Testing the no-hair theorem with black hole ringdowns using TIGER

J. Meidam,1,* M. Agathos,1 C. Van Den Broeck,1 J. Veitch,1,2 and B. S. Sathyaprakash3

1Nikhef—National Institute for Subatomic Physics, Science Park 105, 1098 XG Amsterdam, The Netherlands
2School of Physics and Astronomy, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom
3School of Physics and Astronomy, Cardiff University, 5, The Parade, Cardiff CF24 3AA, United Kingdom

(Received 12 June 2014; published 4 September 2014)

The Einstein Telescope, a proposed third-generation gravitational-wave observatory, would enable tests of the no-hair theorem by looking at the characteristic frequencies and damping times of black hole ringdown signals. In previous work it was shown that with a single 500–1000 $M_\odot$ black hole at a distance $\lesssim 6$ Gpc (or redshift $z \lesssim 1$), deviations of a few percent in the frequencies and damping times of dominant and subdominant modes would be within the range of detectability. Given that such sources may be relatively rare, it is of interest to see how well the no-hair theorem can be tested with events at much larger distances and with smaller signal-to-noise ratios, thus accessing a far bigger volume of space and a larger number of sources. We employ a model-selection scheme called TIGER (Test Infrastructure for GEneral Relativity), which was originally developed to test general relativity with weak binary coalescence signals that will be seen in second-generation detectors, such as Advanced LIGO and Advanced Virgo. TIGER is well suited for the regime of low signal-to-noise ratios, and information from a population of sources can be combined so as to arrive at a stronger test. By performing a range of simulations using the expected noise power spectral density of the Einstein Telescope, we show that with TIGER, similar deviations from the no-hair theorem (such as those considered in previous works) will be detectable with great confidence using $O(10)$ sources distributed uniformly in a comoving volume out to 50 Gpc ($z \lesssim 5$).

DOI: 10.1103/PhysRevD.90.064009 PACS numbers: 04.25.dg, 04.80.Nn, 95.55.Ym, 04.80.Cc

1. INTRODUCTION

The no-hair theorem states that a black hole that has settled down to its final stationary vacuum state is determined only by its mass, spin, and electric charge [1–5]. Astrophysical black holes are thought to be electrically neutral so that only mass and spin need to be considered, leading to the Kerr geometry. When a black hole is formed as a result of the inspiral and merger of two other compact objects, it will undergo “ringdown” as it evolves towards its quiescent state. This process can be modeled by considering linear perturbations of the Kerr metric—or quasinormal modes—which are characterized by frequencies $\omega_{lm}$ and damping times $\tau_{lm}$ [6–9]. Since the underlying Kerr spacetime is only characterized by its mass $M$ and spin $J$, these frequencies and times are constrained by linearized general relativity to only depend on these quantities through specific functional relationships, so that observational tests of these dependences would constitute a test of the no-hair theorem, and hence of general relativity (GR) [10]; this was first hinted at by Detweiler [12], made concrete by Dreyer et al. [13], and further explored in Refs. [14–16].

Recently Gossan, Veitch, and Sathyaprakash [17] investigated the possibility of performing this kind of test using the Einstein Telescope (ET), a proposed third-generation ground-based gravitational-wave detector [18], as well as with the space-based eLISA [19]. These authors evaluated two methods for checking the dependences $\omega_{lm} = \omega_{lm}^{GR}(M, J)$ and $\tau_{lm} = \tau_{lm}^{GR}(M, J)$ predicted by GR: Bayesian parameter estimation and model selection. Specifically, one can write possible deviations from these dependences as

$$\omega_{lm} = \omega_{lm}^{GR}(M, J)(1 + \delta \omega_{lm}), \quad (1)$$

$$\tau_{lm} = \tau_{lm}^{GR}(M, J)(1 + \delta \tau_{lm}), \quad (2)$$

and then (a) calculate how well the dimensionless quantities $\delta \omega_{lm}$, $\delta \tau_{lm}$ can be measured, or (b) compare the evidences for two models: one where the $\delta \omega_{lm}$, $\delta \tau_{lm}$ are free parameters, and another in which they are all identically zero, corresponding to the GR prediction. In practice, the authors of Ref. [17] restricted their attention to the set

$$\{ \delta \omega_{22}, \delta \omega_{33}, \delta \tau_{22} \}.$$ (3)

It was found that for black holes with masses in the range 500–1000 $M_\odot$ at a distance of 6 Gpc, the ET would allow for measurements of $\delta \omega_{22}$, $\delta \omega_{33}$, and $\delta \tau_{22}$ with accuracies of a few percent for the first two parameters, and about 10% for the third. (For comparison, boson stars in the same mass range would cause $\delta \omega_{22}$ and $\delta \tau_{22}$ to be of order 1 [20].) With model selection and assuming a 500 $M_\odot$ black hole, a deviation of a few percent in $\delta \omega_{22}$ could be discriminated.

---

*jmeidam@nikhef.nl
from GR with $\ln \frac{P^{\text{dev}}}{P^{\text{GR}}} > 10$, were $B^{\text{dev}}_{\text{GR}}$ is the Bayes factor, or ratio of evidences, for the model that deviates from GR [with the variables in Eq. (3) as extra free parameters] versus the GR model.

How frequently might one test GR in this way? Coalescence rates of intermediate-mass binary black holes which would give rise to ringdowns with masses in the above range are highly uncertain [21–23]; the ET may see between a few and a few thousand per year [24,25]. Gossan et al. considered single, relatively loud sources, but one will also want to combine information from multiple, possibly weak signals out to large distances so as to maximally exploit the available set of detections. Since deviations from the no-hair theorem may be such that $\delta \omega_{22}, \delta \omega_{33}, \text{and/or} \delta \tilde{\tau}_{22}$ take on different nonzero values for different sources, when doing parameter estimation it will not be possible to combine posterior probability densities from multiple events unless one already assumes that GR is correct. On the other hand, although Bayesian model selection does lend itself quite easily to the utilization of all available detections, if one lets $\{ \delta \omega_{22}, \delta \omega_{33}, \delta \tilde{\tau}_{22} \}$ (and possibly more of the $\delta \omega_{\text{lm}}, \delta \tilde{\tau}_{\text{lm}}$) vary all at the same time, one may be penalized if the corresponding model is insufficiently parsimonious, i.e., if the correct model involves a smaller number of additional parameters.

In Refs. [26–30], a more general algorithm for testing GR was developed, called TIGER (Test Infrastructure for General Relativity). One takes a gravitational waveform model as predicted by GR, and introduces deformations parametrized by dimensionless quantities $\delta \xi_i$, $i = 1, 2, \ldots, N_T$ such that all of the $\delta \xi_i$ being zero corresponds to GR being correct. One can then ask the question, “Do one or more of the $\delta \xi_i$ differ from zero?” Let us denote the corresponding hypothesis by $\mathcal{H}_{\text{modGR}}$, and the GR hypothesis by $\mathcal{H}_{\text{GR}}$. Now, there is no waveform model that corresponds to $\mathcal{H}_{\text{modGR}}$. However, as shown in Ref. [26], one can define logically disjoint “subhypotheses” $\mathcal{H}_{i_1 i_2 \ldots i_k}$, in each of which a fixed set of parameters $\{ \delta \xi_{i_1}, \delta \xi_{i_2}, \ldots, \delta \xi_{i_k} \}$ are nonzero while $\delta \xi_j = 0$ for $j \notin \{ i_1, i_2, \ldots, i_k \}$. There are $2^{N_T} - 1$ such subhypotheses, corresponding to the nonempty subsets of the full set $\{ \delta \xi_1, \delta \xi_2, \ldots, \delta \xi_{N_T} \}$. The $\mathcal{H}_{i_1 i_2 \ldots i_k}$ do have waveform models associated with them that can be compared with the data, and $\mathcal{H}_{\text{modGR}}$ can be expressed as the logical union of all the subhypotheses:

$$\mathcal{H}_{\text{modGR}} = \bigvee_{i_1 < i_2 < \ldots < i_k; k \leq N_T} \mathcal{H}_{i_1 i_2 \ldots i_k}. \quad (4)$$

Given a catalog of detections $d_1, d_2, \ldots, d_N$ and whatever background information $I$ one may possess, one can then compute the odds ratio for $\mathcal{H}_{\text{modGR}}$ against $\mathcal{H}_{\text{GR}}$:

$$O_{\text{modGR}}^{\text{GR}} = \frac{P(\mathcal{H}_{\text{modGR}}|d,I)}{P(\mathcal{H}_{\text{GR}}|d,I)} = \frac{\alpha}{2^{N_T} - 1} \sum_{i_1 < i_2 < \ldots < i_k; k \leq N_T} \prod_{A=1}^{N} B^{(i_1 i_2 \ldots i_k)}_{\text{GR}}. \quad (5)$$

Here $\alpha$ is an unimportant scaling factor which below will be set to unity, and the Bayes factors $B^{(i_1 i_2 \ldots i_k)}_{\text{GR}}$ for a detection $d_A$ are given by

$$B^{(i_1 i_2 \ldots i_k)}_{\text{GR}} = \frac{P(d_A|\mathcal{H}_{i_1 i_2 \ldots i_k},I)}{P(d_A|\mathcal{H}_{\text{GR}},I)}, \quad (6)$$

where $P(d_A|\mathcal{H}_{i_1 i_2 \ldots i_k},I)$ and $P(d_A|\mathcal{H}_{\text{GR}},I)$ are the evidences for $\mathcal{H}_{i_1 i_2 \ldots i_k}$ and $\mathcal{H}_{\text{GR}}$, respectively. For basic assumptions and detailed derivations we refer the reader to Refs. [26–28].

The TIGER formalism has been evaluated extensively in the context of binary neutron star inspirals that will be observed by second-generation detectors, such as Advanced LIGO [31], Advanced Virgo [32], GEO-HF [33], KAGRA [34], and LIGO-India [35]. In Refs. [26,27] it was shown that, thanks to the introduction of the $\mathcal{H}_{i_1 i_2 \ldots i_k}$, the method avoids potential problems due to insufficient parsimony, is well suited to dealing with weak signals, and enables the discovery of a wide range of deviations from GR, including ones that are well outside the particular parametrized waveform family used; moreover, information from multiple sources can trivially be combined.

However, TIGER is not tied to any particular gravitational waveform model and can be applied to testing the no-hair theorem with ringdown signals. Consider the $N_T = 3$ testing parameters of Ref. [17],

$$\delta \xi_1 = \delta \omega_{22}, \quad \delta \xi_2 = \delta \omega_{33}, \quad \delta \xi_3 = \delta \tilde{\tau}_{22}. \quad (7)$$

$\mathcal{H}_{\text{modGR}}$, the hypothesis that one or more of the $\delta \xi_i$ deviate from their GR value, is then the logical union of $2^3 - 1 = 7$ subhypotheses $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3, \mathcal{H}_{12}, \mathcal{H}_{13}, \mathcal{H}_{23}$, and $\mathcal{H}_{123}$. Here $\mathcal{H}_1$ is the hypothesis that $\delta \xi_1 \neq 0$ while $\delta \xi_2 = \delta \xi_3 = 0$, $\mathcal{H}_{12}$ is the hypothesis that both $\delta \xi_1 \neq 0$ and $\delta \xi_2 \neq 0$ but $\delta \xi_3 = 0$, and similarly for the other subhypotheses. In the above language, the model-selection setup of Gossan et al. [17] only involved calculating, for a single source, the Bayes factor

$$B^{123}_{\text{GR}} = \frac{P(d|\mathcal{H}_{123},I)}{P(d|\mathcal{H}_{\text{GR}},I)}. \quad (8)$$

It would be of great interest to see how our ability to discern violations of the no-hair theorem with ringdown signals would improve if the full formalism of TIGER were brought to bear. This will be the main topic of the present paper.

When evaluating the odds ratio $O_{\text{modGR}}^{\text{GR}}$ of Eq. (5) using one or more detected signals, we may find that there is no reason to believe that GR is incorrect. However, in that case it will still be of interest to measure $\delta \omega_{22}, \delta \omega_{33}, \text{and} \delta \tilde{\tau}_{22}$ for each source and combine the resulting posterior density distributions so as to arrive at a joint result for the entire catalog of detections. This we will also do, and as we shall...
see, potentially tight constraints can be set on these parameters.

This paper is structured as follows. In Sec. II we explain our assumptions regarding the Einstein Telescope as well as our waveform models for signals and templates, and the setup of the simulations. In Sec. III we evaluate TIGER’s ability to perform tests of the no-hair theorem. The possibility of precision measurements of the free parameters in the case where we have no reason to doubt GR is discussed in Sec. IV. Section V provides a summary and conclusions.

Throughout this paper we will use units such that $G = c = 1$.

II. DETECTORS, WAVEFORM MODELS, AND SETUP OF THE SIMULATIONS

A. The Einstein Telescope

In 2011, a conceptual design study for the Einstein Telescope was concluded [18,36–38]. The ET is envisaged to consist of three underground detectors arranged in a unilateral triangle with 10 km sides. Each detector is composed of two interferometers: a cryogenic one with improved sensitivity at low frequencies (up to ~40 Hz) due to the suppression of thermal noise, and a noncryogenic interferometer which is more sensitive at high frequencies (up to several kHz) due to higher laser power, which reduces quantum noise.

The combined strain sensitivity as a function of frequency for each detector is the one labeled “ET-D” in Fig. 1 [39]. In the same plot we show the older “ET-B” curve used by Gossan et al. [17], where only a single interferometer was assumed for each of the detectors. Currently ET-D corresponds to the most detailed assessment of the possible noise budget of the Einstein Telescope. We will take the lower cutoff frequency to be 10 Hz, and the dominant mode frequencies considered in this paper will roughly lie between 15 and 100 Hz, a range in which, for the most part, ET-D is less sensitive than ET-B, by up to a factor of 2. We note that in reality it may be possible to achieve a lower frequency cutoff of only a few Hz, and between there and 25 Hz, ET-D is actually more sensitive than ET-B, which would lead to comparably better visibility of higher-mass sources; hence, our assumptions are conservative. For each of the three detectors, stretches of simulated stationary, Gaussian noise were produced with ET-D as an underlying power spectral density. Simulated signals were added coherently to each of the three data streams, taking into account the different detector responses [40].

B. Waveform models

The ringdown signal is given by a superposition of quasinormal modes characterized by triples of integers $(l, m, n)$, where $l \geq 2$ and $m = -l, -l+1, \ldots, l-1, l$; $n \geq 0$ is an overtone index [8,9,41,42]. Here we will only consider the modes with $n = 0$, as overtones with $n > 0$ are not significantly excited and have much shorter damping times [14]. The “plus” and “cross” polarizations read

$$h_+(t) = \frac{M}{D_L} \sum_{l,m>0} A_{l|m|} e^{-t/\tau_{lm}} Y_{l;m}^{lm}(t) \cos(\omega_{lm} t - m\phi),$$

$$h_\times(t) = -\frac{M}{D_L} \sum_{l,m>0} A_{l|m|} e^{-t/\tau_{lm}} Y_{l;m}^{lm}(t) \sin(\omega_{lm} t - m\phi),$$

(9)

where the $Y_{l;m}^{lm}(t)$ can be written in terms of spin-weighted spherical harmonics of weight $-2$:

$$Y_{l;m}^{lm}(t) \equiv (-1)^l Y_{l;m}^{lm}(t, 0) + (-1)^{l+1} Y_{l;m}^{lm+1}(t, 0),$$

$$Y_{l;m}^{lm}(t) \equiv (-1)^l Y_{l;m}^{lm}(t, 0) - (-1)^{l+1} Y_{l;m}^{lm-1}(t, 0).$$

(10)

In the above, $M$ is the observed mass of the black hole, which is related to the intrinsic mass by $M = (1 + z)M_{\text{int}}$ with redshift $z$, $D_L$ is the luminosity distance to the source, $t$ is the angle between the black hole’s spin and the line of sight, and $\phi$ is the azimuth angle of the black hole with respect to the observer. Note that in principle there will be additional phase offsets $\phi_{lm}$ in Eq. (9); since analytic fits for their dependence on progenitor parameters are not yet available, we set them to zero, as was also done in Refs. [17,43]. $\omega_{lm}(M, j)$ and $\tau_{lm}(M, j)$ are the characteristic frequencies and damping times of the modes, respectively, as functions of the mass and of the dimensionless spin $j = J/M^2$.

As in Ref. [17], we only consider the modes $(l, m) = (2, 2), (2, 1), (3, 3), (4, 4)$, which are among the most dominant ones. Analytic expressions for the mode amplitudes $A_{l|m|}$ are not available, but there exist accurate fits to numerical simulations. The authors of Ref. [17] took the progenitor black holes to be nonspinning, in which case...
one can use the approximate expressions for the $A_{lm}$ in terms of the symmetric mass ratio $\nu = m_1 m_2 / (m_1 + m_2)^2$ (where $m_1$, $m_2$ are the progenitor component masses) from Kamaretsos et al. [16,44]. Here we will relax this assumption and include the effect of nonzero progenitor spins in the waveforms, using more recent results. For spinning progenitors, Kamaretsos, Hannam, and Sathyaprakash [46] found that mainly $A_{21}$ is strongly affected, and a good fit for all the relevant amplitudes is given by

$$A_{22}(\nu) = 0.864 \nu,$$

$$A_{21}(\nu) = 0.43 \sqrt{1 - 4\nu} - \chi_{\text{eff}}|A_{22}(\nu)|,$$

$$A_{33}(\nu) = 0.44 (1 - 4\nu)^{0.45} A_{22}(\nu),$$

$$A_{44}(\nu) = [5.4(\nu - 0.22)^2 + 0.04] A_{22}(\nu),$$

where

$$\chi_{\text{eff}} = \frac{1}{2} \left( \sqrt{1 - 4\nu \chi_1 + \chi_-} \right),$$

with

$$\chi_- = \frac{m_1 \chi_1 - m_2 \chi_2}{M_{\text{in}}}.\quad (16)$$

Here $(m_1, m_2)$ and $(\chi_1, \chi_2)$ are, respectively, the progenitor component masses and dimensionless spin magnitudes, and $M_{\text{in}}$ is the initial total mass of the system, which to reasonable approximation can be taken to be equal to the mass of the final black hole.

For the frequencies $\omega_{lm}$ and damping times $\tau_{lm}$ there also exist good fits, which can be expressed through the quality factors $Q_{lm} = \omega_{lm} \tau_{lm}/2$:

$$M \omega = f_1 + f_2 (1 - j) f_3,$$

$$Q = q_1 + q_2 (1 - j) q_3,$$

where for the values of the coefficients $f_1, f_2, f_3, q_1, q_2, q_3$ we refer to Ref. [14]. Finally, there exists a simple fit for the spin $j$ of the final black hole in terms of the component masses $(m_1, m_2)$ and spins $(\tilde{\chi}_1, \tilde{\chi}_2)$ [47,48], for which we refer to Ref. [48].

For the simulated signals, or injections, we choose progenitor spins $\tilde{\chi}_1, \tilde{\chi}_2$ from a distribution with isotropic directions, and a Gaussian distribution for the magnitudes centered on 0.7, with standard deviation 0.2 and hard cutoffs at 0.5 and 0.99 [49]; note that the value of 0.7 roughly corresponds to what one gets from the coalescence of nonspinning, equal-mass binary black holes. The mass $M$ is drawn from a uniform distribution between 500 and 1000 $M_\odot$, and the mass ratio $q = m_1 / m_2$ from a uniform distribution between 0.3 and 1. Amplitudes are computed as in Eqs. (11)–(16), where we take $\chi_{1,2} = [\tilde{\chi}_{1,2}]$, and the final spin $j$ is calculated from component masses and spins using the formula of Barausse and Rezzolla [48]. With these choices for masses and spins, the characteristic frequency $f_{22} = \omega_{22}/(2\pi)$ of the dominant ringdown mode ranges from about 15 to 100 Hz, while the inspiral signal, which ends roughly at $f_{\text{ISO}} = (6^{3/2} \pi M)^{-1}$, stays below the lower cutoff frequency of 10 Hz and hence is never in the sensitive frequency band. Redshifts are taken to be between 1.5 and 5, and sources are placed uniformly in a comoving volume assuming a ΛCDM cosmology with $(\Omega_M, \Omega_\Lambda, h_0) = (0.27, 0.73, 0.70)$, so that luminosity distances approximately range from 10 to 50 Gpc. Since part of the exercise is to stress-test the TIGER framework, we only analyze sources with signal-to-noise ratio (SNR) < 30, corresponding to a minimum angle-averaged distance of 14.97 Gpc ($\bar{z} = 1.90$). Sky positions $(\theta, \phi)$ and orientations $(t, \psi)$ are drawn from uniform distributions on the sphere.

To gauge our sensitivity to deviations in $\omega_{22}(M, j), \omega_{33}(M, j)$, and $\tau_{22}(M, j)$, we introduce constant relative shifts $\delta \omega_{22}, \delta \omega_{33}$, and $\delta \tau_{22}$ as explained in the Introduction.

For the templates, we only take $\chi_{\text{eff}}$ and $j$ to be the spin-related free parameters, as the progenitor component spins $\tilde{\chi}_1$ and $\tilde{\chi}_2$ will not be separately measurable from a ringdown signal alone [50]. The free parameters for the waveform model corresponding to the GR hypothesis $\mathcal{H}_\text{GR}$ are then

$$\tilde{\theta}_\text{GR} = \{M, \nu, j, \chi_{\text{eff}}, D_L, \theta, \phi, \psi, t, \phi, t_0\},$$

where $t_0$ is the time of arrival of the signal at the detector. The prior on $M$ is chosen to be uniform between 300 and 1200 $M_\odot$, and that for the symmetric mass ratio $\nu$ is flat between 0.01 and 0.25; in terms of the mass ratio $q = m_1 / m_2$ this range corresponds to $0.01 \lesssim q \leq 1$. The prior on $j$ is uniform between 0.01 and 0.99, and that on $\chi_{\text{eff}}$ is uniform between $-1$ and 1. Sky positions and orientations are taken to be uniform on the sphere, and the prior on distance is uniform in a comoving volume between 8 and 60 Gpc. $t_0$ is taken to be in a window of width 100 ms.

C. TIGER for ringdown

To apply TIGER in the context of ringdown, we introduce the same parametrized deformations of the waveform as in Ref. [17], namely the ones of Eqs. (1)–(3). The parameter spaces corresponding to the various subhypotheses $\mathcal{H}_{i_1, \ldots, i_n}$ of $\mathcal{H}_\text{modGR}$ are given by

$$H_{1} \leftrightarrow \{\tilde{\theta}_\text{GR}, \delta \omega_{22}\},$$

$$H_{2} \leftrightarrow \{\tilde{\theta}_\text{GR}, \delta \omega_{33}\},$$

$$H_{3} \leftrightarrow \{\tilde{\theta}_\text{GR}, \delta \tau_{22}\},$$

$$H_{12} \leftrightarrow \{\tilde{\theta}_\text{GR}, \delta \omega_{22}, \delta \omega_{33}\},$$

$$H_{13} \leftrightarrow \{\tilde{\theta}_\text{GR}, \delta \omega_{22}, \delta \tau_{22}\},$$

$$H_{23} \leftrightarrow \{\tilde{\theta}_\text{GR}, \delta \omega_{33}, \delta \tau_{22}\},$$

$$H_{123} \leftrightarrow \{\tilde{\theta}_\text{GR}, \delta \omega_{22}, \delta \omega_{33}, \delta \tau_{22}\}.\quad (20)$$
Given a detection $d_A$, the corresponding Bayes factors $(A)B^1_{GR}$, $(A)B^2_{GR}$, $(A)B^3_{GR}$, $(A)B^{12}_{GR}$, $(A)B^{13}_{GR}$, $(A)B^{23}_{GR}$, and $(A)B^{123}_{GR}$ are calculated using

$$
(A)B^{i_1\ldots i_k}_{GR} = \frac{(A)B^{i_1\ldots i_k}_{noise}}{(A)B^{i_1\ldots i_k}_{noise}},
$$

(21)

where $(A)B^{i_1\ldots i_k}_{noise}$ and $(A)B^{GR}$ are, respectively, the Bayes factors for $H^{i_1\ldots i_k}_{noise}$ and $H^{GR}$ against the hypothesis that the data contain only noise. The latter are computed using an appropriate adaptation of the nested sampling algorithm as implemented by Veitch and Vecchio [51–53].

For completeness, we give the expression for the odds ratio $O^{modGR}_{GR}$ in the present context; given a catalog of $N$ ringdown signals, it reads

$$
O^{modGR}_{GR} = \frac{1}{7} \prod_{A=1}^{N} (A)B^1_{GR} + \prod_{A=1}^{N} (A)B^2_{GR} + \prod_{A=1}^{N} (A)B^3_{GR}
$$

$$
+ \prod_{A=1}^{N} (A)B^{12}_{GR} + \prod_{A=1}^{N} (A)B^{13}_{GR} + \prod_{A=1}^{N} (A)B^{23}_{GR}
$$

$$
+ \prod_{A=1}^{N} (A)B^{123}_{GR},
$$

(22)

In practice it is often convenient to work with the logarithm of the odds ratio, $\ln O^{modGR}_{GR}$. If GR is correct, then naively one would expect $O^{modGR}_{GR} < 1$, or $\ln O^{modGR}_{GR} < 0$. However, features in the noise can have a detrimental effect on the measurement of the log odds ratio, and in practice one can obtain slightly positive values of $\ln O^{modGR}_{GR}$ even if no deviation from GR is present. For this reason one usually constructs a background distribution $P(\ln O|H_{GR}, \kappa_{GR}, I)$ [26–30]. Here $\kappa_{GR}$ denotes a large number of (catalogs of) injections with GR waveforms, for each of which one computes $\ln O^{modGR}_{GR}$, whose normalized distribution constitutes $P(\ln O|H_{GR}, \kappa_{GR}, I)$. Given a maximum false alarm probability $\beta$ that one is willing to tolerate, one can use this background to set a threshold $\ln O_{\beta}$ for the measured log odds ratio to overcome; this threshold is defined such that $\beta$ is the fraction of the background distribution that is above $\ln O_{\beta}$,

$$
\beta = \int_{\ln O_{\beta}}^{\infty} P(\ln O|H_{GR}, \kappa_{GR}, I) d\ln O.
$$

(23)

In reality there will only be a single value for the measured log odds ratio, computed from the signals one actually detects. However, if one wants to assess how likely it is that a particular type of deviation from GR, denoted by $H_{non-GR}$, will lead to a log odds ratio above threshold, then one can construct a foreground distribution $P(\ln O|H_{non-GR}, \kappa_{non-GR}, I)$, where this time $\kappa_{non-GR}$ is a set of injections whose waveforms are in accordance with the given GR violation. One can then define the efficiency $\zeta$ as the fraction of the foreground that is above threshold:

$$
\zeta = \int_{\ln O_{\beta}}^{\infty} P(\ln O|H_{non-GR}, \kappa_{non-GR}, I) d\ln O.
$$

(24)

This can be viewed as the probability that the particular kind of deviation from GR considered will be discovered with a false-alarm probability of at most $\beta$.

In what follows, we will consider both the case where only a single ringdown detection is ever made by the ET, so that $N = 1$, and the case where multiple detections are made. As mentioned before, the event rate for ringdowns with masses in the range 500–1000 $M_\odot$ is highly uncertain, but a few tens of detections out to tens of Gpc is consistent with expectations in the literature [21–25]. Below we will show results where injections are randomly combined into catalogs of $O(10)$ sources each.

To evaluate TIGER’s ability to find deviations from the no-hair theorem, we will mostly study its behavior in the following cases:

1. There is a 10% deviation in the dominant-mode frequency $\omega_{22}$, but other mode frequencies as well as the damping times are unaffected, i.e., the injections have $(\delta \omega_{22}, \delta \omega_{33}, \delta \tau_{22}) = (0.1, 0, 0)$.
2. There is a 10% deviation in the (3,3)-mode frequency $\omega_{33}$, but no deviation in other frequencies or in the damping times, i.e., $(\delta \omega_{22}, \delta \omega_{33}, \delta \tau_{22}) = (0, 0, 1)$.
3. There is a 10% deviation in the dominant-mode damping time $\tau_{22}$, but no deviation in other damping times or in the frequencies: $(\delta \omega_{22}, \delta \omega_{33}, \delta \tau_{22}) = (0, 0.1, 0)$.
4. There is a 25% deviation in $\tau_{22}$, but no deviation in other damping times or in the frequencies: $(\delta \omega_{22}, \delta \omega_{33}, \delta \tau_{22}) = (0, 0.0, 25)$.

Note that in the notation introduced above, this means that, in turn, we take $H_1$, $H_2$, and $H_3$ to be the correct hypotheses; the resulting distributions of log odds ratio for single sources as well as catalogs of sources will be our foreground distributions. We also consider the case where the no-hair theorem holds, i.e., $H_{GR}$ is the correct hypothesis and the injections have $\delta \omega_{22} = \delta \omega_{33} = \delta \tau_{22} = 0$. The log odds ratio distributions resulting from the latter—again for single sources and catalogs of sources—will be our backgrounds.

III. TESTING THE NO-HAIR THEOREM WITH TIGER

Let us first focus on background and foreground distributions for the case where only a single ringdown detection is ever made by the ET, i.e., $N = 1$. Results for $\sim 1300$ sources are shown in Fig. 2. We see that in all of the cases, there is significant overlap between background.
and foreground, so that with a single source one has no guarantee that violations of the no-hair theorem at these levels will be picked up. With a maximum tolerable false-alarm probability of \( \beta = 0.05 \), the efficiency \( \zeta \) for a 10% shift in \( \omega_{22} \) is 0.47, and for a 10% shift in \( \omega_{33} \) it is 0.46. Note how the efficiencies for deviations in \( \omega_{22} \) and \( \omega_{33} \) are comparable; with our choice for the injected range of mass ratios (\( 0.3 < q < 1 \), or \( 0.18 \lesssim \nu < 0.25 \)) there will be sources with \( A_{33} > A_{22} \) as well as sources with \( A_{33} < A_{22} \) (see Fig. of Ref. [17]). For a 10% shift in \( \tau_{22} \) we find \( \zeta = 0 \). Thus, even with only a single detection, one will have a reasonable chance of finding a GR violation of the given size in \( \omega_{22} \) and \( \omega_{33} \); however, the same shift in \( \tau_{22} \) will be essentially unobservable.

At least for anomalies in \( \omega_{22} \) and \( \omega_{33} \), the situation changes dramatically if information from multiple detections can be combined. This is shown in Fig. 3, for catalogs of ten sources each. For the same maximum false-alarm probability and the given shifts in \( \omega_{22} \), \( \omega_{33} \), and \( \tau_{22} \), the efficiencies become, respectively, 0.98, 0.98, and 0.13. Thus, there is a very significant improvement in the first two cases, but the shift in \( \tau_{22} \) remains hard to observe.

It is also of interest to see how the efficiencies grow with an increasing number of sources per catalog. This is shown in Fig. 4, for two choices of maximum tolerable false-alarm probability: \( \beta = 0.05 \) and \( \beta = 0.01 \). Due to the finite number of catalogs considered, inevitably the numbers we quote for efficiencies are not exact; in the plot we show medians and 95% confidence intervals obtained for \( \zeta \) when randomly combining the available simulated sources into catalogs of a given size in 1000 different ways. For the cases \( \delta \omega_{22} = 0.1 \) and \( \delta \omega_{33} = 0.1 \), we see that for either value of \( \beta \), the efficiency reaches essentially 100% for \( \sim 20 \) sources per catalog. However, for a GR violation with \( \delta \tau_{22} = 0.1 \) and as many as 50 sources per catalog, even with \( \beta = 0.05 \) the median efficiency is only \( \sim 0.2 \), with a large spread.

One may then wonder how large a deviation in \( \tau_{22} \) needs to be before it becomes detectable with good efficiency, still assuming a few tens of sources per catalog. In Fig. 5, we...
The use of multiple subhypotheses has a significant impact on finding a deviation from GR, as illustrated in Fig. 6. Here we arrange simulated sources in order of increasing SNR, and we consider the Bayes factors and the cumulative number of times that the correct hypothesis comes out on top. However, the advantages of TIGER are not limited to this. The above pertained to single sources. In Fig. 7 we consider, for an example catalog of sources, the evolution of the combined Bayes factors

\[ B_{\text{GR}}^{i_1i_2 \ldots i_k} = \prod_{A=1}^{N} (A) B_{\text{GR}}^{i_1i_2 \ldots i_k}, \]

as well as in \( \text{modGR} \), as information from more and more detections is added; the sources are arranged in order of increasing SNR. We can make two observations:

(i) The hypothesis \( H_{123} \) where all the parameters \{\( \delta \omega_{33}, \delta \tau_{22}, \delta \omega_{12} \)\} are left free does not dominate the log odds ratio, and indeed is deprecated compared with some of the other subhypotheses. This illustrates how one can suffer significant loss in discriminatory power if the non-GR model is insufficiently parsimonious, i.e., has more free parameters compared with the number of additional parameters that is actually needed. TIGER does not have this problem.

(ii) The correct hypothesis—in this case \( H_1 \)—is also not necessarily the dominant one. Indeed, it can happen that detector noise obscures the true nature of the GR violation so that some other hypothesis (in this example \( H_{12} \)) ends up on top. However, what is unlikely to happen is that the noise makes a non-GR signal look like a GR one. In a situation where most signals are weak, one should use TIGER with as many testing parameters \{\( \delta \xi_1, \delta \xi_2, \ldots, \delta \xi_N \)\} as is computationally feasible.
If, when comparing the measured log odds ratio with GR, as well as the log odds ratio \( \ln B_{\text{modGR}}^{i_1 \ldots i_N} \), one can combine PDFs across sources through

\[
p(\lambda_1 | \mathcal{H}, d_1, d_2, \ldots, d_N, I) = p(\lambda_1 | \mathcal{H}, I)^{1-N} \times \prod_{A=1}^{N} p(\lambda_1 | \mathcal{H}, d_A, I),
\]

as was done in a different context in, e.g., Ref. [55].

In the present context one can obtain PDFs for \( \delta \omega_{22} \), \( \delta \omega_{33} \), and \( \delta \tau_{22} \) by, e.g., letting \( \mathcal{H} \) be \( H_1 \), \( H_2 \), and \( H_3 \), respectively. Now, if there is evidence that GR is violated (because of the measured \( \ln B_{\text{modGR}}^{i_1 \ldots i_N} \) being above threshold), then there is no \textit{a priori} reason to assume that the \( \{ \delta \omega_{22}, \delta \omega_{33}, \delta \tau_{22} \} \) will be the same for all sources. Note that although this is the choice we made for injections in the previous section, even if these additional parameters had been nonconstant it would not have been a problem to do \textit{model selection} with multiple sources, since each of the Bayes factors \( B_{\text{GR}}^{i_1 \ldots i_N} \) only gauge whether the hypothesis \( H_{i_1 \ldots i_N} \) is more probable than \( \mathcal{H}_{\text{GR}} \). In doing \textit{parameter estimation} one has to be more careful.
On the other hand, suppose that there is no evidence of GR being incorrect, i.e., the measured $\ln O_{\text{modGR}}^{\text{GR}}$ is well below threshold. Then one can expect that $\hat{\delta} \omega_{22}^2$, $\hat{\delta} \omega_{33}^2$, and $\hat{\delta} \tau_{22}^2$ are all constant, namely, $\hat{\delta} \omega_{22}^2 = \hat{\delta} \omega_{33}^2 = \hat{\delta} \tau_{22}^2 = 0$, and it makes sense to combine PDFs from multiple sources as in Eq. (28). In turn, we let $H$ be $H_1$, $H_2$, and $H_3$, and compute marginalized PDFs for $\hat{\delta} \omega_{22}^2$, $\hat{\delta} \omega_{33}^2$, and $\hat{\delta} \tau_{22}^2$, respectively.

Results are shown in Fig. 8. In the top panels, we consider PDFs both for an example single source at $D_L = 20.69$ Gpc ($z = 2.47$), and for a catalog of 20 sources. For the single source, the spreads of the PDFs are roughly consistent with an extrapolation of the results of Gossan et al. to the given luminosity distance. (For injections and templates with nonspinning progenitors and $D_L < 6$ Gpc, we get uncertainties that are in close agreement with theirs.) As expected, the single-source PDFs are quite wide and uninformative, with standard deviations of 0.10, 0.13, and 0.21, respectively. For $\hat{\delta} \omega_{22}^2$ and $\hat{\delta} \omega_{33}^2$, with 20 sources the PDFs become strongly peaked (with widths of 0.0051 and 0.0066, respectively), and there is very little bias. For $\hat{\delta} \tau_{22}^2$ the combined PDF is not only wide (with a standard deviation of 0.048), but it shows some bias as well (although the correct value of zero is within its support). In the bottom panels of the figure, we show the evolution of medians and 95% confidence intervals of the combined PDF as more and more detections are added. We see that even for $\hat{\delta} \tau_{22}^2$, the 95% confidence interval shrinks to $\sim 0.1$ when $\sim 40$ sources are at our disposal. Hence there is a clear advantage in computing PDFs using all available detections.

V. CONCLUSIONS

We have revisited the problem of testing the no-hair theorem using ringdown signals that will be seen by the Einstein Telescope. In previous work [17], it was shown how deviations of up to 10% in the ringdown mode frequencies $\omega_{22}$, $\omega_{33}$ and the damping time $\tau_{22}$ could be observed out to distances of $\sim 6$ Gpc, both through parameter estimation and model selection. Here we used the TIGER framework that was originally developed to test general relativity with stellar-mass binary inspiral signals in second-generation detectors [26–30]. In this model-selection scheme, parametrized deviations are introduced in the waveforms, and multiple auxiliary hypotheses are tested corresponding to all subsets of the extra free parameters. Information from multiple sources can trivially be combined. A log odds ratio $\ln O_{\text{modGR}}^{\text{GR}}$ is computed, which compares the probability that one or more of the auxiliary hypotheses are correct with the probability that GR is the right theory. Given the expected distribution of $\ln O_{\text{modGR}}^{\text{GR}}$ in the case that GR is correct, violations of GR are searched for by checking whether the measured log odds ratio is above a threshold set by a predetermined maximum false-alarm probability. If this is not the case then there is no reason to doubt GR, and one can calculate bounds on the free parameters, again combining information from all available sources.

Ringdown signals from black holes with masses in the range $500–1000 M_\odot$ can result from coalescences of intermediate-mass binary black holes, but such events...
may be rare \cite{21–25}. On the other hand, they can be seen with the ET out to redshifts of \( z \gtrsim 5 \). We have shown that with \( \mathcal{O}(10) \) sources and using the TIGER framework, deviations of the same size as the ones considered in Ref. \cite{17} can be seen, but for sources at distances up to \( 50 \) Gpc.

Our work illustrates how TIGER is not tied to any particular waveform model (nor even any particular type of source). It is well suited to the regime of low signal-to-noise ratios due to its use of multiple subhypotheses, which increases the chance of finding a GR violation. Because of detector noise, the correct hypothesis (or for that matter, the most inclusive hypothesis) may not yield the largest contribution to the log odds ratio, but it is unlikely that noise will make a GR-violating signal look like one that is in accordance with GR. For concreteness we only considered possible deviations in \( \{\omega_{22}, \omega_{33}, \tau_{22}\} \) (as was also done in Ref. \cite{17}), leading to seven auxiliary hypotheses in the TIGER framework, but in reality one should include as many subhypotheses as is feasible and modeling allows.

TIGER offers an effective way of finding very generic violations of GR. As shown in the context of compact binary coalescence, it can uncover deviations that are not included in any of the waveform models associated with the subhypotheses \( \mathcal{H}_{i_1i_2\ldots i_k} \) \cite{26,27}. We fully expect the same to be true for ringdown; an example could be the appearance of modes with spin weights different from \( -2 \), as in the case of a black hole in certain \( f(R) \) theories that are dynamically equivalent to Einstein-Proca theory \cite{56–58}. An explicit demonstration is left for future work. On the other hand, if the signal waveform is not among any of the template models, then fundamental bias can make it difficult to reliably pinpoint the underlying nature of the violation \cite{26,59–62}. This remains an open problem for ringdown as well.

When the log odds ratio does not indicate a violation of GR, upper limits can be put on deviations in the extra free parameters. A single source at large distance (\( >10 \) Gpc) may only give weak bounds and could show considerable bias. On the other hand, with \( \mathcal{O}(10) \) sources, deviations are well constrained even for the parameter \( \tau_{22} \), for which no meaningful bounds can be obtained with a single source at \( \text{SNR} \sim 20 \).

In this study we deliberately restricted attention to a black hole mass range for which the preceding inspiral signal cannot be seen, but the dominant ringdown mode is visible. In reality one would also expect lighter systems to be seen, for which one would want to utilize information from the inspiral and merger regimes as well. Given appropriate GR waveform models (as are likely to become available on the timescale of the ET) it should be possible to put extremely stringent restrictions on GR violations by using the thousands of stellar-mass binary coalescence events that the ET will plausibly observe. However, as we have shown, even events where only the ringdown can be accessed will separately allow for interesting tests of the strong-field dynamics of GR.

Finally—as was found in Ref. \cite{17} for the case of single systems with \( M \sim 10^6 \, M_\odot \)—eLISA will be able to perform tests of the no-hair theorem at a comparable level of accuracy as the ET with \( M \sim 10^3 \, M_\odot \). Since the detection rate for such sources with eLISA may be on the order of tens per year \cite{63} (i.e., what we assumed for the ET in this paper), results from TIGER—including the combining of information from multiple sources—should also be similar. Detailed investigations are left for future work.

ACKNOWLEDGEMENTS

M. A., J. M., C. V. D. B., and J. V. were supported by the research programme of the Foundation for Fundamental Research on Matter (FOM), which is partially supported by the Netherlands Organisation for Scientific Research (NWO). J. V. was also supported by STFC grant ST/K005014/1. B. S. S. was supported by STFC grants ST/L000962/1 and ST/L000342/1. It is a pleasure to thank E. Berti, V. Cardoso, and I. Kamaretos for useful comments and suggestions.

\begin{thebibliography}{99}
\bibitem{9} E. Berti, V. Cardoso, and A. O. Starinets, \textit{Classical Quantum Gravity} \textbf{26}, 163001 (2009).
\bibitem{10} For the purposes of this paper we will treat the no-hair theorem as if it were a prediction of GR, so that a violation of the theorem implies a GR violation. However, it should be noted that existing proofs require strong assumptions, such as analyticity of spacetime; see Ref. \cite{11} for an overview. For this reason, the no-hair theorem is often referred to as the no-hair conjecture.
\end{thebibliography}