Statistical validation of the detection of a sub-dominant quasi-normal mode in GW190521

Collin D. Capano,[1,2,3] Jahed Abedi,[4,2,3] Shilpa Kastha,[2,3] Alexander H. Nitz,[2,3] Julian Westerweck,[2,3] Yi-Fan Wang,[2,3] Miriam Cabero,[5] Alex B. Nielsen,[4] and Badri Krishnan[6,2,3]

[1]*Department of Physics, University of Massachusetts, Dartmouth, MA 02747, USA*
[2]*Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut), Callinstraße 38, 30167 Hannover, Germany*
[3]*Leibniz Universität Hannover, 30167 Hannover, Germany*
[4]*Department of Mathematics and Physics, University of Stavanger, NO-4036 Stavanger, Norway*
[5]*Department of Physics and Astronomy, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada*
[6]*Institute for Mathematics, Astrophysics and Particle Physics, Radboud University, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands*

arXiv:2209.00640v1 [gr-qc] 1 Sep 2022

One of the major aims of gravitational wave astronomy is to observationally test the Kerr nature of black holes. The strongest such test, with minimal additional assumptions, is provided by observations of multiple ringdown modes, also known as black hole spectroscopy. For the gravitational wave merger event GW190521, we have previously claimed the detection of two ringdown modes emitted by the remnant black hole. In this paper we provide further evidence for the detection of multiple ringdown modes from this event. We analyze the recovery of simulated gravitational wave signals designed to replicate the ringdown properties of GW190521. We quantify how often our detection statistic reports strong evidence for a sub-dominant $(\ell, m, n) = (3, 3, 0)$ ringdown mode, even when no such mode is present in the simulated signal. We find this only occurs with a probability ~ 0.02, which is consistent with a Bayes factor of 56 ± 1 (1σ uncertainty) found for GW190521. We also quantify our agnostic analysis of GW190521, in which no relationship is assumed between ringdown modes, and find that less than 1 in 500 simulated signals without a $(3, 3, 0)$ mode yield a result as significant as GW190521. Conversely, we verify that when simulated signals do have an observable $(3, 3, 0)$ mode they consistently yield a strong evidence and significant agnostic results. We also find that simulated GW190521-like signals with a $(3, 3, 0)$ mode present yield tight constraints on deviations of that mode from Kerr, whereas constraints on the $(2, 2, 1)$ overtone of the dominant mode yield wide constraints that are not consistent with Kerr. These results on simulated signals are similar to what we find for GW190521. Our results strongly support our previous conclusion that the gravitational wave signal from GW190521 contains an observable sub-dominant $(\ell, m, n) = (3, 3, 0)$ mode.

I. INTRODUCTION

Einstein's theory of general relativity (GR) predicts that black holes are stable to perturbations [1]. A distorted black hole should settle down to a stationary Kerr state through the emission of gravitational waves [2]. This applies to the remnant black hole formed in a binary black hole merger event, which is highly distorted on formation, but is expected to eventually settle down to a Kerr black hole due to the emission of gravitational waves. The gravitational waveform in the late stages of a merger event consists of a spectrum of quasi-normal modes with a rich structure of different fundamental modes and overtones [3]. The spectrum consists of a set of complex frequencies (determined by the black hole mass and spin) labeled by three integers (ℓ, m, n), with $\ell \geq 2$, $-\ell \leq m \leq \ell$, and $n \geq 0$. Modes with $n \geq 1$ are known as "overtones". Using black hole spectroscopy, the observation of more than one such ringdown mode can be used to determine if the black hole is consistent with GR [4, 5]. A clear and unambiguous determination of multiple ringdown modes provides one of the strongest tests of the Kerr nature of black holes in our universe and a possible route to discover new physics beyond standard general relativity.

Quasi-normal modes for a Schwarzschild black hole were first identified by Vishveshwara [6, 7], and further studied within black hole perturbation theory by Chandrasekhar and Detweiler [8]. There remain several outstanding theoretical questions regarding black hole quasi-normal modes which are also important for observational studies. The first is the question of the start time for the ringdown. When the remnant black hole is formed, it is initially highly distorted away from a Kerr black hole. The black hole loses these distortions over time, and at some point it can be considered to be a linear perturbation of a Kerr black hole. It is not clear when (and if [9]) this perturbative regime can be distinguished. See [5, 10] for studies of the start time of the ringdown phase, and [11–14] for studies of possible non-linear effects. The different regimes seen in the gravitational wave signal are expected to have counterparts in the strong field dynamical spacetime region near the binary system [15–17]. See e.g. [18–22] for studies of black hole horizon geometry in the post-merger phase and whether a ringdown regime can be identified using the horizon dynamics as well.

It has long been expected that only the most dominant ringdown mode will be observable with the current generation of gravitational wave detectors [23, 24]. Those expectations were based on astrophysical assumptions about the total mass and mass ratio distributions of bi-

nary black hole systems in the observable universe, which in turn determine the amplitudes of various ringdown modes [25]. However, evidence for an overtone of the dominant mode of GW150914 was presented in [26, 27]. There it was shown that it is possible to model the gravitational waveform as a superposition of ringdown modes starting from the merger by using the overtones of the dominant mode. This is a significant result, though there remain several interesting open questions regarding data analysis and theoretical issues. Some of the data analysis issues are discussed in [28, 29]. On the theoretical side, the stability of the overtones under small perturbations raises several interesting open questions; see e.g. [30–34]. Evidence of a second fundamental mode, without using overtones, was first presented for the event GW190521 in [35], henceforth referred to as "Capano et al.", and will be elaborated further in this paper.

With this analysis we address three fundamental questions. Firstly, if a signal explicitly does not contain any sub-dominant ringdown modes, how often does our detection pipeline falsely claim the existence of such modes? Secondly, if one or more sub-dominant modes are present in the data, how often does our pipeline correctly recover them? Thirdly, if our pipeline is used to constrain deviations from Kerr, how well do the resulting inferred parameters match those of the simulated signal? The key results for detection of a second mode are shown in Figs. 5 and 6. Fig. 5 applies our analysis to simulated signals without a $(3, 3, 0)$ mode and shows that the false alarm probability is consistent with expectations from noise. Fig. 6 quantifies the ability of our method to detect the $(3, 3, 0)$ mode when it is present, as a function of the signal strength.

In section II we give additional details of how the data is treated in the analysis of Capano et al. Section III explains how we generate the simulated data sets. In sections IV and V we investigate the statistical significance of detecting two modes versus one using a set of simulated signals. Section IV presents an agnostic analysis that looks at the consistency of the second mode with the first mode. Section V presents an analysis more closely tied to the Kerr hypothesis, analysing the likelihood of two Kerr modes versus just one. In section VI we use our simulated signals to compare the accuracy with which the no hair theorem can be tested using fundamental modes or overtones for an event similar to GW190521.

We conclude this introduction by briefly summarizing some basic properties of the event GW190521, which will be relevant in the rest of this paper.

A. GW190521

The gravitational wave event GW190521 was detected on May 21st 2019 at 03:02:29 UTC by the Advanced LIGO and Advanced Virgo detectors [36]. The most conservative explanation of the signal is the binary merger of two black holes [36, 37], although there are also various other interpretations of this event [39–44].

While the progenitors of the event GW190521 are open to speculation, in most scenarios the final outcome is still likely to be a single black hole. The event GW190521 shows clear evidence of a dominant ringdown mode of a final black hole after the merger [36]. In Capano et al. [35] the ringdown signal was found to contain an additional sub-dominant ringdown mode. The dominant mode is consistent with being the $(\ell, m, n) = (2, 2, 0)$ ringdown mode of a Kerr black hole; the second mode is consistent with the sub-dominant fundamental $(\ell, m, n) = (3, 3, 0)$ mode. As detailed in this paper, under a Kerr hypothesis, the Bayes factor preferring the existence of the $(2, 2, 0)$ and $(3, 3, 0)$ modes over just the $(2, 2, 0)$ or the $(2, 2, 0)$ and $(2, 2, 1)$ modes is estimated to be 56 ± 1 (1σ uncertainty).

If GW190521 is indeed a binary black hole merger, the inferred total mass of the system would make it one of the most massive binary black hole systems observed to date [45, 46]. Other interpretations have found even higher total masses [39]. A high total mass implies that very little of the inspiral phase occurs inside the sensitive band of the detectors and the recorded signal is dominated by the merger and ringdown. Therefore an analysis that focuses solely on the ringdown phase is of interest and avoids some of the modelling issues in the progenitor inspiral phase.

Inferences about the final black hole parameters using the ringdown signal alone are sensitive to the assumed start time of the ringdown [5, 28]. Different starting times can lead to different results [47, 48]. A ringdown-only analysis must explicitly exclude some of the signal that is outside the ringdown phase. In this work we present additional details of the approach used in Capano et al. Parameter estimates for the event GW190521 based on the binary black hole interpretation are shown in Figs. 1 and 2. These estimates come from different authors using different methods [35, 46, 49, 50]. The redshifted final total mass spans a wide range from around 200 $M_\odot$ to nearly 400 $M_\odot$.

The peak gravitational wave strain is expected to occur close to the merger. The GPS time of the peak strain in the Hanford detector was initially estimated using the posterior median with the NRSurPHM waveform model to be $1242442967^{+0.0067}_{-0.0106}$ s. [37]. See [49] for further discussion. As can be seen in Fig. 2, estimates of the merger coalescence time range over some 20 ms depending on the waveform model considered. This is a significant time range since, in geometric units, it corresponds to approximately $13M$ for an object with a mass $M = 300$ $M_\odot$.

We use the open source `PyCBC Inference` library for performing Bayesian inference [38, 51]. For sampling the parameter space we use the `dynesty` nested sampler [52]. We use data for the event GW190521 made publicly available by the Gravitational Wave Open Science Center [53]. We fix the sky location to the values given by the maximum likelihood result of Nitz & Capano [50], although we have obtained similar results using the LVC's maximum

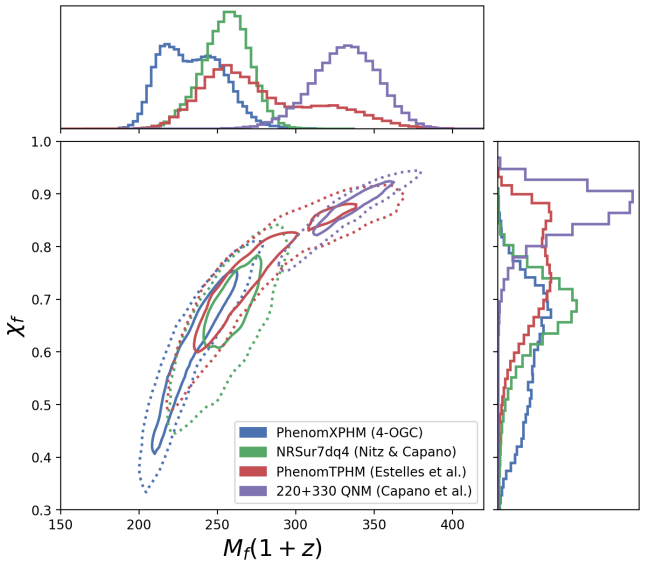

FIG. 1. Comparison of the final mass and spin of GW190521, as estimated by NRSur7dq4 [50], IMRPhenomXPHM [46], IMRPhenomTPHM [49], and a Kerr ringdown with both the $(2,2,0)$ and $(3,3,0)$ modes [35]. The IMRPhenomTPHM results have a second mode in the posterior that is consistent with the Kerr ringdown results.

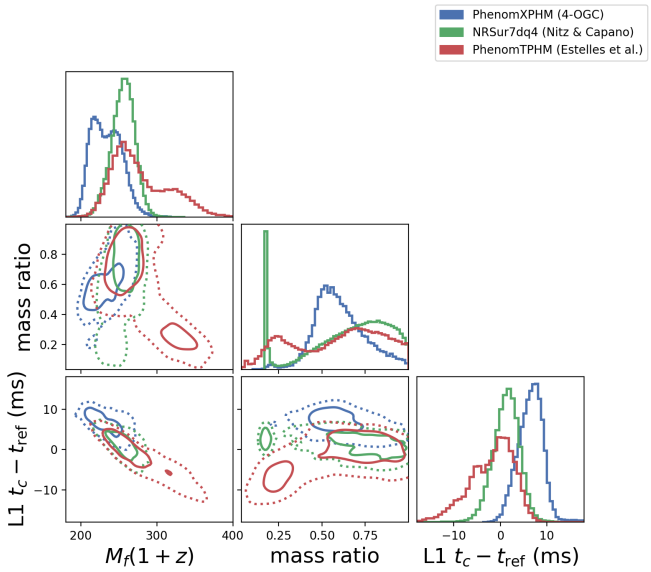

FIG. 2. Comparison of the final mass, mass ratio, and coalescence time in the Livingston detector as estimated by the NRSur7dq4 [50], IMRPhenomXPHM [46], and IMRPhenomTPHM [49] waveform models. The second mode in final mass found by IMRPhenomTPHM, which is consistent with the Kerr ringdown results (cf. Fig. 1), corresponds to a second mode at more asymmetric masses. This mode also yields a coalescence time that is $\sim 6\,$ms earlier than the equal mass mode found by the other approximants. This earlier coalescence time estimate is $\sim 13\,$ms before the time at which the Bayes factor for the $(2,2,0)+(3,3,0)$ mode peaks in Capano et al. [35]. It is also consistent with the time at which the evidence for the $(2,2,0)+(2,2,1)$ Kerr ringdown model peaks.

likelihood sky location [37]. We use a geocentric GPS reference time of $t_{\rm ref} = 1242442967.445$ [50]. With the sky location used in our analyses, this corresponds to the detector GPS reference times $1242442967+0.4259$ at LIGO Hanford, $+0.4243$ at LIGO Livingston and $+0.4361$ at Virgo. Credible intervals in the text are quoted to 90%.

II. BASICS OF RINGDOWN DETECTION AND PARAMETER ESTIMATION

In this section we summarize some essential elements of the data analysis procedure that we employ. Since we analyze exclusively the ringdown which is only a part of the full signal, an important challenge is to identify the portion of the full signal corresponding to the ringdown. Similarly, it is necessary to ensure that the procedure for extracting this portion of the data properly takes into account correlations with neighboring time samples that should be excluded from the analysis.

Let $\vec{s} = \{s_0, ..., s_{N-1}\}$ denote time-ordered samples of the strain data from a gravitational wave detector. The data is sampled every Δt seconds over a duration T, so that the number of samples is $N = \lfloor T/\Delta t \rfloor + 1$. A network of K detectors sampled in this way will produce a set of samples $\vec{s}_{\rm net} = \{\vec{s}_1, ..., \vec{s}_K\}$. The strain data is assumed to be a combination of a possible signal $\vec{h}$ and noise $\vec{n}$

$$\vec{s} = \vec{h} + \vec{n}\,. \tag{1}$$

Let $p(\vec{s}|\vec{\lambda}, H)$ be the likelihood of the data $\vec{s}$ in the presence of a signal with given parameters $\vec{\lambda}$ under background hypotheses H, such as the signal model. The probability of finding a realisation of noise $\vec{n}$ under the hypotheses H is $p(\vec{n}|H)$. Therefore the likelihood for data $\vec{s}$ can be written as

$$p(\vec{s}|\vec{\lambda}, H) = p(\vec{s} - \vec{h}(\vec{\lambda})|H, n)\,, \tag{2}$$

where the right-hand side is under the hypothesis n that no signal is present. In gravitational-wave astronomy, in the absence of a signal, it is common to assume over short times that the detectors output stochastic Gaussian noise which is independent across detectors. With this assumption the probability density function describing the time-ordered noise samples of the detector network $\vec{n}_{\rm net}$ is a product of K $N-$dimensional multivariate normal distributions,

$$p(\vec{n}_{\rm net}) = \frac{\exp\left[-\frac{1}{2}\sum_{d=1}^{K} \vec{n}_d^{\top} \vec{C}_d^{-1} \vec{n}_d\right]}{\sqrt{(2\pi)^{NK} \prod_{d=1}^{K} \det \vec{C}_d}}. \tag{3}$$

Here, $\vec{C}_d$ is the covariance matrix of the noise in detector d, and we drop the hypotheses H in our notation. See [54] for further details.

If the detector's noise is wide-sense stationary and ergodic, which is typically the case for the LIGO and Virgo detectors, the noise likelihood takes a simple form

$$p(\vec{n}_{\rm net}) \propto \exp\left[-\frac{1}{2}\sum_{d=1}^{K}\langle \vec{n}_d, \vec{n}_d\rangle\right]. \tag{4}$$

Here, the inner product $\langle\cdot,\cdot\rangle$ is defined as

$$\langle \vec{u}_d, \vec{v}_d\rangle \equiv 4\Re\left\{\frac{1}{T}\sum_{p=1}^{N/2-1}\frac{\tilde{u}_d^*[p]\tilde{v}_d[p]}{S_n^{(d)}[p]}\right\}, \tag{5}$$

where $\tilde{u}$ is the discrete Fourier transform of the time series $\vec{u}$, an asterisk denotes the complex conjugate, and S_n is the power spectral density of the detector's noise. To obtain the posterior probability density function $p(\vec{\lambda}|\vec{s}, H)$ for the parameters $\vec{\lambda}$, we use Bayes' theorem

$$p(\vec{\lambda}|\vec{s}, H) = \frac{1}{Z}p(\vec{s}|\vec{\lambda}, H)p(\vec{\lambda}|H), \tag{6}$$

where $p(\vec{s}|\vec{\lambda}, H)$ is the likelihood function, $p(\vec{\lambda}|H)$ is the prior, and Z is a normalization constant known as the evidence, depending only on the data. Taking the ratio of evidences Z_A/Z_B for two different models H_A and H_B yields the "Bayes factor". In this work, the signal models will be GW ringdown waveforms with only fundamental modes and overtones. If our prior belief for the validity of the two models is the same, the Bayes factor gives the odds that model A is favoured over model B. Ref. [55] suggested Bayes factors greater than 3.2, 10 and 100 are considered substantial, strong, and decisive, respectively.

The ringdown waveform model takes the following form

$$h_+ + ih_\times = \frac{M_f}{D_L}\sum_{\ell mn} {}_{-2}S_{\ell mn}(\iota, \varphi, \chi_f) A_{\ell mn} e^{i(\Omega_{\ell mn}t + \phi_{\ell mn})}, \tag{7}$$

where $h_{+/\times}$ are the plus/cross polarizations of the wave, M_f is the total mass of the remnant black hole in the detector frame and D_L is the source luminosity distance. The waveform is decomposed with respect to the spin-2 weighted spheroidal basis ${}_{-2}S_{\ell mn}$, which is a function of the remnant black hole's spin χ_f, the inclination angle ι and azimuthal angle φ relative to the observer. The amplitude and phase of the quasi-normal modes are denoted by $A_{\ell mn}$ and $\phi_{\ell mn}$. The complex frequency is $\Omega_{\ell mn} = 2\pi f_{\ell mn} + i/\tau_{\ell mn}$, where the characteristic frequency $f_{\ell mn}$ and decay time $\tau_{\ell mn}$ are solely determined by the mass and spin of the remnant black hole, as predicted by the no-hair theorem in GR. We also consider an agnostic ringdown waveform model in this work, for which we absorb the M_f/D_L term into the amplitude and replace the spheroidal harmonics with arbitrary complex numbers $X_{\ell\pm mn} = e^{i\psi_{\ell\pm mn}}$.

In a standard full-signal analysis, to obtain the likelihood for the signal hypothesis, the noise $\vec{n}_d$ in Eq. 4 is replaced by the residuals $\vec{s}_d - \vec{h}_d$. This requires that $\vec{h}$ is an accurate model of the signal across the entire observation time T, which is not valid for a ringdown-only analysis. Quasi-normal modes only model the gravitational wave from a binary black hole after the merger, when the two component black holes have formed a single, perturbed black hole. Performing Bayesian inference using quasi-normal modes as the signal model therefore requires ignoring times from the data when the ringdown prescription is not valid.

We perform the "gating and in-painting" technique [56] to remove the influence of pre-ringdown data. Define $\vec{n}' = \vec{n}_g + \vec{x}$, where $\vec{n}_g$ is the noise with the pre-merger data zeroed out. We choose $\vec{x}$ such that $\vec{C}^{-1}\vec{n}' = 0$ in the gated region, therefore the likelihood Eq. 3 remains the same outside the gating region while we can still utilize the frequency domain likelihood of Eq. 4.

We use the gated-Gaussian likelihood described above in `PyCBC Inference` [38], which evaluates the noise residuals with $\vec{n}_g = \vec{s}_g - \vec{h}_g$ (i.e., the residual with the gated region zeroed out) and solves for $\vec{x}$ with the following condition

$$\overline{\vec{C}^{-1}}\,\overline{\vec{x}} = -\overline{\vec{C}^{-1}\vec{n}_g}, \tag{8}$$

where the overbar indicates the gating region.

We can then use $\vec{x} + \vec{s}_g - \vec{h}_g$ in the standard likelihood, Eq. 4.

For all analyses we use a gate of two seconds, ending at the start time of the ringdown.

In this work we consider a variety of signal models with different combinations of angular and overtone modes characterized by Eq. 7. The fundamental mode is $(\ell, m, n) = (2, 2, 0)$, and we further consider models with an additional $(2, 2, 1)$ overtone or $(3, 3, 0)$ mode, whose complex frequencies are either predicted by the Kerr hypothesis or treated agnostically as parameters to be determined. We list the priors $p(\vec{\lambda}|H)$ for all parameters used in this work in Table II. In particular, the (2,2,1) amplitude is chosen to be [0,5] times that of the (2,2,0) mode's. This choice is motivated by the numerical relativity fits from [27], and helps to prevent "label switching" in which the (2,2,1) mode template matches to the fundamental mode signal in the data.

For the $(3, 3, 0)$ amplitude we chose a prior that is $[0, 0.5]$ times that of the (2,2,0) mode. This choice is motivated by the numerical simulations of binary black hole mergers in Ref. [25].

When sampling the posterior for the Kerr analysis, we numerically marginalize the polarization angle using a discrete grid of 1000 points. The original motivation was to speed up sampler convergence for the large number of injections analyzed here. However, we found that doing so also led to more robust estimates of the Bayesian evidence, as the sampler was better able to converge on the posterior. Consequently we also reanalyzed GW190521 using the numerical marginalization of the polarization. The effect on the estimation of the Bayes

factor is discussed in more detail in Appendix A. No numerical marginalization is done for the agnostic analysis, as the polarization angle is absorbed into the arbitrary complex numbers used in place of the spheroidal harmonics there.

III. SELECTION OF SIMULATED SIGNALS

In this paper we seek to validate the evidence for the observation of the $(3,3,0)$ mode in GW190521. To do so, we create two sets of simulated signals ("injections"): one set with no $(3,3,0)$ mode in the ringdown (the *Control* set), and another set containing a $(3,3,0)$ mode in the ringdown (the *Signal* set). The Control set is used to measure the rate of false alarms – i.e., to answer the question, how often do we get large evidence for the $(3,3,0)$ mode when the signal contains no $(3,3,0)$ mode? – while the Signal set is used to validate that our pipeline can in fact detect a $(3,3,0)$ mode when it exists in the signal.

For the Control injections we randomly select 500 injections from the NRSurrogate posterior published in Nitz & Capano [50]. This posterior was similar to the posterior published in the initial LIGO/Virgo publication on GW190521 [37]. With the exception of a secondary peak in the posterior around $m_1/m_2 \sim 6$, this NRSurrogate posterior favored approximately equal masses for the binary.[1] It also yielded a merger time for GW190521 only $\sim 6\,\mathrm{ms}$ before the claimed observation time of the $(3,3,0)$ mode in Capano et al. and a relatively low final mass estimate; see Figs. 1 and 2. These results contrast with the claimed observation in Capano et al.: a large $(3,3,0)$ amplitude is not expected for equal-mass binaries [25], and a ringdown model consisting of only fundamental modes is not expected to be a good model for the signal until $\sim 10\,M$ after merger, which for GW190521 would be $\sim 12-16\,\mathrm{ms}$, not $\sim 6\,\mathrm{ms}$. As such, these injections are ideal to test the false alarm rate of our analyses.

To ensure that no $(3,3,0)$ mode exists in the Control injections, we constrain all 500 injections to have mass ratios $m_2/m_1 > 0.5$ and we turn off all but the $\ell = 2$ modes when generating the simulated waveforms. The waveforms are generated using the NRSur7dq4 approximant [57]. We use 500 injections to get a sufficient number of samples at the Bayes factor of GW190521 (56 ± 1); see Sec. V for more details.

To produce the Signal injections we draw random samples from the posterior published in Estelles et al. [49]. This analysis used the IMRPhenomTPHM approximant to analyze GW190521. As with the results presented in Nitz & Capano, Estelles et al. found a bimodal posterior in the component masses for GW190521: one mode favoring nearly equal masses, and one mode favoring mass ratios of $\sim 4:1$. Intriguingly, as shown in Figs. 1 and 2, the second mode yielded a mass and spin estimate for the final black hole that is consistent with the estimate from the ringdown analysis in Capano et al. The estimated merger time for this second mode was also $\sim 5-10\,\mathrm{ms}$ earlier than the NRSurrogate estimate, which is consistent with the peak in the $(2,2,1)$ Bayes factor found in Capano et al. and $\sim 10\,M$ before the peak in the $(3,3,0)$ Bayes factor. The IMRPhenomTPHM waveforms are therefore ideal for our Signal injection set, particularly those from the more asymmetric mass ratio part of the posterior.

To try to ensure that the Signal injections have an observable $(3,3,0)$ mode after the merger, we draw 100 injections from the IMRPhenomTPHM posterior published in Estelles et al. and keep only those that have a $(\ell,m,n) = (3,3,0)$ amplitude > 0.2 after merger. We also require that the signal-to-noise ratio (SNR) of the $(3,3,0)$ mode be at least 4 (the SNR estimated for the $(3,3,0)$ mode in GW190521 in Capano et al.) at some point after merger. To estimate the $(3,3,0)$ SNR we filter each injection in noise with a template consisting only of the $(\ell,m) = (3,3)$ mode, and we gate both the template and signal to remove pre-merger times. Note that here, (ℓ,m) refer to *spherical* harmonics, which is the basis used for IMR models, not the *spheroidal* harmonics used for QNMs. Additionally, many of the posterior samples have large precession. Precession mixes the m modes with the same ℓ in the observer frame. Consequently, an $(\ell,m) = (3,3)$ mode for a IMRPhenomTPHM waveform may consist of a combination of (ℓ,m,n) QNM modes, and not necessarily just the $(3,3,0)$ mode. As such, the estimated SNR may be considered an upper bound on the underlying $(3,3,0)$ QNM.

Applying the SNR cut to the initial 100 draws yields 45 Signal injections. We do not try to generate more Signal injections as they are only used to check that the analysis can recover signals with a $(3,3,0)$ mode and not to estimate small false alarm rates, as we do with the Control injections. We use IMRPhenomTPHM to generate the waveform for the Signal set. Due to the differences between spherical and spheroidal modes, and to try to simulate a realistic signal, we use all available modes in IMRPhenomTPHM when generating the Signal set.

Both sets of injections are added to detector data at random times surrounding the estimated merger time of GW190521. Specifically, an offset time t_{offset} is drawn uniformly in $\pm[4,20]\,\mathrm{s}$ and added to the coalescence time t_c that is drawn from the relevant posterior for each injection. The gap of $\pm 4\,\mathrm{s}$ around GW190521 is to prevent contamination of the data from GW190521. As described below, we perform ringdown analyses on a grid of times surrounding each injection. The widest grid – used in

[1] In Nitz & Capano a prior uniform in $m_1/m_2 \in [1,6]$ was used in the NRSurrogate analysis. If a prior uniform in m_2/m_1 is used (which is approximately the same as a prior uniform in component masses, as done in the LIGO/Virgo analysis), the second mode in the posterior at $m_1/m_2 \sim 6$ is down-weighted, giving further support to the equal-mass portion of the posterior. Here, we draw from the original posterior published in Nitz & Capano, which used a prior uniform in m_1/m_2.

Model	Parameter	Parameter description	Uniform prior range
Agnostic	$f_{A/B/C}$	frequencies of regions A/B/C	$[50, 80]/[80, 256]/[15, 50]$ Hz
	$\tau_{A/B/C}$	decay times of regions A/B/C	$[0.001, 0.1]$ s
	$\log_{10} A_B$	base-10 logarithm of the amplitude of region B	[-24,19]
	$A_{A/C}/A_B$	ratio of amplitudes between region A/C and region B	[0,0.9]
	$\phi_{A/B/C}$	phases of regions A/B/C	$[0, 2\pi]$
	$\psi^{+/-}_{A/B/C}$	phase of the +m and -m modes of the arbitrary complex number in region A/B/C	$[0, 2\pi]$
	$d\beta$	angular difference in amplitudes of +m and -m modes	$[-\pi/4, \pi/4]$
Kerr	M_f	final black hole mass in the detector frame	[100,500] $M_\odot$
	χ_f	final black hole spin	[-0.99,0.99]
	$\log_{10} A_{220}$	base-10 logarithm of the amplitude of (2,2,0)	[-24,-19]
	$A_{330/220}$	ratio of amplitudes between $(3, 3, 0)$ and $(2, 2, 0)$	[0,0.5]
	$A_{221/220}$	ratio of amplitude between $(2, 2, 1)$ and $(2, 2, 0)$	[0,5]
	$\phi_{220/330/221}$	phase of (2,2,0)/(2,2,1)/(3,3,0)	$[0, 2\pi]$
No hair test	δf_{221}	fractional deviation from GR of the (2,2,1) frequency	[-0.16,0.3] with the constraint $f_{221}(1 + \delta f_{221}) > 55$ Hz
	$\delta\tau_{221}$	fractional deviation from GR of the (2,2,1) decay time	[-0.8,0.8]
	δf_{330}	fractional deviation from GR of the $(3, 3, 0)$ frequency	[-0.3,0.3] with the constraint $f_{330}(1 + \delta f_{330}) > 75$ Hz
	$\delta\tau_{330}$	fractional deviation from GR of the $(3, 3, 0)$ decay time	[-0.9,3]
All models	$\cos\iota$	cosine of inclination angle	[-1,1]
	ψ	polarization angle	$[0, 2\pi]$

TABLE I. Prior distributions of sampling parameters for the models used in this work: The agnostic model with the spheroidal harmonics replaced by an arbitrary complex number as discussed in Sec. IV, the Kerr model described by Eq. 7 and discussed in Sec. V, and the testing-GR model discussed in Sec. VI.

the validation of the Kerr Bayes factor (see Sec. V) – is $[-9, 24]$ ms. We therefore draw the $t_{\rm offset}$ such that they are at least 33 ms apart, to ensure that no two analyses analyze exactly the same detector data.

As with the analysis of GW190521 in Capano et al., we use a reference time $t^{\rm inj}_{\rm ref}$ for each injection, around which we construct the grid of times used in the ringdown analyses. For each injection we set the reference time to be $t^{\rm inj}_{\rm ref} = t_{\rm ref} + t_{\rm offset}$, where $t_{\rm ref} = 1242442967.445$ GPS seconds is the estimated geocentric merger time of GW190521, as determined by the maximum likelihood parameters taken from the NRSurrogate analysis in Nitz & Capano [50]. This is the same $t_{\rm ref}$ used in Capano et al. Note that $t^{\rm inj}_{\rm ref}$ is not the injection's coalescence time $t^{\rm inj}_c$; instead, $t^{\rm inj}_c - t^{\rm inj}_{\rm ref}$ follow the same distribution as $t_c - t_{\rm ref}$ (see Fig. 2). In the case of IMRPhenomTPHM, this can mean that some of our Signal injections merge as much as 20 ms before the reference time, well before the grid times used for the analysis.

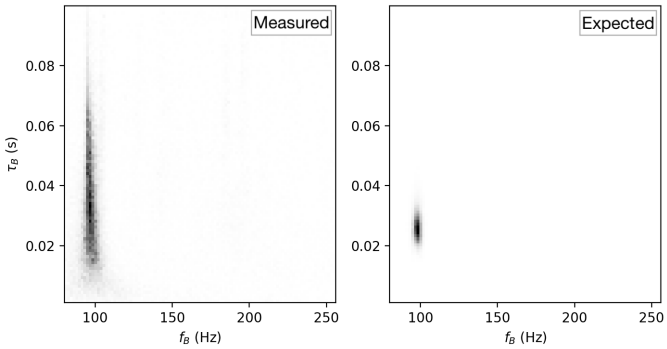

FIG. 3. *Left:* Marginal posterior of the frequency and damping time in frequency range B from the agnostic analysis of GW190521 at $t_{\rm ref} + 6$ ms (same as the heat map in Fig. 1 of Capano et al). *Right:* the expected distribution of the $(3, 3, 0)$ mode assuming the mode observed in frequency range A is the dominant, $(2, 2, 0)$ mode (same as the blue contour in Fig. 1 of Capano et al.) of a Kerr black hole. Darker regions indicate higher probability. The expected distribution is more concentrated due to the larger SNR of the dominant mode. We quantify the agreement between the measured and expected by multiplying the two distributions together and integrating (ζ). The figure and ζ values are produced using 100 bins each in frequency and damping time.

IV. STATISTICAL SIGNIFICANCE OF THE AGNOSTIC ANALYSIS

Two ringdown analyses of GW190521 were presented in Capano et al. [35]: an "agnostic" analysis and a "Kerr" analysis. In the former, the data were analyzed using three QNMs with no assumption made about the relationship between the frequency and damping times of each mode. To prevent all three modes from locking on to the single dominant mode, each mode was assigned a separate frequency range: 50 – 80Hz (range "A"), 80 – 256Hz (range "B"), and 15 – 50Hz (range "C"). Range A

covered the dominant mode, which was clearly visible in the data. This analysis was repeated in intervals of 6ms, between $t_{\rm ref} + [0, 24]$ ms.

A signal with a well-defined posterior was found in Range A, having frequency ~ 63 Hz and damping time ~ 26 ms (see Fig. 1 of Capano et al.). No signal was found in Range C. A second putative mode was found in Range B. This signal was most pronounced at $t_{\rm ref}+6$ ms, at which point it has a frequency of ~ 98 Hz and damping time ~ 32 ms. As shown in Fig.3, these frequencies and damping times were where one would expect the $(3, 3, 0)$ would be assuming GW190521 formed a Kerr black hole, with the signal in Range A being the $(2, 2, 0)$ mode.

Initially, the agnostic analysis was presented as qualitative evidence for the presence of the $(3, 3, 0)$ mode. Here, we repeat the analysis on our two sets of injections and use them to develop a statistic to quantify the statistical significance of the agnostic result.

If an observable $(3, 3, 0)$ mode is truly present in the signal, and it is the only observable mode in Range B, then the measured posterior distribution should peak at the same values as the expected distribution. We expect the measured distribution to be more diffuse than the expected distribution. This is because the expected distribution is derived from the observed dominant mode, which is more accurately measured due to its larger SNR. With these considerations in mind, we quantify the agreement between the measured distribution $p_{\rm meas}(f_B, \tau_B)$ and the expected distribution $p_{330}(f_B, \tau_B)$ using:

$$\zeta \equiv \int p_{\rm meas}(f_B, \tau_B) p_{330}(f_B, \tau_B) {\rm d}f_B {\rm d}\tau_B. \qquad (9)$$

To evaluate this we construct 2D histograms in Range B using 100 bins each in frequency and damping time. This is done at each time step; we then maximize ζ over all the time steps.

We calculate ζ for the Control injections. Since these injections have no $(3, 3, 0)$ mode by construction, the resulting ζ values represent the distribution of false positives. The cumulative distribution of the maximized ζ is shown by the black line in Fig. 4. We also calculate ζ for the Signal injections that have post-merger SNR > 4. The cumulative distribution of the maximized ζ is also shown in Fig. 4 as a blue line. As evident in the figure, we find good separation between the signal and control injections.[2]

Calculating ζ for GW190521, we find that it is at a maximum at $t_{\rm ref} + 6$ ms, with a value of 1.55. This is consistent with our initial qualitative assessment that the observed mode is most consistent with the expected $(3, 3, 0)$ mode at 6 ms. As shown in Fig. 4, GW190521 has a ζ larger than all 500 of our Control injections. We therefore conclude that the probability of obtaining a ζ greater than or equal that of GW190521 by chance from noise (the p-value) is < 1 in 500.

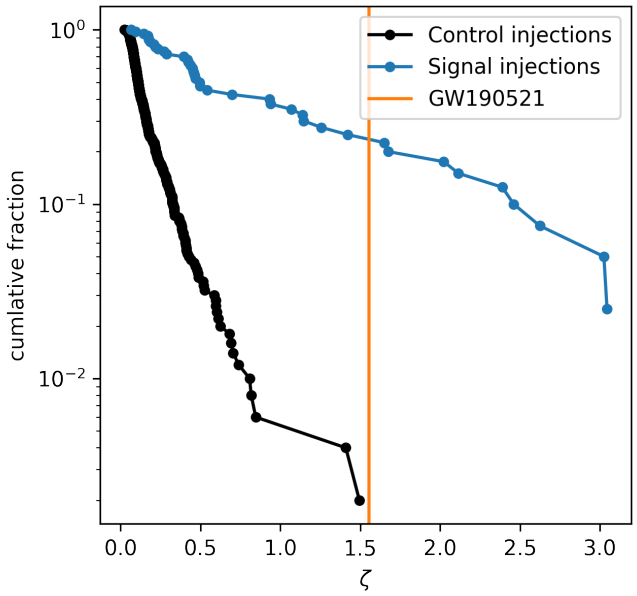

FIG. 4. Cumulative distribution of ζ values for the Control injections (black dots/line) and the Signal injections (blue dots/line). The orange vertical line shows ζ for GW190521. We use the cumulative distribution of Control injections to estimate a p-value for the ζ of GW190521. Since ζ of GW190521 is larger than all Control injections, we estimate its p-value to be < 1 in 500.

V. STATISTICAL SIGNIFICANCE OF THE KERR ANALYSIS

In the Kerr analysis we assume the final black hole is described by the Kerr metric. In this case, the frequencies and damping times of all post-merger QNMs are given uniquely by the mass and spin of the black hole. Each additional mode therefore adds two additional degrees of freedom: one for the amplitude and one for the phase of the mode. The relative amplitudes and phases of the modes can in principle be determined by the pre-merger component masses, their spins, and their relative orientation at merger.

Knowing what amplitude and phase to use for each mode requires detailed knowledge of the pre-merger conditions, which are not easily discernible for events like GW190521 in which the pre-merger signal is short and difficult to observe. Furthermore, models mapping pre-merger properties to post-merger QNMs are limited for highly precessing systems, particularly those with large

[2] Note that ζ would be the Pearson correlation coefficient (without the means subtracted; sometimes referred to as the "reflective correlation") between the measured and expected distributions if we normalized by $\sqrt{\int p^2_{\rm meas}}\sqrt{\int p^2_{330}}$. We in fact tried this, but found poor separation between the Signal and Control injections doing so. This is due to the fact that the expected distribution is more concentrated than the measured distribution, as described above.

($\gtrsim 2$) mass ratios. For these reasons, even when assuming a Kerr model for the post-merger signal, we use uniform priors on the phases and relative amplitudes of the sub-dominant modes with respect to the dominant mode.

Using such broad priors on amplitude and phase makes the analysis susceptible to overfitting. In principle, all modes are present in the signal. However, the vast majority of these modes are negligible compared to the dominant mode. For the types of signals detectable by the current generation of detectors, we expect only a few fundamental modes to have amplitudes that are at most $O(10\%)$ of the dominant mode's [23, 24]. A signal model that contains more than a few modes with such broad priors is effectively unphysical, as it is more likely to fit to noise elements rather than signal (assuming the signal is sufficiently close to GR). To give meaningful results, the signal model should only include the *observable* modes, not the possible ones.

As with the agnostic analysis, the Kerr analysis also needs to determine when the observable modes are present. Before the merger the QNM model is not valid – there is not a single perturbed black hole at this point. During the merger there may be non-linear components to the signal and/or significant contributions from overtones. Too late after the merger, and the signal will have damped away too much to make anything but the dominant mode observable.

We address both challenges through the use of Bayes factors. Given a signal model with observable modes $\vec{X} = \{(2,2,0), ...\}$ at a given time $t - t_{\rm ref}$, we calculate the evidence that the data contain those modes at that time,

$$Z_{\vec{X}}(t) = \int p(\vec{s}|\{\vec{\lambda}_{\vec{X}}; t\}, h) p(\{\vec{\lambda}_{\vec{X}}; t\}|h) \mathrm{d}\vec{\lambda}_{\vec{X}}. \qquad (10)$$

Taking the ratio of this evidence to the evidence for the $(2,2,0)$-only model (Z_{220}) at the same time gives the relative odds (or Bayes factor) that the data favor that model as compared to the $(2,2,0)$-only model.

As discussed above, we do not normalize the likelihood function in our analysis. This means that evidence values at different times cannot be directly compared to each other. However, the likelihood function's normalization factor cancels in the Bayes factor since the normalization only depends on the noise properties and not the signal model. It is therefore possible to compare Bayes factors at different times.

Taking the point that $Z_{\vec{X}}/Z_{220}$ is at a maximum yields the time that the model with modes $\vec{X}$ is the best fit to the data relative to the $(2,2,0)$ model. However, the $(2,2,0)$-only model is known not to be a good model for the signal at merger [27]. As a result, if we find $Z_{\vec{X}}/Z_{220}$ to be large at some time, it is not clear if this is because modes $\vec{X}$ are a good model for the signal, or if the $(2,2,0)$-only model is just a very bad model at that time. Put another way, $Z_{\vec{X}}/Z_{220}$ only tells us whether the $\vec{X}$ modes are a better fit for the data than just the $(2,2,0)$, not whether the $\vec{X}$-modes are truly observable. This problem becomes particularly acute as we get close to merger.

To account for this, we make use of the observation in Refs. [27] that including overtones of the dominant mode better fit the signal close to (or even at) merger than the $(2,2,0)$-mode only. We modify the Bayes factor to be

$$\mathcal{B}(X,t) \equiv \frac{Z_X(t)}{\max\{Z_{220}, Z_{220+221}\}} \qquad (11)$$

for all models $X \neq (2,2,0) + (2,2,1)$ (for the $(2,2,0) + (2,2,1)$ model we simply use $\mathcal{B} = Z_{220+221}/Z_{220}$). This allows us to both identify the most likely observable modes and the time at which they are most observable.

When applying this method to GW190521 we find $\mathcal{B}(220+330)$ to peak at $t_{\rm ref}+6\,$ms with a value of 56 ± 1. This means that the $(2,2,0) + (3,3,0)$ model is 56 times more likely to be true than the $(2,2,0)$-only model, qualifying it as "strong" evidence. Put another way, if the signal did not have an observable $(3,3,0)$ mode, then we should expect to get a $\mathcal{B}$ as large as this from noise only 1 in 56 times.

To test the validity of this observation, we repeat the Kerr analysis on our Control injections. As with GW190521, we repeat the analysis on a grid of times spanning $t_{\rm ref} + [-9, 24]\,$ms, although to reduce computational cost for the large number of analyses involved, we sample in intervals of 3 ms instead of the 1 ms interval used in Capano et al. Since our Control injections contain no $(3,3,0)$ mode by construction, any large $\mathcal{B}$ observed with them is a false alarm. If our analysis assumptions are correct – that the real data is Gaussian and that we are after the merger – then on average we expect to get a $\mathcal{B} \geq 56$ from 8.9 of the 500 injections.

Figure 5 shows the cumulative fraction of Control injections that yield Bayes factors larger than the value given on the x-axis. For larger $\mathcal{B}$, we expect the distribution to follow the line $1/x$. We show two results: one in which we maximize $\mathcal{B}$ over all times tested, $t - t_{\rm ref} \in [-9, 24]\,$ms and one in which we maximize over times $t - t_{\rm ref} \in [0, 24]\,$ms. When maximizing over all times, we find that the injected distribution does not follow the expected distribution of $1/x$; more injections yield large Bayes factors than expected from noise. At the Bayes factor found for GW190521 (56 ± 1), 16 of the injections yield larger Bayes factor, whereas we expect ~ 9.[3] However, when maximizing over times $t - t_{\rm ref} \geq 0$, the injections show remarkable agreement with the expected distribution. Indeed, we find 10 Control injections yield a $\mathcal{B} \geq 56 \pm 1$.

Maximizing over all grid times yields an excess of large Bayes factors because the negative times include times

[3] When maximizing over all time, we use 497 injections instead of 500. This is because one time point failed to converge for three of the injections. For all three injections, this time point was before $t_{\rm ref}$, which is why we are able to use all 500 injections when maximizing over $t \geq t_{\rm ref}$.

before merger for all of the injections (note the distribution of merger times for the NRSurrogate results in Fig. 2). As stated above, before merger the signal is not a superposition of QNMs. This breaks one of our assumptions above. Thought another way – anything not modeled by our signal model is "noise"; in the pre-merger regime the "noise" is not Gaussian, and so larger Bayes factors can be obtained than otherwise expected.

However, this only happens if we sample *before* the merger. By maximizing over $t - t_{\rm ref} \in [0, 24]$ ms we are in the post-merger regime for 174 of the 500 the injections. In this case, we get good agreement with the expectations. We find similarly good agreement if we use our knowledge of the injections' coalescence time to only maximize over grid points that occur after t_c. Doing so introduces complications due to the fact that a different number of grid points is maximized over for each injection; see Appendix B for more details.

In order for the larger-than-expected false alarm rate to apply to GW190521, the time at which the maximum Bayes factor occurred ($t_{\rm ref} + 6$ ms) would have to have been *before* the merger. Of the 500 Control injections only 15 had coalescence times after $t_{\rm ref} + 6$ ms. We therefore conclude this scenario to be unlikely, and use the result when maximizing over $t_{\rm ref} \geq 0$. Given the excellent agreement between expectations and measurement, we conclude that our measured $(3, 3, 0)$ Bayes factor for GW190521 is valid.

To check that our code can recover large Bayes factors when a signal actually has a $(3, 3, 0)$ mode, we repeat this analysis on the Signal injections. The result is summarized in Fig. 6. As expected, the cumulative distribution of Bayes factors for Signal injections does not follow the distribution expected from noise. We also find there to be little difference between maximizing over $t - t_{\rm ref} \in [-9, 24]$ ms and $t \geq t_{\rm ref}$. Of the 45 Signal injections, 22 have Bayes factors larger than GW190521 when maximized over $t \geq t_{\rm ref}$ while 23 have larger Bayes factors when maximized over all times.

That the pipeline recovers large Bayes factors when the signal is present further validates its ability to recover the $(3, 3, 0)$ mode from a signal if it is present.

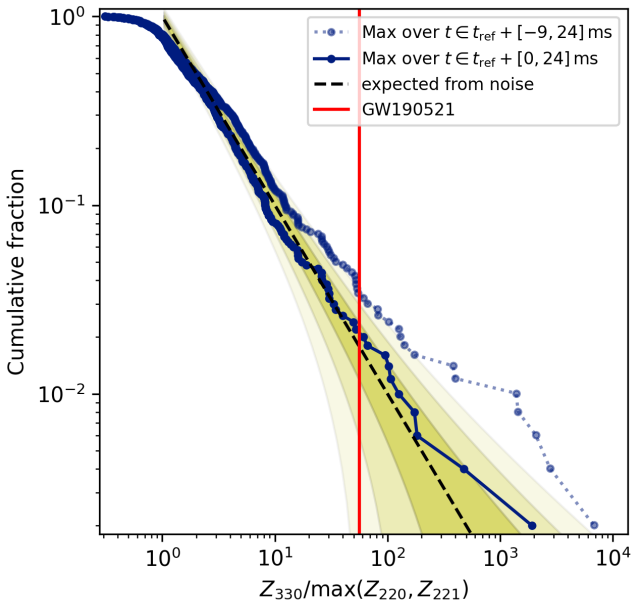

FIG. 5. Cumulative fraction of Control injections (i.e., ones without a $(3, 3, 0)$ mode in the ringdown) versus $(3, 3, 0)$ Bayes factor maximized over time. Since these injections have no $(3, 3, 0)$ mode in the ringdown, we expect the cumulative distribution of Bayes factors $\gtrsim 2$ when calculated after the merger to follow the black-dashed line. Shaded yellow regions show the 1-3σ deviation regions. The dark blue markers/line show the cumulative distribution of Bayes factors for the Control injections when maximized over time steps $\geq t_{\rm ref}$. For all injections $t \geq t_{\rm ref}$ was after the merger. This line follows the expected distribution. Light blue markers/lines show the distribution of Bayes factors when maximized over all times, including before merger. We get an elevated set of Bayes factors in this case, since the ringdown model is no longer valid before merger. The red vertical line shows the maximized Bayes factor for GW190521 (56 ± 1), which occurred at $t_{\rm ref} + 6$ ms. On average, we expect 8.9 out of the 500 injections to have a Bayes factor greater than this; we find 10 when maximized over $t_{\rm ref} \geq 0$. Based on this, we conclude that the quoted Bayes factor for GW190521 is statistically sound.

VI. TESTS OF THE NO-HAIR THEOREM

With more than one ringdown mode, a non-trivial test of the black hole no-hair theorem can be performed [4]. Here we parameterize this test through the deviations $\delta f_{\ell mn}$, $\delta \tau_{\ell mn}$ associated with the measured frequency and the damping time of the sub-dominant mode or the overtone.

The deviation parameters are defined in terms of the measured dominant mode parameters f_{220}, τ_{220} by $f_{\ell mn} = (1 + \delta f_{\ell mn}) f_{\ell mn}(f_{220}, \tau_{220})$ and equivalently for $\tau_{\ell mn}$. The mappings $f_{\ell mn}(f_{220}, \tau_{220})$ and $\tau_{\ell mn}(f_{220}, \tau_{220})$ are given by the relation between the black hole's final mass and spin and each individual mode's frequency and damping time, calculated using the `pykerr` package [58]. We can perform this test either with the $(\ell, m, n) = (3, 3, 0)$ mode or the $(\ell, m, n) = (2, 2, 1)$ overtone of the dominant $(2, 2, 0)$ mode.

The parameter estimation is now performed on the simulated signals of set Control and set Signal as in the Kerr analysis, but adding the deviation parameters ($\delta f_{\ell mn}$ and $\delta \tau_{\ell mn}$) to the set of varied parameters. For the injection set Control, we add $\delta f_{221}, \delta \tau_{221}$ and similarly for injection set Signal we consider $\delta f_{330}, \delta \tau_{330}$ as two additional parameters in the respective analyses. We use uniform prior distributions for the deviation parameters with $\delta f_{221} \in [-0.16, 0.3], \delta \tau_{221} \in [-0.8, 0.8]$ and $\delta f_{330} \in [-0.3, 0.3], \delta \tau_{330} \in [-0.9, 3]$, (see Table II). For the $(\ell, m, n) = (2, 2, 1)$ analysis, we use the Controlinjections. Their simulated signals are generated from the waveform approximant NRSur7dq4, including only the

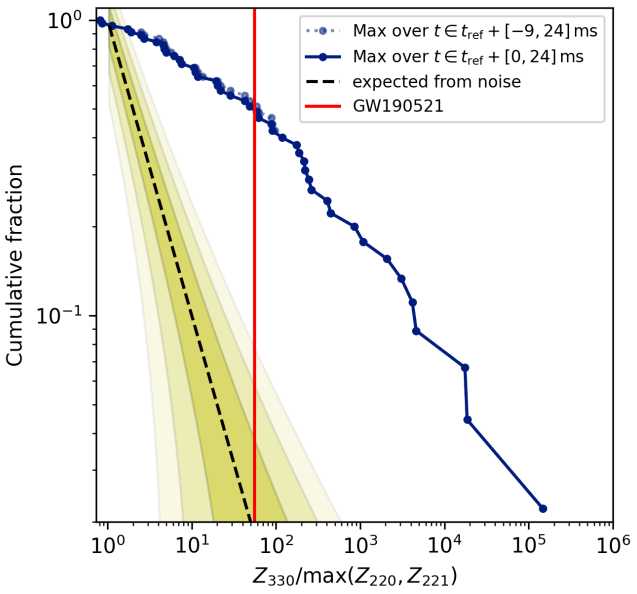

FIG. 6. Cumulative fraction of $(3,3,0)$ Bayes factors (maximized over time) for the Signal injections that have an $\ell, m = (3,3)$ post-merger SNR of at least 4. As in Fig. 5, the black dashed line and shaded regions show the expected distribution from noise. Unlike for the Control injections, we find the Bayes factors of this set of 45 injections to stay substantially above the noise background. The red vertical line shows the maximized Bayes factor for GW190521. When maximized over $t \geq t_{\rm ref}$, 22 of the 45 injections have a Bayes factor greater than GW190521. This indicates that our pipeline is capable of detecting large Bayes factors when a $(3,3,0)$ mode is present.

$\ell = 2$ modes. For the $(\ell, m, n) = (3,3,0)$ analysis, we use the Signalinjections, with simulated signals generated with IMRPhenomTPHM including all its available modes. All injections are thus consistent with GR, and should show no deviation, $\delta f_{\ell mn} = \delta\tau_{\ell mn} = 0$.

We apply the overtone analysis to injections in set Control at the coalescence time t_c. For the injections in set Signal, we perform the test at the time where the largest Bayes factor in favor of the subdominant $(3,3,0)$ mode in addition to the dominant $(2,2,0)$ mode was found in the Kerr analysis in section V, where no deviations were allowed. Our results from the overtone analyses performed on the injections in set Control and the subdominant mode analyses on set Signal are presented in Fig. 7. In the figure we show the 50%-credible regions of the posterior distributions in $\delta f_{\ell mn}$ and $\delta\tau_{\ell mn}$, and the one-dimensional marginalized distributions for several events. We do not include the 90% contours here since they typically cover the whole prior range in the $(2,2,1)$ analysis.

In Fig. 7, for the $(3,3,0)$ mode analyses, results from 44 events with post-merger SNR ≥ 4 are shown[4], while for the $(2,2,1)$ overtone analyses, we draw at random 40 events from those with $\mathcal{B} \geq 100$. These show that the frequency and damping time of the $(3,3,0)$ mode is more accurately recovered than the frequency and damping time of the $(2,2,1)$ overtone, even though wider priors were allowed in the $(3,3,0)$ analysis. The damping time posteriors are broad for both analyses, yet more tightly constrained for the $(3,3,0)$ study. The subdominant $(3,3,0)$ mode's frequency is constrained more tightly compared to that of the $(2,2,1)$ overtone. Additionally, the number of $(3,3,0)$ frequency posteriors centered on the expected value $\delta f_{330} = 0$ increases with the Bayes factors, while for the $(2,2,1)$ frequency posteriors, $\delta f_{221} < 0$ is preferred for large Bayes factors. For several $(2,2,1)$ injections, we observe a bimodal posterior distribution for the overtone frequency deviation. However, the recovery of the $(3,3,0)$ damping time for a single injection shows a bias and the $\delta\tau_{330}$ posterior is centered away from zero even though this injection has a large Bayes factor. We expect this shift in the recovered damping time to be due to the presence of even higher modes, but defer study of this feature to future investigations. We find that tests for deviations of subdominant mode parameters are favored by use of the $(3,3,0)$ mode compared to the $(2,2,1)$ mode for this type of event. We also leave further investigations of full parameter recovery of black hole spectroscopy to subsequent papers.

VII. CONCLUSIONS

Our results here support the conclusion that the ringdown signal of GW190521 contains an observable 33 mode in addition to the dominant 22 mode.

We find a Bayes factor 56 ± 1 for the likelihood of two Kerr ringdown modes over just one mode in the data of GW190521. In the simulated tests of section V, for 500 simulated signals with higher ringdown modes explicitly turned off, only 10 were recovered with a Bayes factor higher than 56 ± 1. Thus a statistic at least as significant as GW190521 occurs once in 50 times. We have in addition demonstrated that the Bayes factor can detect a $(3,3,0)$ mode when it is present in the data.

For future events, our results in section IV for the agnostic test suggest that an improved detection statistic may be better able to detect multiple modes in the data. This involves comparing the consistency of the two-dimensional frequency and damping-time likelihood of the second mode, with that predicted by the frequency and damping-time likelihood from the first mode, assuming that they are the $(2,2,0)$ and $(3,3,0)$ modes of a Kerr black hole.

Our results for the no hair theorem test in VI indicate that using two fundamental modes may perform more re-

[4] In total, 45 injections in set Signal meet the criterion of having post-merger SNR ≥ 4. We show results for 44 of these as the sampler did not converge for one of the injections.

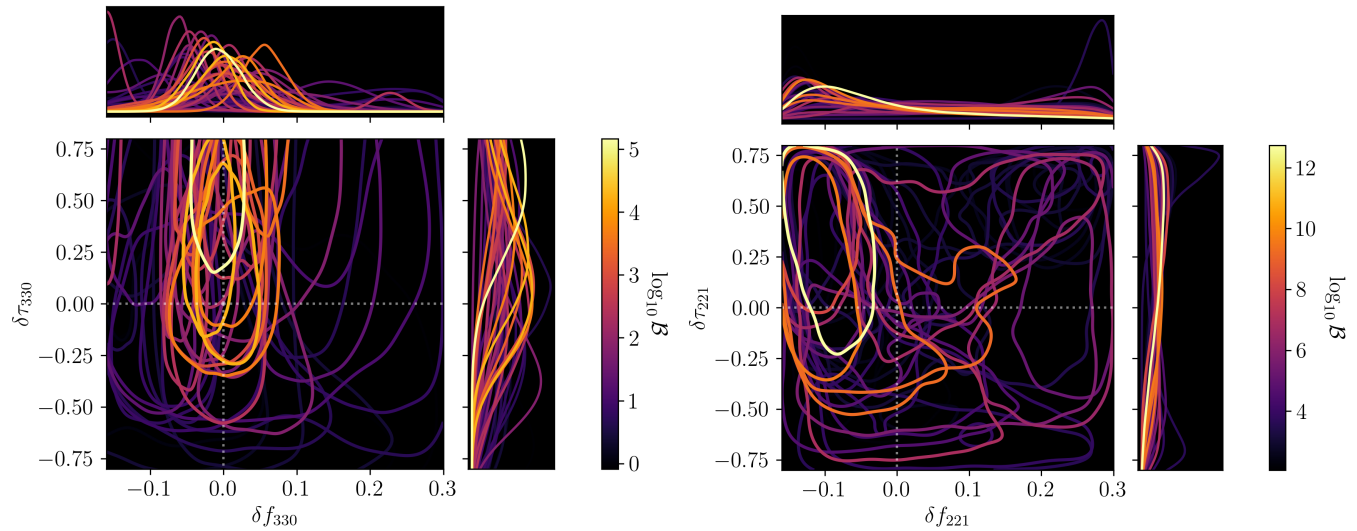

FIG. 7. Each plot shows the contours of the 50% credible regions for the deviation parameters $\delta f_{\ell mn}, \delta\tau_{\ell mn}$, and the corresponding 1-D marginalized distributions for several injections. The colors show the log Bayes-factor $\log_{10}\mathcal{B}$ in favor of the presence of the respective subdominant mode or overtone in addition to the dominant $(2,2,0)$-mode. Results are shown for the analyses considering the $(3,3,0)$ subdominant mode at the time of the highest Bayes factor on the left, for all injections with SNR ≥ 4 in the subdominant mode. The right plot shows the results for the analysis using the $(2,2,1)$ overtone, for 40 randomly chosen injections with $\mathcal{B} \geq 100$. Dotted lines mark the values of the injected signals, $\delta f_{\ell mn} = \delta\tau_{\ell mn} = 0$. Note the axes' ranges, showing the $(2,2,1)$ posteriors often peaking at negative values of δf_{221} and positive $\delta\tau_{221}$, while the $(3,3,0)$ posteriors are typically centered around $\delta f_{330} = 0$.

liably than using the dominant mode and a single overtone. However, there still remain some issues when using two fundamental modes over a very broad range of astrophysical parameters. We leave a full investigation of these features to future work.

[1] L. Andersson, T. Bäckdahl, P. Blue and S. Ma, [arXiv:1903.03859 [math.AP]].
[2] B. F. Whiting, J. Math. Phys. **30** (1989), 1301
[3] E. Berti, V. Cardoso and A. O. Starinets, Class. Quant. Grav. **26** (2009), 163001 [arXiv:0905.2975 [gr-qc]].
[4] O. Dreyer, B. J. Kelly, B. Krishnan, L. S. Finn, D. Garrison and R. Lopez-Aleman, Class. Quant. Grav. **21** (2004), 787-804 [arXiv:gr-qc/0309007 [gr-qc]].
[5] I. Kamaretsos, M. Hannam, S. Husa and B. S. Sathyaprakash, Phys. Rev. D **85** (2012), 024018 [arXiv:1107.0854 [gr-qc]].
[6] C. V. Vishveshwara, Nature **227**, 936-938 (1970)
[7] C. V. Vishveshwara, Phys. Rev. D **1**, 2870-2879 (1970)
[8] S. Chandrasekhar and S. L. Detweiler, Proc. Roy. Soc. Lond. A **344**, 441-452 (1975)
[9] E. Thrane, P. D. Lasky and Y. Levin, Phys. Rev. D **96**, no.10, 102004 (2017) [arXiv:1706.05152 [gr-qc]].
[10] S. Bhagwat, M. Okounkova, S. W. Ballmer, D. A. Brown, M. Giesler, M. A. Scheel and S. A. Teukolsky, Phys. Rev. D **97**, no.10, 104065 (2018) [arXiv:1711.00926 [gr-qc]].
[11] M. Okounkova, [arXiv:2004.00671 [gr-qc]].
[12] K. Mitman, M. Lagos, L. C. Stein, S. Ma, L. Hui, Y. Chen, N. Deppe, F. Hébert, L. E. Kidder and J. Moxon, *et al.* [arXiv:2208.07380 [gr-qc]].
[13] M. Lagos and L. Hui, [arXiv:2208.07379 [gr-qc]].
[14] M. H. Y. Cheung, V. Baibhav, E. Berti, V. Cardoso, G. Carullo, R. Cotesta, W. Del Pozzo, F. Duque, T. Helfer and E. Shukla, *et al.* [arXiv:2208.07374 [gr-qc]].
[15] J. L. Jaramillo, R. P. Macedo, P. Moesta and L. Rezzolla, AIP Conf. Proc. **1458**, no.1, 158-173 (2012) [arXiv:1205.3902 [gr-qc]].
[16] J. L. Jaramillo, R. Panosso Macedo, P. Moesta and L. Rezzolla, Phys. Rev. D **85**, 084030 (2012) [arXiv:1108.0060 [gr-qc]].
[17] V. Prasad, A. Gupta, S. Bose, B. Krishnan and E. Schnetter, Phys. Rev. Lett. **125**, no.12, 121101 (2020) [arXiv:2003.06215 [gr-qc]].
[18] P. Mourier, X. Jiménez Forteza, D. Pook-Kolb, B. Krishnan and E. Schnetter, Phys. Rev. D **103**, no.4, 044054 (2021) [arXiv:2010.15186 [gr-qc]].
[19] X. J. Forteza and P. Mourier, Phys. Rev. D **104**, no.12, 124072 (2021) [arXiv:2107.11829 [gr-qc]].
[20] D. Pook-Kolb, O. Birnholtz, J. L. Jaramillo, B. Krishnan and E. Schnetter, [arXiv:2006.03940 [gr-qc]].
[21] A. Gupta, B. Krishnan, A. Nielsen and E. Schnetter, Phys. Rev. D **97**, no.8, 084028 (2018) [arXiv:1801.07048 [gr-qc]].
[22] Y. Chen, P. Kumar, N. Khera, N. Deppe, A. Dhani, M. Boyle, M. Giesler, L. E. Kidder, H. P. Pfeiffer and M. A. Scheel, *et al.* [arXiv:2208.02965 [gr-qc]].

[23] E. Berti, A. Sesana, E. Barausse, V. Cardoso and K. Belczynski, Phys. Rev. Lett. **117** (2016) no.10, 101102 [arXiv:1605.09286 [gr-qc]].

[24] M. Cabero, J. Westerweck, C. D. Capano, S. Kumar, A. B. Nielsen and B. Krishnan, Phys. Rev. D **101** (2020) no.6, 064044 [arXiv:1911.01361 [gr-qc]].

[25] S. Borhanian, K. G. Arun, H. P. Pfeiffer and B. S. Sathyaprakash, Class. Quant. Grav. **37** (2020) no.6, 065006 [arXiv:1901.08516 [gr-qc]].

[26] M. Isi, M. Giesler, W. M. Farr, M. A. Scheel and S. A. Teukolsky, Phys. Rev. Lett. **123** (2019) no.11, 111102 [arXiv:1905.00869 [gr-qc]].

[27] M. Giesler, M. Isi, M. A. Scheel and S. Teukolsky, Phys. Rev. X **9** (2019) no.4, 041060 [arXiv:1903.08284 [gr-qc]].

[28] R. Cotesta, G. Carullo, E. Berti and V. Cardoso, [arXiv:2201.00822 [gr-qc]].

[29] M. Isi and W. M. Farr, [arXiv:2202.02941 [gr-qc]].

[30] H. P. Nollert, Phys. Rev. D **53**, 4397-4402 (1996) [arXiv:gr-qc/9602032 [gr-qc]].

[31] H. P. Nollert and R. H. Price, J. Math. Phys. **40**, 980-1010 (1999) [arXiv:gr-qc/9810074 [gr-qc]].

[32] J. L. Jaramillo, R. Panosso Macedo and L. Al Sheikh, Phys. Rev. X **11**, no.3, 031003 (2021) [arXiv:2004.06434 [gr-qc]].

[33] J. L. Jaramillo, R. Panosso Macedo and L. A. Sheikh, Phys. Rev. Lett. **128**, no.21, 211102 (2022) [arXiv:2105.03451 [gr-qc]].

[34] K. Destounis, R. P. Macedo, E. Berti, V. Cardoso and J. L. Jaramillo, Phys. Rev. D **104**, no.8, 084091 (2021) [arXiv:2107.09673 [gr-qc]].

[35] C. D. Capano, M. Cabero, J. Westerweck, J. Abedi, S. Kastha, A. H. Nitz, A. B. Nielsen and B. Krishnan, [arXiv:2105.05238 [gr-qc]].

[36] R. Abbott *et al.* [LIGO Scientific and Virgo], Phys. Rev. Lett. **125** (2020) no.10, 101102 [arXiv:2009.01075 [gr-qc]].

[37] R. Abbott *et al.* [LIGO Scientific and Virgo], Astrophys. J. Lett. **900** (2020) no.1, L13 [arXiv:2009.01190 [astro-ph.HE]].

[38] C. M. Biwer, C. D. Capano, S. De, M. Cabero, D. A. Brown, A. H. Nitz and V. Raymond, Publ. Astron. Soc. Pac. **131** (2019) no.996, 024503 [arXiv:1807.10312 [astro-ph.IM]].

[39] J. C. Bustillo, N. Sanchis-Gual, A. Torres-Forné, J. A. Font, A. Vajpeyi, R. Smith, C. Herdeiro, E. Radu and S. H. W. Leong, Phys. Rev. Lett. **126** (2021) no.8, 081101 [arXiv:2009.05376 [gr-qc]].

[40] J. Abedi, L. F. L. Micchi and N. Afshordi, [arXiv:2201.00047 [gr-qc]].

[41] M. Shibata, K. Kiuchi, S. Fujibayashi and Y. Sekiguchi, Phys. Rev. D **103**, no.6, 063037 (2021) doi:10.1103/PhysRevD.103.063037 [arXiv:2101.05440 [astro-ph.HE]].

[42] Y. F. Wang, S. M. Brown, L. Shao and W. Zhao, [arXiv:2109.09718 [astro-ph.HE]].

[43] R. Gamba, M. Breschi, G. Carullo, P. Rettegno, S. Albanesi, S. Bernuzzi and A. Nagar, [arXiv:2106.05575 [gr-qc]].

[44] M. Dall'Amico, M. Mapelli, U. N. Di Carlo, Y. Bouffanais, S. Rastello, F. Santoliquido, A. Ballone and M. A. Sedda, Mon. Not. Roy. Astron. Soc. **508** (2021) no.2, 3045-3054 [arXiv:2105.12757 [astro-ph.HE]].

[45] R. Abbott *et al.* [LIGO Scientific, VIRGO and KAGRA], [arXiv:2111.03606 [gr-qc]].

[46] A. H. Nitz, S. Kumar, Y. F. Wang, S. Kastha, S. Wu, M. Schäfer, R. Dhurkunde and C. D. Capano, [arXiv:2112.06878 [astro-ph.HE]].

[47] M. Cabero, C. D. Capano, O. Fischer-Birnholtz, B. Krishnan, A. B. Nielsen, A. H. Nitz and C. M. Biwer, Phys. Rev. D **97** (2018) no.12, 124069 [arXiv:1711.09073 [gr-qc]].

[48] S. Kastha, C. D. Capano, J. Westerweck, M. Cabero, B. Krishnan and A. B. Nielsen, Phys. Rev. D **105** (2022) no.6, 064042 [arXiv:2111.13664 [gr-qc]].

[49] H. Estellés, S. Husa, M. Colleoni, M. Mateu-Lucena, M. d. Planas, C. García-Quirós, D. Keitel, A. Ramos-Buades, A. K. Mehta and A. Buonanno, *et al.* Astrophys. J. **924** (2022) no.2, 79 [arXiv:2105.06360 [gr-qc]].

[50] A. H. Nitz and C. D. Capano, Astrophys. J. Lett. **907** (2021) no.1, L9 [arXiv:2010.12558 [astro-ph.HE]].

[51] `https://github.com/gwastro/pycbc`

[52] J. S. Speagle, Mon. Not. Roy. Astron. Soc. **493** (2020) no.3, 3132-3158 [arXiv:1904.02180 [astro-ph.IM]].

[53] R. Abbott *et al.* [LIGO Scientific and Virgo], SoftwareX **13** (2021), 100658 [arXiv:1912.11716 [gr-qc]].

[54] L. S. Finn, Phys. Rev. D **46** (1992), 5236-5249 [arXiv:gr-qc/9209010 [gr-qc]].

[55] R. E. Kass, Robert E. and A. E. Raftery, J. Am. Statist. Assoc. **90** (1995), 773

[56] B. Zackay, T. Venumadhav, J. Roulet, L. Dai and M. Zaldarriaga, Phys. Rev. D **104** (2021) no.6, 063034 [arXiv:1908.05644 [astro-ph.IM]].

[57] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, D. Gerosa, L. C. Stein, L. E. Kidder and H. P. Pfeiffer, Phys. Rev. Research. **1** (2019), 033015 [arXiv:1905.09300 [gr-qc]].

[58] C. D. Capano, 2021, pykerr, https://github.com/cdcapano/pykerr, GitHub

[59] A. H. Nitz, C. D. Capano, S. Kumar, Y. F. Wang, S. Kastha, M. Schäfer, R. Dhurkunde and M. Cabero, Astrophys. J. **922**, no.1, 76 (2021) [arXiv:2105.09151 [astro-ph.HE]].

Appendix A: Effect of polarization marginalization on the Bayes factor

In the initial analysis in Capano et al. [35] we used `dynesty` to sample over all parameters for the Kerr analysis listed in Table II. We found a maximum Bayes factor of 44^{+6}_{-5} in favor of the $(2,2,0)+(3,3,0)$ model at $t_{\rm ref}+7\,$ms. However, this method proved time-consuming as the sampler struggled to converge for some mode combinations. The difficulty largely arises from the combination of the phases of the modes and the polarization angle. In particular, for GW190521 the phase of the dominant mode and the polarization are degenerate, as the polarization is not measured well due to the low SNR in the Virgo detector. This results in a banding pattern in the marginal likelihood between these parameters that is a challenge to sample.

Sampling over all parameters would have been unfeasible for the large number of injections we analyzed here. We therefore introduced a modified gating-and-inpainting model that numerically marginalized over the

polarization using 1000 grid points. This marginalization technique was employed in the 3-OGC [59] and 4-OGC analyses [46], where it was found to speed convergence for full IMR templates with sub-dominant modes. We are able to apply the same technique here because the dependence on the polarization is approximately constant over time for a short-duration event like GW190521, and so can be separated from the gating-and-in-painting procedure.

In implementing the polarization marginalization, we discovered that we obtained a larger Bayes factor for GW190521 one ms earlier, at $t_{\rm ref} + 6\,{\rm ms}$. To verify this, we repeated the +6 ms and +7 ms analysis 10 times using different starting seeds. We also repeated each analysis once with double the number of live points. We found consistently larger values at +6 ms. Averaging the Bayes factors over the runs we obtained 56 ± 1 at +6 ms and 45 ± 1 at +7 ms, where the uncertainty is reported with 1σ. We further verified these Bayes factors by using the Savage-Dickey ratio on the $(3, 3, 0)$ amplitude posterior to estimate the Bayes factor, and obtained similar results as reported by `dynesty`'s estimate.

The result at 7 ms was consistent with our initial result in Capano et al., but the result at 6 ms was substantially higher. Our initial estimate for the Bayes factor at +6 ms (without marginalization) was 40^{+5}_{-4}. Evidently, without marginalization, the sampler had not fully converged at 6 ms, yielding an underestimate of the Bayes factor. Marginalization also affected our $(2, 2, 1)$ results: we found the Bayes factor for the $(2, 2, 1)$ mode peaked slightly earlier, at $t_{\rm ref} - 7\,{\rm ms}$ instead of the $t_{\rm ref} - 5\,{\rm ms}$ that we initially estimated.

Given the robustness of the new results under polarization marginalization, we quote the updated Bayes factor at $t_{\rm ref} + 6\,{\rm ms}$ here for GW190521.

Appendix B: Maximizing the Kerr Bayes factor after merger

As discussed in Sec. V, we obtain good agreement between the expected distribution of Bayes factors and the measured distribution if we restrict the maximization interval to be strictly after the Control injections' coalescence time t_c. The result is shown in the left plot of Fig. 8. Above Bayes factors of ~ 20 we find excellent agreement with the background. Indeed, we find that 9 of the 500 injections have a Bayes factor larger than GW190521, exactly the amount expected by chance.

However, for Bayes factors $\lesssim 20$ there is a nearly 3σ downward deviation in the measured background. This deviation is due to the fact that differing numbers of grid points are maximized over when using the injection's coalescence time. For example, the maximization interval spans nine grid points (spanning $t_{\rm ref} + [0, 24]\,{\rm ms}$) for injections that have a $t_c \approx t_{\rm ref}$, whereas the interval is only two grid points for injections with $t_c \approx t_{\rm ref} + 21$. Although grid points are not independent of each other – if a large Bayes factor exists at a particular point in time, there is a higher probability that its neighbors will also have larger Bayes factors – they are not entirely dependent either. Due to the stochastic nature of the noise, there are random fluctuations in Bayes factors across time. Consequently, if a maximization interval covers fewer grid points, there is less opportunities to obtain larger Bayes factors.

Large Bayes factors are not strongly affected by differences in maximization interval, since there is a low probability that a noise fluctuation could produce a larger Bayes factor. This is evident in the left plot of Fig. 8. Conversely, smaller Bayes factors will be affected by this, hence the deviation at lower Bayes factors in that plot.

This issue can be corrected for by multiplying the Bayes factors of each injection by $(\max N_{\rm grid})/N_{\rm grid}$, where $N_{\rm grid}$ is the number of grid points maximized over for the given injection and $\max N_{\rm grid}$ is the largest number of grid points maximized over in the set. Renormalizing the Bayes factors yields the result shown in the right plot of Fig. 8. Now we find good agreement with the expected background and measured distribution at all Bayes factors. With this we find 10 Control injections to have a larger Bayes factor when we expect 9.

Note that the normalization factor implicitly assumes that each grid point is independent of the others. As stated above, this is not the case. Since using this factor tends to overestimate the contribution, this is a conservative error.

Due to these complications we present in the main text the simpler maximization over $t_{\rm ref} \geq 0$.

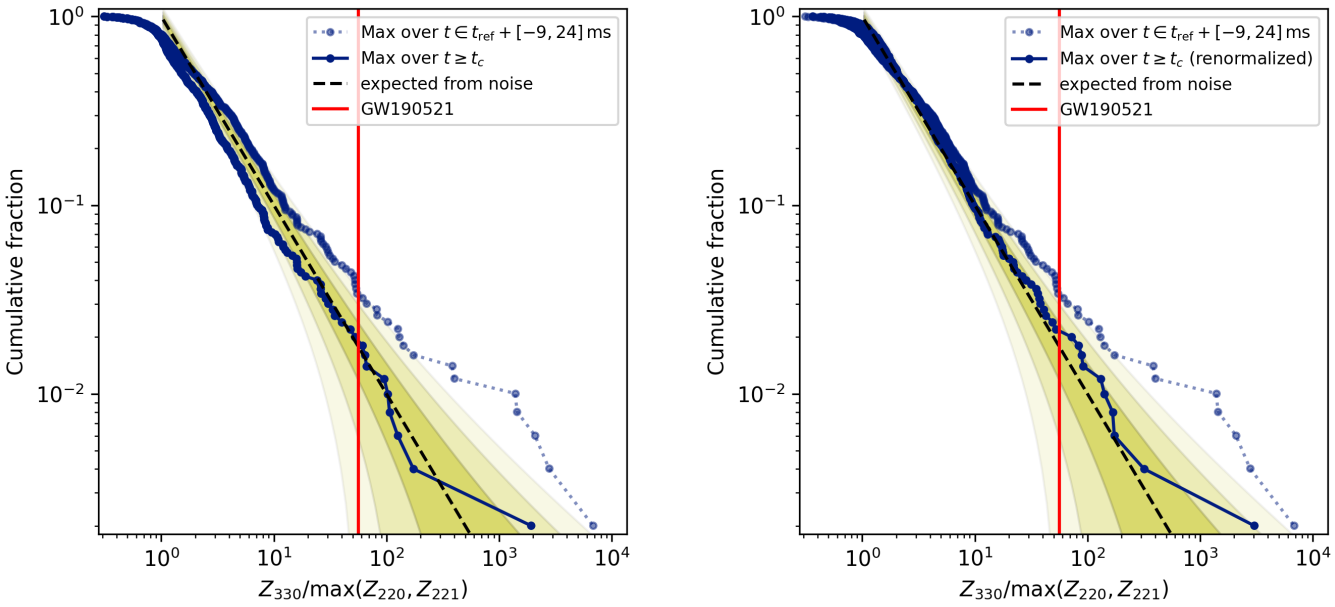

FIG. 8. *Left*: Same as Fig. 5, but with the maximization interval based on the injections' coalescence time t_c (dark blue line). We find excellent agreement with the expected background distribution at large Bayes factors, but a $\sim 3\sigma$ downward deviation in the measured distribution at Bayes factors $\lesssim 20$. This deviation is due to the different maximization range for each injection. *Right:* Renormalized version of the left plot. Here, we've accounted for variations in maximization interval across the Control injections by multiplying their Bayes factor by $(\max N_{\rm grid})/N_{\rm grid}$, where $N_{\rm grid}$ is the number of time points maximized over. We find good agreement with the expected background at both large and small Bayes factors in this case.