

GW231123 ringdown: interpretation as multimodal Kerr signal

Harrison Siegel ^{1,2,3,*} Nicole M. Khusid ^{4,3} Maximiliano Isi ^{5,3} and Will M. Farr ^{4,3}

¹*Perimeter Institute for Theoretical Physics, 31 Caroline St N, Waterloo, ON N2L 2Y5, CA*

²*Department of Physics, Columbia University, 704 Pupin Hall, 538 West 120th Street, New York, New York 10027, USA*

³*Center for Computational Astrophysics, Flatiron Institute, New York NY 10010, USA*

⁴*Department of Physics and Astronomy, Stony Brook University, Stony Brook NY 11794, USA*

⁵*Department of Astronomy and Columbia Astrophysics Laboratory,
Columbia University, Pupin Hall, New York, NY 10027, USA*

GW231123 is a short-duration, low-frequency gravitational wave signal consistent with a binary black hole coalescence and dominated by the merger-ringdown regime due to the high mass of the source. We demonstrate that fits of this ringdown signal using two quasinormal modes are statistically preferred over single-mode fits, for a broad range of fit start times. We also find that two-mode fits give remnant mass and spin measurements consistent with those of the inspiral-merger-ringdown model NRSur7dq4, whereas one-mode fits struggle to do so. Agreement of our fits with those of NRSur7dq4 is achieved by labeling the two quasinormal modes as the $(\ell, m) = (2, 2)$ and $(2, 0)$ Kerr prograde fundamental modes. However, we find some indications that fits with the $(2, 1)$ quasinormal mode instead of the $(2, 0)$ mode may describe the data better, hinting at possible NRSur7dq4 error or other systematics. When fitting at early times near the estimated peak strain, we find that the inclusion of a third mode, an $(\ell, m, n) = (2, 2, 1)$ prograde overtone, improves consistency with fits at later times. Finally, we perform a test of general relativity by searching for deviations from the Kerr frequency spectrum. Setting issues of systematics aside, we validate the Kerr frequency and damping rate spectrum to within $\pm 10\%$ at the 90% credible level using a fundamental mode fit, and we also report $\pm 8\%$ constraints using a model with fundamental modes and an overtone fit at times near the peak strain. Understanding the systematic errors that may be affecting the most accurate analyses of GW231123 is crucial in the context of population and binary formation studies – our $(2, 1)$ mode fits return a significantly higher remnant mass and spin than all available inspiral-merger-ringdown models including NRSur7dq4, and this difference in parameter estimates may have astrophysical implications.

I. INTRODUCTION

GW231123.135430, henceforth GW231123, [1] is a short-duration, low-frequency gravitational wave signal measured by the LIGO-Virgo-KAGRA (LVK) [2–5] collaboration and included in Gravitational-Wave Transient Catalog 4 (GWTC-4) [6]. When the signal is modeled as being sourced by a quasicircular precessing binary black hole (BBH) coalescence using the NRSur7dq4 [7] inspiral-merger-ringdown (IMR) model, the final source-frame black hole remnant mass is found to be $M_f^s = 227^{+18}_{-27} M_\odot$, the primary and secondary binary components have dimensionless spins of $\chi_1 = 0.89^{+0.11}_{-0.20}$ and $\chi_2 = 0.91^{+0.09}_{-0.19}$ respectively, and the matched filter signal-to-noise ratio (SNR) of the signal is $\text{SNR}_{\text{mf}} = 22.6^{+0.3}_{-0.3}$ (reporting median and 90% highest posterior density values, see Table 3 of Ref. [1]).¹ The signal is dominated by its merger-ringdown, due to the high detector-frame mass placing the inspiral regime in the less sensitive lower-frequency band of the LIGO detectors: we estimate the SNR of the signal from the peak of

the strain onwards to be $\text{SNR}_{\text{post-peak}} = 18.0^{+2.0}_{-1.5}$ ² using NRSur7dq4, making the merger-ringdown of GW231123 the loudest of any signals until GW250114 [11, 12].

In the theory of general relativity (GR), the final stage of a BBH coalescence can be described at sufficiently late times by perturbation theory via the Teukolsky equation [13]: this stage is the so-called ringdown. In this perturbative regime, the signal is quickly dominated by quasinormal modes (QNMs), which eventually decay and become subdominant to tails. The QNMs are observed as damped sinusoids possessing frequencies and damping rates determined solely by the mass and spin of the remnant black hole, and having amplitudes and phases related to the binary black hole configuration. Clear observation of more than one QNM in the spectrum emitted by a merger remnant allows for the inference of progenitor properties and validation of the Kerr metric [14–19], and has recently been the subject of intense data analysis

* hs3152@columbia.edu

¹ The IMR posteriors we use in this work come from a version of Ref. [1] with an incorrect likelihood as described in Ref. [8]. The corrected parameter estimates are not sufficiently different to change the main conclusions of our work. See the journal-published version of Ref. [1] for the corrected estimates.

² The post-peak SNR is computed in the time domain, as the distribution of optimal SNRs of 0.2 s duration template segments from their respective peaks onwards generated from 1000 random NRSur7dq4 posterior samples and whitened with the noise autocovariance from our ringdown analysis [9, 10]. Our $\text{SNR}_{\text{post-peak}}$ uses a different noise model from the full-signal SNR_{mf} reported in Ref. [1] (see Sec. II), so direct comparison should not be made, however $\text{SNR}_{\text{post-peak}}$ is the more relevant quantity for our specific ringdown analysis.

efforts [11, 12, 20–34] following the dawn of gravitational-wave astronomy [35–37]. This program of QNM-based inference is often referred to as black hole spectroscopy [38].

Ref. [1] reports in its text on indications that the post-peak signal of GW231123 is better fit by at least two long-lived QNMs as opposed to one. Here we verify these claims with a more complete analysis of the ringdown, using the RINGDOWN code [9, 10, 39] to fit sums of damped sinusoids. Our results are reported with reference to the strain peak time and other parameter estimates given by the NRSur7dq4 IMR model. We show that multi-mode fits are preferred over single-mode fits by statistical goodness of fit metrics over a range of fit start times as late as 20.4 ms after the peak strain (or equivalently $14 t_M$ in units of $t_M = GM_f^d/c^3$, corresponding to the detector-frame remnant mass M_f^d ; see Sec. IA).

We find that two long-lived QNMs are required to get remnant mass and spin measurements in agreement with NRSur7dq4 over almost all times in the signal where QNM fits recover non-zero amplitudes: this mass and spin agreement is achieved by labeling the two long-lived QNMs as the $(\ell, m) = (2, 2)$ and $(2, 0)$ prograde fundamental modes. The amplitudes of the two long-lived modes are comparable. These comparable amplitudes could be attributable to spin-orbit misalignment [17, 40–42] and a preferential viewing angle to the source: but such a configuration would be at least somewhat finely-tuned. Due in part to the sparseness of highly spinning and precessing numerical relativity (NR) BBH simulations, it is hard at present to conclusively determine how astrophysically unlikely it is for the $(2, 0)$ amplitude to be so large.

It is also possible that systematic waveform errors alter the mass and spin found by analysis with the NRSur7dq4 waveform, and therefore the QNM labeling required to match these parameters does not accurately represent the physical modes of the system associated with GW231123. Systematics are more of a concern in GW231123 than in most other gravitational wave signals: in Ref. [1], five different IMR waveforms [7, 43–47] were found to all have significant systematic differences in parameter estimation for GW231123. High precession and heavy masses are known to be sources of systematic error in all current IMR waveforms [1, 7, 48, 49], but we cannot definitively rule out nonstationary noise features or unmodeled physics contributing to the systematic differences between waveforms. Based on the mismatches to NR simulations shown in Refs. [1, 7], NRSur7dq4 most likely outperforms all other available IMR models when fitting GW231123, and NRSur7dq4 is also not itself guaranteed to have systematic errors dominate over its statistical errors in this part of parameter space. For this reason we only explicitly show and make use of NRSur7dq4 parameter estimates, although the findings of this paper broadly hold regardless of which IMR waveform is considered.

To add to concerns of systematic errors, we find some evidence to suggest that GW231123 may actually be bet-

ter fit by the pattern of the Kerr $(\ell, m) = (2, 2)$ and $(2, 1)$ prograde fundamental QNMs, although this model is in tension with the remnant mass and spin estimates of NRSur7dq4 and all other currently available IMR models. The parameter estimates of SEOBNRv5PHM and IMRPhenomTPHM are the closest to these QNM fits, but are still in significant tension. If a case can be more confidently made for the presence of systematic error in even the best-fitting IMR model, this could change the astrophysical interpretation of GW231123, with implications for population and binary formation channel studies using this signal [50–60]. However, more work is still required to make definitive claims regarding NRSur7dq4 systematics for this signal. It is worth noting that GW190521 [61–63], another signal from a heavy and possibly highly precessing BBH, may also show preference for the $(\ell, m) = (2, 2)$ and $(2, 1)$ QNMs as the dominant modes [28]. And it may be the case that precessing systems are more capable of exciting the $(\ell, m) = (2, 2)$ and $(2, 1)$ modes as the dominant QNMs as opposed to the $(\ell, m) = (2, 2)$ and $(2, 0)$ [17], meaning the QNM models which are inconsistent with the IMR models for GW231123 may also be more physically plausible.

In addition to measurements of the two fundamental modes, we also find that the inclusion of an $(\ell, m, n) = (2, 2, 1)$ prograde overtone alongside both fundamental modes may help improve fits by making their parameter estimates and agreement with the Kerr spectrum more consistent over a wider range of fitting times.

In principle, the clear preference for multimodal fits to GW231123 enables tests of the Kerr metric. In practice, the possibility of IMR model systematic error or data quality issues makes interpretation of these tests more difficult. Nonetheless, taking the QNM model which best agrees with the Kerr metric and comparing it at $6 t_M$ to Kerr constraints reported at an equivalent time for GW250114 [11, 12], the best-measured signal to date, we find significantly better constraints: at this fitting time we validate the Kerr frequency and damping rate spectrum of the $(\ell, m) = (2, 1)$ and $(2, 2)$ fundamental modes to within $\pm 10\%$ at the 90% credible level with GW231123, as opposed to the $\pm 30\%$ constraint reported for GW250114. When fitting at earlier times, we find even tighter validations of the Kerr metric which are sub-10% for a model with fundamental and overtone modes.

A key advantage of QNM fits is that their systematics are different from those of the IMR models. The IMR models are built to have rigidly prescribed relationships between inspiral and ringdown [7, 43–46], as well as built-in prescriptions of the possible ringdown mode excitations which are generally phenomenological or derived from fits to the available bank of NR simulations, which may not necessarily span the parameter space of interest [1, 64]: by contrast, our QNM models have the ability to flexibly fit QNM amplitudes and phases without restriction to any specific physical system. So long as the signal is dominated by damped sinusoids with frequencies from the Kerr spectrum (and even some non-Kerr

spectra) our QNM fits should be able to reliably fit all of the signal content. Thus, comparison between the QNM fits we show here and IMR fits can be interpreted as a sanity check of the IMR parameter estimates, especially under the default hypothesis that the signal is from a quasicircular precessing BBH. The tensions we find between QNM and IMR fits highlight the utility of QNM fits as probes of astrophysical parameters in addition to providing tests of general relativity.

The paper is organized as follows. In Sec. I A, we explain our notational conventions for referring to QNMs and times in the signal. In Sec. II we provide a technical description of the process by which we condition the data before analysis. We also address possible concerns of data quality issues affecting GW231123. In Sec. III we report fitting results from a data-driven perspective, for both Kerr and non-Kerr models. In Sec. IV we provide discussion and interpretation of the results, and comment on possible origins of systematic error in analysis of GW231123 including NRSur7dq4 and QNM fitting errors. We conclude in Sec. V. We also include a discussion in Appendix A of QNM fits using the 320 and 321 modes. These models which include the 320 can achieve better goodness-of-fit than the QNM models with the 200 that agree with NRSur7dq4, but they also have seemingly unphysical features which make us disfavor them when compared with the other models in the main text.

A. Notational conventions

In this paper, when referring to individual QNMs we follow the conventions of Ref. [9]. To first order in perturbation theory, individual QNMs can be identified by four indices (p, ℓ, m, n) . The angular structure of the QNMs is described by spin-weighted spheroidal harmonics with angular indices ℓ and m . The radial structure of the QNMs is denoted by the index n , and is also tied to the lifetime of the QNMs: the longest-lived, $n = 0$, QNMs are referred to as fundamental QNMs; and faster-decaying, $n > 0$, QNMs are called overtones. For a given set of (ℓ, m, n) , when $m \neq 0$, the sign of m holds the two polarization degrees of freedom; there are also two distinct QNMs which are labeled by an index $p \equiv \text{sgn}[m \Re(\omega)]$, where $\Re(\omega)$ is the real part of the complex QNM frequency ω , and whose phase fronts are either corotating ($p = +$) or counterrotating ($p = -$) with the black hole. Solutions with $m = 0$ are azimuthally symmetric, so that there is no notion of co- vs counter-rotating fronts and the two possible signs of $\Re(\omega)$ directly encode the two polarization degrees of freedom.

We will frequently make use of the ringdown evolution timescale $t_M = GM_f^d/c^3$, which is defined in units of the final detector-frame remnant mass M_f^d when natural units are taken such that $G = c = 1$. We will use the median detector-frame remnant mass $M_f^d = 296 M_\odot$ inferred by NRSur7dq4 in order to set $t_M = 1.46$ ms. QNM models will be denoted with set notation as comma-

separated lists within braces of all simultaneously fitted modes identified by their $\{\ell, m, n\}$ indices; we only consider prograde QNMs in this work, and so $p = +$ implicitly throughout. Any QNMs given non-Kerr freedoms in a model will be labeled with underlined text. The time at which a given fit is started will be referred to as t_{start} , in units of t_M relative to our estimated peak strain time t_{peak} which is given explicitly in Sec. II.

II. DATA CONDITIONING AND QUALITY

For all results herein our analysis follows a specific set of steps to condition the data we analyze. These steps are carefully chosen to minimize computational expense, without significantly corrupting parameter estimates. The entire conditioning process is outlined below, and a brief consideration of possible data quality issues is also included.

We start from 4096 s long data segments around the time of the detection in both the Hanford and Livingston interferometers, at sample rates of 16384 Hz. From these data segments, we subtract the 60 Hz AC mains noise line from the data [10, 65], a line which overlaps with the central frequencies of the signal. To implement our line subtraction, we use a computationally inexpensive linear algorithm [66]. Because the subtraction involves some filter “warmup” and tapering, this leaves us with a 4080 s long valid data segment. We then downsample this data segment by a factor of 4 to a sample rate of 4096 Hz using our so-called digital filter. When downsampling, a Tukey window is also applied to the 4080 s data segment, which trims the first and last 10% of the data. From this downsampled data we select segments for analysis with durations of 0.2 s as motivated by Ref. [10]. We can analyze relatively short 0.2 s duration segments without losing SNR because of the subtraction of the 60 Hz AC mains line.

Line subtraction for 60 Hz noise was not implemented in production data for GW231123, because the LIGO collaboration’s standard nonlinear line subtraction pipeline was found to elevate noise in the sidebands of the line and this might negatively affect continuous gravitational wave searches. Our particular ringdown analysis experiences significant computational and SNR gains when employing line subtraction, and does not suffer from the noise subtraction in the way that other analyses like the continuous wave searches do, especially since our analysis segment is so short that we do not resolve the sidebands as well as those other longer-duration analyses. Our line subtraction algorithm was tested through injecting damped sinusoid signals into real LIGO data from 1000 to 500 seconds before the actual GW231123 signal, and was found to improve the performance of our analysis [67, 68] (note that these references are to internal LVK documents).

The noise auto-covariance function (ACF) is computed by Fourier transforming a Welch estimate of the PSD.

This Welch estimate uses windowed data segments 16 times the duration of the analysis data segment (i.e. 3.2 s), and is computed over the full 4080 s long data segment. While it has been argued previously in Ref. [69] that long data segments of LIGO noise are subject to PSD estimate drifts, we have not found this empirically to significantly impact our analysis at current SNRs (see e.g. Ref. [68]): this may be due to the short duration and higher SNR of our signals of interest.

The geocenter peak strain GPS time in our analysis is $t_{\text{peak}} = 1384782888.61823 \pm 0.00131$ s, with the median value being used throughout. The sky location of our model determines time delays between the data segments of different interferometers, and is selected by using the sample from the NRSur7dq4 posterior which is closest in time to the median estimated peak strain of the most sensitive interferometer. The sky location in radians for our analysis is fixed at a point estimate of $(\text{ra}, \text{dec}, \text{psi}) = (3.80, 0.64, 1.81)$, when rounded to two decimal places. Note that our t_{peak} is not the same as the time used in Ref. [1]. In that LVK analysis (and also in another GW231123 ringdown analysis, Ref. [70]), the peak time is estimated using the peak of the complex strain at Earth's location on the celestial sphere,

$$t_{\text{peak}}^{\text{LVK}} = \arg \max [h_+^2(t) + h_\times^2(t)]. \quad (1)$$

This time $t_{\text{peak}}^{\text{LVK}}$ is not the one typically considered in theoretical studies of ringdown, which more often consider instead an angle-invariant peak strain time over the whole celestial sphere,

$$t_{\text{peak}}^{\text{invar.}} = \arg \max \sum_{\ell m} H_{\ell m}(t)^2, \quad (2)$$

where $H_{\ell m}(t)$ is related to the complex strain as

$$h_+(t) + ih_\times(t) = \sum_{\ell m} {}_{-2}Y_{\ell m} H_{\ell m}(t), \quad (3)$$

and ${}_{-2}Y_{\ell m}$ are spin-weighted spherical harmonics. Our peak time estimate is obtained by considering the distribution of $t_{\text{peak}}^{\text{invar.}}$ from 1000 random NRSur7dq4 posterior samples.

There were data quality issues throughout the day of the GW231123 detection [71], beyond the non-Gaussian noise features deemed to be inconsequential in Ref. [1]. These additional concerns included notable fluctuations in the binary neutron star observation range throughout the day. The aforementioned injection study which tested our line cleaning algorithm appeared by eye to reliably recover the true parameters of the injected signals in data from this day, suggesting that our analysis is not sensitive to these data quality issues.

See Refs. [9, 10] for more information on our ringdown analysis framework, as well as the data release of this paper [72] for exact analysis settings in configuration files.

III. RESULTS

Here we report on data analysis results, with a focus on data-driven interpretation and minimal reliance on existing understanding of astrophysical QNM amplitudes in BBH coalescences. In Sec. IV we then present stricter physical interpretation of these results *ex post facto*. Given that there is incomplete theoretical knowledge of the regime of validity of QNM fits in BBH coalescences [19, 73–75], we fit our damped sinusoid models starting over a wide range of times around the NRSur7dq4 t_{peak} , both before and after where most works claim the onset of the Kerr perturbative regime to be. We choose fit start times t_{start} in intervals of $2 t_M$ around the estimated peak strain time t_{peak} .

A key tenet of our fitting philosophy is that reasonable QNM models which accurately describe the signal should have consistent parameter estimates when fit at different starting times. We expect that good QNM fits to the data should become self-consistent at some time and remain that way until the SNR decreases to the point of the posteriors being uninformative. We assess this self-consistency predominantly visually, by looking for overlapping posteriors from fits at different times, typically considering 90% credible levels (CLs). We also look for amplitude posteriors for individual QNMs that are consistent with being non-zero within at least the 68% CL, as zero amplitude leads to uninformative posteriors for other associated QNM parameters.

When performing model comparison, we choose to use the leave-one-out cross-validation (LOO) [76, 77].³ For intuition regarding the meaning of the LOO: when using Gaussian likelihood models like ours [9, 10], the LOO is approximately related to the chi-squared test as $\text{LOO} \approx -\frac{1}{2}\chi^2$ plus a penalty term which depends on the leverage of individual data points [78, 79]. Higher LOO values are more preferred. Following Ref. [80], LOO differences greater than 4 are very significant. We deem LOO differences of 1 to be our minimum for noteworthy statistical significance. We choose to use the LOO as opposed to the Bayes factor, which is more common in our research field, because the LOO is in general less sensitive to priors and we prefer this behavior. Regardless of which statistical model selection or goodness of fit criterion is chosen, such quantitative measures only form one part of a larger whole in scientific analysis: we use the LOO here merely as a guide in driving our model explorations and supporting our interpretations and conclusions, rather than as an overwhelming figure of merit [81–83].

³ A technical detail regarding the LOO: we specifically use the Pareto-smoothed importance sampling estimate of the LOO as defined in Eq. 10 of Ref. [76]. The shape parameters we recover for the Pareto distribution suggest that we have reliable estimates of the expected log pointwise predictive density for all fitting times where significant model preferences exist.

The first fitting results are presented in Sec. III A, where we consider a general model made up of damped sinusoids that do not have Kerr frequency and damping rate constraints but are forced to be ordered in real frequency to avoid label switching [84]. This type of “free” or “agnostic” damped sinusoid model is helpful for developing basic exploratory understanding, before honing in on models with physically motivated frequency spectrum constraints. This free model is also closely related to models we use later which allow deviations from the Kerr spectrum of frequencies and damping rates [9, 85].

The free damped sinusoid model has strong statistical preferences for two modes over one, when fitting at start times ranging from before the peak strain to $14 t_M$, as shown in Fig. 1. The statistical uncertainty of the peak time as given by NRSur7dq4 is $\pm 0.9 t_M$. Even when accounting for the statistical uncertainty of the peak time, preferences for multimodal fits persist long after the peak strain. If systematic errors in the peak time estimate dominate over the statistical uncertainty, preferences for multimodal fits may correspond to relative fitting times earlier or even later than those we are using as reference. When fitting two free damped sinusoids, the modes are found to have comparable lifetimes and comparable amplitudes. They have frequencies consistent with the prograde $(\ell, m) = (2, 2)$ and either the $(2, 1)$ or $(2, 0)$ fundamental modes implied by the mass and spin measurements of NRSur7dq4 (Fig. 2).

We then in Sec. III B fit damped sinusoid models with frequency constraints imposed by first-order Kerr metric perturbation theory. When fitting two-mode Kerr models, we find the best remnant mass and spin agreement with NRSur7dq4 by including the $(\ell, m) = (2, 2)$ and $(2, 0)$ fundamental QNMs. The posteriors of this model overlap visually at the 90 % CL with those of NRSur7dq4 when starting QNM fits from $4 t_M$ onward (Fig. 4). However, we do find a goodness of fit preference among two-mode models of at least 1σ for the $\{220, 210\}$ model, until as late as $8 t_M$. This model is also self-consistent over time in both its frequency and damping rate (Fig. 4) and amplitude measurements (Fig. 6), although it is in tension with NRSur7dq4.

We also consider three-mode Kerr models. When fitting a $\{220, 200, 2m1\}$ model, the remnant mass and spin posteriors overlap at the 90% CL with those of NRSur7dq4 starting as early as $0 t_M$. (Fig. 5). Even more strikingly, $\{220, 210, 2m1\}$ models make self-consistent remnant mass and spin parameter estimates going back as early as at least $-8 t_M$, although these models are in tension with NRSur7dq4 remnant mass and spin estimates. The m index when we just referred to the overtones in these models was left intentionally ambiguous, as any m index of the included overtone in the model will achieve qualitatively similar mass and spin posteriors. However, in terms of goodness of fit, there are goodness of fit preferences at some times for $m = 2$ over other overtone m indices: restricting ourselves to $m = 2$ for the overtone, at $-2 t_M$ the $\{220, 210, 221\}$ model is preferred by $\sim 1\sigma$

over all other models we consider, and at earlier times this model is definitively preferred.

Finally, in Sec. III C we perform a test of general relativity using the Kerr models most preferred by statistical goodness of fit and/or in best agreement with NRSur7dq4. Based on this test, it seems that models with the 210 as opposed to the 200 are consistent with the Kerr metric for GW231123 over a broader range of fitting times, and are capable of recovering the Kerr metric with higher precision at the 90% CL than 200 models. Comparing with similar fitting times in GW250114 [11, 12], the loudest signal to date, we find much better constraints and confirm the Kerr spectrum to $\pm 10\%$ using the 210 QNM, and $\pm 40\%$ for the 221 QNM when fit alongside the 210. We report even more precise validation of Kerr when fitting at earlier times. This result, along with some of our observations of goodness of fit and parameter estimate consistency over time for different models, raises questions about whether the NRSur7dq4 model is possibly dominated by systematic error for GW231123, since our analysis seems to have preferences for QNM fits which are in tension with NRSur7dq4. See Table I for a collection of select Kerr constraints.

A. Free damped sinusoid fits

We fit a model composed of free damped sinusoids that are not constrained by the Kerr frequency spectrum but are forced to be ordered in real frequency in order to avoid degenerate sampling. Each damped sinusoid has two polarizations, and flat priors on the frequency (0 to 150 Hz), damping rate (0 to 250 Hz), phase (0 to 2π radians), and amplitude (sampled via the marginalization technique described in Ref. [86], with $a_{\max} = 10^{-19}$).⁴ The damping rate and frequency priors are chosen to encompass both the first overtone and $\ell = 3$ mode parameters suggested by mass and spin measurements of NRSur7dq4.

In Fig. 1 we show model comparisons using the LOO. There is a statistically significant goodness-of-fit preference for two-mode free damped sinusoid fits over one mode fits when starting the fits as late as $14 t_M$ after the peak strain. There is no added support for three modes over two modes using these free damped sinusoid fits.

In frequency and damping rate, the two-mode free damped sinusoid fits overlap at the 90% CL with the prograde $(\ell, m) = (2, 2)$ and either the $(2, 1)$ or $(2, 0)$ fundamental modes of NRSur7dq4, as shown in Fig. 2. Based on this frequency and damping rate behavior as well as

⁴ The flat prior on the amplitude of each mode is placed on the combined amplitude of both polarizations of the mode, not on each individual polarization’s amplitude. See Fig. 3 of Ref. [28] and Fig. 12 of [87] for discussion of this distinction.

the LOO preferences, we are motivated to explore multi-modal Kerr QNM fits with at least two prograde fundamental $\ell = 2$ QNMs, as shown in the next subsection.

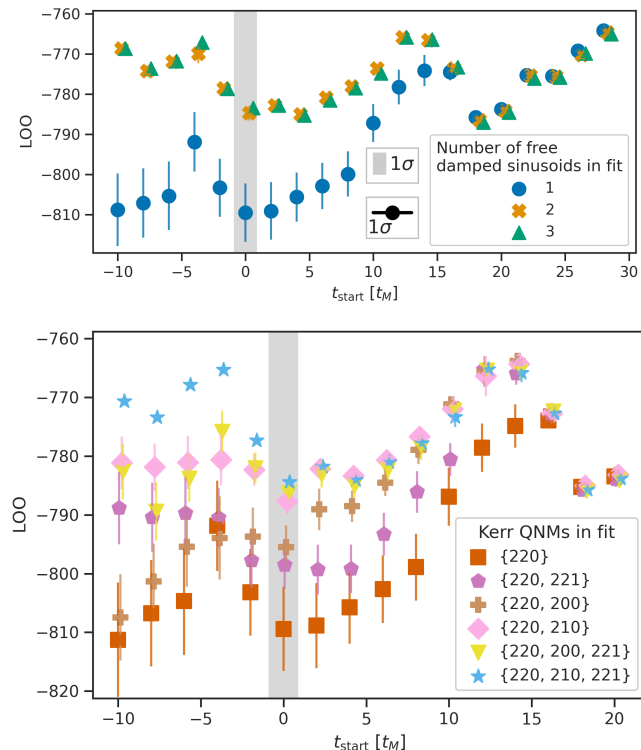


FIG. 1. To quantify statistical goodness of fit, we use the Leave-One-Out Cross-Validation (LOO) [76, 77], which is approximately related to the chi-squared test as $\text{LOO} \approx -\frac{1}{2}\chi^2$ plus a penalty term which depends on the leverage of individual data points. Following Ref. [80], LOO differences greater than 4 are very significant. We deem differences of 1 to be our minimum for noteworthy statistical significance. The error bars indicate differences between LOO of each less-favored model and the top model at a given fit time, and are computed with the `compare` method in the ARVIZ package [88]. Peak strain statistical timing uncertainty from NRSur7dq4 is $\pm 0.9 t_M$, shown as a grey band. *Top*: When fitting free damped sinusoids without Kerr frequency constraints, we find significant preference for two modes over one when the fits start as late as $14 t_M$, and no preference for three modes over two. *Bottom*: LOO of plausible Kerr models which we found to have the best-constrained non-zero amplitudes. Again, we find significant preference for two modes over one. Within two-mode models, LOO differences in favor of the {220, 210} over the {220, 200} model have expectation values greater than 1 and statistical significance of at least 1σ as late as $8 t_M$. When fitting three-mode models with an overtone, there is little LOO improvement if fitting post-peak. However, the {220, 210, 221} model is strongly preferred before the peak strain time. This is at odds with the behavior of the free sinusoid models, which had no preference for three-mode fits. The results when replacing the 221 with the 211 or 201 are similar (not shown), although at certain pre-peak times (in particular $6 t_M$) the 221 is still significantly preferred.

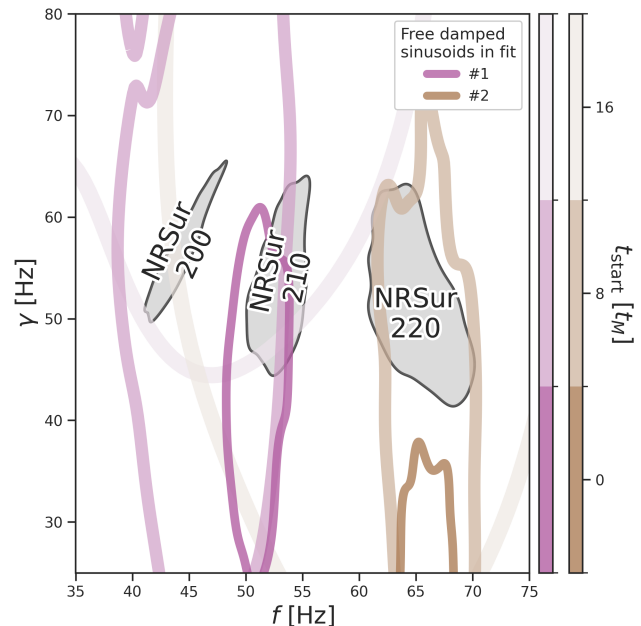


FIG. 2. When simultaneously fitting two free damped sinusoids over a wide range of fit start times, the mode frequencies and damping rates remain broadly consistent with the parameters of $\ell = 2$ fundamental QNMs implied by NRSur7dq4 remnant mass and spin measurements. The frequency and damping rate of modes are shown on the x and y axes respectively. Colors correspond to different fitted sinusoids, and transparency is related to the fit start time as shown in the colorbar. 90% credible contours shown for all posteriors. The NRSur7dq4 QNM distributions are inferred using the QNM package [89].

Our free damped sinusoid model is not the same as the model used in Ref. [1]. The model used in the LVK analysis only includes one of the two polarizations, meaning that parameters recovered by that model may not correspond directly to the content of the physical strain. Any free damped sinusoid analysis with the PYRING [20, 21, 90, 91] code performed for signals from GW231123 and earlier will be subject to this polarization limitation.

The amplitudes of the free damped sinusoids are not shown here. The amplitudes are not as informative as the frequencies and damping rates for this model in our particular analysis. This is because at later fitting times as the SNR decreases, the amplitude uncertainties of our free damped sinusoid fits grow dramatically, likely due to the fact that the priors we use for damping rate and frequency are exceptionally wide.

B. Kerr QNM fits

We fit a model composed of damped sinusoids constrained to follow the Kerr frequency and damping rate spectrum, limited to QNMs from first-order perturba-

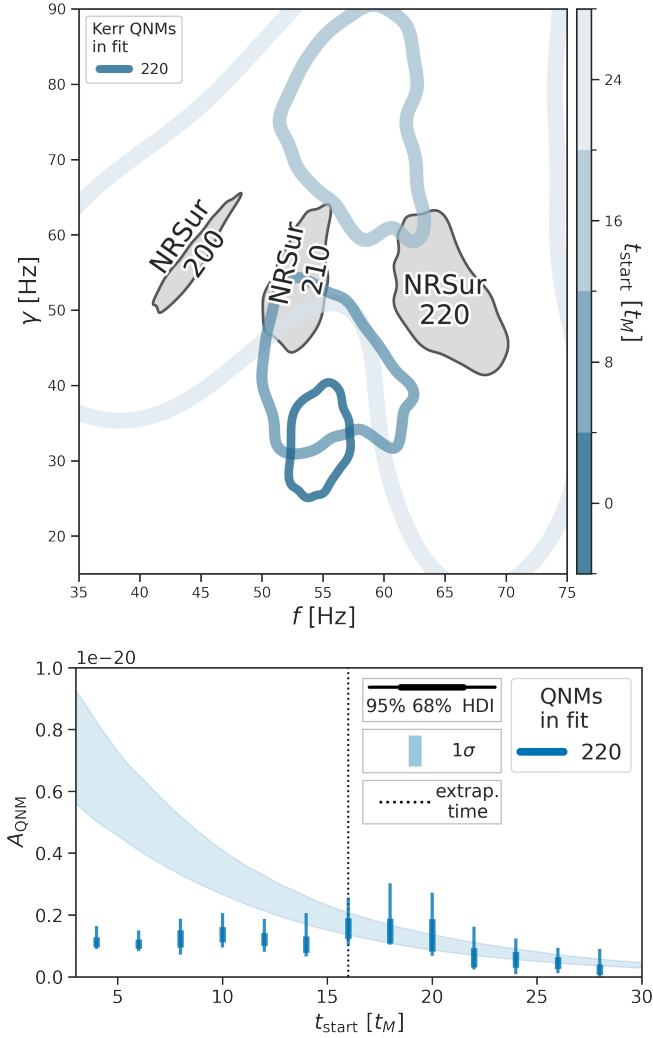


FIG. 3. Frequency and damping rate of single-mode Kerr fits, as well as their amplitudes. *Top*: See Fig. 2 caption for figure conventions. When fitting a single Kerr mode, consistency with NRSur7dq4 is not found until $22 t_M$, after the point at which single-mode fits are equivalent to multi-mode fits in terms of LOO, as shown in Fig. 1. *Bottom*: QNM amplitudes. Errorbars on each scatterpoint indicate 68% and 95% highest density interval (HDI). We extrapolate the amplitude from the fit at $16 t_M$ (as indicated by dashed line), the earliest time where single-mode fits are equivalent to multi-mode fits in terms of LOO, as shown in Fig. 1. We show the 1σ uncertainty of the extrapolated exponential decay of the 220 QNM as a colored band. This extrapolation is only consistent with fit times after $16 t_M$. The amplitude of a single-mode Kerr fit is consistent with being non-zero at 2σ until $28 t_M$.

tion theory. We place flat priors on the detector-frame remnant black hole mass (0.5 to 1.5 times the median detector-frame remnant mass inferred by NRSur7dq4, $M_f^d = 296 M_\odot$) and dimensionless spin (0 to 0.99), mode amplitude (sampled via the marginalization technique described in Ref. [86], with $a_{\text{max}} = 10^{-19}$) and phase (0 to 2π radians).

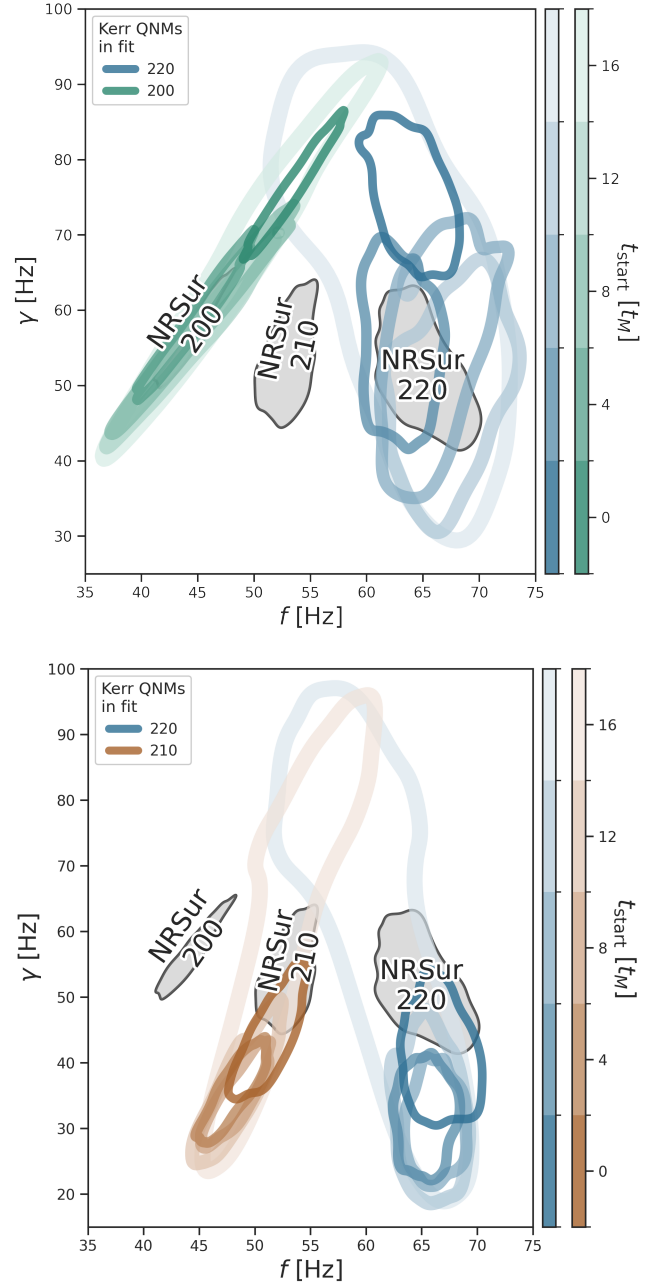


FIG. 4. The frequency and damping rate of fits with two fundamental Kerr modes. See Fig. 2 for figure conventions. *Top*: We find that $\{220, 200\}$ fits are consistent with NRSur7dq4 and also self-consistent over time, starting from $4 t_M$. We do not find another combination of well-measured fundamental QNMs consistent with NRSur7dq4. *Bottom*: The $\{220, 210\}$ model is similarly self-consistent over time but not in agreement with NRSur7dq4.

As shown in Fig. 1, Kerr models with two QNMs give significantly better fits to the data than a Kerr model with one mode, for the same range of times as in the case of the free damped sinusoid fits. When comparing with NRSur7dq4 mass and spin estimates, a single

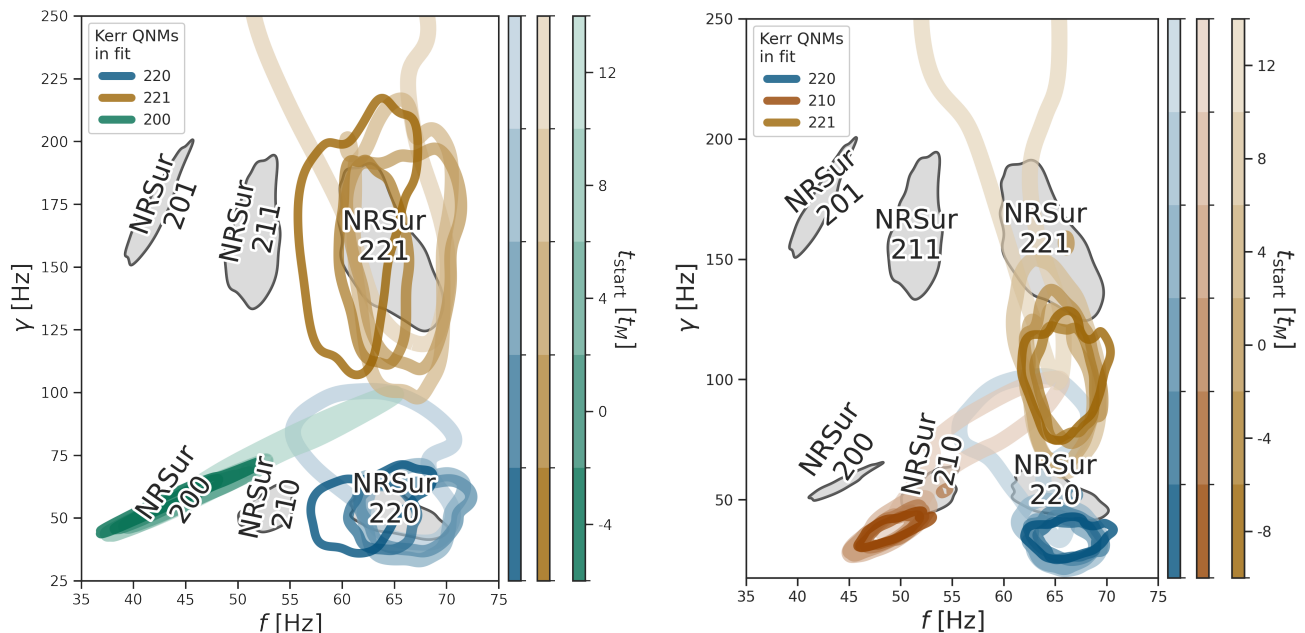


FIG. 5. Frequency and damping rate of Kerr models with overtones. See Fig. 2 caption for figure conventions. *Left:* We find that $\{220, 200, 2m1\}$ fits are consistent with NRSur7dq4 and also self-consistent over time, starting from $0 t_M$. While the above figure shows only the three-mode model with the 221 as the included overtone, similarly consistent results are obtained with the 211 and 201 QNMs. *Right:* While $\{220, 210, 2m1\}$ fits are inconsistent with NRSur7dq4, the added overtone makes this model self-consistent as early as at least $-8 t_M$ depending on which m index overtone is added.

QNM fit does not agree until late times over $20 t_M$ after the peak, as shown in Fig. 3. Also, the expected amplitude decay of these single-mode fits does not arise until at least $16 t_M$, the point at which goodness-of-fit indicates no preference for multimodal fits. The single-mode agreement with NRSur7dq4 comes past the point at which single-mode fits are equivalent in goodness-of-fit to multi-mode fits. Significant non-zero amplitude measurements indicate that a QNM signal persists as late as $28 t_M$. These results are qualitatively similar to those shown in Ref. [1]. We look for non-zero amplitude constraints because an amplitude of zero leads to uninformative posteriors of other parameters.

For ringdown-dominated signals like GW231123, we expect unbiased QNM and IMR remnant mass and spin posteriors to fully encompass each other, as argued in e.g. Appendix B of Ref. [28]. As shown in Fig. 4, between 4 and $16 t_M$ the only two-mode QNM model we find that agrees with the NRSur7dq4 remnant mass and spin is the $\{220, 200\}$. Another physically plausible model [17], $\{220, 210\}$, has frequency and damping rate posteriors which do not overlap with those of NRSur7dq4 within at least the 90 % CL, but this model fits the data equally well or slightly better between as late as $8 t_M$ as shown in Fig. 1. The LOO is greater than 4 in favor of the $\{220, 210\}$ model over the $\{220, 200\}$ model at $0 t_M$ and earlier times, even when taking into account LOO difference uncertainty. At $8 t_M$, the LOO difference between these two models has an expectation of 2 but support for

values greater than 4 within statistical uncertainty. At later times, there is no preference between both models.

As shown in Fig. 6, in both the $\{220, 200\}$ and $\{220, 210\}$ models the amplitudes of the two modes are comparable with the lower frequency mode having a slightly larger amplitude. The amplitude decays appear to broadly follow the expected exponential evolution over time.

While there is no goodness-of-fit preference for fitting more than two free damped sinusoids to the data, this does not necessarily guarantee that the same will be true of the Kerr models: and even if three-mode Kerr models are not more preferred by goodness of fit when compared with two-mode models, this does not mean that Kerr models with three or more modes cannot have any explanatory advantages from the physical point of view. Therefore, there is motivation for pursuing at least three-mode Kerr fits.

As shown in Fig. 5, when fitting a $\{220, 200, 2m1\}$ model (m index intentionally ambiguous, explained immediately below) the mass and spin agrees with NRSur7dq4 as early as $0 t_M$. This is $4 t_M$ earlier than was found for the $\{220, 200\}$ model in Fig. 4. The m index has been left ambiguous here because mass and spin agreement can be achieved by adding to the model any one of the 221, 211, or 201 overtones. Furthermore, all three QNMs in these overtone models are well-constrained to have non-zero amplitudes as late as $4 t_M$, as shown in

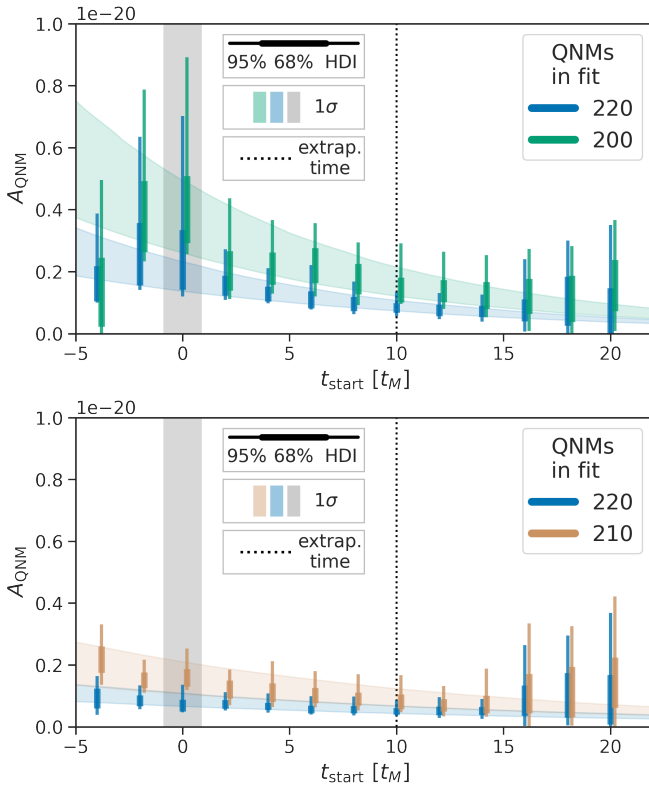


FIG. 6. QNM amplitudes. We find multimodal Kerr fits over time to be consistent with their expected decay at least as early as t_{peak} . Scatterpoints are staggered horizontally for visibility, but each grouping corresponds to a fit at a single time. The 1σ peak time uncertainty is shown as a grey band. Extrapolating from the fit at $10 t_M$ (as indicated by dashed line), we show the 1σ uncertainty of each fitted QNM as colored bands. *Top*: Amplitudes of $\{220, 200\}$ model. The amplitude of the lower-frequency QNM is found to be slightly higher than that of the higher-frequency QNM. *Bottom*: The $\{220, 210\}$ model produces similar amplitudes, although they are overall smaller. The expected exponential decay extrapolated from late times holds earlier in the signal for the 210 model than for the 200 model.

Fig. 7. However, at times especially before t_{peak} there are goodness-of-fit preferences overall for the 221 as opposed to the other overtones in the three-mode models, and we will restrict our attention to the 221 throughout.

In a similar vein, when fitting a $\{220, 210, 221\}$ model, the frequency and damping rate measurements become self-consistent over time going back at least as far as $-8 t_M$ (Fig. 5). This model also has a significantly preferred LOO over all other models when fitting before the peak strain, which is at odds with the free damped sinusoid models which showed no preference for three-mode fits. The $\{220, 210, 221\}$ model makes confident measurements of all three modes as well, albeit at slightly earlier times than the $\{220, 200, 221\}$ model (Fig. 7). Note that similar early-time fit consistency was also found in Ref. [28] for GW190521, another low-frequency signal.

Beyond the Kerr models shown, we also explored two-

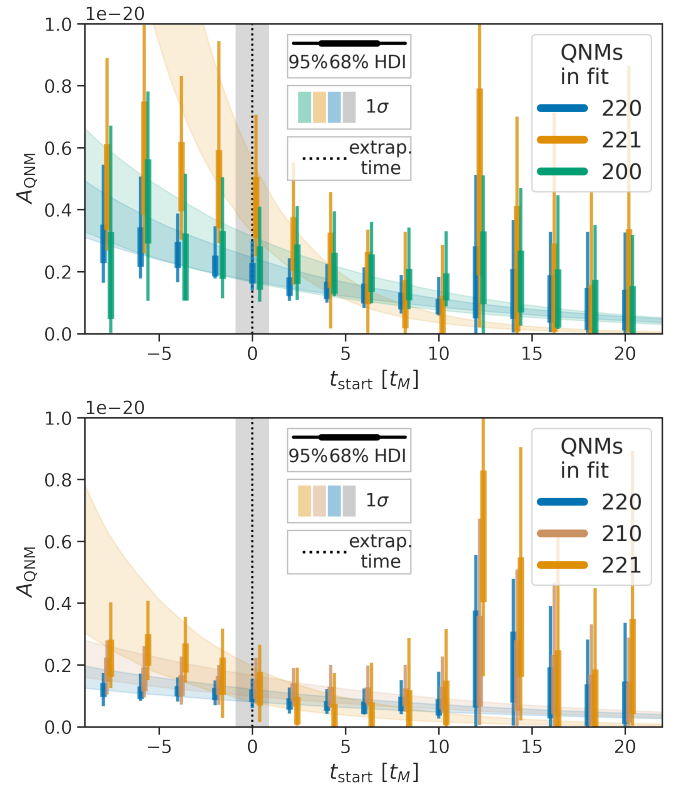


FIG. 7. Amplitudes of $\{220, 200, 221\}$ (top panel) and $\{220, 210, 221\}$ (bottom panel) models. See Fig. 6 for figure conventions. All three amplitudes are confidently non-zero in the 200 (210) model as late as $4 t_M$ ($0 t_M$), and expected to decay for the overtone persist from these late times to as early as $-2 t_M$ ($-8 t_M$). Qualitatively similar results are obtained with other $2m1$ overtones instead of the 221. Instead of extrapolating amplitudes from $10 t_M$ as in Fig. 6, we choose here to extrapolate from a time where the overtone amplitudes are well-constrained to be non-zero, $0 t_M$, so that the extrapolation is not dominated by noise.

and three-mode fits with one $\ell = 3$ mode included, either the 320 or 330. These models confidently constrained the $\ell = 3$ mode amplitudes to be zero both before and after the peak strain, more than QNMs in the other models we have discussed. Because of this, as well as the fact that the free damped sinusoid models did not explore the $\ell = 3$ parameter space estimated through NRSur7dq4 measurements, we do not explicitly show these QNM models. We also found the $\{220, 221\}$ model to be a poor fit to the data as shown in Fig. 1 and inconsistent with NRSur7dq4 (not shown), and for these reasons we do not consider this model further. Lastly, we also fit four modes, $\{220, 221, 210, 211\}$, but since the 211 amplitude was not consistently measured to be non-zero even going back to $-10 t_M$ we did not further explore this model. We did not consider retrograde $p = -$ Kerr QNM fits, because the free damped sinusoids don't support them when comparing with NRSur7dq4 and the estimated remnant spin being high makes retrograde modes unlikely to be

as strongly excited as prograde modes [17, 75, 92].

C. Beyond-Kerr QNM fits

We perform a test of general relativity (TGR) through a search for deviations from the Kerr frequency and damping rate spectrum. To do so, we fit a TGR model in which a subset of all of the QNMs have extra degrees of freedom δf and $\delta\gamma$, such that their frequencies f and damping rates γ are given by [9, 85]

$$\begin{aligned} f_{\text{TGR}} &= f_{\text{Kerr}} \exp(\delta f), \\ \gamma_{\text{TGR}} &= \gamma_{\text{Kerr}} \exp(\delta\gamma). \end{aligned} \quad (4)$$

When referencing specific TGR models in the text, any QNMs with non-Kerr freedoms in the model will be underlined. The default hypothesis of our test is that Kerr should be a good description of the data.

Based on comparisons with NRSur7dq4, we choose to consider the $\{220, \underline{200}\}$ and $\{220, \underline{200}, \underline{221}\}$ TGR models, underlining modes on which the deviation parameter is placed. We also consider TGR models with the 210 instead of the 200, motivated by goodness of fit from the previous subsection as well as knowledge of precessing BBH coalescences [17, 40–42].

We emphasize the measurements of δf over $\delta\gamma$, because of their generally greater precision and also because we find the poorer $\delta\gamma$ measurements to be exacerbated by degeneracies with the remnant mass and spin [28, 93, 94]. TGR parameter priors are chosen to avoid label-switching between QNMs when their Kerr frequencies and damping rates are close in parameter space: this label-switching can cause TGR results to be erroneously inconsistent with GR even in cases where the signal being fit is exactly consistent with GR. A tell-tale sign of such label-switching can be multimodal structure in the TGR parameters. The prior on δf goes from -0.5 to +0.25 (+0.27) for the 210 (200) in any model where those QNMs have TGR freedoms. Tight $\delta\gamma$ priors with an absolute value of 0.1 are also implemented in whichever single direction would cause overtones and fundamental modes to swap places. When quoting δf and $\delta\gamma$ constraints, we will state median values with $\pm 90\%$ HDI uncertainties.

1. Fundamental modes

We begin by considering fundamental mode models only. In Fig. 8 we compare the TGR posterior of the $\{220, \underline{200}\}$ and $\{220, \underline{210}\}$ models. The 210 model appears more consistent with Kerr over most of the time interval where both modes are confidently measured. Before and up to the peak strain time, the measurements of δf and $\delta\gamma$ in both models drift in a manner commensurate with the behavior of Kerr frequency and damping rate measurements shown in Fig. 4. Around 2–4 t_M

TABLE I. TGR model constraints on non-Kerr deviations δf and $\delta\gamma$ (see Eq. (4)) at selected t_{start} . Each QNM with parameterized deviations from Kerr is indicated with underlined indices. The earliest time for each model in the table is the first time at which δf is found to be consistent with 0 at 90% HDI for all non-Kerr QNMs. For comparison with GW250114 [11, 12], we include 6 t_M when possible. See Sec. III C for further discussion. Many of the Kerr constraints should be interpreted with caution, especially for 200 models which are often railing against priors.

Model	t_{start}	$(\delta f, \delta\gamma)$, median $\pm 90\%$ HDI
$\{220, \underline{200}\}$	$8 t_M$	<u>200</u> : $(0.15^{+0.12}_{-0.17}, 0.08^{+0.38}_{-0.34})$
$\{220, \underline{210}\}$	$2 t_M$	<u>210</u> : $(0.05^{+0.08}_{-0.11}, 0.07^{+0.34}_{-0.33})$
	$6 t_M$	<u>210</u> : $(-0.03^{+0.10}_{-0.11}, 0.26^{+0.24}_{-0.27})$
$\{220, \underline{210}, \underline{221}\}$	$-6 t_M$	<u>210</u> : $(0.06^{+0.08}_{-0.08}, -0.03^{+0.13}_{-0.20})$
		<u>221</u> : $(0.07^{+0.28}_{-0.26}, 0.17^{+0.25}_{-0.27})$
	$0 t_M$	<u>210</u> : $(-0.08^{+0.19}_{-0.16}, -0.09^{+0.19}_{-0.28})$
		<u>221</u> : $(-0.23^{+0.24}_{-0.27}, 0.21^{+0.28}_{-0.25})$
	$6 t_M$	<u>210</u> : $(-0.14^{+0.21}_{-0.21}, -0.06^{+0.16}_{-0.26})$
		<u>221</u> : $(-0.12^{+0.44}_{-0.38}, 0.19^{+0.24}_{-0.29})$

and later times, the TGR posteriors appear to settle and gradually expand as the SNR decreases. Incidentally, 2–4 t_M corresponds to the time around which Kerr models with overtones start to allow the overtone amplitude to be zero, as in Fig. 7. At 12 t_M , the TGR posteriors then shift again to become broader, and at times after this become uninformative: this final transition seems to correspond to the point in the signal at which the models start to lose resolution of both modes, based on comparison with the Kerr fit behavior shown in Figs. 4 and 6.

We first focus on the $\{220, \underline{200}\}$ TGR model. At 4 t_M , the earliest time at which we interpret the fundamental mode models to become consistent with their later time fits, the δf_{200} posterior is inconsistent with Kerr at the 90% CL. Not only is it inconsistent, but δf_{200} is railing against the upper prior, and so in principle the GR value may be excluded at even higher CLs: interactions with the prior bounds suggest that CLs should be interpreted with caution. The first time at which the δf_{200} posterior gives consistency with GR at the 90% CL is 8 t_M , where $\delta f_{200}(8 t_M) = 0.15^{+0.12}_{-0.17}$: but again, δf_{200} is still largely cut off by the prior at positive values, and thus even this Kerr consistency should be interpreted with caution. The Kerr consistency persists when fitting up to 16 t_M , although the posterior shifts in these later times from railing against the upper prior bound to railing against the lower prior bound and so again interpretations should be made with caution. At 18 t_M and later, the δf_{200} posterior reverts to the prior.

By contrast, the $\{220, \underline{210}\}$ TGR model is not only more in agreement with Kerr but also achieves this agreement at much earlier times. The earliest

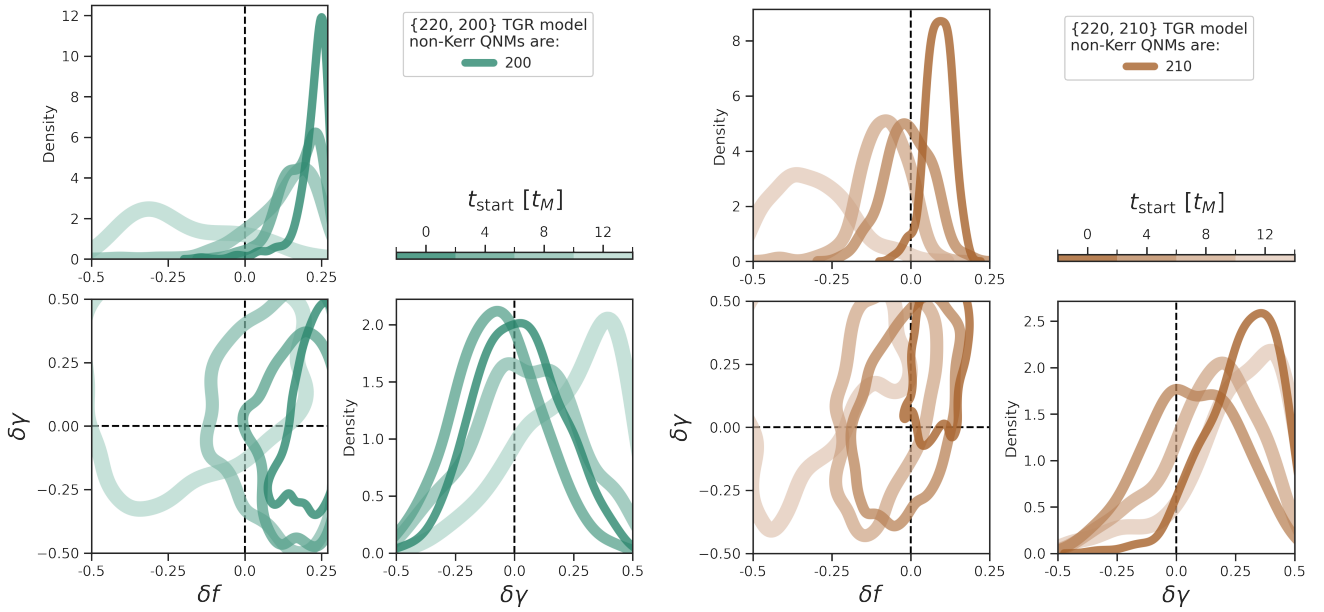


FIG. 8. Test of GR (TGR) as parameterized in Eq. (4), using models with only fundamental modes. Transparency of the posteriors relates to fit start time shown in colorbar, color denotes non-Kerr QNM. Dotted lines indicate GR values. Axis limits correspond to prior boundaries. 90% credible contours shown. *Left:* The $\{220, \underline{200}\}$ model with non-Kerr 200. While the posterior does begin encompassing the Kerr values at 90% confidence at $8 t_M$, it is also railing against the upper δf prior and so this agreement should be interpreted with caution. At later times it becomes difficult to confidently measure both QNMs in the Kerr model as demonstrated in Fig. 6, and there is a clear drift in the TGR posteriors at this time. *Right:* The $\{220, \underline{210}\}$ model with non-Kerr 210. The posterior is noticeably more consistent with Kerr than the 200 model through $8 t_M$. This consistency is best achieved at $4 t_M$, which is near where the Kerr model starts making mass and spin measurements more consistent with itself when fit at later times as shown in Fig. 4.

Kerr-consistent fit start time is $2 t_M$, where we find $(\delta f_{210}(2 t_M), \delta \gamma_{210}(2 t_M)) = (0.05^{+0.08}_{-0.11}, 0.07^{+0.34}_{-0.33})$.

Fitting the $\{220, \underline{210}\}$ TGR model at $6 t_M$ allows for comparison with a similar test of GR for GW250114 reported in Fig. 4 of Ref. [11] and Fig. 2 of Ref. [12], which used a $\{220, 221\}$ model. We find $(\delta f_{210}(6 t_M), \delta \gamma_{210}(6 t_M)) = (-0.03^{+0.10}_{-0.11}, 0.26^{+0.24}_{-0.27})$, significantly improving on the $\pm 30\%$ frequency constraint reported for GW250114 at the equivalent fit start time. The $\delta \gamma_{210}$ posterior is once again somewhat cut off by the upper prior at this time for GW231123 though, and caution in interpretation is warranted.

Fitting the $\{220, \underline{210}\}$ TGR model between $10 t_M$ and $16 t_M$, the δf_{210} posterior drifts to rail against the negative end of the prior, ruling out 0 at the 90% CL from $10 t_M$ to $14 t_M$. At $18 t_M$ and later, the posterior reverts to the prior: this behavior is in line with the evolution of the Kerr fits, which lose resolution of two modes around $16 t_M$ as shown in Figs. 1, 4, 6. The railing may just be a consequence of the decreasing resolution of the two modes with time, or could be a sign of mismodeling.

2. Addition of an overtone

We now consider fitting TGR models with an overtone included. In the previously fit Kerr models of Sec. III B,

overtones are not found to give substantially better fits to the data at times after the peak strain (Fig. 1), but do make remnant mass and spin measurements more consistent over a broader range of times (Fig. 5) and have amplitudes confidently measured to be nonzero as late as $\sim 4 t_M$ after the peak strain (Fig. 7). Therefore, we might expect overtones to change the TGR posteriors of the most dominant modes, even if the overtone TGR parameters themselves are not well-constrained.

Fig. 9 shows representative results of TGR fits with two fundamental modes and one overtone. An interesting feature of these fits is the improved consistency of their TGR parameters with Kerr over a wider range of fitting start times than the fundamental-only models. Notably, the $\{220, \underline{210}, \underline{221}\}$ model is consistent with Kerr as early as $-6 t_M$. A similar behavior of consistency in very early fits was observed in Ref. [28] when fitting both Kerr and TGR quasinormal mode models to GW190521, another low-frequency signal. For GW190521, consistent TGR fits seemed to be achievable as early as $-5 t_M$.

The $\{220, \underline{210}, \underline{221}\}$ TGR model in GW231123 is more consistent overall with Kerr than the $\{220, \underline{200}, \underline{221}\}$ model, but both have strikingly similar overtone δf constraints. Comparing with Fig. 5, these shared overtone constraints are perhaps not so surprising: regardless of which Kerr model is fit, the $(\ell, m) = (2, 2)$ modes in the Kerr models are fit at the same frequency around 65 Hz.

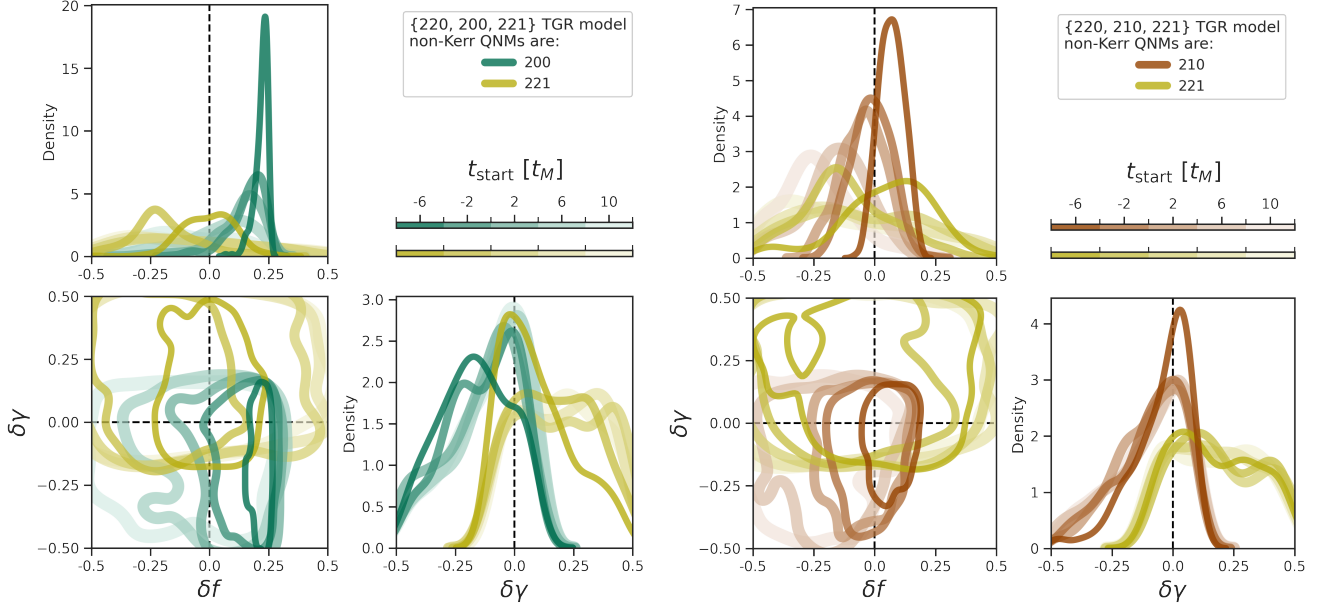


FIG. 9. Test of GR (TGR) using models with fundamental and overtone modes, see Fig. 8 for figure conventions. Comparing with Fig. 8, the overtone models are more consistent with Kerr at all times shown, especially in the case of the $\{220, 210, 221\}$ model. Note that the $\delta\gamma$ parameters in particular have restrictive priors designed to prevent label-switching, which the posteriors consistently rail against. Consistency with Kerr is found at early times before peak strain. *Left:* The $\{220, 200, 221\}$ model with non-Kerr 200 and 221. While this model does have some support for Kerr at most times shown, the δf_{200} posterior is always railing against one of the prior bounds (at 0.27 or -0.5) and thus should be interpreted with caution. At all times from $-6 t_M$ onward, the 221 is in reasonable agreement with Kerr. *Right:* The $\{220, 210, 221\}$ model with non-Kerr 210 and 221. Kerr is clearly supported at all shown fit times, notably both earlier and later in time than for the fundamental-only fits in Fig. 8. At earlier times than those shown, fits start to drift away from Kerr values.

The instrumental noise is lower at higher frequencies in this part of the LIGO band, and correspondingly the $(\ell, m) = (2, 2)$ modes are more precisely measured than the lower frequency $\ell = 2 \neq m$ modes, which implies that any improvements in fit when adding TGR freedoms will likely come from the $\ell = 2 \neq m$ modes in the model.

Given that the overtone model closest to Kerr does not seem to be the model in best agreement with NRSur7dq4, interpretation of the TGR results is difficult. We choose to report on Kerr constraints of the $\{220, 210, 221\}$ TGR model at a few time intervals, since this model displays some possible signs of better consistency with Kerr and its δf posteriors do not rail against the priors as significantly as the $\{220, 200, 221\}$ model at most fitting times. At $6 t_M$, we find Kerr-consistent frequency constraints of roughly $\pm 20\%$ and $\pm 40\%$ for the 210 and 221 respectively. These constraints remain Kerr consistent and improve going back in time as early as $-6 t_M$, to $\pm 8\%$ and $\pm 28\%$ for the 210 and 221 respectively. See Table I for details. Earlier than $-6 t_M$, the $\{220, 210, 221\}$ TGR model begins deviating from Kerr. This time corresponds to an apparent transition in the amplitudes of the Kerr model, as shown in Fig. 7.

It is possible that the improvement in Kerr consistency of the three-mode models with overtones over two-mode models with only fundamental modes is just a function of the additional parameters of the three-mode models

broadening the posteriors. However, given that the Kerr fits with overtones are at least equally preferable if not significantly preferred over fits without overtones in the range of times where Kerr consistent measurements are made (Fig. 1), and the overtone amplitudes are confidently nonzero in most of these times and follow their expected decays (Fig. 7), it seems plausible that the overtones are meaningfully contributing to the fits and parameter estimates.

IV. DISCUSSION

Assessing the best QNM fits to the GW231123 data and determining whether we have successfully validated the QNM frequency and damping rate spectrum of the Kerr metric is complicated not only by our incomplete understanding of the regime of validity of QNM fits in BBH coalescences but also by the possibility of systematics in the IMR model to which we are comparing our fits, NRSur7dq4. Fully determining the conclusiveness of our results will have to rely on a holistic consideration of statistical quantities, astrophysical expectations, and some educated conjecture regarding the unknowns of QNM analysis from both the data and theory points of view.

A. Motivating choices of QNM fit start times

A central challenge in the assessment of our fits is determining what should constitute physically meaningful QNM fit start times relative to the peak strain time in BBH coalescence signals. Debate continues on this topic in the literature, centering around fits to NR which hope to subsequently inform actual data analysis. Much of this NR-based work reports some indications of systematic error when fitting damped-sinusoid models close to (but after) the peak strain [19, 73–75].

However, a key distinction between these NR studies and data analysis of LIGO signals is the noise in both instances. LIGO signals are orders of magnitude quieter than NR, and have a specific noise spectrum morphology which interacts with the signal differently than the imperfectly understood numerical errors associated with most NR simulations in current use. While the findings of systematic errors in QNM fits to NR are important to keep in mind when performing data analysis, we are not aware of any rigorous quantification of these systematic errors. For example, if the extent of these systematics in NR analyses is never more than an error of 1 part in 1000 when fitting after the peak strain (or even before it), that is irrelevant to data analysis of $\text{SNR} = 20$ LIGO signals where statistical errors of Kerr constraints are $\mathcal{O}(10\%)$.

Furthermore, most NR studies perform maximum likelihood fitting whereas we do fully Bayesian analysis for LIGO data, presenting another complication in making direct comparison. Another issue which further complicates comparison is that our analysis of LIGO data does not estimate the location of the peak strain in the signal using QNM fits, but rather relies on alternate models like NRSur7dq4 to give us these estimates. If the peak strain estimates of these alternate models turn out to be significantly systematically biased, our data analysis may not be able to appeal directly to our theoretical understanding of regimes of validity for QNM fits anyway. Nevertheless, we are always allowed to empirically explore which regions of the data appear consistent with QNM content.

B. Consistency of QNM fits over time

A key tenet of our fitting philosophy is that reasonable QNM models which accurately describe the signal should have consistent parameter estimates when fit at different starting times. This philosophy is shared with most NR studies [74, 75, 95, 96]. To confirm consistency between fits at different times, we visually compare credible levels of posteriors, typically the 90% CL but not always: while we could more strictly quantify the level of agreement between posteriors by using a statistical distance, e.g. Kullback-Leibler divergence, there are non-trivial complexities in employing statistical distances [97] and we assert that an eye test is sufficiently accurate for our purposes. We find that fits to GW231123 with two

fundamental modes are capable of achieving consistent parameter estimates over time, as shown in Figs. 4, 6, 8. Three-mode fits are capable of achieving this behavior as well, as shown in Figs. 5, 7, 9. This is one indication that our fits are reasonable.

Our two-mode fits confidently give Kerr constraints at some late times close to $10 t_M$, as shown in Fig. 8. The time $10 t_M$ is often used as approximately where QNM fits should be reliable in general, although we do not assert that this is the earliest possible time for QNM fits to be reliable even in theory. While the two-mode models validate Kerr at the 90% CL at some late times around $10 t_M$, they do not necessarily do so at all late times. This discrepancy in Kerr validation is ameliorated by the inclusion of one overtone in addition to the two fundamental modes in the models, as shown in Fig. 9. The overtone is not itself well-measured at these late times as indicated by its amplitude in Fig. 7, but this is not necessarily a reason to ignore the overtone in the late time fits. The change in posteriors from including the overtone may be an indication of stealth bias in the fundamental mode models [82, 83].

Ultimately, we presume that every possible QNM is excited to some non-zero amplitude in astrophysical BBH coalescences, and in principle a physically faithful QNM model would always include every possible QNM even if the contribution of most of the QNMs to the fit was very small, and would also take into account currently ignored effects like the time-dependent ring-up of QNMs [98]. For such types of all-inclusive models, the priors would then determine which QNMs were most significant in the fit. This type of informative-prior driven analysis is a more desirable approach than what is commonly done in our field (and was done in this paper) wherein different numbers of QNMs are added to different models, in part because the prior-driven approach provides a better foundation for model comparison [77]. That being said, robust informative priors for QNM fits are currently beyond the reach of our field. Efforts are being undertaken to solve this problem [42, 75, 92–94, 99, 100], but more work is still needed. The lack of informative priors and restricted number of QNMs in our current models is a limitation of the program of black-hole spectroscopy as it was originally envisioned [38]: even though we may find consistency with the frequencies and damping rates of the Kerr spectrum using our current models, our fits allow such a wide range of amplitudes and phases that they may be settling on astrophysically unlikely or totally unattainable amplitude and phase combinations. An even more ideal improvement over informative priors for QNM models would be full inspiral-merger-ringdown models [101] in different theories of gravity [102–104] which are both well-posed [105] and well-motivated, but this seems to be even further from our current reach than informative QNM models. Some works argue that enlarging the model parameter space by adding more and more QNMs will make Bayesian inference unfeasible [106]. We contend that if this is an accurate statement, it may only be

accurate in the case of models with uninformative priors. Refs. [107, 108] make qualitatively similar arguments: that being said, we are strictly referring to priors derived from and for direct QNM fitting, while those works seek to use the ringdown stages from existing IMR models which are likely not identical to our proposed priors. In research fields outside of gravitational wave science, it is common to have highly performing models with orders of magnitude more parameters than our current ringdown models [109]: if they can do it, perhaps so can we.

C. QNM fits starting before peak strain

While our GW231123 fits seem to produce consistent parameter estimates over a range of fit start times after the peak strain, they also are consistent in some instances at times before the peak strain. The Kerr constraints at times as early as $-6 t_M$ for the {220, 210, 221} model in particular are much tighter than at late times, as shown in Table I, and still support general relativity. Some theoretical studies of extreme and comparable mass binary systems suggest the presence of QNM and other Kerr perturbative content at times near or even before peak strain [74, 75, 110, 111], but the early time fit behavior we have observed may not necessarily be physical and deserves scrutiny.

As mentioned above, most theoretical studies of QNM fits are done at very different SNRs from actual data, and with different noise. The specific noise PSD of LIGO whitens very low-frequency BBH signals such that their inspiral content is significantly suppressed, and this may be affecting the ringdown analysis of GW231123 in unanticipated ways. One possibility is that a whitened ringdown signal with no discernible inspiral preceding it in time may still be fit accurately for some time interval before the ringdown signal.

Another possibility is that perturbative content like QNMs exists in the signal even before the peak strain but is obscured by other lower-frequency signal content, and the whitening of the PSD works to reveal this perturbative content which would otherwise be hidden. These explanations are purely speculative however.

It is worth noting that a similar early time fit behavior was also observed in GW190521 [28], hinting at the possibility that this is a generic feature of low-frequency signals. In future work we aim to perform thorough studies of NR signals in LIGO noise to determine if the fitting behavior we have observed should be expected. The Kerr constraints at these early times are much better than at later times, and so we should make use of them if they are indeed trustworthy.

D. IMR waveform systematics

Another possible explanation of the early QNM fit behavior is that we have misidentified the peak strain time

and/or the timescale t_M due to NRSur7dq4 systematic error. To assess the systematic error of IMR waveform models, a standard approach is to use the mismatch of the model with NR simulations [1, 7]. Based on the standard method of assessing mismatch and also through empirical observation of fits to NR, when considering highly-spinning NR simulations NRSur7dq4 is the overall best-performing quasicircular precessing BBH IMR model in the GW231123 part of parameter space. Mismatches furthermore indicate that, for many but not all values of parameters in this region and at the SNR of GW231123, we expect NRSur7dq4 to have systematic errors which are subdominant to statistical errors [1, 7]. While in general the mismatches as traditionally computed actually provide a conservative estimate of minimum SNRs at which systematic bias dominates over statistical uncertainty, Ref. [112] notes that incorrect choices of the number of degrees of freedom in the model may lead to the traditional minimum SNR estimate not being a conservative lower bound, and so the mismatches should be interpreted with caution. We cannot definitively rule out the possibility that the NRSur7dq4 measurement errors are dominated by systematics for GW231123.

A good way to determine if NRSur7dq4 is systematically biased is to find a better fit to the data with a physically plausible model. To this end, we note that the QNM fits with the 200 mode, which produce the best agreement with NRSur7dq4 remnant mass and spin estimates (Figs. 4, 5), are not as favored by goodness of fit as 210 QNM fits overall (Fig. 1), even though those 210 QNM fits are in tension with NRSur7dq4 parameter estimates. Furthermore, the NRSur7dq4-consistent 200 QNM fits do not seem to be as in agreement with Kerr as the 210 QNM fits (Figs. 8, 9).

E. Astrophysical plausibility of 210 and 200 QNMs

The 210 fits may also constitute a more physically plausible model than the 200 fits. It is known that the 210 can be excited by binary spin-orbit misalignment [17, 40–42], and while the 200 is also excited by spin-orbit misalignment it is excited to a lesser extent over the angular two-sphere. To have the 200 be so loud as we find it to be in Figs. 6, 7, implies that the 210 should also be similarly loud if not more so over the angular two-sphere: but we do not find strong support for fitting three $\ell = 2$ damped sinusoids as suggested, e.g., by the behavior of free damped sinusoid fits in Figs. 1, 2.

Such a suppression of the 210 would require a rather fine-tuned binary spin configuration. The mechanism by which the 210 could be significantly suppressed is the asymmetric emission of mirror mode QNMs associated with different angular harmonics, as described in e.g. Fig. 3 of Ref. [17], combined with a viewing angle closer to the remnant’s equator than its poles. Based on current knowledge, this fine-tuned configuration by itself might not even be sufficient to create the observations we have

made in GW231123.

It is worth noting that NRSur7dq4 finds the most support for binary spins that are not only in the orbital plane but also maximal, as shown in Fig. 10. It is possible that the aforementioned emission mechanisms are amplified for near-extremally spinning precessing BBHs, a part of parameter space which has not been rigorously studied. Nonetheless, as of now the 210 seems like a more plausible candidate for strong excitation than the 200.

We also comment here on Fig. 10 of Ref. [1], which shows a harmonic mode decomposition of the NRSur7dq4 waveform using posterior samples from analysis of GW231123. The text of the accompanying appendix for this figure claims that it demonstrates an inconsistency between the angular harmonic content of NRSur7dq4 and a relatively large observed 200 QNM amplitude. However, direct comparison between this figure and QNM analyses is difficult to make, since the waveforms in the figure are not in the appropriate frame and consider only combined amplitudes over the 2-sphere of the $\pm m$ harmonic modes in this frame. A version of this figure better-suited for direct comparison with our analysis would show the NRSur7dq4 waveform from a specific point on the celestial sphere with angular harmonics defined according to the remnant frame, and the $\pm m$ modes would be kept separate because they can differ dramatically for precessing systems [17, 113]. In its current form, the figure does not provide direct evidence that NRSur7dq4 does not support a relatively large observed 200 amplitude.

F. QNM fit systematics

One might also worry that our QNM fits are systematically biased rather than the IMR fits. This is possible if the signal we are fitting contains significant content besides damped sinusoids, the data conditioning we perform has corrupted our posteriors [10], or the QNM model we fit is misspecified to the actual signal. For a correctly specified QNM model fit to data without significant non-stationary noise features, we expect to see consistent parameter estimates when starting our QNM fits at different times in the signal. We successfully demonstrate such consistent fitting behavior in Figs. 4, 5, 6, 7, 8, 9.

That being said, because the SNR decreases with time and is moderate even at the peak strain, it is only at early times in the signal where definitive preferences between different QNM models appear. It is difficult at present to determine the earliest time at which QNM fits should be considered valid in this signal, so the aforementioned definitive model preferences at early times may not be meaningful. This model selection ambiguity is a general limitation of our very flexible damped sinusoid models: more informative amplitude and phase priors might help better discriminate between different QNM combinations at later fitting times. We note that the post-peak SNR

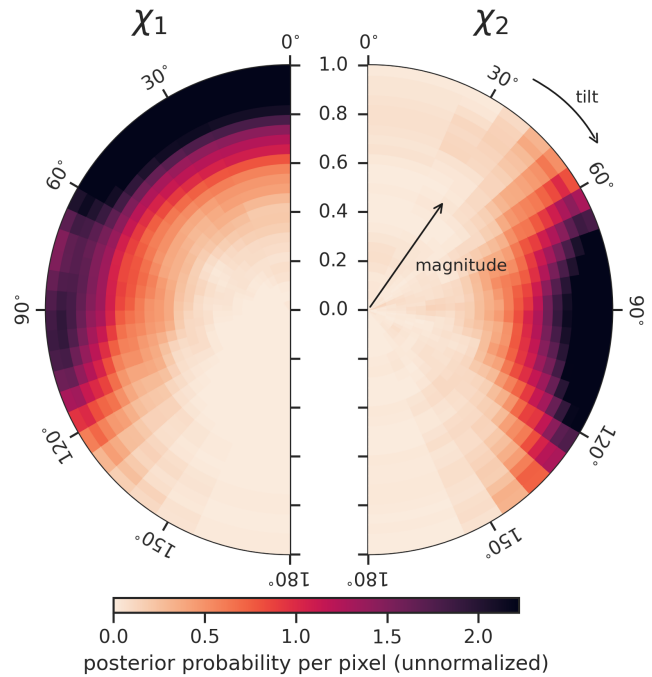


FIG. 10. Spin tilts and magnitudes given by NRSur7dq4 for each of the binary components. The bulk of the inferred distribution lies at maximal spins and spin-orbit misalignment of at least the secondary component, with significant support for the primary being misaligned as well. Spin-orbit misalignment is associated with excitation of both prograde and retrograde $\ell \neq m$ fundamental QNMs [17, 40–42] like those fit throughout this work.

is not so high as to make GW231123 very susceptible to conditioning-induced bias [10], and poor data quality affecting our inferences also seems less likely given the discussion in Sec. II as well as Ref. [1].

G. Interpretation of Kerr and Beyond-Kerr QNM fits

Taking all of the evidence together, there is a case to be made for significant NRSur7dq4 systematic errors affecting inferences of GW231123, although we cannot definitively conclude that this is taking place. Significant systematic error in the IMR parameter estimates would have implications for population and binary formation channel studies using GW231123 [50–60]. If indeed our 210 QNM fits are more accurate, future studies might consider using the remnant mass and spin estimates of these fits in addition to those of IMR models. The 210 QNM fits imply a higher-spinning and higher-mass remnant than NRSur7dq4 and all other available IMR models [7, 43–46]. We provide the parameters of the 210 models in the data release of this paper [72], and for convenience show the mass and spin of the {220, 210} model in Fig. 11.

There are alternative BBH dynamics which we can-

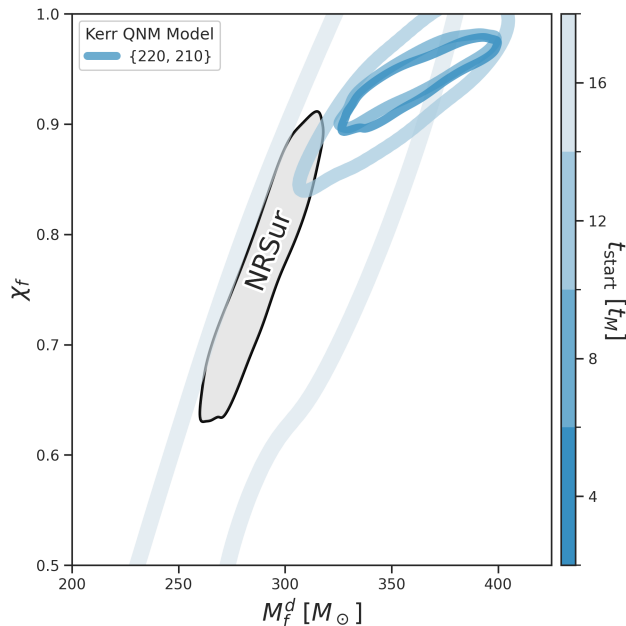


FIG. 11. Detector-frame remnant mass (x-axis) and spin (y-axis) of NRSur7dq4 compared against the model which may be a more accurate fit, the Kerr $\{220, 210\}$ model. The $\{220, 210\}$ model finds higher remnant spins and masses, implying highly spinning but potentially more spin-aligned binary components than the NRSur7dq4 result shown in Fig. 10. If other studies consider using our QNM posteriors for parameter estimates of GW231123, we suggest taking a posterior from one of the fits around the times where the Kerr model is tightest and most self-consistent, somewhere between 4 and 12 t_M .

not conclusively rule out and which may be contributing to the excitation of the QNMs we observe. These dynamics include eccentricity. While we cannot rule out eccentricity, we believe that it is unlikely that eccentricity alone could be responsible for the QNM excitations in GW231123, as eccentricity does not leave a strong imprint in the relative QNM amplitudes unless the eccentricity is so extreme as for the merger to be a nearly head-on collision [114]. Such highly eccentric events are astrophysically unlikely [115], and also have a detection penalty because their gravitational wave emission intensity can be up to roughly an order of magnitude smaller than their quasicircular counterparts [114, 116].

Gravitational lensing might also be capable of altering the QNM amplitudes, although we are not aware of a lensing study which has shown exactly how the QNM amplitudes and phases might be affected. For example, in the case of millilensing, multiple copies of the signal are overlaid with an amplification, a phase shift, and a time shift [117]. In the regime of the signal dominated by damped sinusoids, added copies of damped sinusoids would produce the same number of observed modes as in the unlensed signal and with the same frequencies, but with different amplitudes and phases, see e.g. Eq. B.3 of

Ref. [10].

Regarding our test of general relativity, we note that while the parametrization of Eq. (4) is reasonably motivated by our knowledge of the non-Kerr frequency shifts to QNMs induced in some classes of beyond-GR theories [118, 119], it may not be optimal for all beyond-GR models [120–123]. Also, when allowing non-Kerr deviations of multiple QNMs which are close in frequency and damping rate, avoiding label-switching is a data analysis challenge. One solution for generic label-switching problems is provided in Ref. [84], which we may consider for future works. And finally, we find that it matters which QNMs in our models are given non-Kerr freedoms. Empirically we find that the method which results in the most precise TGR posteriors is to first look at the Kerr model and leave as-is its best-measured QNM, while adding freedoms to all other QNMs in the model. However, to our knowledge the formalism for this choice has not been rigorously derived in the literature.

V. CONCLUSION

By fitting quasinormal modes to GW231123, we confirm strong statistical evidence preferring two-mode fits over one-mode fits for a wide range of fitting times like in Ref. [1], as shown in Fig. 1. We confirm that the two modes are similarly long-lived and have comparable amplitudes, as shown in Fig. 6. Comparison of the posteriors from our QNM fits with the strain peak time of the overall best-performing inspiral-merger-ringdown model, NRSur7dq4, shows that there is preference for two QNMs until 14 t_M after the peak, and evidence of a single QNM as late as 28 t_M . Comparison with NRSur7dq4 also identifies the dominant two QNMs as the $(\ell, m) = (2, 2)$ and $(2, 0)$ fundamental prograde modes, as shown in Fig. 4. However, we find that models with the $(\ell, m) = (2, 2)$ and $(2, 1)$ QNMs instead may be better fits to the data and more consistent with the Kerr frequency and damping rate spectrum, as shown in Figs. 1, 8, even though they are in tension with NRSur7dq4 as shown in Fig. 4. The $(\ell, m) = (2, 2)$ and $(2, 1)$ QNM model may also be more physically plausible than the $(\ell, m) = (2, 2)$ and $(2, 0)$ QNM model, if we assume that GW231123 is sourced by a quasicircular precessing BBH coalescence [17, 40–42]. Lastly, we find confident amplitude measurements and consistent remnant mass and spin inferences when adding an $(\ell, m, n) = (2, 2, 1)$ prograde overtone to our models (Fig. 7), and the overtone is found to improve validations of the Kerr spectrum as shown in Fig. 9.

The fact that we find a QNM model which does not fully agree with NRSur7dq4 (or any of the available IMR models, for that matter) but seems to fit the data better and is more in agreement with the Kerr spectrum raises the possibility of significant systematic errors in the parameter estimates of the NRSur7dq4 model. If NRSur7dq4 is indeed systematically biased and our QNM fits more accurately characterize the remnant mass and

spin of GW231123, this has implications for population and binary formation channel studies making use of GW231123. In the data release for this paper [72], we provide posteriors for the $(\ell, m) = (2, 2)$ and $(2, 1)$ QNM models which may give more accurate estimates of the remnant mass and spin. As shown in Fig. 11, these QNM models suggest that the remnant of GW231123 may have been more highly spinning and heavier than what is found by the inspiral-merger-ringdown models, still implying highly-spinning progenitors but less spin-orbit misalignment. That being said, we cannot definitively rule out the possibility of systematics in our QNM fits, bias caused by non-stationary noise features, or unmodeled physics producing systematic errors.

The possibility of NRSur7dq4 systematics makes interpretation of this signal and its associated constraints of the Kerr metric exceptionally difficult. Nonetheless, if we take the GW231123 QNM model which best fits the signal and is most consistent with the Kerr spectrum, and we compare it at equivalent fitting times relative to the peak strain with QNM fits of GW250114 [11, 12], the loudest signal to date, we recover much better constraints of the Kerr metric: our GW231123 fits constrain the Kerr frequencies to within $\pm 10\%$ at the 90% credible level (shown in Figs. 8, 9, Table I), as opposed to the $\pm 30\%$ reported for GW250114. We also report sub-10% and 30% constraints for one fundamental mode and overtone respectively when fitting at earlier times, although more work is required to definitively confirm whether these constraints are physically meaningful.

The tensions we find between QNM and IMR fits highlight the utility of QNM fits as probes of astrophysical parameters in addition to providing tests of general relativity. The flexibility of QNM models makes them especially useful for probing astrophysical parameters in parts of parameter space where more standard models are known to struggle with systematic errors.

While preparing this manuscript, we became aware of another QNM analysis of GW231123 [70]. Ref. [70] finds preference for two modes in the signal, similar to what is reported in Ref. [1] and our own work, and also reports a test of general relativity. We use different methodologies from both Refs. [70] and [1] and do not reach identical conclusions. We consider a broader range of fitting times than both analyses, and we also consider models with combinations of QNMs that are not explored in either work. While all analyses agree on the preference for multimodal fits to the data, we do not interpret this preference in the same way. Ref. [70] broadly reports similar results as contained in the text of Ref. [1] and concludes that there is a fitting preference for the $\{220, 200\}$ model, whereas our analysis has indications of disfavoring this model in favor of models which use the 210 instead of the 200. Ref. [70] claims that the 200 model is consistent with the Kerr metric, although we do not find the evidence for the Kerr consistency of this model to be as definitive as for the 210 models. Neither Ref. [1] or Ref. [70] uses the QNM analysis to suggest possible

NRSur7dq4 systematics as we do.

In Appendix A, we discuss QNM fits using the 320 and 321 modes. These models which include the 320 can achieve better goodness-of-fit than the QNM models with the 200 that agree with NRSur7dq4, but they also have seemingly unphysical features which make us disfavor them when compared with the other models in the main text.

ACKNOWLEDGMENTS

We thank Alberto Vecchio and Keefe Mitman for insightful conversation. We also thank Lorenzo Pompili for helpful comments during internal LVK review, as well as Luis Lehner.

This material is based upon work supported by NSF’s LIGO Laboratory which is a major facility fully funded by the National Science Foundation. This research has made use of data or software obtained from the Gravitational Wave Open Science Center (gwosc.org), a service of the LIGO Scientific Collaboration, the Virgo Collaboration, and KAGRA. This material is based upon work supported by NSF’s LIGO Laboratory which is a major facility fully funded by the National Science Foundation, as well as the Science and Technology Facilities Council (STFC) of the United Kingdom, the Max-Planck-Society (MPS), and the State of Niedersachsen/Germany for support of the construction of Advanced LIGO and construction and operation of the GEO600 detector. Additional support for Advanced LIGO was provided by the Australian Research Council. Virgo is funded, through the European Gravitational Observatory (EGO), by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale di Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by institutions from Belgium, Germany, Greece, Hungary, Ireland, Japan, Monaco, Poland, Portugal, Spain. KAGRA is supported by Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan Society for the Promotion of Science (JSPS) in Japan; National Research Foundation (NRF) and Ministry of Science and ICT (MSIT) in Korea; Academia Sinica (AS) and National Science and Technology Council (NSTC) in Taiwan. This paper carries LIGO document number P2500629.

This research was supported in part by Perimeter Institute for Theoretical Physics. Research at Perimeter Institute is supported in part by the Government of Canada through the Department of Innovation, Science and Economic Development and by the Province of Ontario through the Ministry of Colleges and Universities.

The Flatiron Institute is a division of the Simons Foundation. H.S. was supported by Yuri Levin’s Simons Investigator Award 827103 while conducting this research.

Software: RINGDOWN [39], ARVIZ [88], PYMC [124], SEABORN [125], MATPLOTLIB [126], JUPYTER [127], NUMPY [128], SCIPY [129], QNM [89], DISPATCH [130],

PANDAS [131], PYTHON3 [132], CHATGPT [133] (used for figure plotting code).

Appendix A: Alternative QNM models better-fitting than {220, 200}

The QNM models with 220 and 210 are not the only ones we find that outperform the {220, 200} in terms of goodness of fit. We also find that the {220, 320} and {220, 320, 321} models perform as well as or better than the 200 models and at least as well as the 210 models, as shown in Fig. 12.

Despite the promising goodness of fit of the 320 models, we suggest that their parameter estimates evolve over time in physically implausible ways which lead us to prefer the 210 models. First, looking at the amplitudes, we can see that the {220, 320} amplitudes do not decay as expected until $10 t_M$, as shown in Fig. 13. The frequency and damping rate measurements of the {220, 320} model are also shifting over time more than the {220, 210} model, as shown in Fig. 14. By $10 t_M$, there are no longer meaningful goodness of fit preferences between the 210, 200, and 320 models, as shown in Fig. 12. So, although at earlier times the {220, 320} models may outperform the {220, 210} models in terms of goodness of fit, this does not seem to be physically meaningful.

When including a 321 overtone in the model, similar parameter time evolutions occur. Again, the frequencies and damping rates of the modes in the model slowly shift over time, as shown in Fig. 14. By contrast, the frequency and damping rate measurements of the {220, 210, 221} and {220, 200, 221} models are more self-consistent over time. Interestingly, when performing our test of the Kerr metric with the {220, 320, 321} model, Kerr consistency can be found as early as $-30 t_M$ (Fig. 15) – despite the fact that the amplitudes are not decaying as damped sinusoids, as shown in Fig. 13.

Ultimately, unusual behavior of these QNM fits aside, we appeal to our knowledge of IMR model systematics to claim that these 320 QNM models should not be considered on equal footing with those in the main text. If NRSur7dq4 is indeed systematically biased, its systematic uncertainty is unlikely to be many times that of the statistical uncertainty for GW231123 given the injection studies in Ref. [1]. If the 320 models were indeed the best description of the signal, it would imply that NRSur7dq4 was strongly dominated by systematic bias, to a degree that seems improbable. The time-dependent parameter behavior we see in these QNM models seems to highlight the difficult nature of our flexible analysis, and shows that consistency with the Kerr frequencies may not by itself be meaningful without consideration of the physicality of the other parameters of the model as motivated by theoretical understanding of black hole ringdown signals. Where the line is drawn as to exactly what constitutes physicality has an element of individual interpretation.

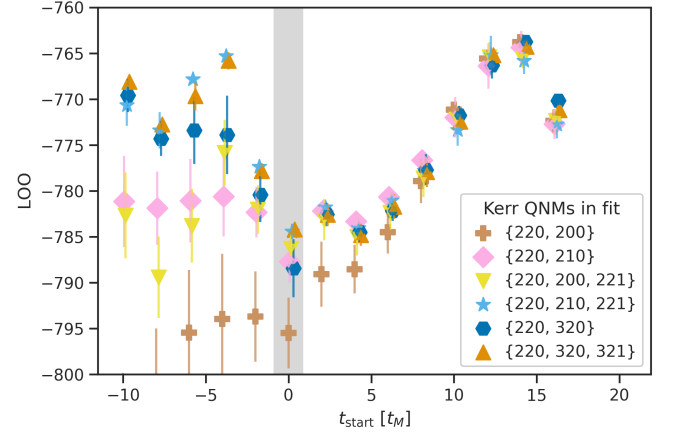


FIG. 12. LOO now including fits with the 320. See Fig. 1 for figure conventions. The {220, 320} model performs as well as {220, 210} from $-2 t_M$ onwards, and performs significantly better at earlier times. The {220, 320, 321} and {220, 210, 221} perform comparably.

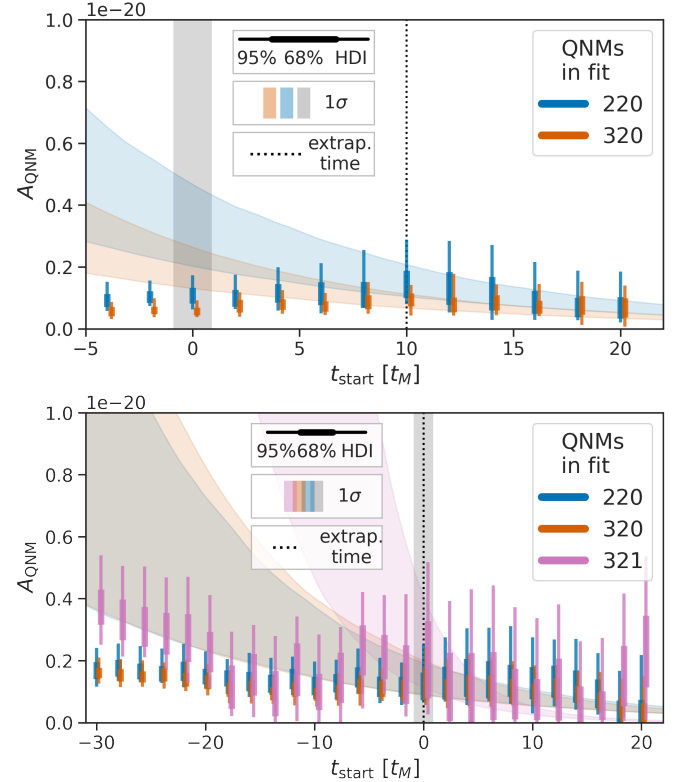


FIG. 13. The amplitudes of {220, 320} and {220, 320, 321} models do not follow physically well-motivated decays as clearly as those of the {220, 210, 221} and {220, 200, 221} models shown in Figs. 6 and 7, despite the model frequencies being consistent with Kerr as shown in Fig. 15. This is one piece of evidence which favors the models of the main text.

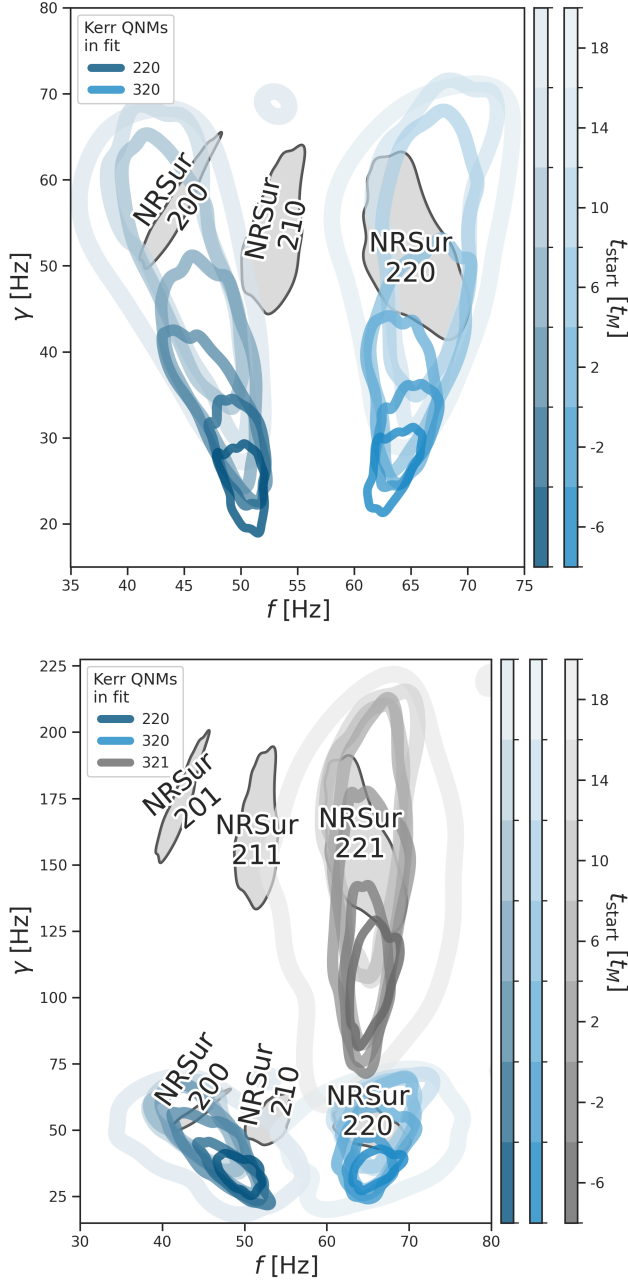


FIG. 14. The frequencies of the 320 models are placed roughly in the same locations as those of the 210 models in the main text (Figs. 4, 5), but for these 320 models the damping rates move more over time. Note that the mass prior for this model is different from those in the main text: it ranges from 250 to $550 M_{\odot}$.

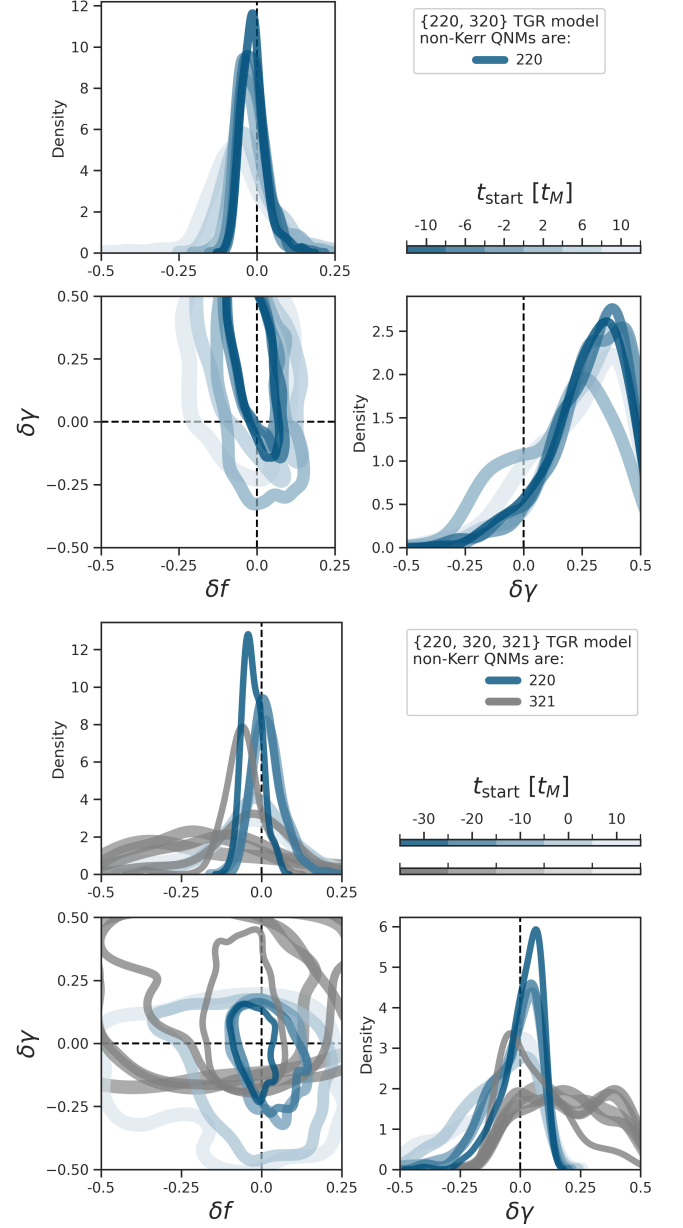


FIG. 15. We find the 320 models to be consistent with the Kerr metric at implausibly early times, and over spans of time where the amplitudes of the Kerr models do not decay as expected of QNMs (as shown in Fig. 13). Validation of the Kerr metric alone does not seem sufficient to guarantee that the model is physically meaningful, as there seems to be evidence suggesting that this model does not have the physical behavior we expect of QNM models and thus is likely a less accurate description of the signal than those considered in the main text.

-
- [1] R. Abbott *et al.* (LIGO Scientific, VIRGO, KAGRA), arXiv eprints (2025), arXiv:2507.08219 [astro-ph.HE].
 - [2] E. Capote *et al.*, Phys. Rev. D **111**, 062002 (2025).
 - [3] F. Acernese *et al.* (Virgo), arXiv eprints (2022), arXiv:2205.01555 [gr-qc].
 - [4] T. Akutsu *et al.*, Progress of Theoretical and Experimental Physics **2021**, 05A101 (2020), <https://academic.oup.com/ptep/article-pdf/2021/5/05A101/37974994/ptaa125.pdf>.
 - [5] J. Aasi *et al.* (LIGO Scientific), Class. Quant. Grav. **32**, 074001 (2015), arXiv:1411.4547 [gr-qc].
 - [6] A. G. Abac *et al.* (LIGO Scientific, VIRGO, KAGRA), GWTC-4.0: Updating the Gravitational-Wave Transient Catalog with Observations from the First Part of the Fourth LIGO-Virgo-KAGRA Observing Run (2025), arXiv:2508.18082 [gr-qc].
 - [7] 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**, 033015 (2019), arXiv:1905.09300 [gr-qc].
 - [8] C. Talbot *et al.*, Inference with finite time series II: the window strikes back (2025), arXiv:2508.11091 [gr-qc].
 - [9] M. Isi and W. M. Farr, arXiv eprints (2021), arXiv:2107.05609 [gr-qc].
 - [10] H. Siegel, M. Isi, and W. M. Farr, Phys. Rev. D **111**, 044070 (2025), arXiv:2410.02704 [gr-qc].
 - [11] A. G. Abac *et al.* (KAGRA, Virgo, LIGO Scientific), Phys. Rev. Lett. **135**, 111403 (2025), arXiv:2509.08054 [gr-qc].
 - [12] arXiv eprints (2025), arXiv:2509.08099 [gr-qc].
 - [13] S. A. Teukolsky, Astrophys. J. **185**, 635 (1973).
 - [14] O. Dreyer, B. J. Kelly, B. Krishnan, L. S. Finn, D. Garrison, and R. Lopez-Aleman, Class. Quant. Grav. **21**, 787 (2004), arXiv:gr-qc/0309007.
 - [15] S. Gossan, J. Veitch, and B. S. Sathyaprakash, Phys. Rev. D **85**, 124056 (2012), arXiv:1111.5819 [gr-qc].
 - [16] E. Berti *et al.*, Class. Quant. Grav. **32**, 243001 (2015), arXiv:1501.07274 [gr-qc].
 - [17] H. Zhu *et al.*, Phys. Rev. D **111**, 064052 (2025), arXiv:2312.08588 [gr-qc].
 - [18] I. Kamaretsos, M. Hannam, and B. Sathyaprakash, Phys. Rev. Lett. **109**, 141102 (2012), arXiv:1207.0399 [gr-qc].
 - [19] E. Berti *et al.*, arXiv eprints (2025), arXiv:2505.23895 [gr-qc].
 - [20] G. Carullo, W. Del Pozzo, and J. Veitch, Phys. Rev. D **99**, 123029 (2019), [Erratum: Phys.Rev.D 100, 089903 (2019)], arXiv:1902.07527 [gr-qc].
 - [21] M. Isi, M. Giesler, W. M. Farr, M. A. Scheel, and S. A. Teukolsky, Phys. Rev. Lett. **123**, 111102 (2019).
 - [22] M. Isi, W. M. Farr, M. Giesler, M. A. Scheel, and S. A. Teukolsky, Phys. Rev. Lett. **127**, 011103 (2021).
 - [23] R. Abbott *et al.* (LIGO Scientific, VIRGO, KAGRA), arXiv eprints (2021), arXiv:2112.06861 [gr-qc].
 - [24] M. Isi and W. M. Farr, arXiv:2202.02941 [gr-qc] (2022).
 - [25] E. Finch and C. J. Moore, Phys. Rev. D **106**, 043005 (2022).
 - [26] R. Cotesta, G. Carullo, E. Berti, and V. Cardoso, Phys. Rev. Lett. **129**, 111102 (2022).
 - [27] M. Isi and W. M. Farr, Phys. Rev. Lett. **131**, 169001 (2023), arXiv:2310.13869 [astro-ph.HE].
 - [28] H. Siegel, M. Isi, and W. M. Farr, Phys. Rev. D **108**, 064008 (2023), arXiv:2307.11975 [gr-qc].
 - [29] C. D. Capano, M. Cabero, J. Westerweck, J. Abedi, S. Kastha, A. H. Nitz, Y.-F. Wang, A. B. Nielsen, and B. Krishnan, Phys. Rev. Lett. **131**, 221402 (2023), arXiv:2105.05238 [gr-qc].
 - [30] C. D. Capano, J. Abedi, S. Kastha, A. H. Nitz, J. Westerweck, Y.-F. Wang, M. Cabero, A. B. Nielsen, and B. Krishnan, Class. Quant. Grav. **41**, 245009 (2024), arXiv:2209.00640 [gr-qc].
 - [31] R. Brito, A. Buonanno, and V. Raymond, Phys. Rev. D **98**, 084038 (2018), arXiv:1805.00293 [gr-qc].
 - [32] L. Pompili, E. Maggio, H. O. Silva, and A. Buonanno, Phys. Rev. D **111**, 124040 (2025), arXiv:2504.10130 [gr-qc].
 - [33] S. Ma, K. Mitman, L. Sun, N. Deppe, F. Hébert, L. E. Kidder, J. Moxon, W. Throwe, N. L. Vu, and Y. Chen, Phys. Rev. D **106**, 084036 (2022), arXiv:2207.10870 [gr-qc].
 - [34] N. Lu, S. Ma, O. J. Piccinni, L. Sun, and E. Finch, Phys. Rev. D **112**, 064047 (2025), arXiv:2505.18560 [gr-qc].
 - [35] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. Lett. **116**, 061102 (2016), arXiv:1602.03837 [gr-qc].
 - [36] M. Bailes *et al.*, Nature Rev. Phys. **3**, 344 (2021).
 - [37] E. E. Flanagan and S. A. Hughes, New J. Phys. **7**, 204 (2005), arXiv:gr-qc/0501041.
 - [38] O. Dreyer, B. J. Kelly, B. Krishnan, L. S. Finn, D. Garrison, and R. Lopez-Aleman, Class. Quant. Grav. **21**, 787 (2004), arXiv:gr-qc/0309007.
 - [39] M. Isi, H. Siegel, W. M. Farr, N. Khusid, A. Hussain, and R. Udall, maxisi/ringdown: v1.0.0 (2024).
 - [40] R. O'Shaughnessy, L. London, J. Healy, and D. Shoemaker, Phys. Rev. D **87**, 044038 (2013), arXiv:1209.3712 [gr-qc].
 - [41] E. Hamilton, L. London, and M. Hannam, Phys. Rev. D **107**, 104035 (2023), arXiv:2301.06558 [gr-qc].
 - [42] F. Nobili, S. Bhagwat, C. Pacilio, and D. Gerosa, Phys. Rev. D **112**, 044058 (2025), arXiv:2504.17021 [gr-qc].
 - [43] H. Estellés, M. Colleoni, C. García-Quirós, S. Husa, D. Keitel, M. Mateu-Lucena, M. d. L. Planas, and A. Ramos-Buades, Phys. Rev. D **105**, 084040 (2022), arXiv:2105.05872 [gr-qc].
 - [44] J. E. Thompson, E. Hamilton, L. London, S. Ghosh, P. Kolitsidou, C. Hoy, and M. Hannam, Phys. Rev. D **109**, 063012 (2024), arXiv:2312.10025 [gr-qc].
 - [45] A. Ramos-Buades, A. Buonanno, H. Estellés, M. Khalil, D. P. Mihaylov, S. Ossokine, L. Pompili, and M. Shiferaw, Phys. Rev. D **108**, 124037 (2023), arXiv:2303.18046 [gr-qc].
 - [46] G. Pratten *et al.*, Phys. Rev. D **103**, 104056 (2021), arXiv:2004.06503 [gr-qc].
 - [47] M. Colleoni, F. A. R. Vidal, C. García-Quirós, S. Akçay, and S. Bera, Phys. Rev. D **111**, 104019 (2025), arXiv:2412.16721 [gr-qc].
 - [48] J. Mac Uilliam, S. Akçay, and J. E. Thompson, Phys. Rev. D **109**, 084077 (2024), arXiv:2402.06781 [gr-qc].
 - [49] A. Dhani, S. Völkel, A. Buonanno, H. Estelles, J. Gair, H. P. Pfeiffer, L. Pompili, and A. Toubiana, (2024), arXiv:2404.05811 [gr-qc].
 - [50] I. Mandel, arXiv eprints (2025), arXiv:2509.05885 [astro-ph.HE].

- [51] F. Kiroğlu, K. Kremer, and F. A. Rasio, arXiv eprints (2025), arXiv:2509.05415 [astro-ph.HE].
- [52] H. Tong *et al.*, arXiv eprints (2025), arXiv:2509.04151 [astro-ph.HE].
- [53] L. Paiella, C. Ugolini, M. Spera, M. Branchesi, and M. A. Sedda, arXiv eprints (2025), arXiv:2509.10609 [astro-ph.GA].
- [54] G.-P. Li and X.-L. Fan, arXiv eprints (2025), arXiv:2509.08298 [astro-ph.HE].
- [55] T. W. Baumgarte and S. L. Shapiro, arXiv eprints (2025), arXiv:2509.04574 [astro-ph.HE].
- [56] S. A. Popa and S. E. de Mink, arXiv eprints (2025), arXiv:2509.00154 [astro-ph.HE].
- [57] O. Gottlieb, B. D. Metzger, D. Issa, S. E. Li, M. Renzo, and M. Isi, arXiv eprints (2025), arXiv:2508.15887 [astro-ph.HE].
- [58] V. Delfavero, S. Ray, H. E. Cook, K. Nathaniel, B. McKernan, K. E. S. Ford, J. Postiglione, E. McPike, and R. O'Shaughnessy, arXiv eprints (2025), arXiv:2508.13412 [gr-qc].
- [59] D. Croon, J. Sakstein, and D. Gerosa, arXiv eprints (2025), arXiv:2508.10088 [astro-ph.HE].
- [60] I. Bartos and Z. Haiman, arXiv eprints (2025), arXiv:2508.08558 [astro-ph.HE].
- [61] R. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. Lett. **125**, 101102 (2020), arXiv:2009.01075 [gr-qc].
- [62] R. Abbott *et al.* (LIGO Scientific, Virgo), Astrophys. J. Lett. **900**, L13 (2020), arXiv:2009.01190 [astro-ph.HE].
- [63] R. Abbott *et al.* (KAGRA, VIRGO, LIGO Scientific), Phys. Rev. X **13**, 041039 (2023), arXiv:2111.03606 [gr-qc].
- [64] M. A. Scheel *et al.*, arXiv eprints (2025), arXiv:2505.13378 [gr-qc].
- [65] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Class. Quant. Grav. **37**, 055002 (2020), arXiv:1908.11170 [gr-qc].
- [66] W. M. Farr, farr/linecleaner (2024).
- [67] H. Siegel and W. M. Farr, Line cleaner review gitlab (2025).
- [68] H. Siegel, Line cleaner injection study dcc slides (2025).
- [69] B. Zackay, T. Venumadhav, J. Roulet, L. Dai, and M. Zaldarriaga, Phys. Rev. D **104**, 063034 (2021), arXiv:1908.05644 [astro-ph.IM].
- [70] H.-T. Wang, S.-P. Tang, P.-C. Li, and Y.-Z. Fan, arXiv eprints (2025), arXiv:2509.02047 [gr-qc].
- [71] K. McGowan *et al.*, Gw231123 data quality report (2024).
- [72] H. Siegel, Data release.
- [73] V. Baibhav, M. H.-Y. Cheung, E. Berti, V. Cardoso, G. Carullo, R. Cotesta, W. Del Pozzo, and F. Duque, Phys. Rev. D **108**, 104020 (2023), arXiv:2302.03050 [gr-qc].
- [74] M. Giesler *et al.*, Phys. Rev. D **111**, 084041 (2025), arXiv:2411.11269 [gr-qc].
- [75] K. Mitman *et al.*, Phys. Rev. D **112**, 064016 (2025), arXiv:2503.09678 [gr-qc].
- [76] A. Vehtari, A. Gelman, and J. Gabry, arXiv e-prints, arXiv:1507.04544 (2015), arXiv:1507.04544 [stat.CO].
- [77] A. Vehtari, Cross-validation faq.
- [78] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed., Chapman & Hall/CRC Texts in Statistical Science Series (CRC, Boca Raton, Florida, 2013).
- [79] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer series in statistics (Springer, 2009).
- [80] Y. McLatchie and A. Vehtari, Statistics and Computing **34**, 132 (2024).
- [81] D. Navarro, Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection (2018).
- [82] C. Cutler and M. Vallisneri, Phys. Rev. D **76**, 104018 (2007), arXiv:0707.2982 [gr-qc].
- [83] N. Cornish, L. Sampson, N. Yunes, and F. Pretorius, Phys. Rev. D **84**, 062003 (2011), arXiv:1105.2088 [gr-qc].
- [84] R. Buscicchio, E. Roebber, J. M. Goldstein, and C. J. Moore, Phys. Rev. D **100**, 084041 (2019), arXiv:1907.11631 [astro-ph.IM].
- [85] S. Gossan, J. Veitch, and B. S. Sathyaprakash, Phys. Rev. D **85**, 124056 (2012), arXiv:1111.5819 [gr-qc].
- [86] W. M. Farr *et al.*, Marginalization prior technical note.
- [87] M. Isi, Class. Quant. Grav. **40**, 203001 (2023), arXiv:2208.03372 [gr-qc].
- [88] R. Kumar, C. Carroll, A. Hartikainen, and O. Martin, Journal of Open Source Software **4**, 1143 (2019).
- [89] L. C. Stein, J. Open Source Softw. **4**, 1683 (2019), arXiv:1908.10377 [gr-qc].
- [90] G. Carullo, W. D. Pozzo, D. Laghi, M. Isi, and J. Veitch, pyring, <https://git.ligo.org/lscsoft/pyring>.
- [91] R. Abbott, T. D. Abbott, S. Abraham, F. Acernese, K. Ackley, A. Adams, C. Adams, R. X. Adhikari, V. B. Adya, C. Affeldt, *et al.*, Phys. Rev. D **103**, 122002 (2021), arXiv:2010.14529 [gr-qc].
- [92] L. Magaña Zertuche, L. C. Stein, K. Mitman, S. E. Field, V. Varma, M. Boyle, N. Deppe, L. E. Kidder, J. Moxon, H. P. Pfeiffer, M. A. Scheel, K. C. Nelli, W. Throwe, and N. L. Vu, Phys. Rev. D **112**, 024077 (2025).
- [93] A. Ghosh, R. Brito, and A. Buonanno, Phys. Rev. D **103**, 124041 (2021), arXiv:2104.01906 [gr-qc].
- [94] V. Gennari, G. Carullo, and W. Del Pozzo, Eur. Phys. J. C **84**, 233 (2024), arXiv:2312.12515 [gr-qc].
- [95] M. H.-Y. Cheung, E. Berti, V. Baibhav, and R. Cotesta, Phys. Rev. D **109**, 044069 (2024), [Erratum: Phys.Rev.D 110, 049902 (2024)], arXiv:2310.04489 [gr-qc].
- [96] T. A. Clarke *et al.*, Phys. Rev. D **109**, 124030 (2024), arXiv:2402.02819 [gr-qc].
- [97] I. M. Romero-Shaw, C. Talbot, S. Biscoveanu, V. D'Emilio, G. Ashton, C. P. L. Berry, S. Coughlin, S. Galaudage, C. Hoy, M. Hübner, K. S. Phukon, M. Pitkin, M. Rizzo, N. Sarin, R. Smith, S. Stevenson, A. Vajpeyi, M. Arène, K. Athar, S. Banagiri, N. Bose, M. Carney, K. Chatziioannou, J. A. Clark, M. Colleoni, R. Cotesta, B. Edelman, H. Estellés, C. García-Quirós, A. Ghosh, R. Green, C.-J. Haster, S. Husa, D. Keitel, A. X. Kim, F. Hernandez-Vivanco, I. Magaña Hernandez, C. Karathanasis, P. D. Lasky, N. De Lillo, M. E. Lower, D. Macleod, M. Mateu-Lucena, A. Miller, M. Millhouse, S. Morisaki, S. H. Oh, S. Ossokine, E. Payne, J. Powell, G. Pratten, M. Pürrer, A. Ramos-Buades, V. Raymond, E. Thrane, J. Veitch, D. Williams, M. J. Williams, and L. Xiao, Monthly Notices of the Royal Astronomical Society **499**, 3295 (2020), <https://academic.oup.com/mnras/article-pdf/499/3/3295/34052625/staa2850.pdf>.

- [98] A. Chavda, M. Lagos, and L. Hui, JCAP **07**, 084, arXiv:2412.03435 [gr-qc].
- [99] C. Pacilio, S. Bhagwat, F. Nobili, and D. Gerosa, Phys. Rev. D **110**, 103037 (2024), arXiv:2408.05276 [gr-qc].
- [100] X. J. Forteza, S. Bhagwat, S. Kumar, and P. Pani, Phys. Rev. Lett. **130**, 021001 (2023), arXiv:2205.14910 [gr-qc].
- [101] F.-L. Julié, L. Pompili, and A. Buonanno, Phys. Rev. D **111**, 024016 (2025), arXiv:2406.13654 [gr-qc].
- [102] D. D. Doneva, L. Aresté Saló, K. Clough, P. Figueras, and S. S. Yazadjiev, Phys. Rev. D **108**, 084017 (2023), arXiv:2307.06474 [gr-qc].
- [103] M. Corman, J. L. Ripley, and W. E. East, Phys. Rev. D **107**, 024014 (2023), arXiv:2210.09235 [gr-qc].
- [104] M. Corman, L. Lehner, W. E. East, and G. Dideron, Phys. Rev. D **110**, 084048 (2024), arXiv:2405.15581 [gr-qc].
- [105] J. Cayuso, N. Ortiz, and L. Lehner, Phys. Rev. D **96**, 084043 (2017), arXiv:1706.07421 [gr-qc].
- [106] S. Bhagwat, X. J. Forteza, P. Pani, and V. Ferrari, Phys. Rev. D **101**, 044033 (2020).
- [107] K. Chandra and J. Calderón Bustillo, arXiv eprints (2025), arXiv:2509.17315 [gr-qc].
- [108] J. Calderón Bustillo, P. D. Lasky, and E. Thrane, Phys. Rev. D **103**, 024041 (2021), arXiv:2010.01857 [gr-qc].
- [109] B. Efron and T. Hastie, *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*, Institute of Mathematical Statistics Monographs (Cambridge University Press, 2021).
- [110] R. H. Price, S. Nampalliwar, and G. Khanna, Phys. Rev. D **93**, 044060 (2016), arXiv:1508.04797 [gr-qc].
- [111] N. Oshita, S. Ma, Y. Chen, and H. Yang, arXiv eprints (2025), arXiv:2509.09165 [gr-qc].
- [112] J. E. Thompson, C. Hoy, E. Fauchon-Jones, and M. Hannam, Phys. Rev. D **112**, 064011 (2025), arXiv:2506.10530 [gr-qc].
- [113] M. Boyle, L. E. Kidder, S. Ossokine, and H. P. Pfeiffer, Gravitational-wave modes from precessing black-hole binaries (2014), arXiv:1409.4431 [gr-qc].
- [114] H. Siegel, K. Mitman, M. Isi, *et al.* (in prep).
- [115] A. Vijaykumar, A. G. Hanselman, and M. Zevin, Astrophys. J. **969**, 132 (2024), arXiv:2402.07892 [astro-ph.HE].
- [116] J. Calderón Bustillo, N. Sanchis-Gual, A. Torres-Forné, and J. A. Font, Phys. Rev. Lett. **126**, 201101 (2021), arXiv:2009.01066 [gr-qc].
- [117] A. Liu, I. C. F. Wong, S. H. W. Leong, A. More, O. A. Hannuksela, and T. G. F. Li, Mon. Not. Roy. Astron. Soc. **525**, 4149 (2023), arXiv:2302.09870 [gr-qc].
- [118] A. Hussain and A. Zimmerman, Phys. Rev. D **106**, 104018 (2022), arXiv:2206.10653 [gr-qc].
- [119] D. Li, P. Wagle, Y. Chen, and N. Yunes, Phys. Rev. X **13**, 021029 (2023), arXiv:2206.10652 [gr-qc].
- [120] F. Crescimbeni, X. J. Forteza, S. Bhagwat, J. Westerweck, and P. Pani, Theory-agnostic searches for non-gravitational modes in black hole ringdown (2024), arXiv:2408.08956 [gr-qc].
- [121] J. Lestingi, G. D’Addario, and T. P. Sotiriou, Phys. Rev. D **112**, 064070 (2025), arXiv:2505.18261 [gr-qc].
- [122] D. Li, A. Hussain, P. Wagle, Y. Chen, N. Yunes, and A. Zimmerman, Phys. Rev. D **109**, 104026 (2024), arXiv:2310.06033 [gr-qc].
- [123] S. Maenaut, G. Carullo, P. A. Cano, A. Liu, V. Cardoso, T. Hertog, and T. G. F. Li, arXiv eprints (2024), arXiv:2411.17893 [gr-qc].
- [124] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, PeerJ Computer Science **2**, e55 (2016).
- [125] M. L. Waskom, Journal of Open Source Software **6**, 3021 (2021).
- [126] J. D. Hunter, Computing in Science & Engineering **9**, 90 (2007).
- [127] T. Kluyver *et al.*, Jupyter notebooks – a publishing format for reproducible computational workflows (2016).
- [128] C. R. Harris *et al.*, Nature **585**, 357 (2020).
- [129] P. Virtanen *et al.*, Nature Methods **17**, 261 (2020).
- [130] F. Institute.
- [131] T. pandas development team, pandas-dev/pandas: Pandas (2020).
- [132] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual* (CreateSpace, Scotts Valley, CA, 2009).
- [133] OpenAI, Chatgpt, <https://chat.openai.com> (2024), large language model.