

# Merger-ringdown consistency: A new test of strong gravity using deep learning

Swetha Bhagwat\* and Costantino Pacilio†

Dipartimento di Fisica, “Sapienza” Università di Roma & Sezione INFN Roma1, Roma 00185, Italy

(Dated: July 13, 2021)

The gravitational waves emitted during the coalescence of binary black holes are an excellent probe to test the behaviour of strong gravity. In this paper, we propose a new test called the *merger-ringdown consistency test* that focuses on probing the imprints of the dynamics in strong-gravity around the black-holes during the plunge-merger and ringdown phase. Furthermore, we present a scheme that allows us to efficiently combine information across multiple ringdown observations to perform a statistical null test of GR using the detected BH population. We present a proof-of-concept study for this test using simulated binary black hole ringdowns embedded in the next-generation ground-based detector noise. We demonstrate the feasibility of our test using a deep learning framework, setting a precedence for performing precision tests of gravity with neural networks.

## I. INTRODUCTION

The detection of gravitational waves (GWs) emitted during the binary black hole (BH) mergers presents us with an unparalleled opportunity to test the behaviour of strong gravity around BHs [1, 2]. GWs are emitted as the BHs slowly spiral in towards the common center of mass (a.k.a. the inspiral phase); this is followed by a rapid plunge and merger (a.k.a. the plunge-merger phase) where the two BHs coalesce forming a remnant BH which then rings down and settles to a final state (a.k.a. the ringdown phase) [3]. While both plunge-merger and ringdown contain imprints of the dynamics in the strong field regime at few times the horizon-length scale, the plunge-merger is dictated by non-linear dynamics and the ringdown is prescribed by linear perturbation theory [4–6].

Ringdown corresponds to the evolution of linear perturbations on the space-time metric of the remnant [7]. Given an underlying theory of gravity, the dynamics in the strong-field regime that sets up the perturbation conditions for ringdown and the properties of the remnant BH are not independent. If GR were to be modified in the strong non-linear regime, one would expect the relative excitation of modes in ringdown as well as the final BH’s mass and spin to be altered [8, 9]. We propose a novel test that checks if the excitation conditions set during the plunge-merger phase are consistent with the properties of the remnant BH formed after the ringdown phase. The proposed test checks for the consistency by simultaneously using the frequency content and the amplitudes and phases of excitation in the ringdown signals. Furthermore, we stack the information from multiple GW observations efficiently to provide a statistical ‘null’ test across a population of binary BH ringdowns. Henceforth, we call it *the merger-ringdown consistency test*.

We present a complementary test to the already existing battery of tests of GR. Although we draw our inspiration from the IMR test, the two tests address conceptually different questions: while IMR test checks for

global consistency of the binary BH evolution [1, 10–12], violation of the merger-ringdown test indicates GR-modifications that alter the perturbations setup for ringdown in a way that is inconsistent with the expected radiated angular momentum and energy in a binary BH coalescence. However, note that such GR modifications (depending on the details of how GR is modified) might also leave imprints on the global evolution of the binary BH signals, and could be picked up by the IMR test. Furthermore, the merger-ringdown test aims at increasing the sensitivity to the strong field dynamics by zooming in solely on the ringdown phase. Comparing the performance of the two tests in distinguishing GR from non-GR signals is non-trivial and depends on the class of modifications in consideration.

Our test is particularly suited for the third-generation (3g) detectors such as the Einstein Telescope (ET) [13], the Cosmic Explorer [14] and LISA [15], where the ringdowns are expected to be loud and the number of detections can be  $\sim 10^3 - 10^4$ /year [16, 17]. Performing prognostic and realistic benchmarking studies on a large number of events with full Bayesian parameter estimation demands for a rapid and computationally efficient inference algorithm. To this aim, we demonstrate the feasibility of our test entirely using a deep learning framework to speed up the parameter inference by orders of magnitude [18–20]. We train a neural network architecture called a conditional variational autoencoder (CVAE) [21–23] to infer posterior distributions of the parameter set  $\{M, \chi_f, q\}$  from a set of simulated ringdown waveforms. Following the deep learning application to GW science — e.g., detection [24–28] and parameter estimation (PE) [18, 19, 29–32],<sup>1</sup> our work also sets a precedence for precision tests of GR using neural networks. Finally, we also demonstrate that deep learning techniques can be efficiently used for population studies for current and next-generation GW detectors.

\* swetha.bhagwat@roma1.infn.it

† costantino.pacilio@uniroma1.it

<sup>1</sup> See [33] for a recent review.

## II. MERGER-RINGDOWN CONSISTENCY TEST

### A. Theory

Ringdown is modelled as a linear superposition of damped sinusoids with characteristic BH frequencies ( $f_{lm}$ ) and damping times ( $\tau_{lm}$ ) known as the quasi-normal-mode (QNM) spectrum. It is generally decomposed in spin-2 weighted spheroidal harmonic basis  $\mathcal{Y}^{lm}(\iota)$ , where ( $\iota \in [0, \pi)$ ) is the inclination angle. Ringdowns take the following analytical form <sup>2</sup>

$$h_+(t) = \frac{M}{D_L} \sum_{l,m>0} \mathcal{Y}_+^{lm}(\iota) A_{lm} e^{-t/\tau_{lm}} \cos(2\pi f_{lm}t - \phi_{lm}), \quad (1a)$$

$$h_\times(t) = \frac{M}{D_L} \sum_{l,m>0} \mathcal{Y}_\times^{lm}(\iota) A_{lm} e^{-t/\tau_{lm}} \sin(2\pi f_{lm}t - \phi_{lm}). \quad (1b)$$

Here  $\{+, \times\}$  are the GW polarizations and  $D_L$  is the luminosity distance of the system. The QNMs are indexed by the angular multipole numbers ( $l, m$ ) and they are determined by the final mass and final spin  $\{M, \chi_f\}$ , i.e.,  $f_{lm} = f_{lm}(M, \chi_f)$  and  $\tau_{lm} = \tau_{lm}(M, \chi_f)$ .  $A_{lm}$  and  $\phi_{lm}$  are the amplitudes and the phases of excitations of QNMs. For a non-spinning binary, the initial system is completely characterized by the total-mass  $M_{tot}$  and the binary mass ratio  $q$ . While  $M_{tot}$  sets the overall amplitude scale,  $q$  determines the relative excitations of QNMs, i.e.,  $A_{lm}/A_{22} = A_{lm}^R(q)$  and  $\phi_{22} - \phi_{lm} = \delta\phi_{lm}(q)$ . Thus, the ringdown waveform can be parameterized by a set of three parameters  $\{M, \chi_f, q\}$  and Eq. (1) can be rewritten as

$$h_+(t) = h_+(t; M, \chi_f, q), \quad h_\times(t) = h_\times(t; M, \chi_f, q). \quad (2)$$

Using the ringdown phase of the GW event one can infer  $\{M, \chi_f, q\}$  by treating them as independent quantities in a Bayesian PE setup.

Next, in GR, a given set of  $\{M_{tot}, q\}$  can be deterministically mapped to  $\{M, \chi_f\}$  for a non-spinning binary BH system.<sup>3</sup> The three ringdown parameters  $\{M, \chi_f, q\}$  are not truly independent and, in particular, we map  $\chi_f$  to  $q$  using the fitting formula presented in [36] (see also [35, 37, 38])

$$\chi_f = 2\sqrt{3}\eta - 3.871\eta^2 + 4.028\eta^3 + \mathcal{O}(\eta^3) \quad (3)$$

<sup>2</sup> For simplicity, we decompose the ringdown signal on to spin-2 weighted spherical harmonics basis instead of the more natural spin-2 weighted spheroidal harmonics basis. This assumption is reasonable as long as the spins are not too high and can be estimated following [34].

<sup>3</sup> The remnant BH can be expressed in terms of the initial binary BH parameters by fitting the numerical relativity simulations. The relationship can be expressed explicitly in approximate analytical forms [35–41], or implicitly using machine learning algorithms [42–44].

where  $\eta = q/(1+q)^2$ .

The test checks if the independent measurements of  $\{M, \chi_f, q\}$  from the ringdowns are consistent with the relation between  $\chi_f$  and  $q$  as predicted by GR. Specifically, we check if the  $\chi_f$  directly measured from the ringdown agrees with the  $\chi_f$  calculated by plugging the measured value of  $q$  in Eq. (3).

### B. Prescription for the merger-ringdown consistency

Let a population of non-spinning binary BHs ringdowns be detected by a GW observatory. Note that the quantities directly *measured* in PE have a superscript ‘meas’ and those *inferred* using Eq. (3) have ‘infer’.

1. Parametrize the ringdown as in Eq. (2) and estimate  $\{M^{\text{meas}}, \chi_f^{\text{meas}}, q^{\text{meas}}\}$  for each event. We used the median of the marginalized posterior distribution as the ‘measured’ value.
2. For each event, compute the  $\chi_f^{\text{infer}}$  from the median value of  $q^{\text{meas}}$  in step 1 using the relation in Eq. (3).
3. Make a scatter plot with  $\{\chi_f^{\text{infer}}, \chi_f^{\text{meas}}\}$  using all ringdown observations. In GR, one expects that all the data should lie along the  $\chi_f^{\text{infer}} = \chi_f^{\text{meas}}$  line in a 2-D scatter plot, with the noise in the data leading to a spread around this line. To perform the merger-ringdown consistency test, we express

$$\chi_f^{\text{meas}} = a + b \chi_f^{\text{infer}} \quad (4)$$

and fit for the parameters  $\{a, b\}$ . If the best-fit parameters for Eq. (4) are compatible with  $\{a = 0, b = 1\}$  the observations are consistent with GR, providing a statistical null test.

### C. Details of Implementation

For simplicity, we restrict our study to non-spinning quasi-circular binary BHs. We compute the QNM spectra  $\{f_{lm}(M, \chi_f), \tau_{lm}(M, \chi_f)\}$  using the data in [45]. Further, we focus our attention on the dominant mode ( $l, m$ ) = (2, 2) and the two most excited subdominant angular modes for the case of non-spinning systems — ( $l, m$ ) = {(2, 1), (3, 3)} [46, 47]. We concentrate solely on the dominant overtone, i.e.,  $n_{\text{overtone}} = 0$  [48, 49]. We use these simplifying assumptions for this proof-of-concept study. However, note that including more angular modes and overtones is a tangible extension to our work.

Key to our study is the expression of the QNM excitation amplitudes and phases, as functions of  $q$ . Following

the prescription in [47], we express  $A_{lm}^R$  and  $\delta\phi_{lm}$  as

$$A_{lm}^R(q) = a_0 + \frac{a_1}{q} + \frac{a_2}{q^2} + \frac{a_3}{q^3}, \quad (5a)$$

$$\delta\phi(q) = b_0 + \frac{b_1}{b_2 + q^2}, \quad (5b)$$

where we use the convention in which  $q > 1$ . An updated list of coefficients  $\{a_i, b_i\}$  is provided in the Supplemental Material. Further, for the dominant mode’s amplitude, we use  $A_{22} = 0.86\eta$  [50]. Lastly, we assume a uniform support in  $\phi_{22} \in [0, 2\pi]$  and generate the waveforms expressed in Eq. (1).

### III. DEEP LEARNING FRAMEWORK

We use a deep learning framework to reconstruct the posteriors for the parameters  $\{M, \chi_f, q\}$  from the waveform. We follow [18–20] and train a CVAE, a neural network architecture well suited to posterior sampling.

#### A. Details on the CVAE implementation

The CVAE acts as an inverse nonlinear map from the ringdown strain  $h = y(x)$  to the posteriors of  $x = \{M, \chi_f, q\}$ .

$$\text{CVAE} : y \rightarrow p(x|y). \quad (6)$$

It has the structure of a variational autoencoder [21, 22]: it is made of two serial neural network units (the ‘encoder’ and the ‘decoder’) separated by a stochastic latent layer. The first neural network (encoder) maps the input  $y$  into the latent layer. The second neural network (decoder) maps the latent representation of the input into the output probability distribution  $p(x|y)$ .

The CVAE is trained by introducing a third neural network unit, called the auxiliary encoder or the ‘guide’, at training time [51]. The training consists in optimising a loss function. For the CVAE, the loss naturally splits into [19, 51]:

1. the Kullback-Leibler (KL) divergence  $\mathcal{L}_{\text{KL}}$ , measuring the similarity between the outputs of the encoder and the guide; it quantifies the ability of the encoder to produce a meaningful mapping of the input into the latent space;
2. the reconstruction loss  $\mathcal{L}_{\text{recon}}$ , measuring the probability that the true values  $x_{\text{true}}$  falls within the decoder distribution.

The total loss to be optimised is

$$\mathcal{L}_{\text{tot}} = \mathcal{L}_{\text{recon}} + \beta\mathcal{L}_{\text{KL}}. \quad (7)$$

When  $\beta = 1$ ,  $\mathcal{L}_{\text{tot}}$  coincides with the standard ELBO loss [21, 22]. The additional parameter  $\beta$  gives flexibility in implementing effective training strategies.

Batch size	512
Epochs	500
Optimizer	Adam
Initial lr	$10^{-4}$
lr decay	$\times 0.5$ every 80 epochs
$\beta$ annealing	$3 \times [10^{-5}, \frac{1}{3}, \frac{2}{3}, 1, 1, 1]$
Validation fraction	10%

TABLE I. Hyperparameters used for training the CVAE.

After the training, the guide is dropped out. At production time, only the encoder and the decoder are used to sample the posteriors  $p(x|y)$ . Fig. 1 contains a flow-diagram of the training and production steps. More details on the neural network architecture and our codes are provided in a dedicated git repository [52].

The hyperparameters which determine the CVAE training are listed in Tab. I. We train the CVAE in batches of 512 waveforms for 500 epochs, i.e., 500 forward-backward passes of the entire training set. The loss is minimized using the Adam optimizer with an initial learning rate set to  $10^{-4}$ . The learning rate is decreased by a factor of 2 every 80 epochs. To monitor the convergence of the loss, we set aside 10% of the training dataset and we use it for validation. Following [18], we use  $\beta$  to implement a cyclic annealing schedule. Annealing improves the efficiency of the training and allows autoencoders to express more meaningful latent variables [53]. We increase  $\beta$  from 0 to 1 in steps of  $[10^{-5}, \frac{1}{3}, \frac{2}{3}, 1, 1, 1]$  and these steps are repeated 3 times. After this,  $\beta$  is definitively fixed to 1.

Next, the training performances improve when the inputs  $y$  are standardized to zero mean and unit variance, and when the outputs  $x$  are normalized to have support in  $[1, 100]$ .  $x$  is then scaled back to the original normalization at production time.

#### B. Network training

To train the network, we simulate a dataset of  $10^5$  ringdowns by sampling the waveform parameters uniformly in the ranges indicated in Tab. II. The ringdown waveforms are sampled at 4096 Hz with a total signal duration of 31.25 ms, thus corresponding to arrays of length 128. Signal-to-noise ratio (SNR) is used to set the waveform scale w.r.t. the noise. The SNR is computed as in [54, 55]. When performing PE, we only estimate posteriors for  $\{M, \chi_f, q\}$ . For simplicity, we only consider the + polarization and fix the inclination angle to  $\iota = \pi/3$ . The ringdowns are embedded in simulated ET-like noise segments [56, 57]. At each training iteration, we assign noise instances randomly to the waveforms to prevent the CVAE from learning spurious correlations between the waveforms and the noise realizations.

Our training takes 84 minutes on a single GPU. Fig. 2 shows the evolution of the reconstruction loss  $\mathcal{L}_{\text{recon}}$  and

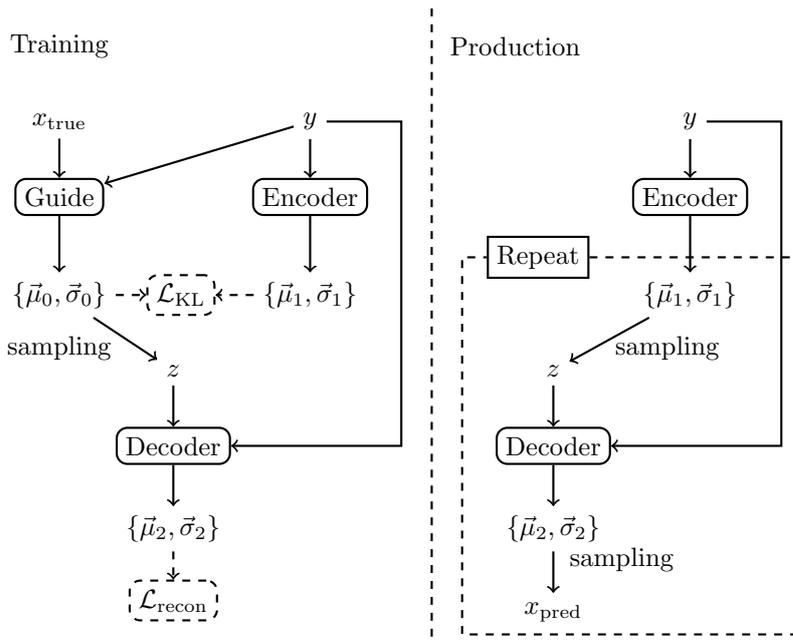


FIG. 1. A schematic representation of the CVAE architecture. On the left, a single training step is represented. First, the signal  $y$  is mapped by the encoder into a latent stochastic distribution, which is a multivariate diagonal Gaussian with means and standard deviations  $\{\vec{\mu}_1, \vec{\sigma}_1\}$ ; similarly, the couple  $(y, x_{\text{true}})$  is mapped by the guide into a second Gaussian with parameters  $\{\vec{\mu}_0, \vec{\sigma}_0\}$ ; the two are then combined into the  $\mathcal{L}_{\text{KL}}$  loss. Next, a latent variable  $z$  is sampled from the guide distribution and mapped by the decoder into a third Gaussian with parameters  $\{\vec{\mu}_2, \vec{\sigma}_2\}$ . This final distribution is eventually used to compute the  $\mathcal{L}_{\text{recon}}$  loss. On the right, a single step at production time is shown. Now, the latent representation is sampled from the encoder and a predicted output  $x_{\text{pred}}$  is sampled from the decoder; this step is repeated  $n_{\text{samples}}$  times to produce an informative posterior distribution for  $x$ ; in text, we fix  $n_{\text{samples}} = 10^4$ . Note the final distribution of  $x_{\text{pred}}$  is not a Gaussian, but is a complex distribution resulting from the convolution of the two serial sampling steps.

Parameter	Symbol	Range
Final BH mass	$M$	$[25, 100] M_{\odot}$
Final BH spin	$\chi_f$	$[0, 0.9]$
Binary mass ratio	$q$	$[1, 8]$
Phase of the (2,2) mode	$\phi_{22}$	$[0, 2\pi]$ rad
Signal-to-noise ratio	SNR	$[40, 80]$

TABLE II. Ranges for the waveform parameters. All the parameters are sampled uniformly. Note we marginalize over the last two parameters.

the KL divergence  $\mathcal{L}_{\text{KL}}$  separately. Further, we show the loss evaluated on the 90% training dataset and on the 10% validation dataset. Notice that the training and validation losses are consistent, substantiating that the network is not overfitting.<sup>4</sup>

To test the network, we generate a new dataset of  $10^3$  simulated ringdown waveforms, whose parameters are sampled again from the ranges indicated in Tab. II.

<sup>4</sup> The initial oscillations which are visible in  $\mathcal{L}_{\text{KL}}$  are due to the cyclic annealing.

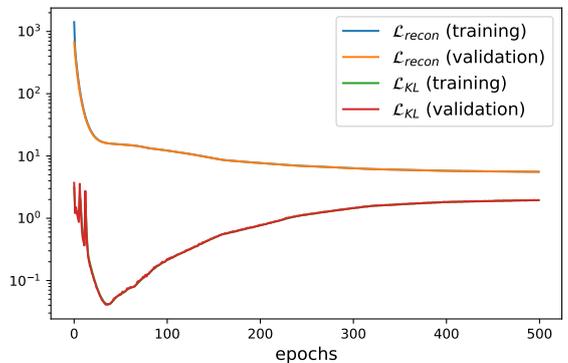


FIG. 2. Evolution of the reconstruction and KL losses across the training epochs.

For each input waveform, the CVAE samples  $10^4$  distinct points to build the posterior. The total time to analyze all the samples is approximately 40 s on a single GPU, corresponding to 40 ms per waveform. For illustration, Fig. 3 shows the contour plot obtained from the PE of one signal from the test dataset. For a quantitative diagnostic of the network performances, we present the P-P

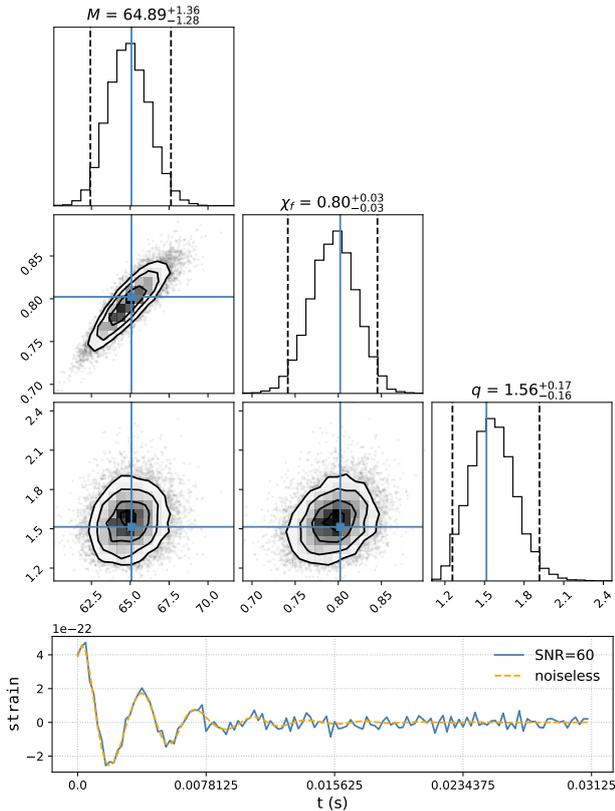


FIG. 3. Contour plot for the PE of the ringdown signal shown in the lower panel, with  $(M, \chi_f, q) = (65, 0.8, 1.5)$  and  $\text{SNR} = 60$ . Blue lines indicate the true values. Dashed lines mark the 95% confidence interval. The plot titles indicate  $1\sigma$  uncertainties.

plot in Fig. 4: the plot shows marginalized cumulative distribution CDF of the true values  $x_{\text{true}}$  as a function of  $p\%$  confidence interval. A diagonal P-P plot means that  $x_{\text{true}}$  is contained  $p\%$  of the times within  $p\%$  confidence interval of the marginalized posteriors for  $x_{\text{pred}}$ . Note that all the CDFs are consistent with the diagonal, demonstrating that our CVAE recovers the posterior samples for  $\{M, \chi_f, q\}$  from the ringdowns reliably.

#### IV. RESULTS

We present our proof-of-concept study in two parts. In section IV A we demonstrate that when our test is applied to a set of ringdowns consistent with GR, these signals satisfy the null test. Next, in section IV B we use 4 sets of non-GR ringdowns and show that with  $\sim 20$ – $50$  events the non-GR signals violate our null test - allowing us to distinguish GR from non-GR ringdown.

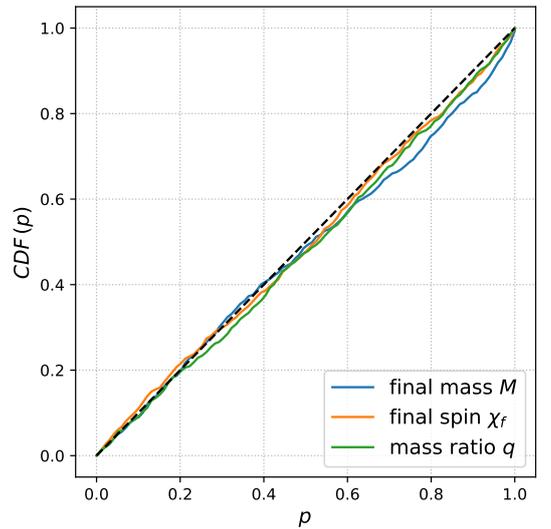


FIG. 4. P-P plot of test dataset. For each observable, the plot shows the cumulative distribution CDF of the true values as a function of the  $p\%$  confidence interval of the posterior distribution.

##### A. Proof-of-concept Merger-Ringdown test: GR signals

We simulate a dataset consisting of  $10^3$  ringdowns with parameter ranges as presented in Tab. II except for  $\chi_f$ . Here,  $\chi_f$  is inferred from  $q$  by imposing the relation (3).

First, the posteriors produced by our CVAE for a single event can be used to check if the event validates the spin relation (3). In Fig. 5, we show the posterior samples for  $\chi_f^{\text{meas}}$  and  $\chi_f^{\text{infer}}$  — where  $\chi_f^{\text{infer}}$  is determined by  $q^{\text{meas}}$  as per Eq. (3), for a randomly chosen event from our dataset. For this event, we see that  $\chi_f^{\text{meas}} = \chi_f^{\text{infer}}$  (blue dashed line) lies within the 68% credible interval of the posteriors, asserting that this event is consistent with GR evolution.

Further, we use the scheme outlined in Sec. II B to combine the information across multiple ringdown observations for a more stringent test of GR, as illustrated in Fig 6. For a noiseless GR ringdown, we expect  $\chi_f^{\text{meas}} = \chi_f^{\text{infer}}$ . However, our inferences are probabilistic and contain noise. This leads to measurement uncertainties that translate as a scatter around the diagonal line.

In Fig. 6, we confirm that our dataset lies around  $\chi_f^{\text{meas}} = \chi_f^{\text{infer}}$ . Also, as expected, lower values of  $q$  give higher values of  $\chi_f$ . A weighted least squared (WLS) fit for Eq. (4) gives  $a \in [-0.014, 0.014]$  and  $b \in [0.963, 1.013]$  at 95% confidence level, showing an agreement with the  $\chi_f^{\text{meas}} = \chi_f^{\text{infer}}$  line. We weighed each event by  $(\sigma_{\chi_f^{\text{meas}}} \sigma_{\chi_f^{\text{infer}}})^{-1/2}$  to emphasize the more confident re-

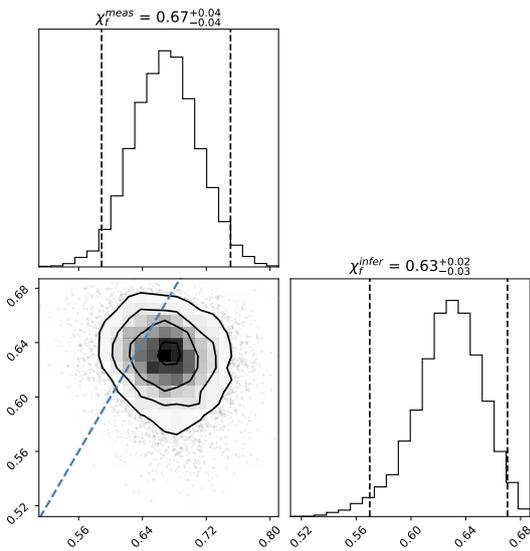


FIG. 5. Contour plot for the posteriors of  $\chi_f^{\text{meas}}$  and  $\chi_f^{\text{infer}}$  for a signal with  $(M, \chi_f, q) = (52.6, 0.66, 1.53)$  and  $\text{SNR} = 60$ . The signal is extracted from the second test dataset, where the relation (3) is enforced. The blue dashed line represents  $\chi_f^{\text{meas}} = \chi_f^{\text{infer}}$  line.

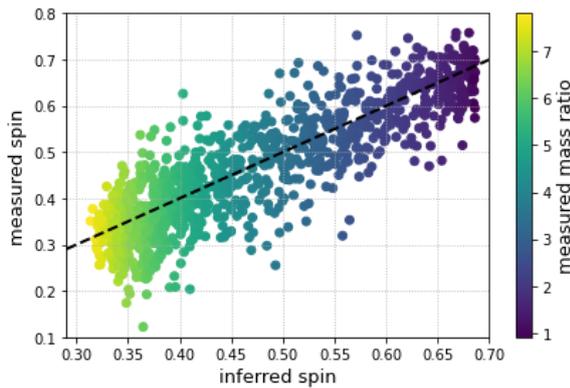


FIG. 6. Scatter plot of  $\chi_f^{\text{meas}}$  vs  $\chi_f^{\text{infer}}$ . The color bar indicates the value  $q^{\text{meas}}$  for each observation. The black dotted line marks the  $\chi_f^{\text{meas}} = \chi_f^{\text{infer}}$ .

covers.<sup>5</sup> Fig. 6 thus observationally validates Eq. (4).

Next, to assess the efficiency of this test with the number of observations, we study the convergence of  $a$  and  $b$ . In Fig. 7, we present the (weighted) best-fit values for  $a$  and  $b$  with their  $2\sigma$  confidence levels as a function of the number of observations.

In Fig. 7, we find that the mean value for  $a$  and  $b$  are  $\sim 0$  and  $1$ , respectively, even for a small number of observations. We see that the uncertainties in the measurements

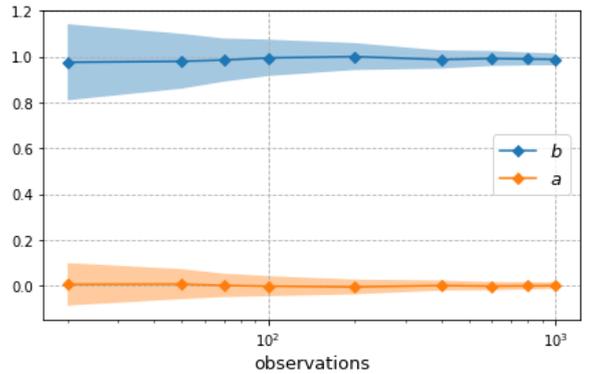


FIG. 7. Evolution of the best-fit values (continuous lines) and  $2\sigma$  contours (shaded regions) for  $b$  and  $a$  in Eq. (4), versus the number of observations. The mean values and the confidence intervals are averaged over 10 random realisations. In GR, the noiseless best-fit corresponds to  $b = 1$  and  $a = 0$ .

of  $a$  and  $b$ , i.e.,  $\{\sigma_b, \sigma_a\}$  decrease with increasing number of observations as a power-law. For instance, with 20 observations we can constrain  $\{\sigma_b(n = 20), \sigma_a(n = 20)\} = \{0.0832, 0.00465\}$  and with 100 observations we have  $\{\sigma_b(n = 100), \sigma_a(n = 100)\} = \{0.0398, 0.0219\}$ . Concretely,  $\sigma_a$  and  $\sigma_b$  scale with the number of observations as

$$\sigma_a(n) = \frac{0.21}{\sqrt{n}}, \quad \sigma_b(n) = \frac{0.41}{\sqrt{n}}, \quad (8)$$

which is consistent with our expectations given our noise model.

Thus, while the merger-ringdown consistency test is powerful when combining a large number of ringdowns, it is also feasible to perform it with just a few observations ( $\sim 20$ ).

## B. Proof-of-concept Merger-Ringdown test: non-GR signals

In this section, we present the performance of the test on a set of non-GR ringdown signals. We consider phenomenological deviations from GR without assuming physical mechanisms responsible for the GR modifications. Specifically, we generate 4 sets of non-GR ringdowns by heuristically modifying amplitudes and phases of mode excitations

$$A_{lm}^R \rightarrow (1 + \Delta A) A_{lm,GR}^R \quad (9a)$$

$$\delta\phi_{lm} \rightarrow (1 + \Delta\delta\phi) \delta\phi_{lm,GR} \quad (9b)$$

with the 4 distinct cases enlisted as

$$\text{Case 1:} \quad \Delta A = 0 \quad \Delta\delta\phi = 0.1 \quad (10a)$$

$$\text{Case 2:} \quad \Delta A = 0 \quad \Delta\delta\phi = -0.1 \quad (10b)$$

$$\text{Case 3:} \quad \Delta A = 0.1 \quad \Delta\delta\phi = 0.1 \quad (10c)$$

$$\text{Case 4:} \quad \Delta A = 0.1 \quad \Delta\delta\phi = -0.1 \quad (10d)$$

<sup>5</sup> We verified that the results of an ordinary (unweighted) least squared fit do not significantly change.

This class of GR-modifications implies an altered relation between the ringdown perturbation conditions and the final properties of the remnant; therefore, we expect our test to be naturally sensitive to these modifications. Furthermore, note that other than the amplitude ratios and phase differences between the modes, the details of the non-GR signal are identical to the GR ones — i.e., the QNM spectrum is unaltered.

The best-fit values of  $a$  and  $b$  in Eq. 4 for each of the non-GR signal-set are presented in Fig. 8. We remind that any departure from  $a = 0$  and  $b = 1$  indicates that the dataset contains ringdown that do not satisfy the GR null test. We note that the merger-ringdown test successfully excludes the  $a = 0$  and  $b = 1$  for a relatively small number of observations. Specifically, GR values are excluded at a  $2\sigma$  level with  $\mathcal{O}(20)$  observations for Case 1-3 and with  $\mathcal{O}(50)$  observations for Case 4.

## V. CONCLUSION AND OUTLOOK

We demonstrated a proof-of-concept study for a novel test of GR called *the merger-ringdown consistency test* that checks for statistical consistency between the plunge-merger-ringdown phase across a set of ringdown detections using a deep learning framework. The test aims at increasing the sensitivity to the merger-ringdown by simultaneously incorporating information on both the amplitude-phase excitations and the QNM frequency spectra in the ringdown. It uses the fact that the QNM complex excitation amplitudes and the spectrum in the ringdown are not independent quantities in GR. The test provides an efficient way of stacking ringdowns. Furthermore, for possible future detections of heavy mass binary systems (that is set precedence by the discovery of [58]), our test can be used to check the consistency of initial BH parameters with the ringdown alone if the inspiral is not well measured.

This work illustrates that Bayesian deep learning methods can be applied to infer the posteriors of ringdown parameters to conduct precision tests of GR. Unlike the traditional Monte-Carlo sampling, neural networks perform the PE in fractions of a second, providing a crucial edge when dealing with the large number of observations that are expected with future GW detectors.

We also highlight that, in its most generic construct, the TIGER infrastructure allows to identify parametric deviations from GR in the waveform by applying a Bayesian hypothesis testing [59, 60]. In testing GR with ringdown signals, the most popular strategy has been to parametrize the deviations in the QNM frequencies and damping times, followed by evaluating the Bayes factor for GR versus non-GR hypothesis [50, 61, 62]. Our test is conceptually different, because we are not explicitly checking whether the QNM spectrum of the remnant is consistent with that of a general relativistic BH. Rather, we focus on the consistency between the relative amplitudes and phases of the ringdown modes with the spin

of the remnant BH. The modifications to which the two tests are sensitive do not necessarily overlap; we expect our test to be efficient in scenarios where the departure from GR influences the relative amplitudes and phase, but the final remnant is similar to a general relativistic BH. Comparing the efficiency of our test to the TIGER implementation is not straightforward and needs further exploration.

Here we have used non-spinning progenitor BHs where the QNM excitation amplitudes and phases are fully parametrizable by its mass ratio  $q$  and the  $\chi_f - q$  relation is approximated by the simple analytical expression in Eq. (3). However, our method can be extended to encompass spinning progenitor BHs where the QNM excitations depend on both  $q$  and  $\chi_{1,2}$ . The dependence of the remnant spin  $\chi_f$  on the binary BH parameters should then be replaced by implicit non-analytical relations such as those in [42, 43]. Our WLS fit strategy does not rely on analytical relations and new parameters can be estimated by increasing the output dimensions of the CVAE.

This study used stellar-mass BH ringdowns targeting the ET-like data. Similar results can be expected to hold for CE. However, LISA will detect ringdowns from super-massive BHs [16, 55], with loud SNRs. We plan to extend our analysis to include LISA-like data in the future. Finally, in this work, we demonstrate the feasibility of a *null* test of GR, by implementing our test on GR as well on a class of phenomenologically constructed non-GR ringdowns. An interesting extension to our work would be an extensive study on non-GR signals in a parameterized framework such as that in ParSpec [63].

## ACKNOWLEDGMENTS

We thank Paolo Pani and Andrea Maselli for their productive comments on an early version of this manuscript and Guido Sanguinetti for his advices on neural networks. We thank Nathan Johnson-McDaniel for useful clarifications about the IMR test. CP is indebted to Enrico Barausse and Luca Heltai for their invaluable support in getting familiar with neural networks and gravitational wave physics. We acknowledge financial support provided under the European Union’s H2020 ERC, Starting Grant agreement no. DarkGRA-757480. We also acknowledge support under the MIUR PRIN and FARE programmes (GW-NEXT, CUP: B84I20000100001), and from the Amaldi Research Center funded by the MIUR program “Dipartimento di Eccellenza” (CUP: B81I18001170001).

*Software.* The PyCBC [64, 65] library was used to generate the noise realizations and LALSuite [66] to generate the ringdowns. The neural network model used in this work was developed using the PyTorch library [67] for Python. The WLS test was performed with the statsmodels package [68] for Python. The corner plots have been produced with the corner package [69] for Python.

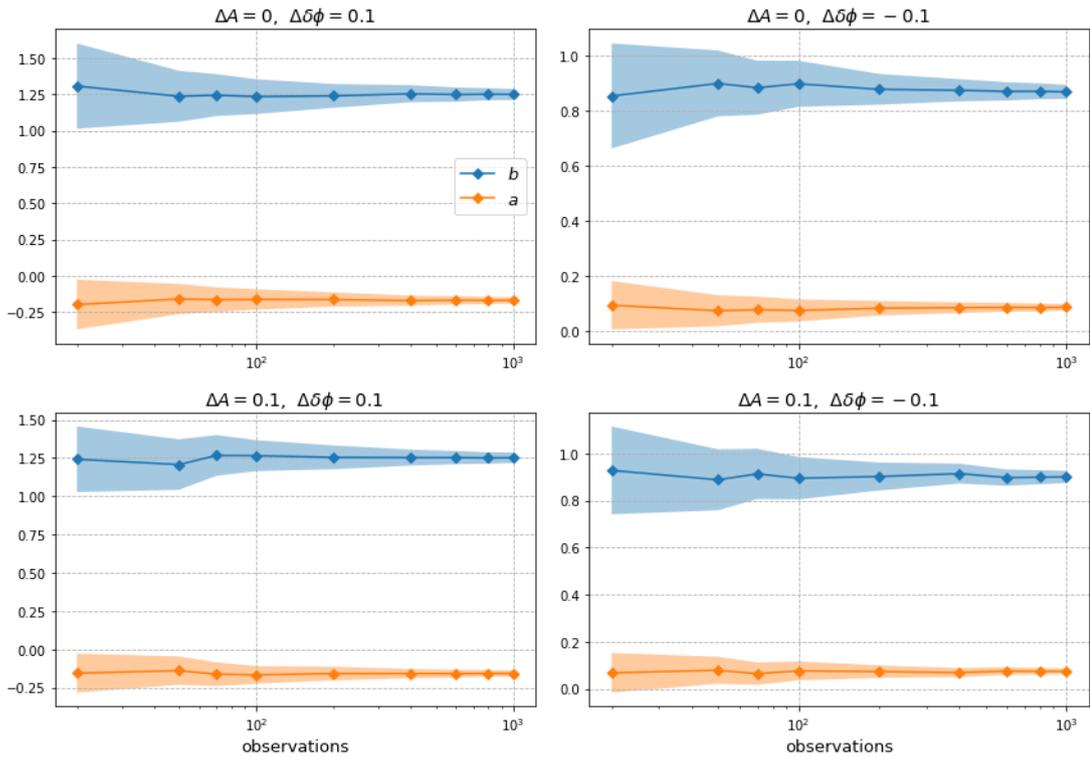


FIG. 8. Similar to Fig. 7 but applied on the set of non-GR test signals described through Case 1-4 in Eq. 10 (depicted as panels, clockwise from the top-left). Each set contains  $10^3$  signals. The mean values and the confidence intervals are averaged over 10 random realisations. Note that for all of the cases we get best fit  $a \neq 0$  and  $b \neq 1$  with 95% credible intervals with less than 50 observations.

*Code availability.* The code used for this paper is made available in a dedicated git repository [52].

and in [47].

### Appendix: Excitation amplitudes and phases

We update the fits presented in [47] by additionally requiring  $A_{lm}/A_{22} \rightarrow 0$  for  $q \rightarrow 1$  [48] and present the coefficients for Eq. (5) in Tab. III. The start of ringdown is chosen at  $t_{peak} + 12M$ . Note that the goodness of the fits do not change significantly between the version here

	(3, 3)	(2, 1)
$a_0$	0.433253	0.472881
$a_1$	-0.555401	-1.1035
$a_2$	0.0845934	1.03775
$a_3$	0.0375546	-0.407131
$b_0$	2.63521	1.80298
$b_1$	8.09316	-9.70704
$b_2$	8.32479	9.77376

TABLE III. Values of the fit coefficients in Eq. (5) for  $(l, m) = (3, 3)$  and  $(2, 1)$ .

- 
- [1] R. Abbott *et al.* (LIGO Scientific, Virgo), (2020), [arXiv:2010.14529 \[gr-qc\]](#).
  - [2] A. H. Nitz, T. Dent, G. S. Davies, S. Kumar, C. D. Capano, I. Harry, S. Mozzon, L. Nuttall, A. Lundgren, and M. Tápai, *The Astrophysical Journal* **891**, 123 (2020).
  - [3] M. Maggiore, *Gravitational Waves. Vol. 1: Theory and Experiments*, Oxford Master Series in Physics (Oxford University Press, 2007).
  - [4] K. D. Kokkotas and B. G. Schmidt, *Living Rev. Rel.* **2**, 2 (1999), [arXiv:gr-qc/9909058](#).
  - [5] R. Konoplya and A. Zhidenko, *Rev. Mod. Phys.* **83**, 793 (2011), [arXiv:1102.4014 \[gr-qc\]](#).
  - [6] E. Berti, V. Cardoso, and A. O. Starinets, *Class. Quant. Grav.* **26**, 163001 (2009), [arXiv:0905.2975 \[gr-qc\]](#).
  - [7] P. Pani, *Int. J. Mod. Phys. A* **28**, 1340018 (2013), [arXiv:1305.6759 \[gr-qc\]](#).
  - [8] I. Kamaretsos, M. Hannam, and B. Sathyaprakash, *Phys. Rev. Lett.* **109**, 141102 (2012), [arXiv:1207.0399 \[gr-qc\]](#).

- [9] S. A. Hughes, A. Apte, G. Khanna, and H. Lim, *Phys. Rev. Lett.* **123**, 161101 (2019), arXiv:1901.05900 [gr-qc].
- [10] A. Ghosh *et al.*, *Phys. Rev. D* **94**, 021101 (2016), arXiv:1602.02453 [gr-qc].
- [11] B. Abbott *et al.* (LIGO Scientific, Virgo), *Phys. Rev. Lett.* **116**, 221101 (2016), [Erratum: Phys.Rev.Lett. 121, 129902 (2018)], arXiv:1602.03841 [gr-qc].
- [12] A. Ghosh, N. K. Johnson-Mcdaniel, A. Ghosh, C. K. Mishra, P. Ajith, W. Del Pozzo, C. P. Berry, A. B. Nielsen, and L. London, *Class. Quant. Grav.* **35**, 014002 (2018), arXiv:1704.06784 [gr-qc].
- [13] M. Maggiore *et al.*, *JCAP* **03**, 050 (2020), arXiv:1912.02622 [astro-ph.CO].
- [14] D. Reitze *et al.*, *Bull. Am. Astron. Soc.* **51**, 035 (2019), arXiv:1907.04833 [astro-ph.IM].
- [15] P. Amaro-Seoane *et al.* (LISA), (2017), arXiv:1702.00786 [astro-ph.IM].
- [16] E. Berti, A. Sesana, E. Barausse, V. Cardoso, and K. Belczynski, *Phys. Rev. Lett.* **117**, 101102 (2016), arXiv:1605.09286 [gr-qc].
- [17] S. Bhagwat, D. A. Brown, and S. W. Ballmer, *Phys. Rev. D* **94**, 084024 (2016), [Erratum: Phys.Rev.D 95, 069906 (2017)], arXiv:1607.07845 [gr-qc].
- [18] S. R. Green, C. Simpson, and J. Gair, *Phys. Rev. D* **102**, 104057 (2020), arXiv:2002.07656 [astro-ph.IM].
- [19] H. Gabbard, C. Messenger, I. S. Heng, F. Tonolini, and R. Murray-Smith, (2019), arXiv:1909.06296 [astro-ph.IM].
- [20] T. S. Yamamoto and T. Tanaka, (2020), arXiv:2002.12095 [gr-qc].
- [21] D. P. Kingma and M. Welling, arXiv preprint arXiv:1312.6114 (2013).
- [22] C. Doersch, arXiv preprint arXiv:1606.05908 (2016).
- [23] G. Erdogan, “Variational autoencoder,” [http://www2.bcs.rochester.edu/sites/jacobslab/cheat\\_sheet/VariationalAutoEncoder.pdf](http://www2.bcs.rochester.edu/sites/jacobslab/cheat_sheet/VariationalAutoEncoder.pdf).
- [24] H. Gabbard, M. Williams, F. Hayes, and C. Messenger, *Phys. Rev. Lett.* **120**, 141103 (2018), arXiv:1712.06041 [astro-ph.IM].
- [25] D. George and E. Huerta, *Phys. Rev. D* **97**, 044039 (2018), arXiv:1701.00008 [astro-ph.IM].
- [26] D. George and E. Huerta, *Phys. Lett. B* **778**, 64 (2018), arXiv:1711.03121 [gr-qc].
- [27] A. Iess, E. Cuoco, F. Morawski, and J. Powell, (2020), arXiv:2001.00279 [gr-qc].
- [28] W. Wei, A. Khan, E. Huerta, X. Huang, and M. Tian, *Phys. Lett. B* **812**, 136029 (2021), arXiv:2010.15845 [gr-qc].
- [29] H. Shen, E. Huerta, Z. Zhao, E. Jennings, and H. Sharma, (2019), arXiv:1903.01998 [gr-qc].
- [30] A. J. Chua, C. R. Galley, and M. Vallisneri, *Phys. Rev. Lett.* **122**, 211101 (2019), arXiv:1811.05491 [astro-ph.IM].
- [31] A. J. Chua and M. Vallisneri, *Phys. Rev. Lett.* **124**, 041102 (2020), arXiv:1909.05966 [gr-qc].
- [32] S. R. Green and J. Gair, *Mach. Learn. Sci. Tech.* **2**, 03LT01 (2021), arXiv:2008.03312 [astro-ph.IM].
- [33] E. Cuoco, J. Powell, M. Cavaglia, K. Ackley, M. Beger, C. Chatterjee, M. Coughlin, S. Coughlin, P. Easter, R. Essick, *et al.*, *Machine Learning: Science and Technology* **2**, 011002 (2020), arXiv:2005.03745 [astro-ph.HE].
- [34] E. Berti and A. Klein, *Physical Review D* **90** (2014), 10.1103/physrevd.90.064012.
- [35] E. Barausse and L. Rezzolla, *Astrophys. J. Lett.* **704**, L40 (2009), arXiv:0904.2577 [gr-qc].
- [36] Y. Pan, A. Buonanno, M. Boyle, L. T. Buchman, L. E. Kidder, H. P. Pfeiffer, and M. A. Scheel, *Phys. Rev. D* **84**, 124052 (2011), arXiv:1106.1021 [gr-qc].
- [37] E. Barausse, V. Morozova, and L. Rezzolla, *Astrophys. J.* **758**, 63 (2012), [Erratum: Astrophys.J. 786, 76 (2014)], arXiv:1206.3803 [gr-qc].
- [38] F. Hofmann, E. Barausse, and L. Rezzolla, *Astrophys. J. Lett.* **825**, L19 (2016), arXiv:1605.01938 [gr-qc].
- [39] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. Jiménez Forteza, and A. Bohé, *Phys. Rev. D* **93**, 044006 (2016), arXiv:1508.07250 [gr-qc].
- [40] X. Jiménez-Forteza, D. Keitel, S. Husa, M. Hannam, S. Khan, and M. Pürrer, *Phys. Rev. D* **95**, 064024 (2017), arXiv:1611.00332 [gr-qc].
- [41] J. Healy and C. O. Lousto, *Phys. Rev. D* **95**, 024037 (2017), arXiv:1610.09713 [gr-qc].
- [42] V. Varma, D. Gerosa, L. C. Stein, F. Hébert, and H. Zhang, *Phys. Rev. Lett.* **122**, 011101 (2019), arXiv:1809.09125 [gr-qc].
- [43] L. Haegel and S. Husa, *Class. Quant. Grav.* **37**, 135005 (2020), arXiv:1911.01496 [gr-qc].
- [44] 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].
- [45] E. Berti, “Kerr qnm frequencies (s=-2),” <https://pages.jh.edu/~eberti2/ringdown/>.
- [46] S. Bhagwat, M. Okounkova, S. W. Ballmer, D. A. Brown, M. Giesler, M. A. Scheel, and S. A. Teukolsky, *Phys. Rev. D* **97**, 104065 (2018), arXiv:1711.00926 [gr-qc].
- [47] X. Jiménez Forteza, S. Bhagwat, P. Pani, and V. Ferrari, *Phys. Rev. D* **102**, 044053 (2020), arXiv:2005.03260 [gr-qc].
- [48] I. Kamaretsos, M. Hannam, S. Husa, and B. Sathyaprakash, *Phys. Rev. D* **85**, 024018 (2012), arXiv:1107.0854 [gr-qc].
- [49] S. Bhagwat, X. J. Forteza, P. Pani, and V. Ferrari, *Physical Review D* **101** (2020), 10.1103/physrevd.101.044033.
- [50] S. Gossan, J. Veitch, and B. Sathyaprakash, *Phys. Rev. D* **85**, 124056 (2012), arXiv:1111.5819 [gr-qc].
- [51] F. Tonolini, J. Radford, A. Turpin, D. Faccio, and R. Murray-Smith, *Journal of Machine Learning Research* **21**, 1 (2020).
- [52] [https://github.com/cpacilio/Merger\\_Ringdown\\_Test](https://github.com/cpacilio/Merger_Ringdown_Test).
- [53] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin, arXiv preprint arXiv:1903.10145 (2019).
- [54] E. Berti, V. Cardoso, and C. M. Will, *Phys. Rev. D* **73**, 064030 (2006), arXiv:gr-qc/0512160.
- [55] V. Baibhav and E. Berti, *Phys. Rev. D* **99**, 024005 (2019), arXiv:1809.03500 [gr-qc].
- [56] S. Hild *et al.*, *Class. Quant. Grav.* **28**, 094013 (2011), arXiv:1012.0908 [gr-qc].
- [57] <http://www.et-gw.eu/index.php/etsensitivities>.
- [58] B. Abbott, R. Abbott, T. Abbott, S. Abraham, F. Acernese, K. Ackley, A. Adams, C. Adams, R. Adhikari, V. Adya, and *et al.*, *Physical Review D* **100** (2019), 10.1103/physrevd.100.064064.
- [59] T. G. F. Li, W. Del Pozzo, S. Vitale, C. Van Den Broeck, M. Agathos, J. Veitch, K. Grover, T. Sidery, R. Sturani, and A. Vecchio, *Phys. Rev. D* **85**, 082003 (2012).

- [60] M. Agathos, W. Del Pozzo, T. G. F. Li, C. Van Den Broeck, J. Veitch, and S. Vitale, *Phys. Rev. D* **89**, 082001 (2014).
- [61] J. Meidam, M. Agathos, C. Van Den Broeck, J. Veitch, and B. Sathyaprakash, *Physical Review D* **90** (2014), 10.1103/physrevd.90.064009.
- [62] R. Brito, A. Buonanno, and V. Raymond, *Phys. Rev. D* **98**, 084038 (2018), arXiv:1805.00293 [gr-qc].
- [63] A. Maselli, P. Pani, L. Gualtieri, and E. Berti, *Phys. Rev. D* **101**, 024043 (2020), arXiv:1910.12893 [gr-qc].
- [64] T. Dal Canton *et al.*, *Phys. Rev. D* **90**, 082004 (2014), arXiv:1405.6731 [gr-qc].
- [65] S. A. Usman *et al.*, *Class. Quant. Grav.* **33**, 215004 (2016), arXiv:1508.02357 [gr-qc].
- [66] LIGO Scientific Collaboration, “LIGO Algorithm Library - LALSuite,” free software (GPL) (2018).
- [67] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019) pp. 8024–8035.
- [68] <https://github.com/statsmodels/statsmodels>.
- [69] D. Foreman-Mackey, *The Journal of Open Source Software* **1**, 24 (2016).