Supplement to "Black Hole Spectroscopy and Tests of General Relativity with GW250114 "

(Dated: November 24, 2025)

Significance estimation for a non-negative quasi-normal mode (QNM) amplitude

In the main text we quote the significance with which the amplitude A of a given QNM is shown to be > 0. We use the one-dimensional highest posterior density (HPD) credible regions based on samples drawn from the posterior density. The one dimensional p HPD is the shortest interval that contains a fraction p of the samples. It can be computed by considering the $\lfloor (1-p)\,N \rfloor$ intervals in the sorted samples that each include $\lceil pN \rceil$ samples, where N is the total number of samples available, and choosing the shortest one (i.e. the one with the highest estimated density). Using sample-based HPD intervals in this way can produce significant differences between the true and estimated interval when $p \simeq 1/N$ [1], but in all cases in this paper, we have sufficient samples so that $pN \gg 1$, and this is not an issue.

To convert a p HPD into a significance level, $x\sigma$, we relate these quantities through the Gaussian distribution, via

$$p = \int_{-x}^{x} \mathrm{d}y\, \phi(y), \tag{1}$$

where $\phi(y)$ is the standard normal density. For example, a 3σ interval contains a fraction $p = 0.9973$ of the samples.

Validity regimes of QNM models

Analyses of numerical-relativity (NR) simulations of binary black hole (BBH) mergers, like GW250114, with Eq. (1) in the main text show that the time interval in which fits at consecutive start times yield exponentially consistent amplitudes varies based on the mode content [2–4]. To determine when Eq. (1) in the main text is valid for the models we consider, we study the equal-mass, non-spinning NR BBH simulation `SXS:BBH:3617` [5, 6], whose intrinsic parameters lie within the 90% credible region of GW250114. For each QNM model, we perform a linear least-squares fit to the $(2,2)$ mode with the start times $t_> \in [0, 40] t_{M_\mathrm{f}}$ in steps of $0.1 t_{M_\mathrm{f}}$ and measure the complex amplitude of each QNM at t_peak [7]. We define a stability window of size $10 t_{M_\mathrm{f}}$ as a region in which the QNMs in the model have amplitudes whose fractional variation is comparable to our 50% credible level measurement uncertainties in GW250114 (14% for the 220 at $t_> = 11 t_{M_\mathrm{f}}$ with the 220 model, 24% for the 221 at $t_> = 6 t_{M_\mathrm{f}}$ with the 220+221 model, and 40% for the 222 at $t_> = 3 t_{M_\mathrm{f}}$ with the 220+221+222 model; these times are chosen based on the value used in Abac *et al.* [1] and the times identified in Fig. 1) and whose values agree with the most stable value to the same uncertainty. We define the stability regime as the union of all such windows. For the 220, 220+221, 220+221+222 models, we find these regimes to be approximately $[11, 40] t_{M_\mathrm{f}}$, $[6, 34] t_{M_\mathrm{f}}$, and $[3, 22] t_{M_\mathrm{f}}$.

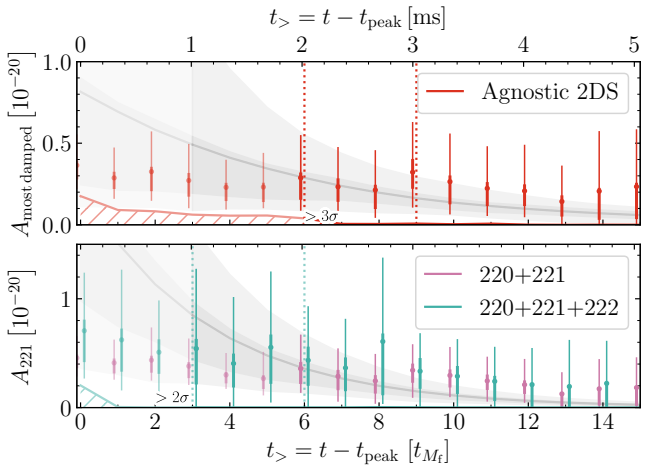

FIG. 1. Consistency of a NR-simulation's post-merger data with two QNMs. Identical to Fig. 1 in the main text, but instead for a simulated signal using the equal-mass, non-spinning NR BBH simulation `SXS:BBH:3617` with Gaussian noise; fits are performed with `ringdown`.

Comparison to NR-informed predictions

In Fig. 1 in the main text, we have found that at least two agnostic damped sinusoids are required to explain the post-merger signal, and when assuming a Kerr remnant, we have observed that the 221 amplitudes remain non-zero at $> 3\sigma$ up to $9 t_{M_\mathrm{f}}$ after the peak. Here, we find similar results when employing a simulated signal of the NR simulation `SXS:BBH:3617` [5, 6]. The signal is injected into Gaussian noise generated using the power spectral density estimated from GW250114, with extrinsic parameters set to the preliminary-reference maximum-likelihood values inferred using `NRSur7dq4` [1]. We generate 10 different noise realizations and choose the one such that the analysis of the simulated signal most closely resembles that obtained for GW250114. Like in Fig. 1 in the main text, we find (see upper panel in Fig. 1) that at around $t_> \gtrsim 9 t_{M_\mathrm{f}}$ the amplitude of the more rapidly decaying damped sinusoid becomes consistent with zero at the 90% level. Furthermore, at around $t_> \lesssim 6 t_{M_\mathrm{f}}$ (lower panel), the amplitude of the 221 mode in the 220+221 model also starts to deviate from its extrapolated values; but, by adding the 222 mode to the fit one can again recover amplitude consistency at early times around $t_> \gtrsim 3 t_{M_\mathrm{f}}$.

We perform another test to verify that the relative amplitudes and phases of the 220 and 221 QNMs are consistent with predictions from NR. This requires mapping the detector QNM amplitudes in Eq. (1) in the main text to the remnant-

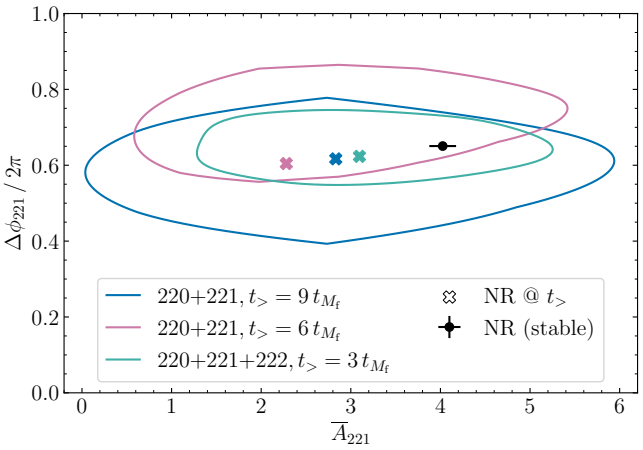

FIG. 2. Consistency of the 220 and 221 QNMs from GW250114 with NR predictions. The 50% credible regions of the two-dimensional posteriors on $\Delta\phi_{221}$ and $\overline{A}_{221}$ for GW250114, at varying fit start times, as measured by `ringdown`. Colored crosses represent fits to the equal-mass, non-spinning NR BBH simulation `SXS:BBH:3617`. For $t_> = t - t_{\rm peak}[t_{M_f}] \in \{6, 9\} t_{M_f}$ the 220+221 model is used, while for $t_> = 3t_{M_f}$, the 220+221+222 model is used. The black marker represents a stable value of these quantities over a window ranging from $t_> \in [12, 22]\, t_{M_f}$, with the data point the mean and the error bars the 1σ variation over said window.

frame QNM amplitudes, with which the strain over the two-sphere in the remnant frame can be written as

$$h \equiv h_+ - i h_\times = \sum_{\substack{\ell\geq 2 \\ -\ell\leq m\leq \ell \\ n\geq 0}} C_{\ell mn} e^{-i\omega_{\ell mn} t}\, {}_{-2}S_{\ell mn}(M_{\rm f}\chi_{\rm f}\omega_{\ell mn}, \theta_{JN}, \varphi), \quad (2)$$

where $\omega_{\ell mn} \equiv 2\pi f_{\ell mn} - i/\tau_{\ell mn}$ and ${}_{-2}S_{\ell mn}(M_{\rm f}\chi_{\rm f}\omega_{\ell mn}, \theta_{JN}, \varphi)$ is the spin-weight -2 ℓmn spheroidal harmonic with oblateness $M_{\rm f}\chi_{\rm f}\omega_{\ell mn}$ and evaluation point on the two-sphere (θ_{JN}, φ). The complex amplitudes $C_{\ell mn}$ in Eq. (2) are related to the complex amplitudes in Eq. (1) in the main text via [8]

$$C_{\ell+|m|n}\, {}_{-2}S_{\ell+|m|n}(\theta_{JN}, \varphi) = A^{\rm R}_{\ell mn}, \quad (3a)$$

$$C_{\ell-|m|n}\, {}_{-2}S_{\ell-|m|n}(\theta_{JN}, \varphi) = A^{\rm L}_{\ell mn}. \quad (3b)$$

However, because $A^{\rm R}_{220}$ is measured to be larger than $A^{\rm L}_{220}$ for this event, from here on we focus exclusively on the $\ell|m|n$ QNMs. Using the QNM amplitudes over the two-sphere, we can compare the relative amplitudes and phases of the 220 and 221 QNMs via

$$\overline{A}_{221} = \left|\frac{C_{221}}{C_{220}}\right|, \quad (4a)$$

$$\Delta\phi_{221} = \arg\left[\frac{C_{221}}{C_{220}}\right]. \quad (4b)$$

The quantities $\overline{A}_{\ell mn}$ and $\Delta\phi_{\ell mn}$ should be interpreted as the amplitude ratio and phase difference between the 220 and 221 QNMs on the two-sphere extrapolated to $t_{\rm peak}$. Each contour in Fig. 2 shows the 50% credible region for the fit performed at the start time indicated in the legend. For $t_> \in \{6, 9\} t_{M_f}$ the 220+221 model is used, while for $t_> = 3t_{M_f}$, the 220+221+222 model is used. The corresponding crosses represent the values extracted from fits to the NR simulation `SXS:BBH:3617` performed using a linear least-squares fit to the $(2, 2)$ mode [7]. The marker with error bars is obtained by fitting the 220+221+222 model over a $10t_{M_f}$ window in which these quantities have stabilized. That is, we fit the NR waveform at times $t_> \in [12, 22]\, t_{M_f}$—the latest $10t_{M_f}$ contained in the regime of validity for the 220+221+222 model, measure the mean and standard deviation of C_{220} and C_{221} over said $10t_{M_f}$ window, and then compute $\overline{A}_{221}$ and $\Delta\phi_{221}$ accordingly. As can be seen, from $t_> \in [3, 9]\, t_{M_f}$ the amplitudes and phases extracted from the data are broadly consistent with the values predicted. More specifically, they are consistent with the most stable value to $\geq 38\%$ credibility. Their values evolve because the amplitudes of the exponentially damped-sinusoid model that are used are insufficient to recover particularly stable overtone amplitudes for these $t_>$ values at effectively infinite signal-to-noise ratio (SNR).

Overall, these results indicate that the 221 amplitudes and phases measured from GW250114 when using different ringdown models at different starting times are consistent with expectations from a BBH in general relativity (GR).

Final mass and spin consistency

In Fig. 3 we show the mass and spin posteriors obtained from the 220, 220+221, and 220+221+222 model fits at $t_> = 11\, t_{M_f}$, $t_> = 6\, t_{M_f}$, and $t_> = 3\, t_{M_f}$, and from the 220 and 440 `pSEOBNR` model. For all these fits, the inferred mass and spin are consistent with that from the full IMR analysis at 90% credibility.

The fact that including multiple overtones allows for a consistent measurement of the remnant mass and spin at earlier times has been established in several studies [9–13]. However, the physical significance of higher-order overtones close to the peak remains an active area of research [2, 14].

Detectability of quadrupolar first overtone via model selection

In the main text, we have assessed QNM detectability based on the posterior amplitude support away from zero. A complementary criterion is represented by the Bayes factor, which quantifies the ratio of the evidences of competing models, with each evidence defined as the integral of the likelihood weighted by the prior [15]. In Fig. 4, we show the $\log_{10}$ Bayes factors of the two-mode model compared to the one-mode model over time. The results from `pyRing` are obtained from the `cpnest` nested sampler used in inference [16], while the ones from `ringdown` are estimated from the 221 amplitudes using the Savage–Dickey ratio [17]. For nested models, these

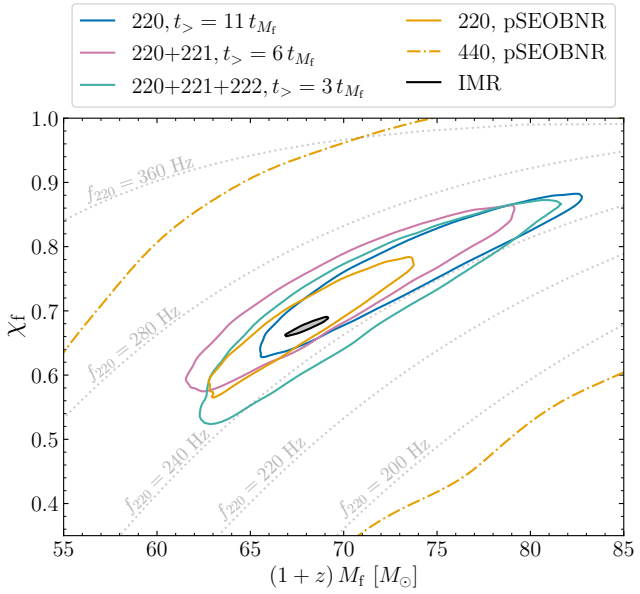

FIG. 3. Final mass and spin consistency. The 90% credible regions of the remnant mass and spin inferred from a series of ringdown fits with `pyRing`: using the 220 model at $t_> = 11\,t_{M_f}$, the 220+221 model at $t_> = 6\,t_{M_f}$, and the 220+221+222 model at $t_> = 3\,t_{M_f}$. We also show results for the 220 and 440 modes with the `pSEOBNR` analysis, as well as the IMR results of the full signal.

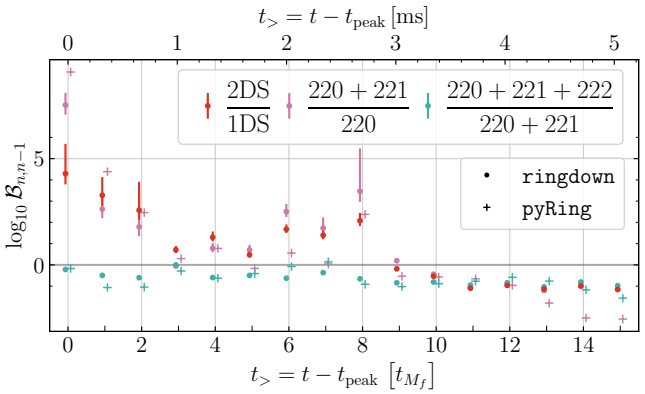

FIG. 4. Consistency of post-merger data with two QNMs via model selection. The Bayes factors comparing the analyzed QNM model to a nested model with one fewer mode. Dots indicate the estimated `ringdown` Bayes factor and plus the `pyRing` Bayes factor.

two approaches are equivalent [18]. For Savage–Dickey, we derive the amplitude maximum prior from the `pyRing` maximum priors on the right- and left-handed polarized contributions of the modes (between $[0, 5\times10^{-20}]$) combined with the spin-weighted spherical harmonics [1]. The reported values correspond to the median and 90% credible interval of 1000 Bayes factor estimates with bootstrap resamples of the amplitude samples. The `pyRing` uncertainties are not displayed since they are comparable to the marker size; they represent the half-width of the 90% credible interval estimated from the four nested sampling chains.

We find positive evidence for the presence of the 221 as late as $8t_{M_f}$ ($9t_{M_f}$) from `pyRing` (`ringdown`), with $\log_{10}$ Bayes factors of $0.56^{+0.26}_{-0.26}$ ($2.51^{+0.36}_{-0.26}$) at $6t_{M_f}$, $2.38^{+0.31}_{-0.31}$ ($3.47^{+2.02}_{-0.50}$) at $8t_{M_f}$ and $-0.53^{+0.31}_{-0.31}$ ($0.20^{+0.09}_{-0.08}$) at $9t_{M_f}$. A similar trend is observed using agnostic sinusoids. The presence of a second overtone 222 is not significantly preferred at any time. These results are in agreement with the QNM rational filter (`QNMRF`) detection statistics in Fig. 3 in the main text, as well as with the non-zero amplitude consistency in Fig. 1 in the main text, as discussed in the main text. They are also in accord with predictions from analysis of similarly loud simulated signals [19].

The difference between the two pipelines is due to different prior volumes and data length used. In fact, we are able to obtain the same values within the Bayes factor uncertainty when analyzing with `pyRing` 0.6 s of data with similar priors to `ringdown`, and when running `ringdown` on the same `pyRing` settings. For a complete description of the differences in settings between the codes, see Supplemental Material in Abac *et al.* [1].

QNM rational filter

The `QNMRF` analysis [20] is designed to isolate and remove specific complex-valued QNMs from the ringdown signal in the frequency domain. The resulting residual is then compared to pure colored Gaussian noise [21, 22]. The filter is constructed for a given set of QNMs, corresponding to specific black hole (BH) masses and spins.

`QNMRF` computes the mode detection statistic, $\mathcal{D}$, defined as a comparison between two ringdown model hypotheses: $\mathcal{H}$, which includes an additional mode, and $\mathcal{H}'$, which excludes it. $\mathcal{D}$ is analogous to a logarithmic Bayes factor, yet differs from the Bayes factors used in other time-domain ringdown analyses [23]. To assess statistical significance, we then take a frequentist approach to estimate false-alarm probabilitys (FAPs) due to the background noise, and determine a threshold, $\mathcal{D}_{1\%}$, corresponding to a 1% FAP. Given the data d, the statistic is expressed as:

$$\mathcal{D}(\mathcal{H}:\mathcal{H}') = \log_{10}\frac{\mathcal{Z}(d|\mathcal{H})}{\mathcal{Z}(d|\mathcal{H}')}, \tag{5}$$

where $\mathcal{Z}(d|\mathcal{H})$ and $\mathcal{Z}(d|\mathcal{H}')$ denote the evidences under hypotheses $\mathcal{H}$ and $\mathcal{H}'$.

In Fig. 3 in the main text, the pink crosses represent the offset of $\mathcal{D}[\mathcal{H}(220+221):\mathcal{H}'(220)]$ relative to the 1% FAP threshold evaluated on background noise, while the green plus indicate the offset of $\mathcal{D}[\mathcal{H}(220+221+222):\mathcal{H}'(220+221)]$ relative to its corresponding threshold.

For each ringdown model hypothesis, we also compute the joint posterior quantile $p(M_f^{\rm IMR},\chi_f^{\rm IMR})$ of the remnant BH mass and spin, inferred from the full IMR analysis, using `NRSur7dq4`. This credible contour represents the region on which the inferred parameters $(M_f^{\rm IMR},\chi_f^{\rm IMR})$ lie, serving as

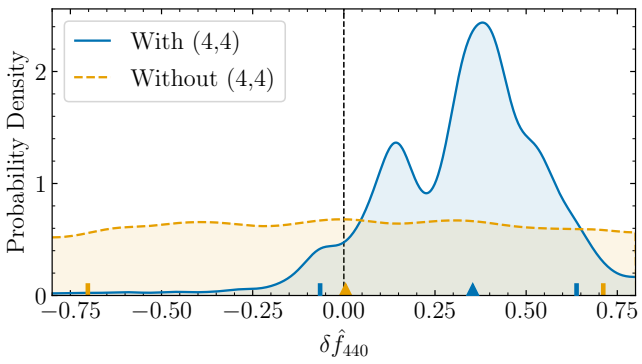

FIG. 5. Validation of the 440-mode constraint using simulated signals. Posterior distributions for the fractional deviation in the 440 QNM frequency, $\delta\hat{f}_{440}$, obtained from simulated NR signals consistent with GW250114. The blue curve shows results for a simulated signal that includes the $(\ell, |m|) = (4, 4)$ multipoles, while the orange curve corresponds to a simulated signal with the modes removed. Triangles mark the median values and vertical bars the symmetric 90% credible interval.

a consistency check with the IMR results [23]. A lower p value indicates a better match between the IMR analysis and the given QNM hypothesis. Among all possible additional modes added to the nested model, one at a time, the mode that (i) yields a detection statistic $\mathcal{D}$ above the threshold and simultaneously (ii) results in the greatest reduction in $p(M_{\rm f}^{\rm IMR}, \chi_{\rm f}^{\rm IMR})$ is considered confidently identified for a given starting time. In Fig. 3 in the main text, when the two-mode model $\mathcal{H}(220 + 221)$ is favored over the single-mode model $\mathcal{H}'(220)$, i.e. when $\mathcal{D}[\mathcal{H}(220 + 221) : \mathcal{H}'(220)] > \mathcal{D}_{1\%}$, we find that the corresponding joint posterior quantile $p(M_{\rm f}^{\rm IMR}, \chi_{\rm f}^{\rm IMR})$ also decreases across starting times from $4\,t_{M_{\rm f}}$ to $10\,t_{M_{\rm f}}$, indicating improved consistency with the IMR-inferred remnant parameters. Similarly, when the three-mode model $\mathcal{H}(220 + 221 + 222)$ is preferred over $\mathcal{H}'(220 + 221)$ with $\mathcal{D} > \mathcal{D}_{1\%}$ from starting times $1\,t_{M_{\rm f}}$ to $5\,t_{M_{\rm f}}$, we observe a reduction in $p(M_{\rm f}^{\rm IMR}, \chi_{\rm f}^{\rm IMR})$. These results suggest that these additional modes enhance the fit to the data and also improve the IMR consistency.

Validation of the $(4, 4)$ fundamental-mode frequency constraint

To validate the constraint on the 440 QNM frequency, we perform targeted simulated-signal studies, recovering the signals using the `pSEOBNR` model. We analyze a synthetic signal generated using the equal-mass, non-spinning NR simulation `SXS:BBH:3617` [5, 6, 24], with extrinsic parameters compatible with GW250114, and injected into Gaussian noise. Several features observed in the real GW250114 data, particularly those associated with the 440 QNM, can be qualitatively reproduced in this Gaussian noise injection. As shown in Fig. 5, the 440 QNM frequency is recovered with comparable precision to the real-signal analysis, while the damping time remains largely unconstrained.

In both the real event and Gaussian-noise injections, the posterior for $\delta\hat{f}_{440}$ can exhibit a multimodal structure, especially under a wide, uninformative prior for $\delta\hat{\tau}_{440}$. By contrast, an analogous injection into zero-noise data yields posteriors peaked at $\delta\hat{f}_{440} = 0$ with no significant substructure. The observed multimodality in $\delta\hat{f}_{440}$ is primarily driven by samples with large values of $\delta\hat{\tau}_{440}$, corresponding to long-lived, nearly sinusoidal modes. Indeed, for a purely sinusoidal signal, the likelihood is expected to exhibit secondary maxima with regular spacing in frequency [25]. When using an extended prior allowing for $\delta\hat{\tau}_{440}$ values up to 4, multimodality in $\delta\hat{f}_{440}$ is clearly associated with large $\delta\hat{\tau}_{440}$ samples. The maximum-likelihood parameters lie near the GR-consistent central mode, indicating that the secondary peaks are not favored by the data. Given the low SNR in the 440 mode, we do not expect to constrain deviations to its damping time. For the main results shown in Fig. 4 in the main text, we adopt a uniform prior on $\delta\hat{\tau}_{440} \in [-0.8, 0.8]$ consistent with that used for the ringdown constraints on the 221 mode.

Finally, we repeat the Gaussian-noise injection, using the same NR simulation and parameters, but with the $(\ell, |m|) = (4, 4)$ multipoles explicitly removed from the signal. In this case, the resulting constraint on $\delta\hat{f}_{440}$ is uninformative (orange curve in Fig. 5) suggesting that the posteriors recovered in the full injections are driven by the presence of the $(4, 4)$ multipoles in the data. These findings reinforce the interpretation that the constraint on $\delta\hat{f}_{440}$ obtained from GW250114 reflects a genuine physical feature of the signal rather than an artifact of the analysis or noise.

Principal component analysis

A limitation of the Flexible Theory Independent (`FTI`) and Test Infrastructure for GEneral Relativity (`TIGER`) results presented above is that individual PN deformation parameters are varied one at a time with all other parameters being fixed to the GR baseline. Whilst robust [26], single-parameter tests do not probe correlations across multiple PN orders, potentially missing more complex departures from GR. An alternative scheme was proposed in [27, 28], in which six PN deformation parameters are simultaneously varied, taken to be the 1.5PN to the 3.5PN parameters. The −1PN, 0PN, 0.5PN, and 1PN terms are fixed to their GR values. The approach is to estimate the joint posterior for the standard binary parameters plus the 6 PN deformation parameters. Then, one marginalizes over the GR parameters to yield a six-dimensional posterior for the PN deformations that captures correlated deviations [27–30], though strong parameter correlations can often render the posteriors uninformative or weakly constrained. Priors on the deformation parameters are taken to be uniform, such that $\delta\hat{\varphi}^{(i)}_{\rm prior} \sim \mathcal{U}(-20, 20)$.

To mitigate against this potential shortcoming, we apply a Principal-Component Analysis (`PCA`) to the six-dimensional posteriors, diagonalize the covariance matrix and identify the

linear combination of PN deformation parameters that are best constrained by the data [27, 28]. The new basis, $\delta\hat{\varphi}^{(i)}_{\rm PCA}$, provides orthogonal directions that minimize posterior widths. The PCA analysis is applied to the TIGER framework, using IMRPhenomXPHM_SpinTaylor [31, 32], and the FTI framework, using SEOBNRv5HM_ROM [33]. We find that the leading two PCA parameters are informative using the TIGER pipeline, and the leading three when using FTI due to its stricter constraints on individual PN coefficients, with all tests being consistent with zero. The 90% credible bounds on the leading PCA parameter $\delta\hat{\varphi}^{(1)}_{\rm PCA}$ for GW250114 are $-0.01^{+0.02}_{-0.02}$ (FTI) and $0.02^{+0.05}_{-0.05}$ (TIGER) respectively. The leading and sub-leading PCA parameters can be re-expressed as weighted combinations of the PN deformation coefficients. From the FTI analysis of GW250114, we find

$$\begin{aligned}\delta\hat{\varphi}^{(1)}_{\rm PCA} &= 0.7505\,\delta\hat{\varphi}_3 - 0.1336\,\delta\hat{\varphi}_4 + 0.4895\,\delta\hat{\varphi}_{5l} \\ &\quad - 0.3779\,\delta\hat{\varphi}_6 + 0.0241\,\delta\hat{\varphi}_{6l} + 0.1897\,\delta\hat{\varphi}_7, \quad \text{(6a)}\\ \delta\hat{\varphi}^{(2)}_{\rm PCA} &= 0.6358\,\delta\hat{\varphi}_3 + 0.0141\,\delta\hat{\varphi}_4 - 0.4144\,\delta\hat{\varphi}_{5l} \\ &\quad + 0.5414\,\delta\hat{\varphi}_6 + 0.0239\,\delta\hat{\varphi}_{6l} - 0.3608\,\delta\hat{\varphi}_7. \quad \text{(6b)}\end{aligned}$$

The PCA coefficients are dominated by the 1.5PN, 2.5PN log, and 3PN terms, in broad agreement with the individual PN coefficient analysis, as seen in Fig. 5 in the main text. This analysis demonstrates that even when allowing for correlated deviations across multiple post-Newtonian (PN) orders, the deviations away from GR inferred from GW250114 alone are constrained to be negligible.

Bounds on the black hole area theorem

A key outcome of Abac *et al.* [1] was a precision constraint on Hawking's area theorem [34], a fundamental consequence of the second law of BH mechanics stating that the horizon area of a BH cannot decrease over time. In practice, this implies that for BH mergers, the area of the final remnant must exceed the combined area of the two progenitor BHs [34]. Analogously to Abac *et al.* [1], we test this prediction by independently estimating the initial and final BH areas using different portions of the signal; however, differently from Abac *et al.* [1], we employ the entire signal, whereas in the other analysis the data around merger are excluded. Our approach closely follows the IMR consistency test. We constrain the masses and spins of the BHs in the inspiral and post-inspiral phases, which we directly map to the initial and final areas.

The areas are calculated using the Kerr formula [35]

$$\mathcal{A}(m,\chi) = 8\pi\left(\frac{Gm}{c^2}\right)^2\left(1+\sqrt{1-\chi^2}\right), \quad (7)$$

where m and χ are the BH mass and dimensionless spin. For the initial area $\mathcal{A}_{\rm i}$, we infer the individual BH masses and spins from the inspiral, and the total area is calculated as $\mathcal{A}_{\rm i} = \mathcal{A}_1 + \mathcal{A}_2$. For the final area $\mathcal{A}_{\rm f}$, we employ NR calibrated fits to estimate the remnant BH mass and spin from the progenitor parameters [36, 37], emphasizing that the initial BH source properties used in this calculation are inferred exclusively from post-inspiral data. In Fig. 6, we show the fractional difference between the final and initial areas, $(\mathcal{A}_{\rm f} - \mathcal{A}_{\rm i})/\mathcal{A}_{\rm i}$. We find that GW250114 is consistent with the area theorem at the $4.8\sigma_{\rm IMRCT}$ credibility level. Here, the significance $X\sigma_{\rm IMRCT}$ is calculated from the difference in areas and defined as the ratio of the difference in means to the standard deviation of the differences [1],

$$X = \frac{\mu_{\rm f} - \mu_{\rm i}}{\sqrt{\sigma_{\rm f}^2 + \sigma_{\rm i}^2}}, \quad (8)$$

which expresses how many standard deviations the observed mean deviates from zero. Here $\mu_{\rm i}$ and $\mu_{\rm f}$ denote the means of the initial and final areas respectively, while $\sigma_{\rm i}$ and $\sigma_{\rm f}$ are their corresponding standard deviations. As discussed in Abac *et al.* [1], this estimate is less sensitive to sampling errors in the distribution tails since it relies only on the first two cumulants. This bound is slightly more stringent than that presented in Abac *et al.* [1], due to stronger GR assumptions and use of the complete signal. Moreover, the test performed here splits the data in the frequency-domain, which is not equivalent to the time-domain analysis done in Abac *et al.* [1]. Using the fractional difference to calculate the significance, we find $3.7\sigma_{\rm IMRCT}$, with differences being driven by uncertainty in the initial area normalization. In Fig. 6, we also show the 90% credible interval from the full-signal analysis using NRSur7dq4 [1], which coherently describe the complete signal assuming both GR and the area theorem. It yields the most stringent bound because it employs the full SNR of GW250114, instead of using a smaller portion associated either to the inspiral or the post-inspiral phases.

Residuals test

The residuals test [38] is a statistical analysis that checks for the presence of excess coherent power remaining in the detector network after subtracting the best-fit waveform from the data [39, 40]. Significant residual power could indicate the presence of additional physical effects that are not captured by current BBH models, modeling systematics, or unaccounted instrumental noise artifacts.

We perform the residual data by subtracting from the original data the maximum-likelihood NRSur7dq4 waveform model. If the model adequately captures the gravitational-wave (GW) signal, the resulting residuals should be consistent with stationary Gaussian noise. The residual data is then analyzed using BayesWave [41], and the 90% credible upper limit on the network SNR ρ_{90} is calculated. To compare this ρ_{90} with its expected distribution, segments of data around the signal (with no injected signal) are also analyzed and the probability of obtaining an ρ_{90} higher than that of the residual data is calculated and reported as the p-value = $P(\rho^n_{90} \geq \rho_{90})$, where ρ^n_{90} is the 90% credible upper limit on the coherent SNR of the background, noise-only segments.

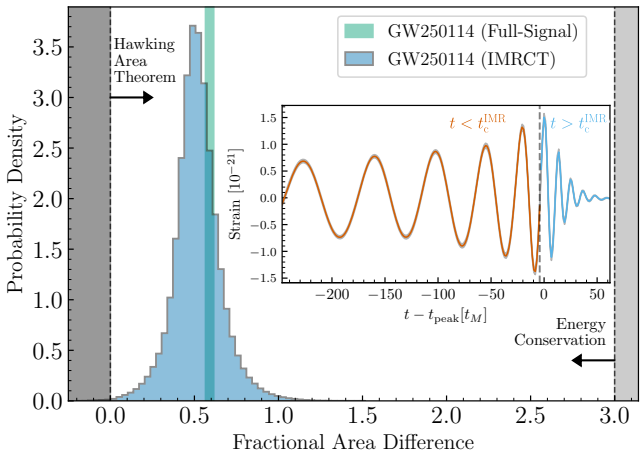

FIG. 6. BH area-law test using the entirety of GW250114. The fractional difference between the area of the final BH $\mathcal{A}_{\rm f}$ and the total area of the initial BH's $\mathcal{A}_{\rm i}$ as calculated using the IMR consistency test on GW250114. The grey shaded region on the left marks the region in which the area theorem is violated, $(\mathcal{A}_{\rm f} - \mathcal{A}_{\rm i})/\mathcal{A}_{\rm i} < 0$. The grey shaded region on the right highlights the region excluded by energy conservation $M_{\rm f} \leq m_1 + m_2$. The vertical green band is the 90% credible interval inferred from the full-signal analysis in Abac *et al.* [1]. The inset schematically shows a reconstructed (grey) signal in LIGO Livingston using the full-signal analysis [1], along with proxies for the inspiral (orange) and post-inspiral (light blue) regions used in the IMR consistency test. The time $t_{\rm c}^{\rm IMR}$ separating the two curves corresponds to the transition frequency $f_{\rm c}^{\rm IMR}$, and does not represent an actual transition time since the data is split in the frequency-domain, not the time-domain.

A higher p-value suggests that the residual power is more consistent with instrumental noise, indicating insufficient evidence to reject the null hypothesis that the residual power originates from noise. For a single event, we also expect the p-value to be a random draw from a uniform distribution on the interval (0,1]. Furthermore, the goodness of fit of the GR template for the signal in the data can be quantified by calculating the 90% credible lower limit on the fitting factor (FF), given by:

$$\mathrm{FF}_{90} = \frac{\rho_{\rm GR}}{\sqrt{\rho_{\rm GR}^2 + \rho_{90}^2}}, \tag{9}$$

where $\rho_{\rm GR}$ is the optimal network SNR for the maximum-likelihood waveform [42].

For GW250114, we find that $\rho_{90} = 6.86$ with a p-value of 0.34. The calculated FF_{90} is 0.996. Based on this, we do not find any significant coherent power beyond what is expected from noise.

[1] A. G. Abac *et al.*, GW250114: testing Hawking's area law and the Kerr nature of black holes, arXiv:2509.0000 [gr-qc] (2025).

[2] V. Baibhav, M. H.-Y. Cheung, E. Berti, V. Cardoso, G. Carullo, R. Cotesta, W. Del Pozzo, and F. Duque, Agnostic black hole spectroscopy: Quasinormal mode content of numerical relativity waveforms and limits of validity of linear perturbation theory, Phys. Rev. D **108**, 104020 (2023), arXiv:2302.03050 [gr-qc].

[3] M. H.-Y. Cheung, E. Berti, V. Baibhav, and R. Cotesta, Extracting linear and nonlinear quasinormal modes from black hole merger simulations, Phys. Rev. D **109**, 044069 (2024), [Erratum: Phys.Rev.D 110, 049902 (2024)], arXiv:2310.04489 [gr-qc].

[4] K. Mitman *et al.*, Probing the ringdown perturbation in binary black hole coalescences with an improved quasi-normal mode extraction algorithm, arXiv:2503.09678 [gr-qc] (2025).

[5] SXS Collaboration, SXS Gravitational Waveform Database, `http://www.black-holes.org/waveforms/`.

[6] M. A. Scheel *et al.*, The SXS Collaboration's third catalog of binary black hole simulations, (2025), arXiv:2505.13378 [gr-qc].

[7] G. B. Cook, Aspects of multimode Kerr ringdown fitting, Phys. Rev. D **102**, 024027 (2020), arXiv:2004.08347 [gr-qc].

[8] M. Isi and W. M. Farr, Analyzing black-hole ringdowns, arXiv:2107.05609 [gr-qc] (2021).

[9] A. Buonanno, G. B. Cook, and F. Pretorius, Inspiral, merger and ring-down of equal-mass black-hole binaries, Phys. Rev. D **75**, 124018 (2007), arXiv:gr-qc/0610122.

[10] V. Baibhav, E. Berti, V. Cardoso, and G. Khanna, Black Hole Spectroscopy: Systematic Errors and Ringdown Energy Estimates, Phys. Rev. D **97**, 044048 (2018), arXiv:1710.02156 [gr-qc].

[11] I. Ota and C. Chirenti, Overtones or higher harmonics? Prospects for testing the no-hair theorem with gravitational wave detections, Phys. Rev. D **101**, 104005 (2020), arXiv:1911.00440 [gr-qc].

[12] S. Bhagwat, X. J. Forteza, P. Pani, and V. Ferrari, Ringdown overtones, black hole spectroscopy, and no-hair theorem tests, Phys. Rev. D **101**, 044033 (2020), arXiv:1910.08708 [gr-qc].

[13] M. Giesler, M. Isi, M. A. Scheel, and S. Teukolsky, Black Hole Ringdown: The Importance of Overtones, Phys. Rev. X **9**, 041060 (2019), arXiv:1903.08284 [gr-qc].

[14] M. Giesler *et al.*, Overtones and nonlinearities in binary black hole ringdowns, Phys. Rev. D **111**, 084041 (2025), arXiv:2411.11269 [gr-qc].

[15] R. D. Morey, J.-W. Romeijn, and J. N. Rouder, The philosophy of bayes' factors and the quantification of statistical evidence, Journal of Mathematical Psychology **72**, 6 (2016).

[16] W. Del Pozzo and J. Veitch, `CPNest`: an efficient python parallelizable nested sampling algorithm, `https://github.com/johnveitch/cpnest` (2025).

[17] J. M. Dickey, The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters, Ann. Math. Statist. **42**, 204 (1971).

[18] A. Zimmerman, C.-J. Haster, and K. Chatziioannou, On combining information from multiple gravitational wave sources, Phys. Rev. D **99**, 124044 (2019), arXiv:1903.11008 [astro-ph.IM].

[19] A. Akyüz, A. Correia, J. Garofalo, K. Kacanja, L. Roy, K. Soni, H. Tan, V. J. Y, A. H. Nitz, and C. D. Capano, Potential science with GW250114 – the loudest binary black hole merger detected to date, arXiv eprints (2025), arXiv:2507.08789 [gr-qc].

[20] S. Ma, K. Mitman, L. Sun, N. Deppe, F. Hébert, L. E. Kidder, J. Moxon, W. Throwe, N. L. Vu, and Y. Chen, Quasinormal-mode filters: A new approach to analyze the gravitational-wave ringdown of binary black-hole mergers, Phys. Rev. D **106**, 084036 (2022), arXiv:2207.10870 [gr-qc].

[21] S. Ma, L. Sun, and Y. Chen, Black Hole Spectroscopy by Mode Cleaning, Phys. Rev. Lett. **130**, 141401 (2023), arXiv:2301.06705 [gr-qc].

[22] S. Ma, L. Sun, and Y. Chen, Using rational filters to uncover the first ringdown overtone in GW150914, Phys. Rev. D **107**, 084010 (2023), arXiv:2301.06639 [gr-qc].

[23] N. Lu, S. Ma, O. J. Piccinni, L. Sun, and E. Finch, Statistical identification of ringdown modes with rational filters, (2025), arXiv:2505.18560 [gr-qc].

[24] M. Boyle, K. Mitman, M. Scheel, and L. Stein, The SXS package (2025).

[25] C. Xin, M. Isi, W. M. Farr, and Z. Haiman, Identifying Compact Chirping SMBHBs in LSST using Bayesian Analysis, arXiv eprints (2025), arXiv:2506.10846 [astro-ph.HE].

[26] J. Meidam *et al.*, Parametrized tests of the strong-field dynamics of general relativity using gravitational wave signals from coalescing binary black holes: Fast likelihood calculations and sensitivity of the method, Phys. Rev. D **97**, 044033 (2018), arXiv:1712.08772 [gr-qc].

[27] M. Saleem, S. Datta, K. G. Arun, and B. S. Sathyaprakash, Parametrized tests of post-Newtonian theory using principal component analysis, Phys. Rev. D **105**, 084062 (2022), arXiv:2110.10147 [gr-qc].

[28] P. Mahapatra *et al.*, Confronting General Relativity with Principal Component Analysis: Simulations and Results from GWTC-3 Events, (2025), arXiv:2508.06862 [gr-qc].

[29] A. Pai and K. G. Arun, Singular value decomposition in parametrised tests of post-Newtonian theory, Class. Quantum Grav. **30**, 025011 (2013), arXiv:1207.1943 [gr-qc].

[30] A. A. Shoom, P. K. Gupta, B. Krishnan, A. B. Nielsen, and C. D. Capano, Testing the post-Newtonian expansion with GW170817, Gen. Rel. Grav. **55**, 55 (2023), arXiv:2105.02191 [gr-qc].

[31] G. Pratten *et al.*, Computationally efficient models for the dominant and subdominant harmonic modes of precessing binary black holes, Phys. Rev. D **103**, 104056 (2021), arXiv:2004.06503 [gr-qc].

[32] M. Colleoni, F. A. R. Vidal, C. García-Quirós, S. Akçay, and S. Bera, Fast frequency-domain gravitational waveforms for precessing binaries with a new twist, Phys. Rev. D **111**, 104019 (2025), arXiv:2412.16721 [gr-qc].

[33] L. Pompili *et al.*, Laying the foundation of the effective-one-body waveform models SEOBNRv5: Improved accuracy and efficiency for spinning nonprecessing binary black holes, Phys. Rev. D **108**, 124035 (2023), arXiv:2303.18039 [gr-qc].

[34] S. W. Hawking, Gravitational radiation from colliding black holes, Phys. Rev. Lett. **26**, 1344 (1971).

[35] J. M. Bardeen, B. Carter, and S. W. Hawking, The Four laws of black hole mechanics, Commun. Math. Phys. **31**, 161 (1973).

[36] X. Jiménez-Forteza, D. Keitel, S. Husa, M. Hannam, S. Khan, and M. Pürrer, Hierarchical data-driven approach to fitting numerical relativity data for nonprecessing binary black holes with an application to final spin and radiated energy, Phys. Rev. D **95**, 064024 (2017), arXiv:1611.00332 [gr-qc].

[37] F. Hofmann, E. Barausse, and L. Rezzolla, The final spin from binary black holes in quasi-circular orbits, Astrophys. J. Lett. **825**, L19 (2016), arXiv:1605.01938 [gr-qc].

[38] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Tests of general relativity with GW150914, Phys. Rev. Lett. **116**, 221101 (2016), [Erratum: Phys.Rev.Lett. 121, 129902 (2018)], arXiv:1602.03841 [gr-qc].

[39] N. Cornish, L. Sampson, N. Yunes, and F. Pretorius, Gravitational Wave Tests of General Relativity with the Parameterized Post-Einsteinian Framework, Phys. Rev. D **84**, 062003 (2011), arXiv:1105.2088 [gr-qc].

[40] M. Vallisneri, Testing general relativity with gravitational waves: a reality check, Phys. Rev. D **86**, 082001 (2012), arXiv:1207.4759 [gr-qc].

[41] N. J. Cornish, T. B. Littenberg, B. Bécsy, K. Chatziioannou, J. A. Clark, S. Ghonge, and M. Millhouse, BayesWave analysis pipeline in the era of gravitational wave observations, Phys. Rev. D **103**, 044006 (2021), arXiv:2011.09494 [gr-qc].

[42] A. G. Abac *et al.*, GWTC-4.0: Tests of General Relativity, arXiv:2509.0000 [gr-qc] (2025).