Hierarchical Test of General Relativity with Gravitational Waves

Maximiliano Isi,[1,2,*] Katerina Chatziioannou,[1,†] and Will M. Farr[1,3,‡]
[1]Center for Computational Astrophysics, Flatiron Institute, 162 5th Ave, New York, New York 10010, USA
[2]LIGO Laboratory and Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[3]Department of Physics and Astronomy, Stony Brook University, Stony Brook, New York 11794, USA

(Received 5 May 2019; published 16 September 2019)

We propose a hierarchical approach to testing general relativity with multiple gravitational wave detections. Unlike existing strategies, our method does not assume that parameters quantifying deviations from general relativity are either common or completely unrelated across all sources. We instead assume that these parameters follow some underlying distribution, which we parametrize and constrain. This can be then compared to the distribution expected from general relativity, i.e., no deviation in any of the events. We demonstrate that our method is robust to measurement uncertainties and can be applied to theories of gravity where the parameters beyond general relativity are related to each other, as generally expected. Our method contains the two extremes of common and unrelated parameters as limiting cases. We apply the hierarchical model to the population of 10 binary black hole systems so far detected by LIGO and Virgo. We do this for a parametrized test of gravitational wave generation, by modeling the population distribution of beyond-general-relativity parameters with a Gaussian distribution. We compute the mean and the variance of the population and show that both are consistent with general relativity for all parameters we consider. In the best case, we find that the population properties of the existing binary signals are consistent with general relativity at the ∼1% level. This hierarchical approach subsumes and extends existing methodologies and is more robust at revealing potential subtle deviations from general relativity with increasing number of detections.

DOI: 10.1103/PhysRevLett.123.121101

Introduction.—The ever-increasing number of binary coalescences [1] detected by LIGO [2] and Virgo [3] has opened up avenues for rich new tests of general relativity (GR) [4–9]. This includes precision probes of strong-field orbital dynamics, the nature of the remnant object, and the properties of gravitational-wave (GW) propagation [4,9]. With the new data, however, comes the problem of properly interpreting constraints in a way that does not apply only to specific modified theories of gravity and that is not biased by hidden assumptions [9,10].

In particular, there is an outstanding challenge to adequately combine information from different GW observations into a single statement about agreement with GR. Existing approaches [6,11–19] rely on strong assumptions about the space of potential GR deviations and their effect on the observable events, rendering them too restrictive [10]. As a result, we might soon have a wealth of measurements from different techniques and events but no cohesive picture that brings them together.

In this Letter, we present a flexible and robust solution to this problem by framing it in the language of hierarchical inference. The result is an easy-to-interpret null test of GR that can incorporate multiple measurements from different events, without strong restrictions to specific theories of gravity or subclasses of events, and without the need to explicitly weigh events based on their significance. We demonstrate that our method can produce strong combined constraints on deviations from GR. If deviations are present, it can detect them even if they affect our measurements nontrivially, e.g., by altering waveforms in ways that depend on the properties of each source. We apply our method to GW detections from the GWTC–1 catalog of compact binaries [1,9], using publicly available posterior samples for parameters controlling waveform deviations from the GR prediction [20]. We obtain joint constraints on deviations from GR that apply to generic theories of gravity and find the data to be in agreement with Einstein's theory up to the ∼1% level.

Method.—Waveform models for quasicircular compact binaries so far exist only within GR. They are generically parametrized by 15 parameters that describe the signal observed by an interferometric detector: component masses, component spins, location, orientation, and phase. To date no parametrized waveform model exists that describes the inspiral, merger, and ringdown of generic binaries in any beyond-GR theory. For this reason, and guided by the desire for model-idependent tests that do not conform to specific theories, many studies are based on parametrized deviations away from the GR waveform. In these tests, new degrees of freedom $\delta\hat{p}_i$ are introduced, with $\delta\hat{p}_i = 0$ corresponding to

GR. These parameters are introduced to modify different aspects of the waveform's frequency or amplitude evolution and, together with the 15 GR parameters, define a generalized waveform model. For more details on these parametric tests, see the Supplemental Material [21], Ref. [9] and references therein.

Unless they are somehow fixed to a constant by the true theory of gravity, we should generally expect the $\delta\hat{p}_i$'s to vary across different GW events. For instance, the GW deformation could depend on the binary mass ratio or other properties of the system, and different combinations of $\delta\hat{p}_i$'s could come into play under different circumstances. Without assuming a theory of gravity, it is not possible to constrain the functional form of the $\delta\hat{p}_i$'s, making it difficult to combine measurements from different events [9,10]. To tackle this problem, we follow [10] and employ a hierarchical formalism wherein we assume that the true value of the beyond-GR parameters for each of the events is drawn from some common unknown distribution [22]. If there are P parameters measured for N events, this amounts to $P \times N$ random variables, which we denote $\delta\hat{p}_i^{(j)}$ for $i = 1, \ldots, P$ and $j = 1, \ldots, N$. Then each set of N variables corresponding to a given $\delta\hat{p}_i$ should follow a shared distribution, implicitly determined by the underlying theory of gravity and the source population properties. The goal of the hierarchical approach (vividly named "extreme deconvolution" by some [23]) is to infer the properties of the underlying distributions based on imperfect measurements from a population of events.

The first step is to select a functional form for the distribution of $\delta\hat{p}_i$, which is in principle nontrivial. Given the small number of detections, here we only attempt to measure its mean μ_i and standard deviation σ_i. Higher moments, such as the skewness, could become measurable with an increasing number of detections. In our case, and under a minimum-information assumption, we can thus model the population distribution with a Gaussian; i.e., we will take the population distribution to be $\delta\hat{p}_i \sim \mathcal{N}(\mu_i, \sigma_i)$. A more complex function could be chosen as needed, with little impact on the method. This potentially includes explicitly considering correlations among different $\delta\hat{p}_i$'s, although we demonstrate below that this is not strictly necessary.

With the above choice of likelihood and appropriate values of σ_i, our method reduces to traditional nonhierarchical approaches for combining events [10]. Setting $\sigma_i = 0$ amounts to assuming that all systems share the same beyond-GR parameter $\delta\hat{p}_i^{(j)} = \mu_i$. The results are equivalent to multiplying the likelihood functions of the $\delta\hat{p}_i^{(j)}$ for all detections j. On the opposite extreme, letting $\sigma_i \to \infty$, the $\delta\hat{p}_i^{(j)}$ are drawn from an effectively flat distribution and, as a result, measurement of one does not inform the others. This corresponds to testing a theory of gravity in which each system is described by its own fundamental constant [10]. The results are equivalent to multiplying the Bayes factors from individual detections (assuming that a flat prior is imposed on each beyond-GR parameter). However, both these assumptions can lead to incorrect conclusions if they do not apply to the true theory of gravity [9,10].

In its general form, our hierarchical method is not limited by those assumptions and provides a robust way of detecting a deviation from GR even when the non-GR parameters are not trivially related to each other. If GR is correct, then both hyperparameters, μ_i and σ_i, are expected to be consistent with zero. If we find a nonzero μ_i, this is an obvious deviation from GR or a systematic error in the analysis of one or several of the events under consideration. Alternately, the true $\delta\hat{p}_i^{(j)}$ could be symmetrically distributed around $\mu_i = 0$. In this case, the inferred μ_i will be consistent with 0, but the σ_i posterior will peak away from zero, signaling that the scatter in $\delta\hat{p}_i^{(j)}$ is larger than expected from statistical measurement errors, again revealing a beyond-GR effect (or modeling error).

Simulation: GR is right.—Given the long history of GR's experimental success [24], it is unavoidable to imagine that GW observations may also fail to reveal any shortcomings of the theory. Accordingly, we begin by demonstrating our method on simulated signals that obey GR. For simplicity, we take the measurement of each beyond-GR parameter to be summarized by a Gaussian likelihood with mean $\tilde{\mu}_i^{(j)}$ and standard deviation $\tilde{\sigma}_i^{(j)}$, i.e., $p(\text{data}^{(j)}|\delta\hat{p}_i^{(j)}) = \mathcal{N}(\tilde{\mu}_i^{(j)}, \tilde{\sigma}_i^{(j)})$. Such a likelihood is hardly realistic, especially for weak signals, but it suffices to illustrate our method and its scaling with the number of detections. Note that $\tilde{\mu}_i^{(j)}$ and $\tilde{\sigma}_i^{(j)}$ describe the idealized measurement of parameter $\delta\hat{p}_i$ in the jth event, while μ_i and σ_i define the distribution of true values of $\delta\hat{p}_i$ across events.

We simulate a population of N observations as follows: first, we assign a random signal-to-noise ratio (SNR) to each event j with the expected probability $\text{SNR}^{(j)} \sim 1/\text{SNR}^4$ [25]; then, for each $\delta\hat{p}_i^{(j)}$, we assign a value of $\tilde{\sigma}_i^{(j)}$ proportional to $1/\text{SNR}^{(j)}$; finally, we choose a value of $\tilde{\mu}_i^{(j)}$ consistent with $\tilde{\sigma}_i^{(j)}$ by drawing it from $\mathcal{N}(0, \tilde{\sigma}_i^{(j)})$, mimicking the expected scatter due to noise in the detector. For concreteness, we consider only three non-GR parameters $\delta\hat{\varphi}_i$, $i = 0, 1, 2$. These are defined as in [9] and are related to the parametrized post-Einsteinian (ppE) framework of [26], as discussed in the Supplemental Material [21]. We set the overall scale of the $\tilde{\sigma}_i^{(j)}$'s based on the uncertainty of measurements from GW150914 data, namely 68%-level widths of 0.06, 0.3, and 0.2 for $\delta\hat{\varphi}_0$, $\delta\hat{\varphi}_1$, and $\delta\hat{\varphi}_2$, respectively [20].

Figure 1 shows the projected constraints on μ_i (top) and σ_i (bottom) for the ppE-like coefficients $\delta\hat{\varphi}_0$, $\delta\hat{\varphi}_1$, and $\delta\hat{\varphi}_2$ as the number of detections grows. Colored bands represent the 1σ variation over 200 simulated populations. The dashed line is proportional to $1/\sqrt{N}$ and demonstrates

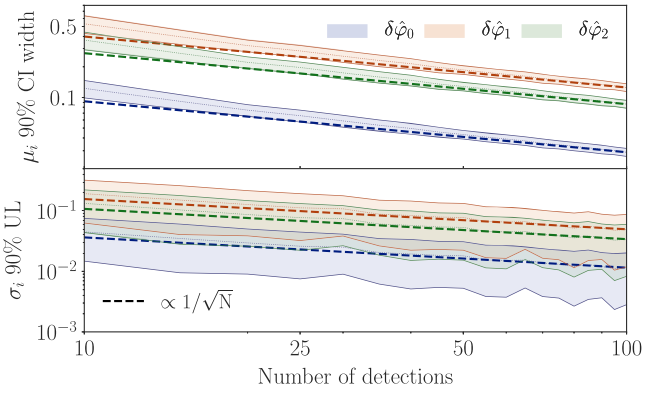

FIG. 1. Expected behavior of the population hyperparameters vs number of detections. We show the width of the 90% credible interval for μ_i (top) and the 90% upper limit on σ_i (bottom). In both panels, we average over 200 population realizations and shaded regions correspond to 1σ uncertainty. The dotted line show the mean over populations. The dashed line is proportional to $1/\sqrt{N}$; the bounds follow the expected scaling with the number of detections.

that bounds scale with the number of detections as expected. Our method improves with increasing number of signals at a rate similar to the simple approach of multiplying the likelihoods, in spite of the presence of an additional parameter, σ_i. This is because μ_i and σ_i are uncorrelated, so we can safely add σ_i to our model without affecting the $1/\sqrt{N}$ scaling of μ_i and vice versa.

Simulation: GR is wrong.—We now turn to the tantalizing scenario that GR disagrees with experiment. In such a case, we should generally expect the deviation from GR to manifest itself in multiple $\delta\hat{p}_i$'s, even if it intrinsically occurs at a specific post-Newtonian (PN) order [17,27]. This is because the phenomenological effect of modifications at different PN orders are not necessarily orthogonal, introducing degeneracies in our measurement. Consequently, a deviation from GR affecting a given $\delta\hat{p}_i$ could be measured through the μ_i and σ_i of multiple parameters, not just the one that is actually modified by the theory.

To demonstrate this effect, we construct a simple mock alternative theory of gravity that differs from GR at the 1PN order, affecting all binaries equally. This intrinsic waveform correction is independent of source parameters, making it amenable to multiplication of the individual parameter likelihoods. Generally, of course, this is not the case [5,28]. Even with this simplifying assumption, the *measured* $\delta\hat{p}_i$'s may vary in a nontrivial way with source properties as signals with different frequency contents may be affected by the same deviation differently.

Following [17], we assume that the measured non-GR parameters $\delta\hat{p}_i$ depend nontrivially on the true values $\delta\hat{p}_i^{\rm true}$. Generally, such relation could always be expressed via some measurement matrix M, such that $\delta\hat{p}_i = M\delta\hat{p}_i^{\rm true}$, where the components of M could depend on the specific properties of each system. For our example, we again consider the three ppE-like parameters $\delta\hat{p}_i = (\delta\hat{\varphi}_0, \delta\hat{\varphi}_1, \delta\hat{\varphi}_2)$ and we imagine $\delta\hat{p}_i^{\rm true} = (0, 0, 0.1)$; i.e., the only parameter in which the modified theory deviates from GR is $\delta\hat{\varphi}_2$. As an illustration, we arbitrarily pick a matrix M that yields $\delta\hat{p}_i = (1.1 - 2q, 0, 0.1)$, where q is the mass ratio of the system. This is inspired by the degeneracy between high- and low-order PN corrections demonstrated in [17]. Quantitative results will be highly dependent on the true measurement matrix, though we only wish to demonstrate the qualitative effect here.

We simulate a population of observations by drawing q uniformly from [0.1, 1] and using those values to produce the measured parameters $\delta\hat{p}_i$. To simulate the corresponding posteriors, we draw the event SNRs and add a scatter due to noise as in the previous section. As a result of the nontrivial dependence on q, the resulting population of each $\delta\hat{p}_i$ is not normally distributed. In spite of this, we demonstrate that our simple Gaussian model can detect the deviation from GR.

Figure 2 shows the posteriors for μ_i and σ_i for a population of 100 events. As expected, we find that the posterior for μ_2 peaks at the injected value of 0.1 and excludes GR at the 96% credible level. Additionally, we find that σ_0 is not consistent with GR at the $\gtrsim$99.99% credible level. This means that the scatter in $\delta\hat{\varphi}_0^{(j)}$ is too large to be accounted for by statistical noise. Indeed, part of the scatter in $\delta\hat{\varphi}_0^{(j)}$ is caused by the deviation from GR. This illustrates that, even if we did not take $\delta\hat{\varphi}_2$ into account, we would have detected this deviation from GR solely through the lower PN order coefficient. Additionally, the σ_0 posterior is farther from GR than the μ_2 one, suggesting that this deviation could be detected first with a lower PN-order parameter. We emphasize that these results are illustrative only: the properties of the posteriors in a real

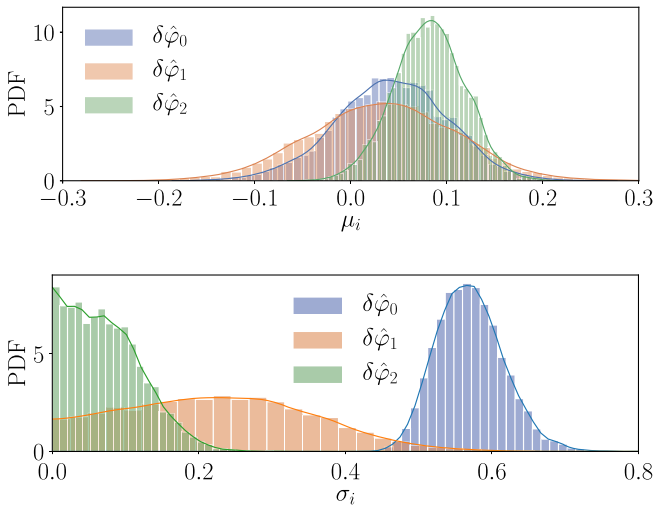

FIG. 2. Example hyperparameter posteriors when GR is not the correct theory of gravity. The deviation is only present at $\delta\hat{p}_2^{(j)} = 0.1$, but it is recovered both in μ_2 and σ_0. All other hyperparameters are consistent with GR.

analysis will depend on the nature of the true measurement matrix, which is generally unknown.

Real events.—We now apply our hierarchical model to the confident binary black hole detections presented in GWTC–1 [1]. As a starting point, we use posterior samples for all $\delta\hat{p}_i$ parameters from [9,20], obtained with the IMRPhenomPv2 waveform model [29,30]. This study did not perform both sets of tests on all detected signals, but rather imposed certain thresholds on the SNR of the signals to determine whether to look for deviations in the inspiral or postinspiral regime, or both. As a result, five signals were analyzed for inspiral deviations and nine for postinspiral ones. See [9] for details.

As demonstrated in the previous sections, for each ppE-like coefficient $\delta\hat{p}_i$, we obtain a posterior distribution for the corresponding hyperparameters μ_i and σ_i. We find that the population of the analyzed BBHs is consistent with GR both in terms of μ_i and σ_i for all beyond-GR parameters. All μ_i posteriors are consistent with 0 at the 0.5σ level or better, while all σ_i posteriors peak at 0. This is a novel test, sensitive to generic beyond-GR effects in a population of detections. With more signals we expect this analysis to either result in improved bounds (Fig. 1), or to reveal deviations from the ensemble properties expected from binaries in GR (Fig. 2).

From the hyperparameter posteriors, we also compute the inferred population distributions for the $\delta\hat{p}_i$'s, formally defined in Eq. (5) of the Supplemental Material [21] and plotted in Fig. 3. These distributions, $dN/d(\delta\hat{p}_i)$, represent our best knowledge of the population from which the allowed deviations from GR, $\delta\hat{p}_i^{(j)}$, were drawn for each binary black hole signal. Because all of these distributions contain zero with high probability, their width indicates the level to which we can constrain our measurements of the ppE-like coefficients to agree with GR. In the best case, for $\delta\hat{\varphi}_{-2}$, we find consistency with GR at the $\sim$1% level with 68% credibility. In the future, and assuming GR is correct, we should find that the distributions tend to a δ function at $\delta\hat{p}_i = 0$ as we accumulate more observations (cf. Fig. 1). We emphasize that, unlike Fig. 3 in [9], Fig. 3 here does not show the inferred posterior on the ppE-like parameters assuming all signals share the same value. Instead, Fig. 3 summarizes our inference for the distributions from which the potentially unequal ppE-like parameters of each signal were drawn [31].

These results are subject to the thresholds imposed in [9] that determine which GW events are subject to each test. They would thus be vulnerable to the same potential selection effects. This includes the requirement that signals be sufficiently loud and akin to GR, such that they are detectable by matched-filtering procedures looking for GR signals. Reference [9] argues that both types of potential selection effects are partially mitigated by the fact that more generic searches are also employed alongside matched-filter ones. With this caveat in mind, we find no evidence of any deviation from GR.

Conclusions.—We use a hierarchical approach to test GR with GWs by assuming that beyond-GR parameters in each event are drawn from a common underlying distribution. This approach is flexible and powerful, as it can encompass generic population distributions even if the chosen parametrization is inaccurate. It can trivially incorporate future detections and can be applied to different kinds of tests of GR, including searches for modified dispersion relations [7,32] or inspiral-merger-ringdown consistency checks [16,18]. We apply this method to the current 10 confident binary black hole detections [1], measuring posterior distributions for the mean and standard deviation of the population of ppE-like parameters $\delta\hat{p}_i$ [20]. This is a conceptually new test that examines the ensemble properties of GW signals rather than the properties of individual events; we find the set of measurements to be consistent with GR (Fig. 3).

Parametrized tests, such as the ones studied here, are powerful probes of beyond-GR effects. Yet, it has long been appreciated that their interpretation demands caution:

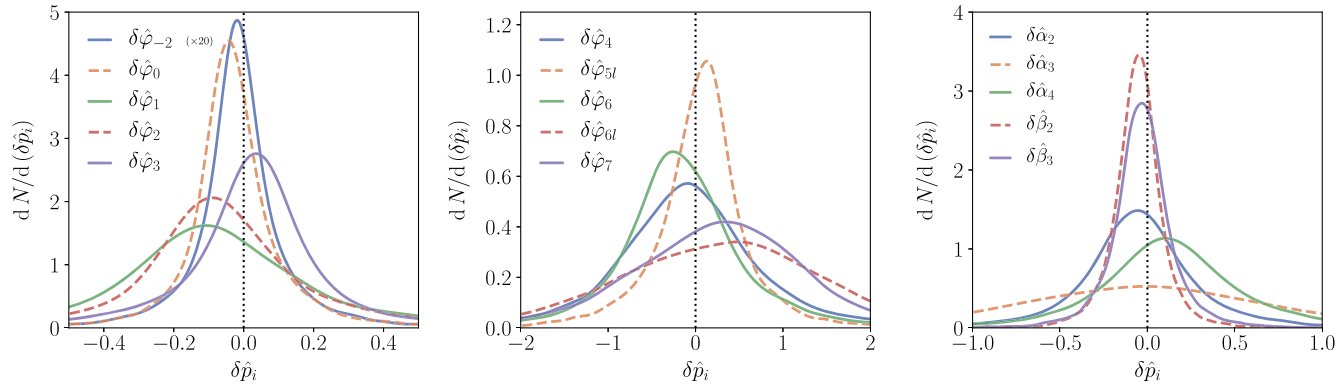

FIG. 3. The inferred population distribution $dN/d(\delta\hat{p}_i)$ for the beyond-GR parameters $\delta\hat{p}_i$'s given the 10 confirmed binary black hole signals observed to date. (The scale of $\delta\hat{\varphi}_{-2}$ has been expanded by a factor of 20.) All population distributions are consistent with $\delta\hat{p}_i = 0$, the GR prediction. With more observed signals, and under the assumption that they will obey GR, the population distributions are expected to become more narrowly centered around the origin, approximating a δ function. These distributions are defined in Eq. (5) of the Supplemental Material [21].

correlations between parameters make it necessary to have a consistent model to characterize a detected deviation. Our method provides a framework to execute a null test of GR with several detections, largely without the need for specific models of potential deviations. It improves with increasing number of signals at a rate similar to simpler approaches (Fig. 1). Furthermore, hierarchical methods could exploit degeneracies in our measurements to detect otherwise inaccessible deviations from GR, e.g., because they intrinsically occur at a higher PN order than can be directly probed (Fig. 2).

The framework presented here is not restricted to tests of GR with GWs, but can be generalized to include information from other observations. For example, the measured likelihood for $\delta\hat{\varphi}_{-2}$ from GWs could be combined with corresponding constraints from binary pulsar measurements. Our hierarchical method not only unifies the signals seen by ground-based detectors, but also offers a way to consider multiple tests of GR simultaneously.

We thank Aaron Zimmerman and Carl-Johan Haster for useful discussions. We thank Nathan Johnson-McDaniel for comments on the draft. Samples from the μ_i and σ_i posteriors were drawn with STAN [33], and plots were produced with MATPLOTLIB [34]. M. I. is supported by NASA through the NASA Hubble Fellowship Grant No. #HST–HF2–51410.001–A awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under Contract No. NAS5–26555. The Flatiron Institute is supported by the Simons Foundation. This Letter carries LIGO document number LIGO–P1900109.

*maxisi@mit.edu
†kchatziioannou@flatironinstitute.org
‡will.farr@stonybrook.edu

[1] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), arXiv:1811.12907.
[2] J. Aasi *et al.* (LIGO Scientific Collaboration), Classical Quantum Gravity **32**, 115012 (2015).
[3] F. Acernese *et al.* (Virgo Collaboration), Classical Quantum Gravity **32**, 024001 (2015).
[4] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **116**, 221101 (2016); **121**, 129902 (E) (2018).
[5] N. Yunes, K. Yagi, and F. Pretorius, Phys. Rev. D **94**, 084002 (2016).
[6] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. X **6**, 041015 (2016); **8**, 039903(E) (2018).
[7] B. P. Abbott *et al.* (LIGO Scientific and VIRGO Collaborations), Phys. Rev. Lett. **118**, 221101 (2017); **121**, 129901 (E) (2018).
[8] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **123**, 011102 (2019).
[9] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), arXiv:1903.04467.
[10] A. Zimmerman, C.-J. Haster, and K. Chatziioannou, Phys. Rev. D **99**, 124044 (2019).
[11] 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).
[12] 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, J. Phys. Conf. Ser. **363**, 012028 (2012).
[13] M. Agathos, W. Del Pozzo, T. G. F. Li, C. Van Den Broeck, J. Veitch, and S. Vitale, Phys. Rev. D **89**, 082001 (2014).
[14] J. Meidam, M. Agathos, C. Van Den Broeck, J. Veitch, and B. S. Sathyaprakash, Phys. Rev. D **90**, 064009 (2014).
[15] W. Del Pozzo, J. Veitch, and A. Vecchio, Phys. Rev. D **83**, 082002 (2011).
[16] A. Ghosh *et al.*, Phys. Rev. D **94**, 021101(R) (2016).
[17] J. Meidam *et al.*, Phys. Rev. D **97**, 044033 (2018).
[18] A. Ghosh, N. K. Johnson-McDaniel, A. Ghosh, C. K. Mishra, P. Ajith, W. Del Pozzo, C. P. L. Berry, A. B. Nielsen, and L. London, Classical Quantum Gravity **35**, 014002 (2018).
[19] R. Brito, A. Buonanno, and V. Raymond, Phys. Rev. D **98**, 084038 (2018).
[20] LIGO Scientific and Virgo Collaborations, Data release for testing GR with GWTC-1, 2019, https://dcc.ligo.org/LIGO-P1900087/public.
[21] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.123.121101 for an overview of hierarchical inference and parametrized tests of GR.
[22] T. J. Loredo, AIP Conf. Proc. **735**, 195 (2004).
[23] J. Bovy, D. W. Hogg, and S. T. Roweis, Astrophys. J. **700**, 1794 (2009).
[24] C. M. Will, Living Rev. Relativity **17**, 4 (2014).
[25] H.-Y. Chen and D. E. Holz, arXiv:1409.0522.
[26] N. Yunes and F. Pretorius, Phys. Rev. D **80**, 122003 (2009).
[27] L. Sampson, N. Cornish, and N. Yunes, Phys. Rev. D **87**, 102001 (2013).
[28] N. Cornish, L. Sampson, N. Yunes, and F. Pretorius, Phys. Rev. D **84**, 062003 (2011).
[29] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Phys. Rev. Lett. **113**, 151101 (2014).
[30] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. J. Forteza, and A. Bohé, Phys. Rev. D **93**, 044007 (2016).
[31] This is not the same as the most likely such distribution, which by construction would be a Gaussian with mean and variance given by the peak of the posterior on μ_i and σ_i. Rather, Fig. 3 *marginalizes* over μ_i and σ_i, which is why the $dN/d(\delta\hat{p}_i)$ distributions in Fig. 3 are not Gaussians. See Supplemental Material at [21] for details.
[32] S. Mirshekari, N. Yunes, and C. M. Will, Phys. Rev. D **85**, 024041 (2012).
[33] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, J. Stat. Software, Articles, **76**, 1 (2017).
[34] J. D. Hunter, Comput. Sci. Eng. **9**, 90 (2007).