dal Canton, T. Dent, S. Fairhurst *et al.*, The PyCBC search for gravitational waves from compact binary coalescence, Classical Quantum Gravity **33,** 215004 (2016).

[37] N. J. Cornish and T. B. Littenberg, Bayeswave: Bayesian inference for gravitational wave bursts and instrument glitches, Classical Quantum Gravity **32,** 135012 (2015).

[38] T. B. Littenberg and N. J. Cornish, Bayesian inference for spectral estimation of gravitational wave detector noise, Phys. Rev. D **91,** 084034 (2015).

[39] B. Zackay, T. Venumadhav, J. Roulet, L. Dai, and M. Zaldarriaga, Detecting gravitational waves in data with non-Gaussian noise, Phys. Rev. D **104,** 063034 (2021).

[40] C. Pankow, K. Chatziioannou, E. A. Chase, T. B. Littenberg, M. Evans, J. McIver, N. J. Cornish, C.-J. Haster, J. Kanner, V. Raymond, S. Vitale, and A. Zimmerman, Mitigation of the instrumental noise transient in gravitational-wave data surrounding GW170817, Phys. Rev. D **98,** 084016 (2018).

[41] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Simple Model of Complete Precessing Black-Hole-Binary Gravitational Waveforms, Phys. Rev. Lett. **113,** 151101 (2014).

[42] W. B. Davenport, W. L. Root *et al.*, *An Introduction to the Theory of Random Signals and Noise* (McGraw-Hill, New York, 1958), Vol. 159.

[43] J. D. Romano and N. J. Cornish, Detection methods for stochastic gravitational-wave backgrounds: A unified treatment, Living Rev. Relativity **20,** 2 (2017).

[44] P. Whittle, *Hypothesis Testing in Time Series Analysis* (Almqvist & Wiksells boktr., Stockholm, 1951), Vol. 4.

[45] S. Khan, K. Chatziioannou, M. Hannam, and F. Ohme, Phenomenological model for the gravitational-wave signal from precessing binary black holes with two-spin effects, Phys. Rev. D **100,** 024059 (2019).

[46] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. J. Forteza, and A. Bohé, Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era, Phys. Rev. D **93,** 044007 (2016).

[47] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. J. Forteza, and A. Bohé, Frequency-domain gravitational waves from nonprecessing black-hole binaries. I. New numerical waveforms and anatomy of the signal, Phys. Rev. D **93,** 044006 (2016).

[48] C. Cutler and E. E. Flanagan, Gravitational waves from merging compact binaries: How accurately can one extract the binary's parameters from the inspiral waveform?, Phys. Rev. D **49,** 2658 (1994).

[49] F. J. Harris, On the use of windows for harmonic analysis with the discrete fourier transform, Proc. IEEE **66,** 51 (1978).

[50] B. P. Abbott, R. Abbott, T. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. Adhikari, V. Adya *et al.*, GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral, Phys. Rev. Lett. **119,** 161101 (2017).

[51] 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.*, Multi-messenger observations of a binary neutron star merger, Astrophys. J. **848,** L12 (2017).

[52] J. McIver, M. Millhouse, G. Ashton, D. Davis, J. Veitch, and T. Littenberg, O3a catalog mitigated data review results and status spreadsheet, Report No. LIGO-T2000260.

[53] 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).

[54] Rico K. L. Lo (private communication).

[55] The event GW1902828_065509 has a full-signal optimal SNR of 9.9—weakest of all events considered for parameterized tests of GR in GWTC-1 and GWTC-2 combined [21,22].

[56] Calibrated, cleaned data from H1 and L1 are taken from the strain channel DCS-CALIB_STRAIN_CLEAN_C01.

Reproduced data from V1 are taken from the strain channel Hrec_hoft_V1O3Repro1A_16384 Hz [57].

[57] R. Abbott *et al.*, Open data from the first and second observing runs of Advanced LIGO and Advanced Virgo, SoftwareX **13,** 100658 (2021).

[58] D. Davis, J. S. Areeda, B. K. Berger, R. Bruntz, A. Effler, R. C. Essick, R. P. Fisher, P. Godwin, E. Goetz, A. F. Helmling-Cornell *et al.*, LIGO detector characterization in the second and third observing runs, Classical Quantum Gravity **38,** 135014 (2021).

[59] K. Chatziioannou, N. Cornish, M. Wijngaarden, and T. B. Littenberg, Modeling compact binary signals and instrumental glitches in gravitational wave data, Phys. Rev. D **103,** 044013 (2021).