Fortifying gravitational-wave tests of general relativity against astrophysical assumptions

Ethan Payne,[1,2,a] Maximiliano Isi,[3,b] Katerina Chatziioannou,[1,2,c] and Will M. Farr[3,4,d]

[1] *Department of Physics, California Institute of Technology, Pasadena, California 91125, USA*
[2] *LIGO Laboratory, California Institute of Technology, Pasadena, California 91125, USA*
[3] *Center for Computational Astrophysics, Flatiron Institute, 162 5th Ave, New York, New York 10010, USA*
[4] *Department of Physics and Astronomy, Stony Brook University, Stony Brook NY 11794, United States*

Most tests of general relativity with gravitational-wave observations rely on inferring the degree to which a signal deviates from general relativity in conjunction with the astrophysical parameters of its source, such as the component masses and spins of a compact binary. Due to features of the signal, measurements of these deviations are often highly correlated with the properties of astrophysical sources. As a consequence, prior assumptions about astrophysical parameters will generally affect the inferred magnitude of the deviations. Incorporating information about the underlying astrophysical population is necessary to avoid biases in the inference of deviations from general relativity. Current tests assume that the astrophysical population follows an unrealistic fiducial prior chosen to ease sampling of the posterior—for example, a prior flat in component masses—which is is inconsistent with both astrophysical expectations and the distribution inferred from observations. We propose a framework for fortifying tests of general relativity by simultaneously inferring the astrophysical population using a catalog of detections. Although this method applies broadly, we demonstrate it concretely on massive graviton constraints and parameterized tests of deviations to the post-Newtonian phase coefficients. Using observations from LIGO-Virgo-KAGRA's third observing run, we show that concurrent inference of the astrophysical distribution strengthens constraints and improves overall consistency with general relativity. We provide updated constraints on deviations from the theory, finding that, upon modeling the astrophysical population, the 90%-credible upper limit on the mass of the graviton improves by 25% to $m_g \leq 9.6 \times 10^{-24}\,\mathrm{eV}/c^2$ and the inferred population-level post-Newtonian deviations move $\sim 0.4\sigma$ closer to zero.

I. MOTIVATION

Gravitational-wave observations from compact binary mergers have provided a unique laboratory to test Einstein's theory of gravity in the strong-field regime [1–7]. These individual detections by the Advanced LIGO [8] and Advanced Virgo [9] detectors allow for various tests—such as inspiral-merger-ringdown consistency [10, 11], parameterized inspiral deviations [12–14], gravitational-wave dispersion [15, 16], birefringence [17, 18] and nontensorial polarizations [19–23], among many more; see Ref. [7] for recent results—to both target specific properties of general relativity (GR) as well as broadly explore its consistency with observations. Beyond analyzing events individually, the ensemble of detections can be analyzed collectively to study the possibility of deviations from GR at the population level [6, 7, 24, 25]. Hierarchical population tests rely on inferring the distribution of deviation parameters across all events and confirming that it is consistent with a globally vanishing deviation [24, 26, 27].

In this study we explore the systematic impact of astrophysical population assumptions on these studies, show that they already come into play for current catalogs due to the increasing number of detections, and offer a solution under the framework of hierarchical population modeling.

In inferences about deviations from GR, there are strong likelihood-level correlations between the deviation parameters and the astrophysical parameters of the source, such as the masses and spins of compact binaries [1, 28, 29]. Therefore, any inference of deviations from GR signals from black hole coalescences will be affected by assumptions about the distribution of binary black-hole masses and spins in the Universe—otherwise known as the astrophysical population distribution [30]. This is true at both the individual-event and catalog levels, regardless of the specific assumptions made in combining deviation parameters across events, whether the analysis is hierarchical or not. Even when astrophysical parameters do not explicitly appear in the catalog-level test of GR, assumptions about these parameters are implicitly encoded in the individual-event deviation posteriors through the prior. As the catalog of gravitational-wave observations grows and the precision of the measurements improves, these systematic effects become more important.

In presence of correlations between deviation and astrophysical parameters, we must simultaneously model the astrophysical population distribution in conjunction with testing GR. By not explicitly doing so, as has been the case in previous tests of GR [1–7, 24], the astrophysical population is typically implicitly assumed to be uni-

[a] epayne@caltech.edu
[b] misi@flatironinstitute.org
[c] kchatziioannou@caltech.edu
[d] wfarr@flatironinstitute.org

form in detector-frame masses and uniform in spin magnitude. This fiducial sampling prior is adopted to ensure broad coverage of the sampled parameter-space, and not to represent a realistic astrophysical population. In reality, the primary-black hole mass population more closely follows a decreasing power-law with an excess of sources at $\sim 35\,M_\odot$, and preferentially supports low spins [30, 31]. This mismatch can lead to biased inference regarding deviations from GR. Simultaneously modeling the astrophysical and deviation distributions will not eliminate the influence of the former on the latter, but it will ensure that this interplay is informed by the data and not arbitrarily prescribed by analysis settings.

While this insight applies to all tests of GR, for concreteness we devote our attention to constraints on the mass of the graviton [15, 16] and deviations in parameterized post-Newtonian (PN) coefficients [12, 13, 32–35]. A massive graviton would affect the propagation of a gravitational wave over cosmological distances; this leads to a frequency-dependent dephasing of the gravitational wave which is related to the mass of the graviton, m_g, and the propagated distance. The PN formalism describes the Fourier-domain phase of an inspiral signal under the stationary phase approximation through an expansion in the orbital velocity of the binary system; each $k/2$ PN expansion order can then be modified by a deviation parameter, $\delta\varphi_k$, which vanishes in GR. See App. A for further details about both calculations. We focus on these tests as they target the signal inspiral phase, which also primarily informs astrophysical parameters such as masses and spins; we leave other tests [5–7, 10, 11, 15, 16, 19–23] to future work.

As motivation, Fig. 1 shows how inference on the 0PN coefficient of a real event (GW191216_213338) depends on astrophysical assumptions. This figure compares measurements with (blue) and without (red) a simultaneous measurement of the population of black hole masses and spins (see Sec. II). The observed binary black-hole population shows a preference for systems with comparable masses; as a consequence of the strong correlation between the 0PN deviation coefficient and the mass ratio of GW191216_213338, this preference then "pulls" the system towards more equal masses and a more negative deviation coefficient. This is a direct manifestation of the fact that tests of GR are contingent on our astrophysical assumptions. Higher PN orders are expected to display similar correlations as in Fig. 1 with these and other parameters. For example, spins are known to be correlated with the coupling constant of dynamical Chern-Simons gravity which modifies the phase at the 2PN order [36–39]. While we have constructed the posterior informed results here, it is more robust to simultaneously infer the astrophysical population while also testing GR. Fixing the prior to one astrophysical population realization or marginalizing over possible distributions from other analyses will not capture any correlated structure between the inferred deviation parameters and the astrophysical distributions. The above example serves only to illustrate the impact of the arbitrary choices previously made.

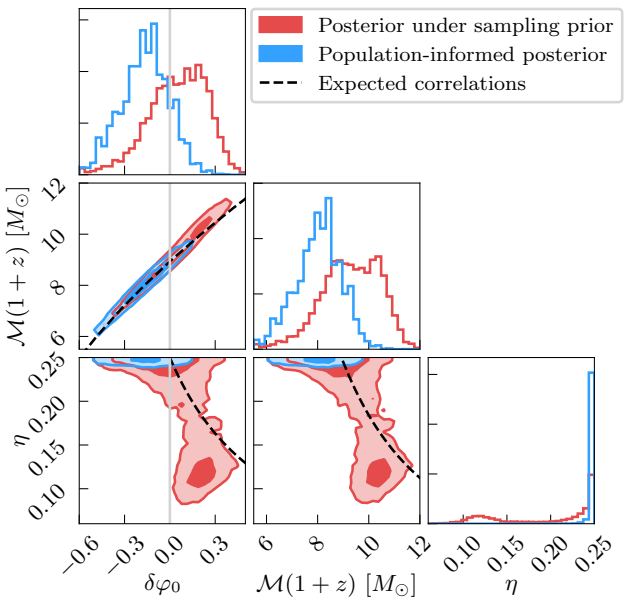

FIG. 1. Posterior distributions for the 0PN deviation coefficient $\delta\varphi_0$, detector-frame chirp mass $\mathcal{M}(1+z)$, and symmetric mass ratio η for the gravitational-wave event GW191216_213338 [6, 40], as inferred by a modified SEOBNRv4 waveform [14, 41–44]. Posteriors are conditioned on two different astrophysical assumptions: the broad prior used during parameter estimation (red), and the astrophysical population inferred by the data using the model in Sec. II B (blue). The black dashed curves show the expected correlation (App. B). Due to the correlations between astrophysical and deviation parameters, different astrophysical populations lead to different posteriors for $\delta\varphi_0$.

The remainder of the manuscript focuses on combining information from many observations to simultaneously infer the astrophysical population while testing GR; it is structured as follows. We first introduce our hierarchical analysis framework, as well as astrophysical and GR deviation models, in Sec. II. We then demonstrate the impact of incorporating astrophysical information by constraining the graviton mass and inferring the PN deviation properties with an ensemble of gravitational-wave observations in Sec. III. We analyze events from LIGO-Virgo-KAGRA (LVK)'s third observing run with individual-event results from Ref. [7] (the posterior samples are available in Ref. [45]) — a subset of the events in GWTC-3 [46]. The simultaneous modeling of the astrophysical population while testing GR tightens the graviton mass upper limit by 25%, and improves consistency with GR on the PN coefficients by $\sim 0.4\sigma$, when using a modified SEOBNRv4 waveform [14, 41–44]. Finally, we conclude in Sec. IV, where we summarize the case for jointly modeling the astrophysical population when testing GR in order to avoid biases and hidden assumptions, and comment on how the same is true for gravitational-wave studies of cosmology or nuclear matter.

II. POPULATION ANALYSES

In this section, we introduce the fundamentals of inferring a population distribution from individual observations and discuss the population models we employ. We also outline the implementation and importance of observational selection effects in accounting for the events used within the analysis.

A. Preliminaries

We infer the astrophysical population distribution and deviations from GR (see Refs. [47–49] for a discussion of hierarchical inference in the context gravitational-wave astronomy). This framework has already been extensively applied to tests of GR and astrophysical population inference separately [6, 7, 24, 25, 30, 31, 50–66]. Here we focus on combining both methods to jointly infer the astrophysical population while testing GR.

Our approach is based on a *population likelihood*, $p(\{d\}|\Lambda)$, for the ensemble of observations, $\{d\}$, given population hyperparameters, $\Lambda = \{\Lambda_{\rm astro}, \Lambda_{\rm nGR}\}$. We separate the hyperparameters into the parameters describing the astrophysical population distribution, $\Lambda_{\rm astro}$, and parameters describing the deviation to GR, $\Lambda_{\rm nGR}$. The hyperparameters encode the shape of the population distribution, $\pi(\theta|\Lambda)$, where θ are parameters of a single event; we describe our population models in the following subsections. This hierarchical approach allows us to test GR while concurrently inferring the astrophysical population from the data. Given the likelihoods of individual events, $p(d_i|\theta_i)$, the population likelihood is

$$p(\{d\}|\Lambda) = \frac{1}{\xi(\Lambda)^N} \prod_{i=1}^{N} \int {\rm d}\theta_i \, p(d_i|\theta_i) \, \pi(\theta_i|\Lambda) \,, \tag{1}$$

where d_i and θ_i are respectively the data and parameters for the ith event, and $\{d\}$ is the collection of data for the ensemble of N observations[1]. We address the technical aspects of the likelihood calculation in App. C.

In Eq. (1), $\xi(\Lambda)$ is the detectable fraction of observations given a set of population hyperparameters and accounts for selection biases [47]. It is defined as

$$\xi(\Lambda) = \int {\rm d}\theta \, p_{\rm det}(\theta) \, \pi(\theta|\Lambda) \,. \tag{2}$$

Here $p_{\rm det}(\theta)$ is the probability of detecting a binary black-hole system with parameters θ. The selection factor in Eq. (2) accounts for both the intrinsic selection bias of a gravitational-wave detector (e.g., heavier binaries are more detectable), as well as selection thresholds used when deciding which gravitational-wave events to analyze. The detected fraction can also be framed as a "normalizing factor," which relaxes the need for normalizable population distributions (so long as the integrals in Eqs. 1 and 2 are finite) [67]. This correction will become important in Sec. II C when discussing the selection criteria for events to be included in the analysis.

In theory, the selection factor should account for the effect of both astrophysical and deviation parameters. However, we ignore the latter here, the effect of which is subject of ongoing research [68]. For the former, we compute the detectable fraction, $\xi(\Lambda)$, from a set of recovered injections,

$$\xi(\Lambda) = \frac{1}{N_{\rm inj}} \sum_{i=1}^{N_{\rm rec}} \frac{\pi(\theta_i|\Lambda)}{\pi_{\rm draw}(\theta_i)} \,, \tag{3}$$

where $N_{\rm inj}$ is the number of injected signals, $N_{\rm rec}$ is the number of recovered signals, and $\pi_{\rm draw}(\theta_i)$ is the distribution from which the injected signals were drawn (for more details see Refs. [30, 31, 47–50]). The subset of injected signals that are recovered is determined by the particular thresholds used to determine which gravitational-wave observations to use within the hierarchical analysis. To avoid biases, the criteria on the threshold for the detectable fraction calculation must match that of the observed signals. We address the specifics of the relevant criteria for our analysis in Sec. II C.

Finally, Eq. (1) explicitly shows the need for jointly modeling the astrophysical population when testing GR. While the astrophysical population may be separable from the deviation distribution so that $\pi(\theta|\Lambda) = \pi(\theta_{\rm astro}|\Lambda_{\rm astro}) \, \pi(\theta_{\rm nGR}|\Lambda_{\rm nGR})$, this factorization cannot be undertaken for individual event likelihoods, as the deviations are often correlated with astrophysics (see Fig. 1), i.e. $p(\{d_i\}|\theta) \neq p(\{d\}|\theta_{\rm nGR}) \, p(\{d\}|\theta_{\rm astro})$. Therefore, the integrals of Eq. (1) do not separate and tests of GR cannot be undertaken in isolation from the astrophysics.

From the hyperposterior distribution on the population parameters, we can construct the individual event population-informed posteriors following Refs. [69–71] (and references therein). Such distributions represent our best inference about the properties of a given event in the context of the entire catalog of observed signals. These calculations are subtle as they avoid "double-counting" the gravitational-wave events which also used to infer the population distribution.

B. Population models

In this subsection, we outline the population models for both the GR deviations and the astrophysical population. While many astrophysical population models have been

[1] Equation (1) assumes a prior on the rate of observations as $\pi(R) \propto 1/R$, which was analytically marginalized [62].

proposed [30, 31, 50–66] as a product of the increasing number of observations [30, 46], in this work we restrict ourselves to standard parameterized models motivated by previous analyses.

1. GR deviation population models

There are two typical approaches to combining posteriors on GR deviation parameters obtained from different gravitational-wave observations, each stemming from different assumptions behind the deviations (see, e.g., discussions in[6, 7]). The first, more general approach is to assume that the population describing deviations from GR is, to the lowest order, a Gaussian distribution with a mean, μ, and standard deviation, σ [24, 26]. In the limit that all observations are consistent with GR, $(\mu, \sigma) \to (0, 0)$ and the inferred distribution approaches a Dirac delta function at the origin. Since a Gaussian distribution encapsulates the lowest order moments of more complicated distributions, given enough events any deviation from a delta function at the origin will be identified as a violation of GR, even if the exact shape of the deviation distribution is not captured by a Gaussian [24, 27]. This approach is now routinely applied to post-Newtonian deviations tests, inspiral-merger-ringdown consistency tests and ringdown analyses [6, 7, 24], but it can be naturally extended to any analysis that recovers GR in the limit of some vanishing parameter. This method provides a null test in cases where the exact nature of the deviation is unknown.

The second approach assumes all observations share the same value of the deviation parameter [5, 12–14, 44, 72–76]. This is the limiting case of the aforementioned Gaussian model when $\sigma \to 0$. This model (in the absence of astrophysical information) is equivalent to simply multiplying the marginal likelihoods of the deviation parameter obtained from the individual events. The assumption of a shared parameter is only suitable in the context of specific theories or models, in which case the expected degree of deviation for each event can be predicted exactly as a function system specific parameters (e.g., BH masses and spins) and universal, theory-specific parameters (e.g., coupling constants), the second of which can be measured jointly from a catalog of detections by multiplying likelihoods. In practice, the lack of complete waveform models beyond GR means that this approach has so far only been well-suited for measurements such as the mass of the graviton, and features of the propagation of gravitational waves whose observational signatures are independent of specific source properties by construction [5–7].

2. Astrophysical population models

Following Refs. [30, 31, 50], we model the primary black-hole mass (m_1) distribution as a power-law whose slope is given by an index α, with a sharp cut-off governed by the minimum mass, $m_{\rm min}$, and a higher-mass Gaussian peak,

$$\pi(m_1|\Lambda) = (1 - f_{\rm peak})\,\mathcal{P}[\alpha, m_{\rm min}](m_1) + f_{\rm peak}\,\mathcal{N}[\mu_{\rm peak}, \sigma^2_{\rm peak}](m_1)\,. \quad (4)$$

Here, $f_{\rm peak}$ is the fraction of binaries in the Gaussian peak, the powerlaw is given by

$$\mathcal{P}[\alpha, m_{\rm min}](m_1) \propto \begin{cases} m_1^{-\alpha}, & m_1 \geq m_{\rm min} \\ 0, & m_1 < m_{\rm min}\,, \end{cases} \quad (5)$$

and $\mathcal{N}[\mu, \sigma^2](x)$ is the probability density function for a Gaussian with mean μ and variance σ^2. We fix $m_{\rm min} = 5\,M_\odot$ for simplicity. Unlike other studies [30, 31, 53], we do not infer much structure in the Gaussian peak as higher mass features become unresolvable when looking at the light binary systems that provide constraints of PN coefficients (see Sec. II C).

We parameterize the distribution of mass ratios, $q \equiv m_2/m_1$, as a conditional power-law, with index β, and a sharp cut-off imposed by $m_{\rm min}$, such that

$$\pi(q|m_1; \Lambda) \propto \begin{cases} q^\beta, & 1 \geq q \geq m_{\rm min}/m_1 \\ 0, & q \leq m_{\rm min}/m_1\,. \end{cases} \quad (6)$$

Here β can take any value without leading to a singularity due to the lower bound on the mass ratio.

We adopt a truncated Gaussian population model for the component spins with a mean, μ_χ, and standard deviation, σ_χ, bounded between zero and one, assuming both spins are drawn independently from the same population distribution. This differs from standard Beta distribution utilized in many recent analyses [30, 31, 50, 54, 77]. as it allows for non-zero support at the edges of the spin-magnitude domain [78]. Furthermore, adopting a Gaussian model allows for efficient computation of the population likelihood via analytic integration (see App. C). For individual-event analyses where the spins are assumed to be aligned with the orbital angular momentum (as is the case for posteriors using a modified SEOBNRv4 waveform [14, 41–44]), this model treats the measured spin along the orbital angular momentum as the total spin magnitude.

For analyses where the individual event inferences also possess information about the spin-precession degrees of freedom, we adopt a model for the spin tilts, $\cos\theta_{1/2}$, whereby the population is parameterized as a mixture of isotropically distributed and preferentially aligned spins [54],

$$\pi(\cos\theta_1, \cos\theta_2|\Lambda) = \frac{f_{\rm iso}}{4} + (1 - f_{\rm iso})\times \mathcal{N}[1, \sigma^2_\theta](\cos\theta_1)\,\mathcal{N}[1, \sigma^2_\theta](\cos\theta_2)\,, \quad (7)$$

where $f_{\rm iso}$ is the mixing fraction, and σ_θ is the standard deviation of the preferentially aligned Gaussian component. This model is only relevant for analyses with precessing spins. In this manuscript, this includes the massive graviton constraints (Sec. III A), and PN deviation

tests with the IMRPhenomPv2 [34, 35, 79] waveform (App. D).

Finally, we also adopt a power-law model for the merger-rate density as a function of redshift [62],

$$\pi(z|\Lambda) \propto \frac{1}{1+z}\frac{\mathrm{d}V_c}{\mathrm{d}z}(1+z)^{\lambda}\,, \tag{8}$$

where $\mathrm{d}V_c/\mathrm{d}z$ denotes the evolution of the comoving volume with redshift, and λ is the power-law index. When $\lambda = 0$, the binary black-hole population is uniformly distributed within the source-frame comoving volume.

C. Selection criteria and observations

We limit ourselves to binary black-hole observations made during LIGO-Virgo-KAGRA's third observing run [46] with false-alarm-rates of less than 10^{-3} per year[2]. This mirrors the selection criteria chosen for the tests of GR within Refs. [5–7], and therefore we do need not reanalyze any individual gravitational-wave observations [45, 80]. The events that pass these criteria are listed in Table IV of Ref. [6] and Table V of Ref. [7]. In future studies, the false-alarm-rate threshold could be raised to increase the number of included gravitational-wave events. This would likely improve inference of the astrophysical population and GR deviation constraints due to the larger catalog of observations. In our analyses, we exclude GW190814 [81] as it is an outlier from the binary black-hole population [31] and GW200115_042309 since it is a black hole-neutron star merger [82]. It is straightforward to extend this analysis to additionally incorporate binary neutron star and neutron star-black hole mergers by adopting a mixture model of the different source classifications (see Ref. [30] for one example). We then use all events except GW200316_215756[3] when inferring the mass of the graviton, mirroring the analysis in Ref. [7]. When constraining the PN deviation coefficients, we include the additional criterion that signal-to-noise ratios (SNRs) during the binaries' inspiral must be greater than 6, again mirroring previous analyses [6, 7].

We use posteriors for the graviton's mass inferred using a modified IMRPhenomPv2 [34, 35, 79] waveform, whereas we use both modified SEOBNRv4 [14, 41–44] (for results in Sec. III B) and modified IMRPhenomPv2 [12, 13, 34, 35, 73, 76, 79] (for results in App. D)[4] waveform models when inferring the PN deviations. We summarize these events and their relevant properties in Tab. I. We do not include gravitational-wave events from the first and second LIGO-Virgo observing runs, as a semi-analytic approximation was used to estimate the sensitivity of the detector network during that time [5]. This approximation does not compute a false-alarm rate and therefore cannot be unambiguously incorporated into this methodology.

As described in Sec. II, selection effects are estimated through an injection campaign. While we know the total network SNR of the individual injections, part of our selection criteria is based on the inspiral network SNR. We approximate the inspiral SNR from the total SNR by constructing a linear fit to their ratio as a function of detector-frame total mass (Fig. 2). This fit is constructed by inferring the slope and offset of the line, as well as the uncertainty on the data points. We assume identical uncertainties on all SNR ratios, and marginalize over this parameter to fit the line. We validate this approximation by computing the detection probability $p_{\mathrm{det}}(\theta)$ with different draws of the linear fit. We find that different realizations of the approximation do not change the detection probability, and so we consider this approximation to be sufficiently accurate for our purposes. Future injection campaigns may also opt to compute the inspiral SNR directly.

III. RESULTS

In this section we simultaneously infer the astrophysical population while testing GR and quantify the impact of fixing the population distribution to the sampling prior. Throughout, we use the nomenclature "fixed" and "inferred" to refer to whether the analysis uses the fixed sampling prior or infers the distribution from data, respectively. We implement the analyses using NumPyro [84, 85] and JAX [86], leveraging AstroPy [87–89] and SciPy [90] for additional calculations, and matplotlib [91], arViz [92] and corner [93] for plotting purposes. The code for the hierarchical tests is available in Ref. [94].

A. Massive graviton constraints

We begin by demonstrating that astrophysical assumptions are crucial even in the simplest scenarios, where a global deviation parameter is shared across events. This is the case for the mass of the graviton, m_g [15, 16] (see App. A 1), for which we produce an updated upper limit by simultaneously inferring the astrophysical distribution.

We combine results from individual-event likelihoods under the assumption of a shared deviation parameter as described in Sec. II B. In practice, we compute this as the

[2] For comparison, the population analyses presented Ref. [30] used a false alarm rate threshold of 1 per year. A more stringent false-alarm-rate threshold is often adopted when testing GR to avoid contaminating from false detections.

[3] GW200316_215756 was excluded from propagation tests within Ref. [7] due to poor sampling convergence.

[4] Single-event results with IMRPhenomPv2 were only produced during the first half of the third observing run [6, 7].

TABLE I. Observations from the LIGO-Virgo-KAGRA's third observing run that pass our selection criteria [6, 7, 46, 83]. The different columns outline the gravitational-wave event, the detector-frame chirp mass, the total and inspiral *maximum a posteriori* SNRs ($\rho_{\rm tot}$ and $\rho_{\rm insp}$ respectively), and whether it was included in the graviton constraint calculation (m_g) or the post-Newtonian deviation tests (PN). Horizontal lines split events from the two halves of the third observing period. While we use all events marked under "PN" in Sec. III B, we are limited to the first half of observing run when using IMRPhenomPv2 posterior samples in App. D.

Event	$(1+z)\mathcal{M}$ $[M_\odot]$	$\rho_{\rm tot}$	$\rho_{\rm insp}$	m_g	PN
GW190408_181802	$23.7^{+1.4}_{-1.7}$	15.0	8.3	✓	✓
GW190412	$30.1^{+4.7}_{-5.1}$	19.1	15.1	✓	✓
GW190421_213856	$46.6^{+6.6}_{-6.0}$	10.4	2.9	✓	-
GW190503_185404	$38.6^{+5.3}_{-6.0}$	13.7	4.3	✓	-
GW190512_180714	$18.6^{+0.9}_{-0.8}$	12.8	10.5	✓	✓
GW190513_205428	$29.5^{+5.6}_{-2.5}$	13.3	5.1	✓	-
GW190517_055101	$35.9^{+4.0}_{-3.4}$	11.1	3.4	✓	-
GW190519_153544	$65.1^{+7.7}_{-10.3}$	15.0	0.0	✓	-
GW190521_074359	$39.8^{+2.2}_{-3.0}$	25.4	9.7	✓	✓
GW190602_175927	$72.9^{+10.8}_{-13.7}$	13.1	0.0	✓	-
GW190630_185205	$29.4^{+1.6}_{-1.5}$	16.3	8.1	✓	✓
GW170706_222641	$75.1^{+11.0}_{-17.5}$	12.7	0.0	✓	-
GW190707_093326	$9.89^{+0.1}_{-0.09}$	13.4	12.2	✓	✓
GW190708_232457	$15.5^{+0.3}_{-0.2}$	13.7	11.1	✓	✓
GW190720_000836	$10.4^{+0.2}_{-0.1}$	10.5	9.2	✓	✓
GW170727_060333	$44.7^{+5.3}_{-5.7}$	12.3	2.0	✓	-
GW190728_064510	$10.1^{+0.09}_{-0.08}$	12.6	11.4	✓	✓
GW190828_063405	$34.5^{+2.9}_{-2.8}$	16.2	6.0	✓	✓
GW190828_065509	$17.4^{+0.6}_{-0.7}$	9.9	6.3	✓	✓
GW190910_112807	$43.9^{+4.6}_{-3.6}$	14.4	3.3	✓	-
GW190915_235702	$33.1^{+3.3}_{-3.9}$	13.1	3.7	✓	-
GW190924_021846	$6.44^{+0.04}_{-0.03}$	12.2	11.8	✓	✓
GW191129_134029	$8.49^{+0.06}_{-0.05}$	14.1	12.8	✓	✓
GW191204_171526	$9.70^{+0.05}_{-0.05}$	18.0	16.3	✓	✓
GW191215_223052	$24.9^{+1.5}_{-1.4}$	10.6	5.5	✓	-
GW191216_213338	$8.94^{+0.05}_{-0.05}$	17.9	15.6	✓	✓
GW191222_033537	$51.0^{+7.2}_{-6.5}$	13.1	3.1	✓	-
GW200129_065458	$32.1^{+1.8}_{-2.6}$	25.7	10.4	✓	✓
GW200202_154313	$8.15^{+0.05}_{-0.05}$	11.1	10.5	✓	✓
GW200208_130117	$38.8^{+5.2}_{-4.8}$	9.9	3.0	✓	-
GW200219_094415	$43.7^{+6.3}_{-6.2}$	11.2	2.8	✓	-
GW200224_222234	$40.9^{+3.5}_{-3.8}$	19.4	4.7	✓	-
GW200225_060421	$17.7^{+1.0}_{-2.0}$	12.9	6.8	✓	✓
GW200311_115853	$32.7^{+2.7}_{-2.8}$	17.5	6.5	✓	✓
GW200316_215756	$10.7^{+0.1}_{-0.1}$	11.5	10.7	-	✓

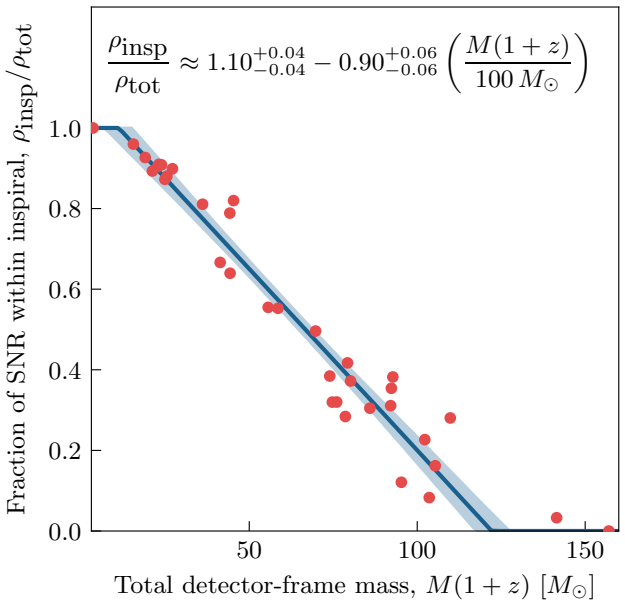

FIG. 2. Ratio between the network *maximum a posteriori* gravitational-wave inspiral and the total SNRs as a function of detector-frame total mass, $M(1+z) \equiv (m_1+m_2)(1+z)$, for all gravitational-wave observations in the LIGO-Virgo-KAGRA third observing run [6, 7, 46, 83] with a false-alarm rate less than 10^{-3}/yr. The solid blue line is the median best-fit line to the observations, with the band representing the 90%-credible uncertainty. While computing this fit, we also estimate the uncertainty in the individual data points. We use this fit to compute the inspiral SNR for the injections used to estimate the detection probability, $p_{\rm det}(\theta)$, as described in Sec. II C.

limit of a vanishing standard deviation of the hierarchical analysis described in Sec. II. For technical reasons, we assume a uniform prior distribution on $\log_{10}(m_g)$ when combining observations, which differs from Refs. [5–7] which applied a uniform prior prior on m_g itself; this is to avoid poor convergence when reweighting between individual-event posterior distributions. In the end, we reweight the shared graviton mass inference to a uniform prior to report upper limits on m_g. We compare this to results obtained assuming the sampling prior for the astrophysical parameters.

The one-dimensional marginal distributions of the shared mass of the graviton are shown in Fig. 3. The inclusion of astrophysical information changes the inferred distributions of the graviton's mass increasing support for $m_g = 0$. When using the sampling prior for the astrophysical population (and thereby assuming the incorrect distribution), the graviton's mass is constrained to be $m_g \leq 1.3\times10^{-23}\,{\rm eV}/c^2$ at the 90% level[5]; however, upon inferring the astrophysical population the graviton's mass

[5] This constraint differs from the 90% upper limit of 1.27×10^{-23} eV/c^2 calculated in Ref. [7], which is determined by addition-

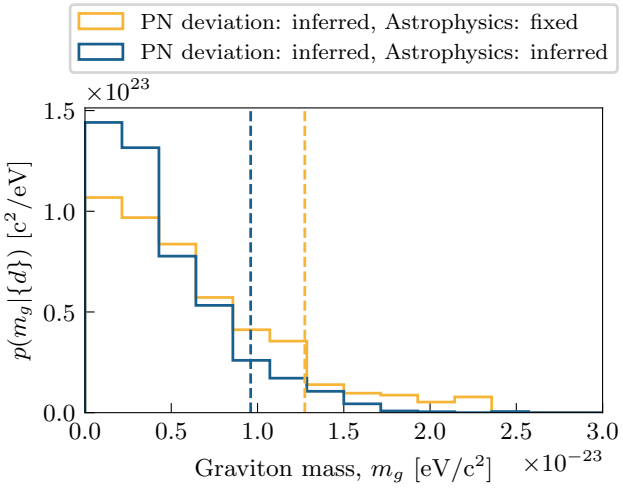

FIG. 3. Marginal one-dimensional posterior distributions for the mass of a massive graviton. In practice, we compute the shared value of graviton mass by assuming a shared deviation parameter $\log_{10}(m_g c^2/\mathrm{eV})$ then reweighting to a uniform graviton mass prior. The dashed lines correspond to the 90% upper limits from the two analyses. We compare the result when astrophysical information is not included, equivalent to multiplying individual event likelihood functions (yellow), to also modeling the astrophysical population (dark blue). The result shifts towards smaller values of m_g if simultaneously modelling the astrophysical population and the graviton's mass.

becomes more constrained, with $m_g \leq 9.6 \times 10^{-24}\,\mathrm{eV}/c^2$ at the 90% credible level. Under the expectation that GR is correct and $m_g = 0$, a reduced constraint is generically expected as we have included the correct information regarding the astrophysical population. This highlights the effect of unreasonable astrophysical assumptions, which are inconsistent with the observed population, on tests of GR.

B. Hierarchical post-Newtonian deviation constraints from SEOBNRv4

We repeat the population analysis, this time measuring the hierarchical PN deviation distribution with a mean, μ_{PN}, and standard deviation, σ_{PN}, for all PN orders. This is corresponds to ten separate analyses where only one PN deviation coefficient is allowed to vary. To compare with the default approach (which implicitly assumes a flat-in-detector-frame mass, uniform mass ratio, uniform spin-magnitude aligned spin, and comoving volume redshift distributions), we also fit the GR deviation in isolation under the assumption of the (astrophysically unrealistic) sampling prior [6, 7].

Figure 4 shows the two-dimensional posterior distribution of the deviation hyperparameters for −1 through to 3.5 PN orders. The standard results implicitly using the sampling prior are shown in yellow, while the results from the simultaneous modeling of the astrophysical and deviation populations are shown in dark blue. When concurrently modeling the astrophysical distribution, in all PN deviation parameters the inferred mean resides closer to zero, i.e., the expected value from GR, while there is no clear trend in σ_{PN}. Overall, $(\mu_{\mathrm{PN}}, \sigma_{\mathrm{PN}}) = (0, 0)$ is retained with greater significance for almost all PN orders.

We quantify this improvement by comparing the two-dimensional credible level[6] at which the expected GR value, $(\mu_{\mathrm{PN}}, \sigma_{\mathrm{PN}}) = (0, 0)$, resides in Fig. 5. A lower value for the credible region implies that the value of hyperparameters expected from GR resides closer to the bulk of the distribution. In all but one PN order, jointly inferring the astrophysical and PN deviation distributions moves the inferred distribution to be more consistent with GR. For the 0.5PN deviation coefficient, $\delta\varphi_1$, there is little change in the credible level at which GR is recovered. Generally, inference of the astrophysical population allows our inferences of GR deviations to be more consistent with GR, with an average improvement of 0.4σ.

To shed further light on the interaction between the GR and astrophysics parameters, we focus on two specific deviation parameters. In particular, we draw attention to the 3PN coefficient (which shows the largest tightening of the supported hyperparameter space in Fig. 4) and the 0PN coefficient (where the PN deviation is most inconsistent with GR in Fig. 5).

1. *Example: 3PN deviation coefficient, $\delta\varphi_6$*

To understand the origin of the improved measurement for $\delta\varphi_6$ when modeling astrophysics in Fig. 4, we show an expanded corner plot in Fig. 6 with an additional subset of the hyperparameter posterior distributions. The top left corner reproduces the corresponding panel in Fig. 4, wherein the yellow posterior distribution is obtained under the assumption of the astrophysical population given by the sampling priors, while the dark blue is obtained by simultaneously inferring the astrophysical-population and the GR deviation parameters.

ally incorporating observations from the first and second LIGO-Virgo-KAGRA observing periods [5, 95]. We do not include these observations due to the ambiguity in the detector network sensitivity during these periods.

[6] This "displacement" is the quantile, $\mathcal{Q}_{\mathrm{GR}}$, reported in Refs. [6, 7] as $(\mathrm{displacement})^2 = -2\ln(1-\mathcal{Q}_{\mathrm{GR}})\,\sigma^2$. The quantile is computed by integrating over all regions of the hyperposterior distribution which are at a higher probability than $(\mu_{\mathrm{PN}}, \sigma_{\mathrm{PN}}) = (0, 0)$. We report values in terms of the standard deviation in two dimensions, 1σ and 2σ correspond to ∼39.3% and ∼86.5% credibility, respectively.

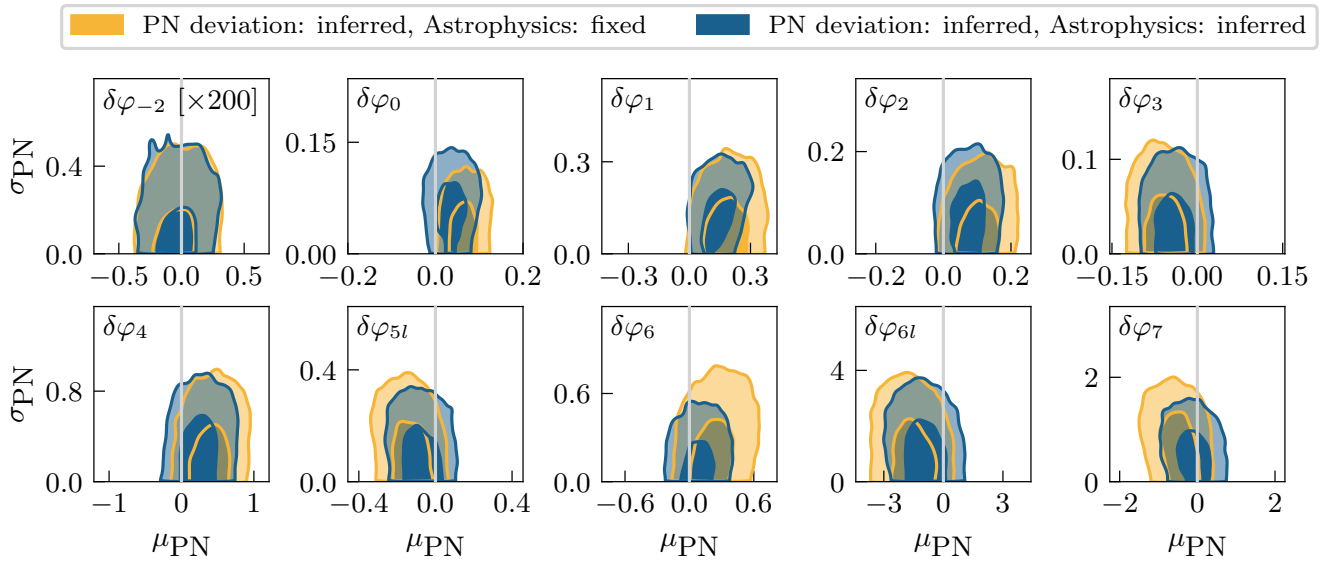

FIG. 4. Two-dimensional marginal posterior distributions for the hyperparameters of the Gaussian PN deviation distribution informed by the 20 events in the third LIGO-Virgo-KAGRA observing run passing the selection criteria, analysed with a modified SEOBNRv4 [14, 41–44] waveform. The contours indicate the 50% and 90% credible regions. Each panel corresponds to a separate analysis where the coefficient varied was at a different PN order. The analysis was undertaken with an implicitly assumed, astrophysically-unrealistic population (yellow), and a model which simultaneously infers the astrophysical population model (dark blue). Modelling both the astrophysical population and the PN deviation population systematically shifts the inferred mean, $\mu_{\rm PN}$, closer to zero.

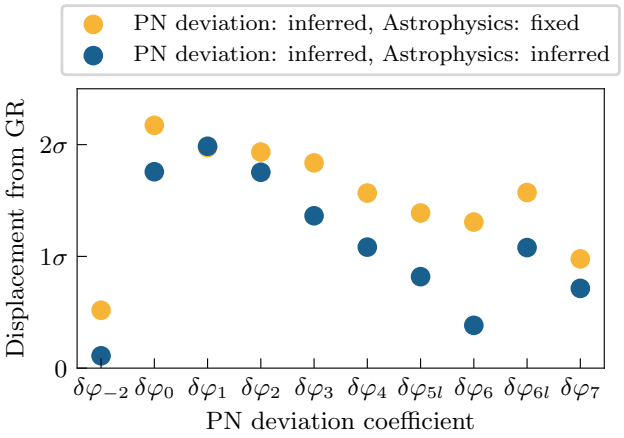

FIG. 5. Displacement of the deviation parameter distribution from GR for each PN deviation coefficient. The displacement corresponds to the credible levels at which the hyperparameter values corresponding to GR, $(\mu_{\rm PN}, \sigma_{\rm PN}) = (0, 0)$, reside for two different models as shown in Fig. 4. This quantity is indicative of the relative position of the posterior to the GR value. Incorporating the astrophysical population as well as the hierarchical model for the PN deviation leads to an inferred result more consistent with GR for most cases.

Additionally, we use the same set of individual-event posterior samples to *separately* infer the astrophysical population independently of the PN deviation parameters, which amounts to assuming a uniform distribution of deviations across events (solid green). This differs from standard astrophysical population inference, which assumes that GR is correct *a priori* and thus starts from individual-event posteriors conditioned on $\delta\varphi = 0$ [30, 31, 50]. Finally, we also compute the astrophysical population under the assumption that GR is correct, $(\mu_{\rm PN}, \sigma_{\rm PN}) = (0, 0)$ (dashed green). The result assuming GR is correct is computed by fixing $(\mu_{\rm PN}, \sigma_{\rm PN}) \to (0, 0)$ to ensure equivalent samples are used between analyses, and is consistent with the usual population inference modulo model choices at the individual-event and population levels [30, 31, 50].

From the two-dimensional marginal distributions, the most apparent feature is that inferring the astrophysical population under the assumption of a broad uniform GR deviation population (shown in solid green) leads to inferences consistent with broad spin populations (large σ_{χ_z}) and populations favoring uneven mass ratios ($\beta < 0$). This can be straightforwardly explained by the presence of correlated structure between $\delta\varphi_6$, mass ratio, and the component spins at the individual-event level.

To demonstrate this, Fig. 7 shows four different posteriors for GW191216_213338 under different priors. The four distributions shown are the posterior obtained with the sampling priors (red), the one informed by the GR deviation population only analysis (yellow), the one informed by the astrophysical population only analysis (green), and the one informed by the jointly-inferred GR

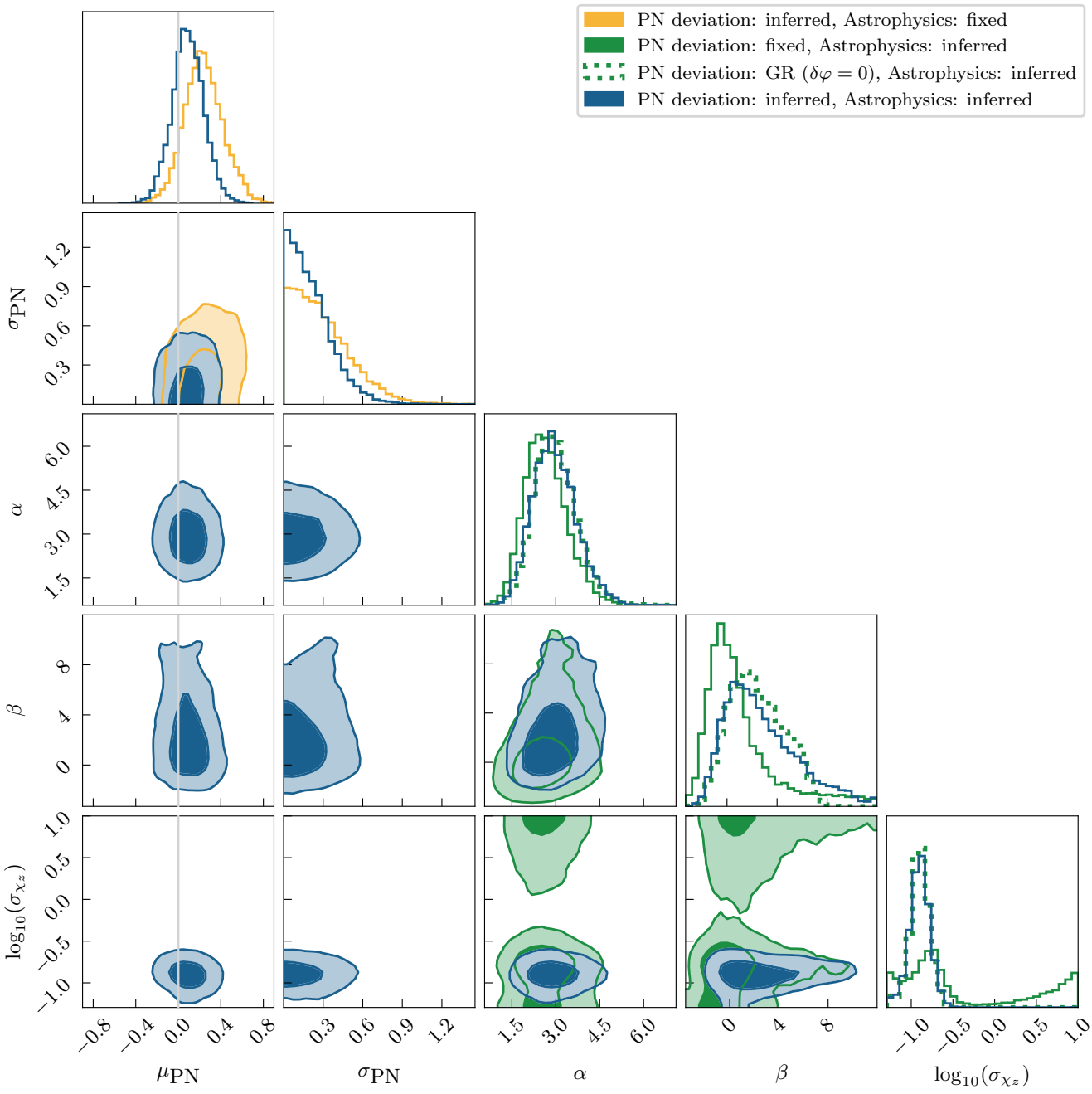

FIG. 6. Marginal one- and two-dimensional posterior distributions for the $\delta\varphi_6$ PN deviation and a subset of astrophysical population hyperparameters. Contours correspond to the 50% and 90% credible regions. Results from four analyses are shown — population inference using the PN deviation population only with the "default" sampling prior astrophysical population (yellow), astrophysical population only (green), astrophysical population under the assumption that GR is correct (dashed green), and the joint analysis inferring the post-Newtonian deviation and astrophysical populations simultaneously (dark blue). No strong correlations exist between either the mean or standard deviation of the deviation Gaussian and astrophysical population parameters. The starkest difference is that inferring the population when the PN deviation population is ignored leads to broad spin magnitude populations.

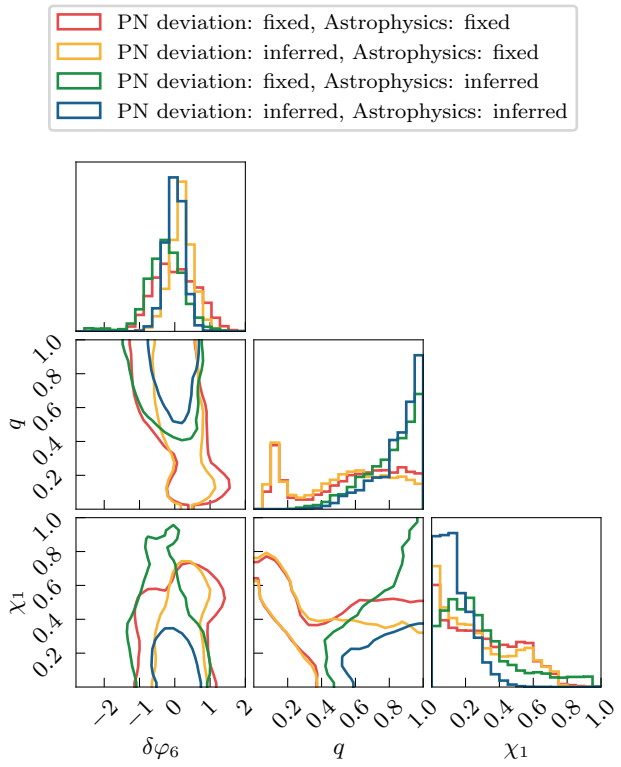

FIG. 7. One- and two-dimensional posterior distributions for the 3PN deviation parameter, the mass ratio, and the primary black-hole spin for GW191216_213338 under four different assumptions: broad sampling priors (red), informed by the GR deviation population analysis (yellow), informed by the astrophysical population (green), informed by the joint inference of PN deviation and astrophysical populations (dark blue). Contours indicate the 90% credible region. Evidence for both a low mass ratio and larger primary spins is strongly contingent upon the astrophysical assumptions. Broad priors such as those used while sampling the posterior distribution have significant support for lower mass ratios. Inclusion of information from both the deviation population and the astrophysics leads to an inferred result with both low primary spin and high mass ratio.

deviation and astrophysical populations (blue). The posteriors which involve information from inferred populations are computed following Ref. [69], and do not double-count the data from GW191216_213338, as discussed in Sec. II A.

Under the sampling astrophysical prior, posteriors exhibit a low-q, high-χ_1 mode. Since the inferred astrophysical population is inconsistent with low mass ratios and high spin magnitudes, the astrophysical-population-informed posteriors have reduced support for unequal masses (compare the red contour to the green one). Additionally incorporating the GR deviation information (blue), the population-informed posterior further reduces support for high-spinning systems. The similarity of the results under the sampling prior (red) with those in which only the GR deviation population is inferred (yellow), suggests that inferring small GR deviations is on its own not enough to significantly affect the inference of the astrophysical parameters in this case.

The tightening of the σ_{χ_z} hyperposterior distribution (i.e., inferring a more narrow spin population) when jointly inferring the GR deviation and astrophysical populations is precisely what we observe at the population level in Fig. 6 comparing the dark blue and green contours. Additionally, when enforcing that $\delta\varphi_6 = 0$ for all events (dashed green), we no longer recover support for broad spin populations. Interestingly, the astrophysical population inferred jointly with the GR deviation population is very similar to the result obtained when fixing $\delta\varphi_6 = 0$. This illustrates that, if we allow the model to infer that the scale of GR deviations is small, we will recover similar inferences overall as if we had fixed $\delta\varphi = 0$ *a priori*: we are learning *both* that spins are small *and* that any GR deviation must be small at this PN order. Conversely, an assumption of a broad GR deviation population leads to unrealistic astrophysical populations to account for the far-fetched astrophysical systems such analyses allow. We can also use this example to understand why inferring the deviation population in the absence of astrophysical modelling leads to a different deviation population with a larger inferred mean. Figure 7 shows that q and $\delta\varphi_6$ are correlated at the individual-event level, and therefore a broader q distribution will lend more support to the higher values of $\delta\varphi_6$. This correlation then systematically pulls the mean of the PN deviation distribution to higher values.

2. *Example: 0PN deviation coefficient, $\delta\varphi_0$*

We now turn to $\delta\varphi_0$, for which the standard analysis with a fixed astrophysical prior finds the least consistency with GR, at the 2.2σ credible level (yellow circle for $\delta\varphi_0$ in Fig. 5), driven by a displacement away from $\mu_{\rm PN} = 0$ (Fig. 4). Since this parameter is strongly correlated with the chirp mass and mass ratio (Fig. 1), we expect improvements when jointly modeling the astrophysical and deviation distributions; indeed that is the case, with GR recovered at the 1.6σ level (blue circle in Fig. 5). This analysis infers a $\sigma_{\rm PN}$ distribution that peaks slightly away from zero.

We can understand this behavior with Fig. 8, where we plot a subset of the two-dimensional marginal population posterior distributions in the same color scheme as Fig. 6. The structure of the PN deviation distribution is directly correlated with the mass ratio power-law index, β: steeper power-laws correspond to more variance in the GR deviation (larger β, larger $\sigma_{\rm PN}$). This is also manifested in the fact that when the PN deviation is assumed to be uniformly distributed (solid green), the astrophysical inference prefers steeper mass ratio power-laws (larger β), and that the analysis with deviations fixed to zero (dashed green) leads to a shallower slope

($\beta \lesssim 6$). There is also a correlation between $\sigma_{\rm PN}$ and the width of the spin distribution, σ_{χ_z}, by which a narrower spin distribution demands for a greater spread in deviation parameters within the population.

Such correlations highlight precisely why we need to account for the astrophysical population when testing GR. By assuming a particular, fixed model for the astrophysical population, the hyperparameter correlations will not be captured in the marginal posterior for the GR deviation population. The analysis assuming the sampling prior for the astrophysical population (yellow), infers a value of $\sigma_{\rm PN}$ which peaks at zero. Among other hyperparameters, the sampling prior corresponds to a uniform ($\beta = 0$) mass-ratio distribution. Fixing the astrophysical population in such a way will lead to the hyperparameter posterior peaking at $\sigma_{\rm PN} = 0$, as seen in Fig. 8.

IV. CONCLUSIONS

In this study, we have shown the importance of modeling the astrophysical population when testing GR with gravitational waves. Current tests do not explicitly model the astrophysical population, and therefore implicitly treat the prior used for sampling the posterior distribution as the assumed astrophysical population. Due to the presence of correlations between many GR deviations and astrophysical parameters, inappropriate astrophysical population choices will bias the test of GR. Like other sources of systematics, including waveform modeling [96–99], the severity of this bias increases with the number of detections. We have shown that the effect of this bias is already being felt in the present catalog. This issue can only be fully addressed by simultaneously modelling both the astrophysical population in addition to the GR deviations.

We demonstrate the effect of inappropriate astrophysical models using constraints of the graviton's mass and tests of PN deviations as concrete examples. We show that jointly modeling the astrophysical population distribution while testing GR leads to results more consistent with GR. Furthermore, for some deviations at various PN orders there are correlations between hyperparameters governing the astrophysical and deviation populations. The impact of the astrophysical distribution is not just important for these parameters and these hierarchical models: any test of GR should accurately account for the astrophysical population. In fact, this problem is not unique to tests of GR— attempts to infer cosmological properties [100] or the equation of state of dense nuclear matter [101] are also impacted by these same considerations.

We can generically understand the impact of folding in the astrophysical population as follows. The standard sampling prior is chosen to broadly cover the parameter range of interest, and not to accurately represent the true astrophysical population. The actual population distribution will then typically provide support on a more narrow region of parameter space than the sampling prior. As a result, population-informed posteriors will not only avoid systematic biases but will also provide more stringent constraints on GR due to the additional information from the associated narrower population.

This posterior shrinkage is illustrated in Fig. 9, which shows the 0PN deviation parameter and detector frame chirp mass for the 20 events considered in our study (Table I). The three sets of distributions correspond to the posteriors under different priors: fixed sampling priors (light red), fixed astrophysical prior and an inferred PN deviation population (yellow), as well as the case where both PN-deviation and astrophysics distributions are inferred (blue). As more information about the GR deviation distribution is included, the inferred posterior of 0PN deviation parameter and the detector-frame chirp mass is more constrained. The posteriors are then constrained further still as additional information regarding the astrophysical population is included.

There are a number of directions in which to extend our work. The first would be to account for selection effects on the hyperparameters of the GR deviation distribution; this is to be addressed in upcoming work [68]. Additionally, here we have assumed a strongly parameterized model for the astrophysical population, with a power law and a Gaussian peak. This model is currently flexible enough given the number of events, with the primary mass Gaussian peak not impacting the inferred PN deviations with the selection of events considered. As the number of events used with these tests increases, and subtle features in the astrophysical population reveal themselves, we will likely need more flexible models [63–66] to further avoid biases from misspecified population models [102–104]. Furthermore, in the case of PN coefficients, one would ideally constrain all orders simultaneously, in addition to the astrophysical parameters [1, 105–109].

Concurrently modeling the astrophysical population when testing GR is inevitable. Models that do not include a parameterized astrophysical population are implicitly assuming the sampling prior as the fixed population model. Such an assumption may induce systematic biases, cause false detections of GR violations, or incorrectly claim a stronger confirmation of GR than is warranted by the data. Moreover, even when accounting for the astrophysical population, correlations between GR deviation and astrophysical hyperparameters suggest that a true deviation could be absorbed into an unphysical inferred astrophysical population, a case that can only be noticed in studying the hyperposterior relating astrophysical to deviation parameters. Hierarchically modeling the astrophysical population while testing GR provides the solution to the implicit bias of assuming a fixed astrophysical population, and allows us to explore correlations between astrophysical parameters and deviations from GR, with fewer hidden assumptions.

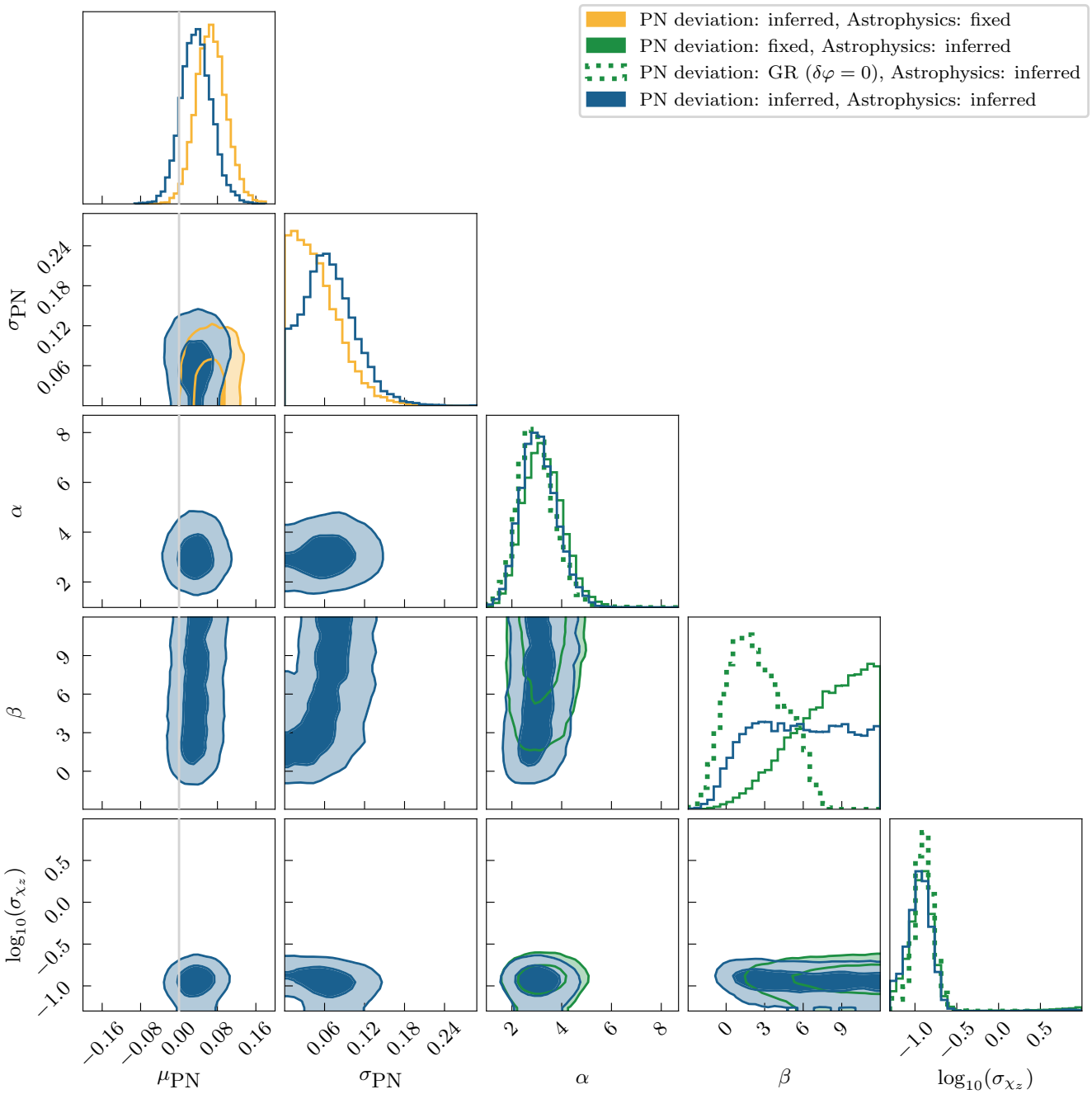

FIG. 8. Similar to Fig. 6, one- and two-dimensional posterior distributions for the $\delta\varphi_0$ deviation and a subset of astrophysical population hyperparameters. A strong correlation is found between the width of the inferred post-Newtonian deviation population and the index of the mass ratio power-law when jointly inferring the deviation and astrophysical population models. There is also a less pronounced correlation between the deviation and spin population standard deviations. In the absence of modelling the astrophysical population, the inferred PN population is pulled to a higher mean with a reduced width.

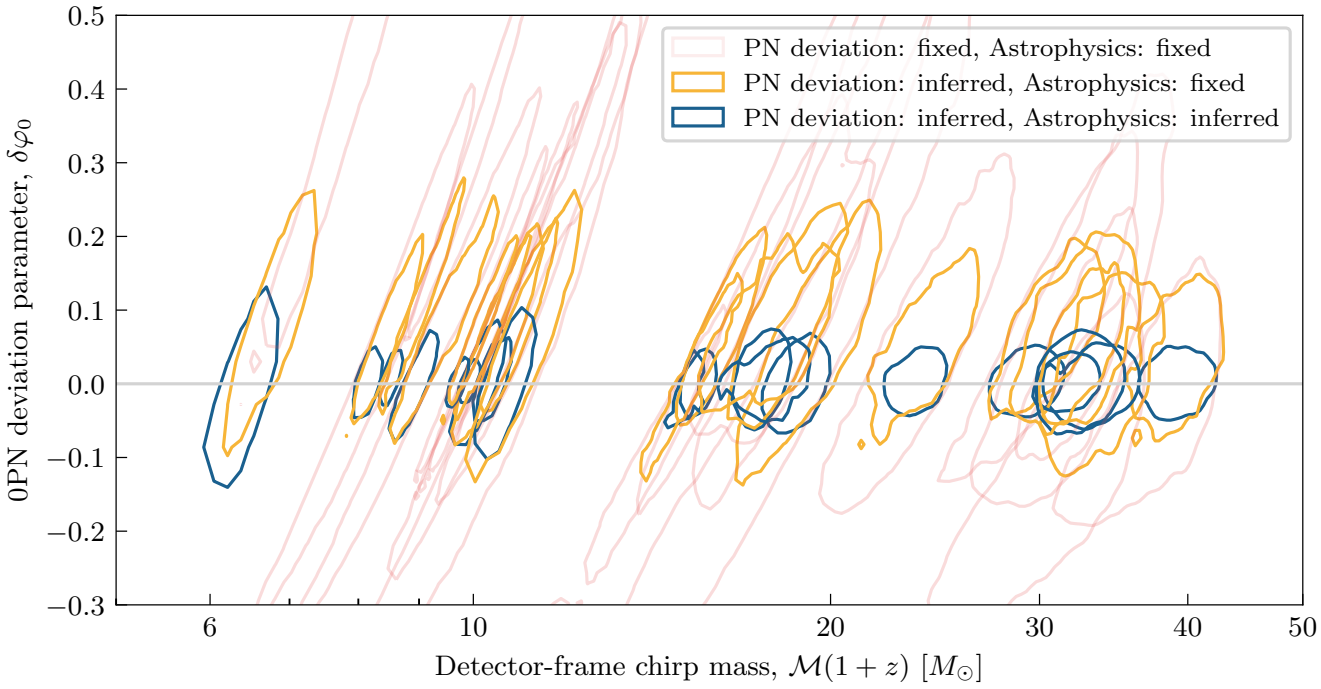

FIG. 9. Marginal two-dimensional posterior distributions for the 0PN deviation coefficient and the detector-frame chirp mass for the events analyzed under the broad prior assumptions (light red), informed PN deviation population only (yellow), and informed by the jointly inferred deviation and astrophysical populations (dark blue). Contours indicate the 90% credible regions. This result demonstrates that as additional information is incorporated into the population distribution, more stringent constraints on the deviation parameters are placed on an individual event level. In the case demonstrated here, this pulls the inferred value towards $\delta\varphi = 0$ for all events.

V. ACKNOWLEDGEMENTS

We thank Jacob Golomb and Alan Weinstein for insightful discussions, and Carl-Johan Haster for useful comments on the manuscript. Computing resources were provided by the Flatiron Institute. The Flatiron Institute is funded by the Simons Foundation. EP was supported by NSF Grant PHY-1764464. KC was supported by NSF Grant PHY-2110111. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. This research has made use of data, software and/or web tools obtained from the Gravitational Wave Open Science Center (https://www.gw-openscience.org), a service of LIGO Laboratory, the LIGO Scientific Collaboration and the Virgo Collaboration. Virgo is funded by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale della Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by Polish and Hungarian institutes. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants PHY-0757058 and PHY-0823459. This manuscript carries LIGO Document Number #P2300292.

Appendix A: Formulation of parameterized tests of general relativity

In this appendix we outline the calculations required to constrain the graviton's mass (App. A 1) and infer the PN deviation parameters (App. A 2).

1. Massive graviton measurements

The impact of a massive graviton on the propagation of gravitational waves has been studied in Refs. [15, 16] and references therein. A graviton with mass m_g modifies the dispersion relation of the gravitational wave. In a cosmological background, $g_{\mu\nu}$,

$$g_{\mu\nu}p^\mu p^\nu = -m_g^2 \tag{A1}$$

where p^μ is the 4-momentum of the graviton. This leads to a dephasing of the gravitational wave, $\delta\Phi(f)$, that scales with the distance over which the signal propagates,

$$\delta\Phi(f) = -\frac{\pi(1+z)D_L^2 m_g^2 c^3}{D_0 h^2} f^{-1}\,, \tag{A2}$$

where D_L is the luminosity distance, h is Planck's constant, and

$$D_0 = \frac{c(1+z)}{H_0}\int_0^z \mathrm{d}z' \, \frac{(1+z')^{-2}}{\sqrt{\Omega_m(1+z')^3+\Omega_\Lambda}}\,. \tag{A3}$$

Here, $H_0 = 67.9\,\mathrm{km\ s^{-1}\,Mpc^{-1}}$ is the Hubble constant, and $\Omega_m = 0.3065$ and $\Omega_\Lambda = 0.6935$ are the matter and dark energy density parameters, respectively, adopting the values used in previous analyses [7, 46, 110].

2. Post-Newtonian deviation tests

Current parameterized PN tests are constructed by single-parameter modifications to the post-Newtonian description of the inspiral gravitational-wave phase in the frequency domain. This is given by [34, 111]

$$\Phi(f) = 2\pi f t_c - \phi_c - \frac{\pi}{4} + \frac{3}{128}\times \sum_{k=0}^{7}\frac{1}{\eta^{k/5}}\left(\varphi_k + \varphi_{k,l}\ln\tilde{f}\right)\tilde{f}^{(k-5)/3}\,. \tag{A4}$$

Here, $\Phi(f)$ is the frequency-domain gravitational-wave phase under the stationary-phase approximation, $\tilde{f} = \pi G\mathcal{M}(1+z)f/c^3$, where $\mathcal{M}(1+z)$ is the redshifted chirp mass, $\mathcal{M} = (m_1m_2)^{3/5}/(m_1+m_2)^{1/5}$ is the source-frame chirp mass, $\eta = m_1m_2/M^2$ is the symmetric mass ratio, t_c and ϕ_c are the coalescence time and phase of the binary; finally, k indexes the $k/2$ PN order, and φ_k and $\varphi_{k,l}$ are the PN coefficients. The logarithmic coefficients, $\varphi_{k,l}$ only enter at 2.5 and 3.5 PN orders and otherwise vanish [112, 113]. In GR, the coefficients are functions of the intrinsic parameters of the binary, their masses and spins. From this prescription, modifications to GR are incorporated by modifying [12, 13, 32]

$$\varphi_k \to (1+\delta\varphi_k)\,\varphi_k\,, \tag{A5}$$

except for k's for which $\varphi_k = 0$ in GR ($k = -2, 1$); in these cases, the modification is $\varphi_k \to \delta\varphi_k$, and $\delta\varphi_k$ is an absolute deviation [114].

In practice, modifications to IMRPhenomPv2 [12, 13, 34, 35, 73, 76, 79] and SEOBNRv4 [14, 41–44] waveforms are computed differently, then the latter is transformed to the former. For the modified SEOBNRv4 waveform, the deviation is applied as above [14]. While, IMRPhenomPv2 is modified to only apply the deviation is onto the nonspinning portion of the PN coefficient [12, 13]. We translate all inferred deviation parameters to the IMRPhenomPv2 deviation parameter $\delta\varphi_k^{\rm IMR}$ for consistency,

$$\delta\varphi_k^{\rm IMR} = \delta\varphi_k\frac{\varphi_k}{\varphi_k^{\rm NS}}\,, \tag{A6}$$

where $\varphi_k^{\rm NS}$ is the nonspinning value of the PN coefficient — calculated by setting the spins to zero for a particular set of compact binary masses. Additionally, care needs to be taken when translating to a uniform prior on $\delta\varphi_k^{\rm IMR}$, as the appropriate Jacobian,

$$\frac{\mathrm{d}\delta\varphi_k^{\rm IMR}}{\mathrm{d}\delta\varphi_k} = \frac{\varphi_k}{\varphi_k^{\rm NS}}\,, \tag{A7}$$

is necessary. If the original prior is uniform on $\delta\varphi_k$, then the $\delta\varphi_k^{\rm IMR}$ must be weighted by the Jacobian to be effectively translated to another uniform prior.

Appendix B: Computing expected parameter correlations

Correlations between GR deviation and astrophysical parameters can be analytically approximated by identifying regions of the parameter space that lead to a similar frequency evolution [28] and signal duration. The dominant correlation is the one between the detector-frame chirp mass, $\mathcal{M}(1+z)$, and the symmetric mass ratio, η. The duration of a gravitational-wave signal is related to the detector-frame chirp mass and some fiducial cut-off frequency [115],

$$T \propto \mathcal{M}^{5/3}(1+z)^{5/3} f_{\rm cut}^{-8/3}\,. \tag{B1}$$

If we relate the final frequency to the innermost stable orbit or any cut-off which scales inversely with the binary's total mass, then $T \propto \eta^{-8/5}\mathcal{M}^{13/3}(1+z)^{13/3}$. A constant duration then imlies

$$\mathcal{M}(1+z) \propto \eta^{-24/65}\,. \tag{B2}$$

Here we have ignored both the contributions of a spin-induced "hang-up" effect [116] and GR deviations.

Correlations between astrophysical parameters and GR deviations can then be computed at lowest order [28] by enforcing that the second-order derivative of the phase evolution as a function of frequency be constant. As an example, for the correlation in Fig. 1, we compare the phase evolution when $\delta\varphi_0 = 0$ and when varying $\delta\varphi_0$ at the leading PN order, resulting in

$$\mathcal{M}_0^{-5/3}(1+z_0)^{-5/3} \sim (1+\delta\varphi)\mathcal{M}^{-5/3}(1+z)^{-5/3}\,. \tag{B3}$$

Here $\mathcal{M}_0$ and z_0 are the values of the chirp mass and redshift when there is no deviation. We find the 0PN deviation coefficient to only be directly correlated with the detector frame chirp mass,

$$\delta\varphi_0 \sim \left(\frac{\mathcal{M}(1+z)}{\mathcal{M}_0(1+z_0)}\right)^{5/3} - 1\,. \tag{B4}$$

This calculation can be repeated for higher PN orders as well, however care needs to be taken as lower PN orders need to be retained when computing higher PN deviation coefficient correlations.

Appendix C: Population likelihood approximation

In practice, we carry out single-event parameter estimation with a fiducial sampling prior, $\pi(\theta)$, before the hierarchical population analysis. We therefore do not possess representations of the individual event likelihoods, $p(d|\theta)$, but rather samples drawn from the fiducial posterior distribution $p(\theta|d) \propto p(d|\theta)\,\pi(\theta)$. Therefore, it is common to instead reformulate the integral within Eq. (1) as an average over samples drawn from each event's posterior distribution [47–49],

$$p(\{d\}|\Lambda) \propto \frac{1}{\xi(\Lambda)^N} \prod_{i=1}^{N} \frac{1}{M_i} \sum_{k=1}^{M_i} \frac{\pi(\theta_{i,k}|\Lambda)}{\pi(\theta_{i,k})}\,, \tag{C1}$$

where M_i is the number of posterior samples for the ith event. It is possible for this Monte Carlo integration to not converge—particularly if the population distribution $\pi(\theta|\Lambda)$ is narrower than posterior distributions for individual events [48, 61, 67, 78, 117, 118]. This is particularly important in our scenario, since the inferred population of deviations from GR is typically narrower than marginal measurements from many individual events. This leads to a dearth of samples within the inferred GR deviation population, which subsequently leads to unreliable Monte Carlo integration in Eq. (C1).

To address this issue, we use Gaussian kernel density estimates to represent the individual-event posteriors in a number of parameters, and simplify the calculation analytically by leveraging Gaussian population models. Dividing the parameters into the subset described by the Gaussian population distributions, $\theta^{\rm G}$, and the non-Gaussian distributions, $\theta^{\rm NG}$, we can analytically integrate over the former without resorting to Eq. (C1). The Gaussian population parameters are the GR deviation parameter and the binary-hole spin magnitudes, whereas the black-hole primary mass and mass ratio, redshift, and spin tilts (for the analysis in App. D) are included in the non-Gaussian set of parameters. For the kernel density estimation, we determine the corresponding covariance matrix for each individual event's distribution using Scott's rule [119],

$$\Sigma_{BW,i} \approx \frac{\Sigma_i}{n_{{\rm eff},i}^{2/(d+4)}}\,, \tag{C2}$$

where Σ_i is the weighted covariance matrix of the parameters being estimated, d is the number of Gaussian dimensions, and $n_{\rm eff}$ is the effective number of samples [120, 121],

$$n_{{\rm eff},i} = \frac{\left(\sum_{k=1}^{M_i} w(\theta^G_{i,k})\right)^2}{\sum_{k=1}^{M_i} w(\theta^G_{i,k})^2}\,, \tag{C3}$$

with the weights, $w(\theta^G_{i,k}) = 1/\pi(\theta^G_{i,k})$.

Since the integrand in the $\theta^{\rm G}$-space is a product of Gaussian distributions, the resulting integral is also a Gaussian [122]. This leads to the straightforward expression for the likelihood function

$$p(\{d\}|\Lambda) \propto \frac{1}{\xi(\Lambda)^N} \prod_{i=1}^{N} \frac{1}{M_i} \sum_{k=1}^{M_i} \frac{\pi(\theta^{\rm NG}_{i,k}|\Lambda)}{\pi(\theta_{i,k})} \times \mathcal{N}[\mu(\Lambda), \Sigma_{BW} + \Sigma(\Lambda)](\theta^{\rm G}_{i,k})\,, \tag{C4}$$

where $\mu(\Lambda) = (\mu, \mu_\chi, \mu_\chi)$ and $\Sigma(\Lambda) = {\rm diag}(\sigma^2, \sigma^2_\chi, \sigma^2_\chi)$, though more complicated structure can be imposed on the population model. Since this integral is computed analytically, we empirically find improved convergence.

Appendix D: Constraints from IMRPhenomPv2

While we have focused on results from SEOBNRv4 [14, 41–44], these analyses do not include precessing degrees of freedom. However, evidence for precession has been found at the population level within gravitational-wave observations [30, 31]. Therefore, to explore if there are any major changes when incorporating precession effects, we use the 12 events from the first half of the third observing run analysed with IMRPhenomPv2 [12, 13, 34, 35, 73, 76, 79] which meet our selection criteria [6]. There are no equivalent results from the second half of the third observing run [7]. We show the summary of the marginal two-dimensional posterior distribution for the Gaussian population hyperparameters with and without the inclusion of astrophysical information in Fig. 10. Generally, these results are less constrained due to the smaller number of events, though we still witness a similar shift in the means of the Gaussian populations as in Fig. 4. We also summarize the quantiles at which the expectation from GR presides in Fig. 11. Generally, the IMRPhenomPv2 results are more consistent with GR than the equivalent SEOBNRv4 results presented in Sec. III B. This could be a product of this waveform model incorporating precession, or simply that fewer events were analyzed, leading to a decrease in precision.

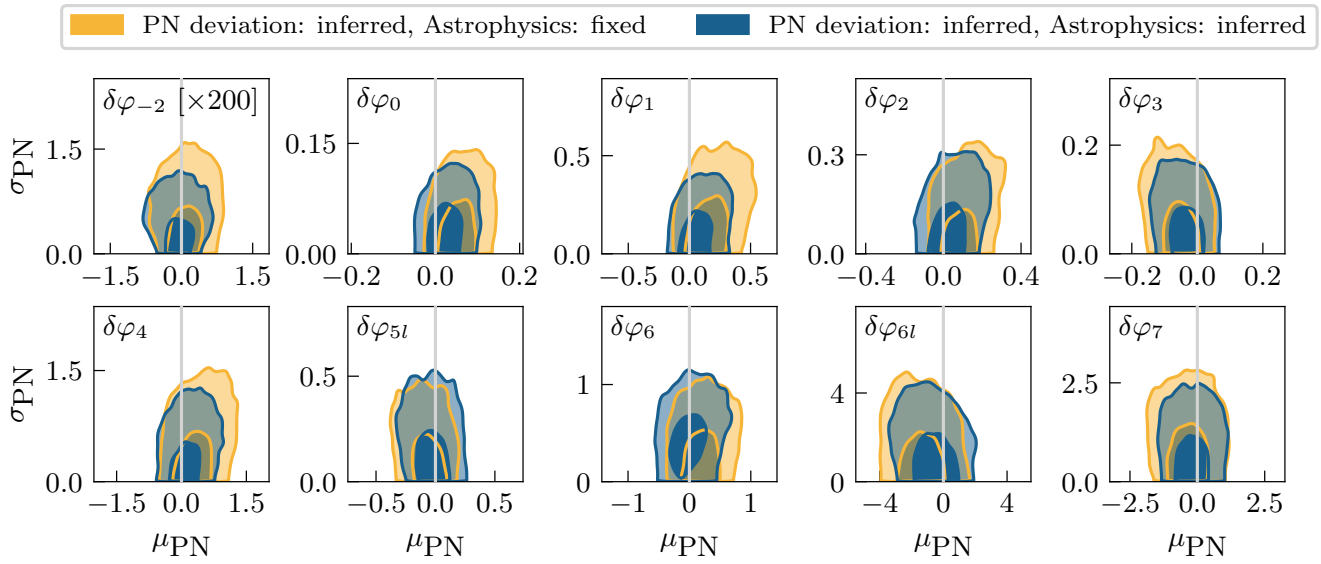

FIG. 10. Same figure as Fig. 4 but using 12 events from the first half of the third LIGO-Virgo-KAGRA observing run, with individual event posterior distributions constructed with IMRPhenomPv2. We generally observe similar structure to the results with SEOBNRv4, although parameters are less constrained—likely due to fewer observations incorporated.

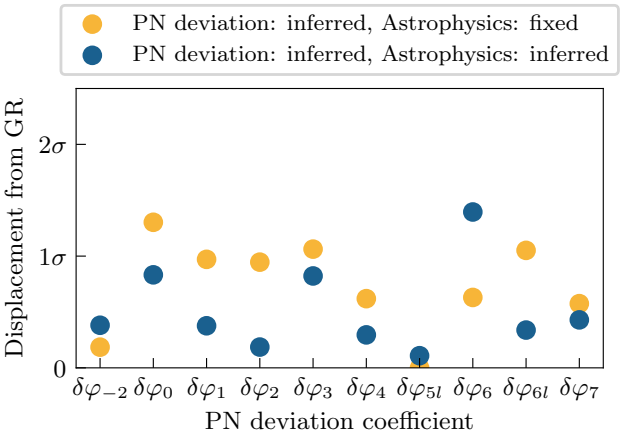

FIG. 11. Same as Fig. 5, for the results from the IMRPhenomPv2 analysis. As seen throughout the manuscript, inclusion of the astrophysical population model in general leads to improved consistency with GR. Furthermore, the posterior distributions sit closer to GR for IMRPhenomPv2 than SEOBNRv4, likely as a result of analyzing fewer events.

[1] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Tests of General Relativity with GW150914, Phys. Rev. Lett. **116**, 221101 (2016), arXiv:1602.03841 [gr-qc].

[2] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), GW170104: Observation of a 50-Solar-Mass Binary Black Hole Coalescence at Redshift 0.2, Phys. Rev. Lett. **118**, 221101 (2017), arXiv:1706.01812 [gr-qc].

[3] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), GW170814: A Three-Detector Observation of Gravitational Waves from a Binary Black Hole Coalescence, Phys. Rev. Lett. **119**, 141101 (2017), arXiv:1709.09660 [gr-qc].

[4] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Tests of General Relativity

with GW170817, Phys. Rev. Lett. **123**, 011102 (2019), arXiv:1811.00364 [gr-qc].
[5] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Tests of general relativity with the binary black hole signals from the LIGO-Virgo catalog GWTC-1, Phys. Rev. D **100**, 104036 (2019), arXiv:1903.04467 [gr-qc].
[6] R. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Tests of general relativity with binary black holes from the second LIGO-Virgo gravitational-wave transient catalog, Phys. Rev. D **103**, 122002 (2021), arXiv:2010.14529 [gr-qc].
[7] R. Abbott *et al.* (LIGO Scientific Collaboration and the Virgo Collaboration and the KAGRA Collaboration), Tests of General Relativity with GWTC-3, arXiv e-prints , arXiv:2112.06861 (2021), arXiv:2112.06861 [gr-qc].
[8] J. Aasi *et al.* (LIGO Scientific), Advanced LIGO, Class. Quant. Grav. **32**, 074001 (2015), arXiv:1411.4547 [gr-qc].
[9] F. Acernese *et al.* (VIRGO), Advanced Virgo: a second-generation interferometric gravitational wave detector, Class. Quant. Grav. **32**, 024001 (2015), arXiv:1408.3978 [gr-qc].
[10] A. Ghosh *et al.*, Testing general relativity using golden black-hole binaries, Phys. Rev. D **94**, 021101 (2016), arXiv:1602.02453 [gr-qc].
[11] 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, Testing general relativity using gravitational wave signals from the inspiral, merger and ringdown of binary black holes, Class. Quant. Grav. **35**, 014002 (2018), arXiv:1704.06784 [gr-qc].
[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, Towards a generic test of the strong field dynamics of general relativity using compact binary coalescence, Phys. Rev. D **85**, 082003 (2012), arXiv:1110.0530 [gr-qc].
[13] M. Agathos, W. Del Pozzo, T. G. F. Li, C. Van Den Broeck, J. Veitch, and S. Vitale, TIGER: A data analysis pipeline for testing the strong-field dynamics of general relativity with gravitational wave signals from coalescing compact binaries, Phys. Rev. D **89**, 082001 (2014), arXiv:1311.0420 [gr-qc].
[14] A. K. Mehta, A. Buonanno, R. Cotesta, A. Ghosh, N. Sennett, and J. Steinhoff, Tests of general relativity with gravitational-wave observations using a flexible theory-independent method, Phys. Rev. D **107**, 044020 (2023), arXiv:2203.13937 [gr-qc].
[15] C. M. Will, Bounding the mass of the graviton using gravitational wave observations of inspiralling compact binaries, Phys. Rev. D **57**, 2061 (1998), arXiv:gr-qc/9709011.
[16] S. Mirshekari, N. Yunes, and C. M. Will, Constraining Generic Lorentz Violation and the Speed of the Graviton with Gravitational Waves, Phys. Rev. D **85**, 024041 (2012), arXiv:1110.2720 [gr-qc].
[17] T. Zhu, W. Zhao, J.-M. Yan, C. Gong, and A. Wang, Tests of modified gravitational wave propagations with gravitational waves, arXiv:2304.09025 [gr-qc] (2023).
[18] T. C. K. Ng, M. Isi, K. W. K. Wong, and W. M. Farr, Constraining gravitational wave amplitude birefringence with GWTC-3 (2023), arXiv:2305.05844 [gr-qc].
[19] D. M. Eardley, D. L. Lee, A. P. Lightman, R. V. Wagoner, and C. M. Will, Gravitational-wave observations as a tool for testing relativistic gravity, Phys. Rev. Lett. **30**, 884 (1973).
[20] D. M. Eardley, D. L. Lee, and A. P. Lightman, Gravitational-wave observations as a tool for testing relativistic gravity, Phys. Rev. D **8**, 3308 (1973).
[21] M. Isi and A. J. Weinstein, Probing gravitational wave polarizations with signals from compact binary coalescences, arXiv e-prints , arXiv:1710.03794 (2017), arXiv:1710.03794 [gr-qc].
[22] P. T. H. Pang, R. K. L. Lo, I. C. F. Wong, T. G. F. Li, and C. Van Den Broeck, Generic searches for alternative gravitational wave polarizations with networks of interferometric detectors, Phys. Rev. D **101**, 104055 (2020), arXiv:2003.07375 [gr-qc].
[23] K. Chatziioannou, M. Isi, C.-J. Haster, and T. B. Littenberg, Morphology-independent test of the mixed polarization content of transient gravitational wave signals, Phys. Rev. D **104**, 044005 (2021), arXiv:2105.01521 [gr-qc].
[24] M. Isi, K. Chatziioannou, and W. M. Farr, Hierarchical Test of General Relativity with Gravitational Waves, Phys. Rev. Lett. **123**, 121101 (2019), arXiv:1904.08011 [gr-qc].
[25] M. Saleem, N. V. Krishnendu, A. Ghosh, A. Gupta, W. Del Pozzo, A. Ghosh, and K. G. Arun, Population inference of spin-induced quadrupole moments as a probe for nonblack hole compact binaries, Phys. Rev. D **105**, 104066 (2022), arXiv:2111.04135 [gr-qc].
[26] A. Zimmerman, C.-J. Haster, and K. Chatziioannou, On combining information from multiple gravitational wave sources, Phys. Rev. D **99**, 124044 (2019), arXiv:1903.11008 [astro-ph.IM].
[27] M. Isi, W. M. Farr, and K. Chatziioannou, Comparing Bayes factors and hierarchical inference for testing general relativity with gravitational waves, Phys. Rev. D **106**, 024048 (2022), arXiv:2204.10742 [gr-qc].
[28] D. Psaltis, C. Talbot, E. Payne, and I. Mandel, Probing the black hole metric: Black hole shadows and binary black-hole inspirals, Phys. Rev. D **103**, 104036 (2021), arXiv:2012.02117 [gr-qc].
[29] N. E. Wolfe, C. Talbot, and J. Golomb, Accelerating Tests of General Relativity with Gravitational-Wave Signals using Hybrid Sampling, arXiv e-prints , arXiv:2208.12872 (2022), arXiv:2208.12872 [gr-qc].
[30] R. Abbott *et al.* (The LIGO Scientific Collaboration and the Virgo Collaboration and the KAGRA Collaboration), The population of merging compact binaries inferred using gravitational waves through GWTC-3, arXiv e-prints , arXiv:2111.03634 (2021), arXiv:2111.03634 [astro-ph.HE].
[31] R. Abbott *et al.*, Population Properties of Compact Objects from the Second LIGO-Virgo Gravitational-Wave Transient Catalog, Astrophys. J. **913**, L7 (2021).
[32] N. Yunes and F. Pretorius, Fundamental Theoretical Bias in Gravitational Wave Astrophysics and the Parameterized Post-Einsteinian Framework, Phys. Rev. D **80**, 122003 (2009), arXiv:0909.3328 [gr-qc].
[33] L. Blanchet, Gravitational Radiation from Post-Newtonian Sources and Inspiralling Compact Binaries, Living Reviews in Relativity **17**, 2 (2014), arXiv:1310.1528 [gr-qc].

[34] K. G. Arun, B. R. Iyer, B. S. Sathyaprakash, and P. A. Sundararajan, Parameter estimation of inspiralling compact binaries using 3.5 post-Newtonian gravitational wave phasing: The nonspinning case, Phys. Rev. D **71**, 084008 (2005), arXiv:gr-qc/0411146 [gr-qc].

[35] K. Chatziioannou, A. Klein, N. Yunes, and N. Cornish, Constructing gravitational waves from generic spin-precessing compact binary inspirals, Phys. Rev. D **95**, 104004 (2017), arXiv:1703.03967 [gr-qc].

[36] N. Loutrel, T. Tanaka, and N. Yunes, Spin-Precessing Black Hole Binaries in Dynamical Chern-Simons Gravity, Phys. Rev. D **98**, 064020 (2018), arXiv:1806.07431 [gr-qc].

[37] S. E. Perkins, R. Nair, H. O. Silva, and N. Yunes, Improved gravitational-wave constraints on higher-order curvature theories of gravity, Phys. Rev. D **104**, 024060 (2021), arXiv:2104.11189 [gr-qc].

[38] N. Loutrel and N. Yunes, Parity violation in spin-precessing binaries: Gravitational waves from the inspiral of black holes in dynamical Chern-Simons gravity, Phys. Rev. D **106**, 064009 (2022), arXiv:2205.02675 [gr-qc].

[39] M. Okounkova, M. Isi, K. Chatziioannou, and W. M. Farr, Gravitational wave inference on a numerical-relativity simulation of a black hole merger beyond general relativity, Phys. Rev. D **107**, 024046 (2023), arXiv:2208.02805 [gr-qc].

[40] R. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo during the First Half of the Third Observing Run, Physical Review X **11**, 021053 (2021), arXiv:2010.14527 [gr-qc].

[41] A. Bohé, L. Shao, A. Taracchini, A. Buonanno, S. Babak, I. W. Harry, I. Hinder, S. Ossokine, M. Pürrer, V. Raymond, T. Chu, H. Fong, P. Kumar, H. P. Pfeiffer, M. Boyle, D. A. Hemberger, L. E. Kidder, G. Lovelace, M. A. Scheel, and B. Szilágyi, Improved effective-one-body model of spinning, nonprecessing binary black holes for the era of gravitational-wave astrophysics with advanced detectors, Phys. Rev. D **95**, 044028 (2017), arXiv:1611.03703 [gr-qc].

[42] R. Cotesta, A. Buonanno, A. Bohé, A. Taracchini, I. Hinder, and S. Ossokine, Enriching the symphony of gravitational waves from binary black holes by tuning higher harmonics, Phys. Rev. D **98**, 084028 (2018), arXiv:1803.10701 [gr-qc].

[43] R. Cotesta, S. Marsat, and M. Pürrer, Frequency-domain reduced-order model of aligned-spin effective-one-body waveforms with higher-order modes, Phys. Rev. D **101**, 124040 (2020), arXiv:2003.12079 [gr-qc].

[44] R. Brito, A. Buonanno, and V. Raymond, Black-hole spectroscopy by making full use of gravitational-wave modeling, Phys. Rev. D **98**, 084038 (2018), arXiv:1805.00293 [gr-qc].

[45] LIGO Scientific Collaboration and Virgo Collaboration and KAGRA Collaboration, Data release for tests of general relativity with gwtc-3 (2022).

[46] R. Abbott *et al.*, GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run, (2021), arxiv/2111.03606.

[47] I. Mandel, W. M. Farr, and J. R. Gair, Extracting distribution parameters from multiple uncertain observations with selection biases, Mon. Not. R. Ast. Soc. **486**, 1086 (2019).

[48] E. Thrane and C. Talbot, An introduction to Bayesian inference in gravitational-wave astronomy: parameter estimation, model selection, and hierarchical models, Pub. Astron. Soc. Aust. **36**, E010 (2019).

[49] S. Vitale, D. Gerosa, W. Farr, and S. Taylor, *Inferring the Properties of a Population of Compact Binaries in Presence of Selection Effects*, Handbook of Gravitational Wave Astronomy (Springer, Singapore, 2022).

[50] R. Abbott *et al.*, Binary black hole population properties inferred from the first and second observing runs of advanced LIGO and advanced virgo, Astrophys. J. **882**, L24 (2019).

[51] J. Roulet, H. S. Chia, S. Olsen, L. Dai, T. Venumadhav, B. Zackay, and M. Zaldarriaga, Distribution of effective spins and masses of binary black holes from the LIGO and virgo o1–o3a observing runs, Phys. Rev. D **104**, 083010 (2021).

[52] W. M. Farr, S. Stevenson, M. C. Miller, I. Mandel, B. Farr, and A. Vecchio, Distinguishing spin-aligned and isotropic black hole populations with gravitational waves, Nature **548**, 426 (2017).

[53] C. Talbot and E. Thrane, Measuring the binary black hole mass spectrum with an astrophysically motivated parameterization, Astrophys. J. **856**, 173 (2018).

[54] C. Talbot and E. Thrane, Determining the population properties of spinning black holes, Phys. Rev. D **96**, 023012 (2017).

[55] T. A. Callister, C. Haster, K. K. Y. Ng, S. Vitale, and W. M. Farr, Who ordered that? unequal-mass binary black hole mergers have larger effective spins, Astrophys. J. Lett. **922**, L5 (2021).

[56] M. Fishbach, C. Kimball, and V. Kalogera, Limits on hierarchical black hole mergers from the most negative $\chi_{\rm eff}$ systems, Astrophys. J. Lett. **935**, L26 (2022).

[57] S. Biscoveanu, T. A. Callister, C. Haster, K. K. Y. Ng, S. Vitale, and W. M. Farr, The binary black hole spin distribution likely broadens with redshift, Astrophys. J. Lett. **932**, L19 (2022).

[58] S. Vitale, R. Lynch, R. Sturani, and P. Graff, Use of gravitational waves to probe the formation channels of compact binaries, Class. Quantum Grav. **34**, 03LT01 (2017).

[59] S. Stevenson, C. P. L. Berry, and I. Mandel, Hierarchical analysis of gravitational-wave measurements of binary black hole spin–orbit misalignments, Mon. Not. R. Ast. Soc. **471**, 2801 (2017).

[60] S. Miller, T. A. Callister, and W. M. Farr, The low effective spin of binary black holes and implications for individual gravitational-wave events, Astrophys. J. **895**, 128 (2020).

[61] S. Galaudage, C. Talbot, T. Nagar, D. Jain, E. Thrane, and I. Mandel, Building better spin models for merging binary black holes: Evidence for nonspinning and rapidly spinning nearly aligned subpopulations, Astrophys. J. Lett. **921**, L15 (2021).

[62] M. Fishbach, D. E. Holz, and W. M. Farr, Does the Black Hole Merger Rate Evolve with Redshift?, Astrophys. J. Lett. **863**, L41 (2018), arXiv:1805.10270 [astro-ph.HE].

[63] B. Edelman, Z. Doctor, J. Godfrey, and B. Farr, Ain't no mountain high enough: Semiparametric modeling of LIGO–virgo's binary black hole mass distribution, Astrophys. J. **924**, 101 (2022).

[64] B. Edelman, B. Farr, and Z. Doctor, Cover Your Basis: Comprehensive Data-driven Characterization of the Binary Black Hole Population, Astrophys. J. **946**, 16 (2023), arXiv:2210.12834 [astro-ph.HE].

[65] J. Golomb and C. Talbot, Searching for structure in the binary black hole spin distribution, arXiv e-prints , arXiv:2210.12287 (2022), arXiv:2210.12287 [astro-ph.HE].

[66] T. A. Callister and W. M. Farr, A Parameter-Free Tour of the Binary Black Hole Population, arXiv e-prints , arXiv:2302.07289 (2023), arXiv:2302.07289 [astro-ph.HE].

[67] W. M. Farr, Accuracy Requirements for Empirically Measured Selection Functions, Research Notes of the American Astronomical Society **3**, 66 (2019), arXiv:1904.10879 [astro-ph.IM].

[68] R. Magee *et al.*, Selection biases in tests of general relativity with gravitational waves (2023), in prep.

[69] C. J. Moore and D. Gerosa, Population-informed priors in gravitational-wave astronomy, Phys. Rev. D **104**, 083008 (2021), arXiv:2108.02462 [gr-qc].

[70] W. M. Farr and T. A. Callister, *Re-Weighting Existing Samples to a Population Analysis*, Tech. Rep. (2021).

[71] T. A. Callister, *Reweighting Single Event Posteriors with Hyperparameter Marginalization*, Tech. Rep. LIGO-T2100301 (2021).

[72] W. Del Pozzo, J. Veitch, and A. Vecchio, Testing general relativity using Bayesian model selection: Applications to observations of gravitational waves from compact binary systems, Phys. Rev. D **83**, 082002 (2011), arXiv:1101.1391 [gr-qc].

[73] J. Meidam, M. Agathos, C. Van Den Broeck, J. Veitch, and B. S. Sathyaprakash, Testing the no-hair theorem with black hole ringdowns using TIGER, Phys. Rev. D **90**, 064009 (2014), arXiv:1406.3201 [gr-qc].

[74] A. Ghosh, A. Ghosh, N. K. Johnson-McDaniel, C. K. Mishra, P. Ajith, W. Del Pozzo, D. A. Nichols, Y. Chen, A. B. Nielsen, C. P. L. Berry, and L. London, Testing general relativity using golden black-hole binaries, Phys. Rev. D **94**, 021101 (2016), arXiv:1602.02453 [gr-qc].

[75] A. Ghosh, N. K. Johnson-McDaniel, A. Ghosh, C. Kant Mishra, P. Ajith, W. Del Pozzo, C. P. L. Berry, A. B. Nielsen, and L. London, Testing general relativity using gravitational wave signals from the inspiral, merger and ringdown of binary black holes, Classical and Quantum Gravity **35**, 014002 (2018), arXiv:1704.06784 [gr-qc].

[76] J. Meidam, K. W. Tsang, J. Goldstein, M. Agathos, A. Ghosh, C.-J. Haster, V. Raymond, A. Samajdar, P. Schmidt, R. Smith, K. Blackburn, W. Del Pozzo, S. E. Field, T. Li, M. Pürrer, C. Van Den Broeck, J. Veitch, and S. Vitale, Parametrized tests of the strong-field dynamics of general relativity using gravitational wave signals from coalescing binary black holes: Fast likelihood calculations and sensitivity of the method, Phys. Rev. D **97**, 044033 (2018), arXiv:1712.08772 [gr-qc].

[77] D. Wysocki, J. Lange, and R. O'Shaughnessy, Reconstructing phenomenological distributions of compact binaries via gravitational wave observations, Phys. Rev. D **100**, 043012 (2019).

[78] T. A. Callister, S. J. Miller, K. Chatziioannou, and W. M. Farr, No evidence that the majority of black holes in binaries have zero spin, (2022), arxiv/2205.08574.

[79] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. J. Forteza, and A. Bohé, Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era, Phys. Rev. D **93**, 044007 (2016), arXiv:1508.07253 [gr-qc].

[80] LIGO Scientific Collaboration and Virgo Collaboration and KAGRA Collaboration, Tests of general relativity with binary black holes from the second ligo–virgo gravitational-wave transient catalog - full posterior sample data release (2021).

[81] R. Abbott *et al.* (LIGO Scientific, Virgo), GW190814: Gravitational Waves from the Coalescence of a 23 Solar Mass Black Hole with a 2.6 Solar Mass Compact Object, Astrophys. J. Lett. **896**, L44 (2020), arXiv:2006.12611 [astro-ph.HE].

[82] R. Abbott *et al.* (LIGO Scientific, KAGRA, VIRGO), Observation of Gravitational Waves from Two Neutron Star–Black Hole Coalescences, Astrophys. J. Lett. **915**, L5 (2021), arXiv:2106.15163 [astro-ph.HE].

[83] R. Abbott *et al.*, GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run, Phys. Rev. X **11**, 021053 (2021).

[84] D. Phan, N. Pradhan, and M. Jankowiak, Composable effects for flexible and accelerated probabilistic programming in numpyro, arXiv preprint arXiv:1912.11554 (2019).

[85] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. A. Szerlip, P. Horsfall, and N. D. Goodman, Pyro: Deep universal probabilistic programming, J. Mach. Learn. Res. **20**, 28:1 (2019).

[86] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, JAX: composable transformations of Python+NumPy programs (2018).

[87] T. P. Robitaille *et al.*, Astropy: A community Python package for astronomy, Astron. & Astrophys. **558**, A33 (2013), arXiv:1307.6212 [astro-ph.IM].

[88] A. M. Price-Whelan *et al.*, The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package, Astron. J. **156**, 123 (2018), arXiv:1801.02634 [astro-ph.IM].

[89] A. M. Price-Whelan *et al.*, The Astropy Project: Sustaining and Growing a Community-oriented Open-source Project and the Latest Major Release (v5.0) of the Core Package, apj **935**, 167 (2022), arXiv:2206.14220 [astro-ph.IM].

[90] P. Virtanen *et al.*, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods **17**, 261 (2020).

[91] J. D. Hunter, Matplotlib: A 2d graphics environment, Computing in Science & Engineering **9**, 90 (2007).

[92] R. Kumar, C. Carroll, A. Hartikainen, and O. Martin, Arviz a unified library for exploratory analysis of bayesian models in python, Journal of Open Source Software **4**, 1143 (2019).

[93] D. Foreman-Mackey, corner.py: Scatterplot matrices in python, The Journal of Open Source Software **1**, 24 (2016).

[94] E. Payne, M. Isi, K. Chatziioannou, and W. M. Farr, Code release for "fortifying gravitational-wave tests of general relativity against astrophysical assumption"

(2023).
[95] B. P. Abbott *et al.* (LIGO Scientific, Virgo), GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs, Phys. Rev. X **9**, 031040 (2019), arXiv:1811.12907 [astro-ph.HE].
[96] C. J. Moore, E. Finch, R. Buscicchio, and D. Gerosa, Testing general relativity with gravitational-wave catalogs: The insidious nature of waveform systematics, iScience **24**, 102577 (2021).
[97] Q. Hu and J. Veitch, Accumulating Errors in Tests of General Relativity with Gravitational Waves: Overlapping Signals and Inaccurate Waveforms, Astrophys. J. **945**, 103 (2023), arXiv:2210.04769 [gr-qc].
[98] P. Saini, M. Favata, and K. G. Arun, Systematic bias on parametrized tests of general relativity due to neglect of orbital eccentricity, Phys. Rev. D **106**, 084031 (2022), arXiv:2203.04634 [gr-qc].
[99] S. A. Bhat, P. Saini, M. Favata, and K. G. Arun, Systematic bias on the inspiral-merger-ringdown consistency test due to neglect of orbital eccentricity, Phys. Rev. D **107**, 024009 (2023), arXiv:2207.13761 [gr-qc].
[100] R. Abbott *et al.* (LIGO Scientific, Virgo,, KAGRA, VIRGO), Constraints on the Cosmic Expansion History from GWTC–3, Astrophys. J. **949**, 76 (2023), arXiv:2111.03604 [astro-ph.CO].
[101] D. Wysocki, R. O'Shaughnessy, L. Wade, and J. Lange, Inferring the neutron star equation of state simultaneously with the population of merging neutron stars, (2020), arXiv:2001.01747 [gr-qc].
[102] I. M. Romero-Shaw, E. Thrane, and P. D. Lasky, When models fail: an introduction to posterior predictive checks and model misspecification in gravitational-wave astronomy, Pub. Astron. Soc. Aust. **39**, E025 (2022).
[103] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis, Third Edition*, Chapman & Hall/CRC Texts in Statistical Science (Taylor & Francis, 2013).
[104] E. Payne and E. Thrane, Model exploration in gravitational-wave astronomy with the maximum population likelihood, Phys. Rev. Res. **5**, 023013 (2023).
[105] A. Gupta, S. Datta, S. Kastha, S. Borhanian, K. G. Arun, and B. S. Sathyaprakash, Multiparameter tests of general relativity using multiband gravitational-wave observations, Phys. Rev. Lett. **125**, 201101 (2020), arXiv:2005.09607 [gr-qc].
[106] S. Datta, A. Gupta, S. Kastha, K. G. Arun, and B. S. Sathyaprakash, Tests of general relativity using multiband observations of intermediate mass binary black hole mergers, Phys. Rev. D **103**, 024036 (2021), arXiv:2006.12137 [gr-qc].
[107] A. A. Shoom, P. K. Gupta, B. Krishnan, A. B. Nielsen, and C. D. Capano, Testing the post-Newtonian expansion with GW170817, General Relativity and Gravitation **55**, 55 (2023), arXiv:2105.02191 [gr-qc].
[108] S. Perkins and N. Yunes, Are parametrized tests of general relativity with gravitational waves robust to unknown higher post-Newtonian order effects?, Phys. Rev. D **105**, 124047 (2022), arXiv:2201.02542 [gr-qc].
[109] S. Datta, M. Saleem, K. G. Arun, and B. S. Sathyaprakash, Multiparameter tests of general relativity using principal component analysis with next-generation gravitational wave detectors, arXiv e-prints , arXiv:2208.07757 (2022), arXiv:2208.07757 [gr-qc].
[110] P. A. R. Ade *et al.* (Planck), Planck 2015 results. XIII. Cosmological parameters, Astron. Astrophys. **594**, A13 (2016), arXiv:1502.01589 [astro-ph.CO].
[111] B. S. Sathyaprakash and S. V. Dhurandhar, Choice of filters for the detection of gravitational waves from coalescing binaries, Phys. Rev. D **44**, 3819 (1991).
[112] L. Blanchet and G. Schafer, Gravitational wave tails and binary star systems, Classical and Quantum Gravity **10**, 2699 (1993).
[113] L. Blanchet and B. S. Sathyaprakash, Signal analysis of gravitational wave tails, Classical and Quantum Gravity **11**, 2807 (1994).
[114] C. M. Will, The Confrontation between General Relativity and Experiment, Living Reviews in Relativity **17**, 4 (2014), arXiv:1403.7377 [gr-qc].
[115] C. Cutler and E. E. Flanagan, Gravitational waves from merging compact binaries: How accurately can one extract the binary's parameters from the inspiral wave form?, Phys. Rev. D **49**, 2658 (1994), arXiv:gr-qc/9402014.
[116] M. Campanelli, C. O. Lousto, and Y. Zlochower, Spinning-black-hole binaries: The orbital hang up, Phys. Rev. D **74**, 041501 (2006), arXiv:gr-qc/0604012.
[117] R. Essick and W. Farr, Precision Requirements for Monte Carlo Sums within Hierarchical Bayesian Inference, arXiv e-prints , arXiv:2204.00461 (2022), arXiv:2204.00461 [astro-ph.IM].
[118] C. Talbot and J. Golomb, Growing Pains: Understanding the Impact of Likelihood Uncertainty on Hierarchical Bayesian Inference for Gravitational-Wave Astronomy, arXiv e-prints , arXiv:2304.06138 (2023), arXiv:2304.06138 [astro-ph.IM].
[119] D. W. Scott, On optimal and data-based histograms, Biometrika **66**, 605 (1979).
[120] L. Kish, *Survey Sampling*, 3rd ed. (Wiley-Interscience, Oxford, England, 1995).
[121] V. Elvira, L. Martino, and C. P. Robert, Rethinking the Effective Sample Size, arXiv e-prints , arXiv:1809.04129 (2018), arXiv:1809.04129 [stat.CO].
[122] D. W. Hogg, A. M. Price-Whelan, and B. Leistedt, Data Analysis Recipes: Products of multivariate Gaussians in Bayesian inferences, arXiv e-prints , arXiv:2005.14199 (2020), arXiv:2005.14199 [stat.CO].