Impact of selection biases on tests of general relativity with gravitational-wave inspirals

Ryan Magee,[1,2,*] Maximiliano Isi,[3,†] Ethan Payne,[1,2,‡] Katerina Chatziioannou,[1,2,§] Will M. Farr,[3,4,‖] Geraint Pratten,[5,¶] and Salvatore Vitale[6,7,**]

[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 Avenue, New York, New York 10010, USA*
[4]*Department of Physics and Astronomy, Stony Brook University, Stony Brook, New York 11794, USA*
[5]*School of Physics and Astronomy and Institute for Gravitational Wave Astronomy, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom*
[6]*LIGO Laboratory, Massachusetts Institute of Technology, 185 Albany Street, Cambridge, Massachusetts 02139, USA*
[7]*Department of Physics and Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA*

(Received 13 November 2023; accepted 15 December 2023; published 8 January 2024)

Tests of general relativity with gravitational-wave observations from merging compact binaries continue to confirm Einstein's theory of gravity with increasing precision. However, these tests have so far been applied only to signals that were first confidently detected by matched-filter searches assuming general relativity templates. This raises the question of selection biases: What is the largest deviation from general relativity that current searches can detect, and are current constraints on such deviations necessarily narrow because they are based on signals that were detected by templated searches in the first place? In this paper, we estimate the impact of selection effects for tests of the inspiral phase evolution of compact binary signals with a simplified version of the GSTLAL search pipeline. We find that selection biases affect the search for very large values of the deviation parameters, much larger than the constraints implied by the detected signals. Therefore, combined population constraints from confidently detected events are mostly unaffected by selection biases, with the largest effect being a broadening at the ∼10% level for the −1 PN term. These findings suggest that current population constraints on the inspiral phase are robust without factoring in selection biases. Our study does not rule out a disjoint, undetectable binary population with large deviations from general relativity or stronger selection effects in other tests or search procedures.

DOI: 10.1103/PhysRevD.109.023014

I. INTRODUCTION

Gravitational-wave (GW) signals detected by LIGO [1] and Virgo [2] have provided otherwise-inaccessible constraints on deviations from general relativity (GR) in the dynamical and strong-field regimes [3–7]. When considered in aggregate, the set of detected binary black hole (BBH) signals is fully consistent with the null hypothesis of quasicircular mergers in vacuum GR. However, existing constraints apply only to signals that have been confidently detected and identified as compact binaries by pipelines based on GR. Even though generic searches exist [8–12], all current BBH signals have been detected with search pipelines that are based on templates produced within Einstein's theory. It remains possible that there exist binaries whose signals depart from GR but have been selected against by searches [13–15]. This raises two interrelated questions: (i) What is the largest deviation from GR that current searches can detect? (ii) Are current constraints on deviations from GR artificially narrow because they are based on signals that were detected in the first place?

Answering these questions amounts to quantifying the *selection biases* that modulate the probability of signal detection as a function of its parameters. The impact of regular binary parameters within GR—such as black hole (BH) masses or spins—can be approximated through their influence on the expected signal-to-noise ratio (SNR) of a

*rmmagee@caltech.edu
†misi@flatironinstitute.org
‡epayne@caltech.edu
§kchatziioannou@caltech.edu
‖wfarr@flatironinstitute.org
¶g.pratten@bham.ac.uk
**salvo@mit.edu

given signal [16,17] or, more robustly, by assessing the performance of the search pipeline on simulated signals [18]. The resulting selection function is an indispensable ingredient in inferring the astrophysical distributions of the detected events [16–18]. While this effect is well understood for GR parameters, the selection on beyond-GR parameters is currently largely unknown and generally unquantified. Nevertheless, studies under specific models suggest searches have non-negligible selection for sufficiently large deviations [13–15].

In the absence of a quantified selection function for tests of GR, current constraints are restricted to assessing agreement of the population properties of detected events with GR. Such an analysis can be performed without reference to any specific alternative theory of gravity by inferring the general shape of the population of deviations using hierarchical inference [19–22]. This procedure can detect anomalies in a collection of signals even if the deviation manifests differently for each individual event [23–25]. However, without selection effects, this procedure does not infer the *intrinsic* population of deviations, which could contain undetectable signals [6,7,24]. Furthermore, if selection biases are strong, these population constraints do not formally correspond to the *detected population* either on account of detector noise [26]. This concern also extends to cases in which events can be combined by simply multiplying likelihoods for a shared deviation parameter.

In this paper, we study the selection function within template-based search pipelines for parameterized tests of the inspiral phasing parameters [27–30]. Among the wide array of possible GR tests, we focus on post-Newtonian (PN) modifications to the waveform phasing, $\varphi(f)$, due to anomalous dynamics [30–42], which could arise from corrections to the theory or due to exotic sources following other nonstandard physics, such as BH mimickers. We use the deviation parameters $\delta\varphi_i$, where $i/2$ denotes the associated PN order. We focus on PN modifications as they are one of the flagship tests of GR with LIGO, Virgo, and KAGRA [43], and their effect is to modify the full inspiral, which dominates the detectability of all but the most massive systems. The latter can more easily be detected by theory-agnostic burst pipelines, potentially reducing the expected impact of selection biases induced by deviations from GR.

We generate simulated signals (also called *injections*) and recover them with a simplified version of the GSTLAL pipeline [44–47] in Sec. II. Rather than evaluating the computationally expensive likelihood ratio that would normally be computed by GSTLAL as a detection statistic, we approximate detection efficiency with a proxy ranking statistic based on the recovered SNR and an autocorrelation-based consistency check. In Sec. III we find that, under these circumstances, selection biases affect the detectability of signals only for very large values of the deviation parameters. These values are significantly higher than the precision achieved by current tests; we therefore expect that incorporating selection effects in population inference will have a minimal impact on the resulting constraints.

Armed with the results from our injection campaign, we confirm this expectation by enhancing existing hierarchical tests of GR [7] with a selection factor and compute the resulting astrophysical distribution of deviation parameters in Sec. IV. We parametrize the deviation population with a Gaussian and infer its mean and standard deviation while taking into account selection effects. Following [48], we simultaneously model the astrophysical distribution of the binary component masses. For most phase deviation terms we consider, the inferred astrophysical distributions for beyond-GR parameters are identical to those obtained by ignoring the GR selection effects. We recover the strongest impact for the -1 PN term, where incorporating selection effects widens the inferred population distribution by 10%. We therefore conclude that the quantitative impact of ignoring selection effects in tests of GR with GW inspirals is small.

This conclusion may be surprising given the crucial role of selection effects in estimating, for example, the mass distribution of BBHs. The crucial difference between deviation parameters and BBH masses is that the former population is inferred to be intrinsically very narrow as all events are consistent with a vanishing deviation. Indeed, after a dozen high-significance BBHs, the population for all deviation parameters inferred from LIGO-Virgo data is already narrower than the impact of selection effects. As more events are detected (and assuming they remain consistent with GR), the inferred deviation population will continue to narrow, making selection effects even less relevant. In other words, selection effects do exist in the population, but their impact is only appreciable for deviation values that are already ruled out. Other population distributions, such as those for the mass and spin, are not inherently narrow and selection effects remain important no matter how many events are detected. These considerations suggest that our conclusions only apply under the assumption that all events come from a narrow, unimodal population of deviation parameters. They do not rule out a disjoint population with deviations large enough to remain hidden to searches; such extreme non-GR signals can only be ruled out with a dedicated search [13–15]. We further this argument in our concluding remarks, Sec. V.

II. ESTIMATING THE MATCHED-FILTER SELECTION FUNCTION FOR SIGNALS WITH GR DEVIATIONS

In this section, we describe the procedure for quantifying the effect of GR deviations on the GW selection function. In summary, we follow the standard practice of estimating detection efficiency by simulating a large set of signals

(Sec. II A), analyzing them with a detection pipeline (Sec. II B), and determining which signals are detectable (Sec. II B).

A. Injection set

We start with the publicly available set of 156878 BBH injections associated with GWTC-3, which target only GR parameters [49]; we leave detailed explorations of binary neutron stars and neutron-star–black-hole binaries to future work. In this injection set, the primary and secondary binary masses are distributed as $p(m_1) \propto m_1^{-2.35}$ and $p(m_2|m_1) \propto m_2$ and bounded, in the source frame, such that $2M_\odot < m_2 \leq m_1 < 100M_\odot$; the BH spins are isotropically distributed with uniformly distributed magnitudes $|\chi_{1,2}| \leq 0.998$. Further specifics of the within-GR population are described in Table XII of [50]. The simulations are generated using a baseline IMRPHENOMPV2 waveform approximant [51–53], which includes the effects of spins misaligned with the orbital angular momentum. We implement deviations from GR using the TIGER framework [28–30], as in [7].

To reduce the computational burden on the original GWTC-3 analysis [50], these injections have already been selected against a minimum optimal network SNR threshold of 6. The network SNR was calculated by adding the LIGO Livingston and LIGO Hanford SNRs in quadrature. Systems with a lower optimal network SNR are considered "hopeless" for detection. To further enhance computational efficiency, we only consider BBHs that have optimal LIGO Livingston SNRs ≥ 6 and redshifted total masses below $300M_\odot$. For our purposes, restricting the total mass injected has negligible effect due to the additional inspiral SNR selection criterion typically applied in PN tests of GR [7,48]; we return to this in Sec. IV. These initial cuts result in 84119 injections.

To measure the selection bias against beyond-GR populations, we perturb the inspiral phasing of the injections and recover them with an approximation of the GSTLAL-based inspiral pipeline described in Sec. II B. Following the standard parametrized post-Einsteinian test [27], we perturb each PN order and repeat the analysis separately. Each simulation is assigned a random fractional[1] deviation drawn from a uniform distribution with bounds ± 0.1, ± 1, ± 5, ± 3, ± 2, ± 15, ± 5, ± 10, ± 50, and ± 30 for the $\delta\varphi_{-2}$, $\delta\varphi_0$, $\delta\varphi_1$, $\delta\varphi_2$, $\delta\varphi_3$, $\delta\varphi_4$, $\delta\varphi_{5l}$, $\delta\varphi_6$, $\delta\varphi_{6l}$, and $\delta\varphi_7$, respectively, where the "l" subscript denotes the logarithmic phase terms. The bounds are chosen such that the inferred deviations from individual events are entirely covered by the selection. We only vary one coefficient at a time to match the analysis usually applied to actual data [7]. This procedure results in one BBH injection set per PN order, each containing the same number of BBHs with identical GR parameters, differing only in the order and strength of the random GR deviations. After specifying injection parameters, we generate a corresponding waveform using the IMRPHENOMPV2 approximant and add it to the data stream of a single detector. We space the simulated signals 7 s apart through a single stretch of data collected in the LIGO Livingston detector during April of 2019 with global-positioning-system times in the range [1239641219 s, 1240334066 s] [54].

B. Detection criterion and efficiency

We analyze the injection sets with a simplified infrastructure based on GSTLAL, one of the matched-filter-based search pipelines presently used to search for GWs from compact binaries [44,45,47,55–64]. Matched-filter-based search pipelines discretely sample the GR-based signal manifold to create template banks of possible signals. The discretization results in a 1%–3% loss of SNR over the parameter space covered by the bank [65,66]. Pipelines presently restrict their searches to emission from sources with spin angular momenta aligned with the orbital angular momenta and, therefore, neglect the impact of precession or higher-order angular modes; the signal loss incurred for these systems is, therefore, larger. We specifically consider the GSTLAL-based matched-filtering pipeline for its signal consistency check and because it most densely sampled the signal space in LIGO-Virgo's third observing run (O3) and, thus, had the minimum expected SNR loss from discreteness. For BBHs, the GSTLAL bank used an effective-one-body model of the GW emission, SEOBNRv4_ROM [67]. The specific structure and maximum SNR loss of GSTLAL's template bank is described in Table II of the GWTC-2 publication [68].

Pipelines correlate waveforms from the template bank with the data collected in each detector to produce an SNR time series. Peaks in the SNR time series, called triggers, are checked for coincidence across detectors and are then ranked according to the pipeline's detection statistic. GSTLAL's ranking statistic is the likelihood ratio $\mathcal{L}$, defined in [47,69], which relates the probability of observing a set of parameters under the signal hypothesis to that of the instrumental-noise hypothesis. This quantity is a function of a number of factors: the set of instruments participating in a detection, the matched-filter SNR, a signal-based-veto parameter, the event time and phase in the frame of each detector, and the masses and spins of the identifying template. In general, it is computationally expensive to accurately estimate the background of the search and recover simulated signals via $\mathcal{L}$. Since no background for O3 is publicly available, and to minimize the analysis cost, we instead employ an approximate detection statistic $\bar{\rho}$ that weights the measured SNR by a signal consistency check [70,71], namely,

[1]In GR, the coefficients corresponding to the -1PN and 0.5PN terms are exactly zero. $\delta\varphi_{-2}$ and $\delta\varphi_1$ therefore represent absolute deviations.

$$\bar{\rho} = \frac{\rho}{[\frac{1}{2}(1+\max(1,\xi^2)^3)]^{1/5}}, \tag{1}$$

where ρ is the matched-filter SNR and ξ^2 is a signal consistency test defined from the autocorrelation as

$$\xi_j^2 = \frac{\int_{-\delta t}^{\delta t} \mathrm{d}t |z_j(t) - z_j(0) R_j(t)|^2}{\int_{-\delta t}^{\delta t} \mathrm{d}t (2 - 2|R_j(t)|^2)}, \tag{2}$$

where z_j and R_j denote the complex SNR and autocorrelation of template j, respectively, and the integrand in the denominator is the expectation value in Gaussian noise [44]. We compute a value of ξ^2 for each trigger by integrating Eq. (2) over a small window of time $\pm\delta t$, centered about the trigger. We use $\delta t = 0.17$ s ($\delta t = 0.34$ s) for templates with chirp masses greater (less) than $15 M_\odot$, which was also done in production by the full GSTLAL pipeline. When the observed strain data closely match the template j, then $\bar{\rho} = \rho$. For each BBH injection, we compute the matched-filter SNR ρ and ξ^2 value against the GSTLAL template bank. Since signals generally match with multiple templates in a bank, we perform the same data reduction clustering as the GSTLAL pipeline does in GWTC-3. We discard triggers within 0.1 s of other triggers with a larger $\bar{\rho}$ value, breaking ties by ρ.

Since we consider the response in only a single detector, we conservatively set a detection threshold of $\bar{\rho} \geq 10$. This choice is motivated by the fact that significant candidates from GWTC-2 and GWTC-3 were identified for network SNR $\rho_{\mathrm{net}} \gtrsim 10$, which typically corresponded to events with single detector SNRs $\rho_{\mathrm{H}} \sim \rho_{\mathrm{L}} \sim 7$. As we only filter a single detector, we assert that a signal in a single detector with $\rho = 10$ will have approximately the same significance as a signal observed in multiple detectors with $\rho_{\mathrm{net}} = 10$. We further assert that our proxy detection statistic threshold is approximately equivalent to the false-alarm-rate (FAR) threshold of $\mathcal{O}(10^{-3}/\mathrm{yr})$ adopted in past tests of GR [5–7]. This choice is conservative for our study in that a weaker detection criterion could only reduce the *detection* bias; i.e., it could only increase the fraction of signals that are detected by the pipeline.

Although we use an abbreviated version of the detection pipeline, we argue that the resulting selection function is a good approximation for the full selection effect for the following reasons.

(1) The threshold of $\bar{\rho} > 10$ selects triggers that are disjoint from the background typically collected by the search. Triggers that meet this criterion exist in the shaded contour shown in Fig. 1, which is cleanly off a representative background observed by the search. In other words, $\bar{\rho} > 10$ implies vanishing support from the background.

(2) In addition to the background, $\mathcal{L}$ contains a signal term that we do not explicitly take into account here. This is justified because, in the $\bar{\rho} > 10$ region, the noise distribution varies significantly more rapidly than the signal distribution (see Figs. 9 and 10 in [44]). Therefore, the contribution of the signal term to $\mathcal{L}$ is approximately constant over this region, and the FAR is mostly determined by the noise distribution.

(3) Finally, although $\mathcal{L}$ depends on parameters beyond ρ and ξ, namely, the event time, phase, mass, and spin, those should be minimally affected by the kinds of GR deviations that we consider here. Since the polarizations are unaffected by phasing corrections and the signals still propagate at the speed of light, the expected distribution of time delays and phase differences across detectors will remain the same. Regarding masses and spins, it is possible for non-GR signals to be identified by GR templates with masses and spins that differ from the source. Though this would change the population model's contribution, the model itself is broad (see Sec. IV B of the GWTC-2 publication [68]) and contributes weakly to the overall value of $\mathcal{L}$.

These three reasons justify our $\bar{\rho}$ criterion as a proxy for detecting signals with high significance.

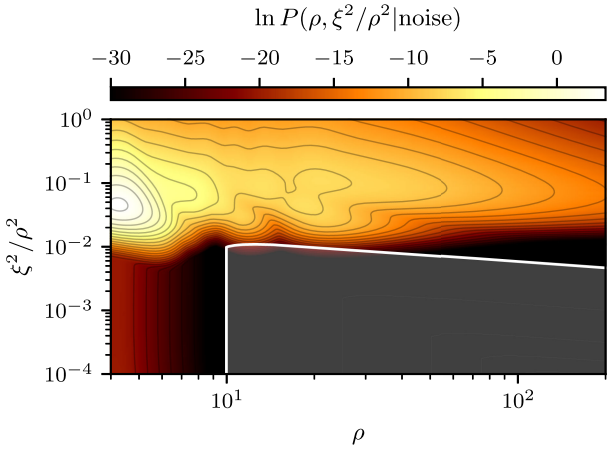

FIG. 1. A representative background distribution for BBHs collected for the LIGO Livingston detector. The background is parametrized in ξ^2/ρ^2 vs ρ space. Regions with high $\ln P$ indicate where noise is most likely (brighter color). The shaded contour enclosed by a white edge corresponds to our detection criterion, $\bar{\rho} \geq 10$. This region is largely separate from the collected background.

III. IMPACT ON DETECTION EFFICIENCY

To develop intuition for how deviations in the PN parameters affect the detection statistic, $\bar{\rho}$ in Eq. (1), Fig. 2 shows the SNR and autocorrelation time series with (right) and without (left) a deviation applied to the -1PN coefficient, $\delta\varphi_{-2}$, for a high (top) and low (bottom) injected SNR. We examine these two ingredients of the total

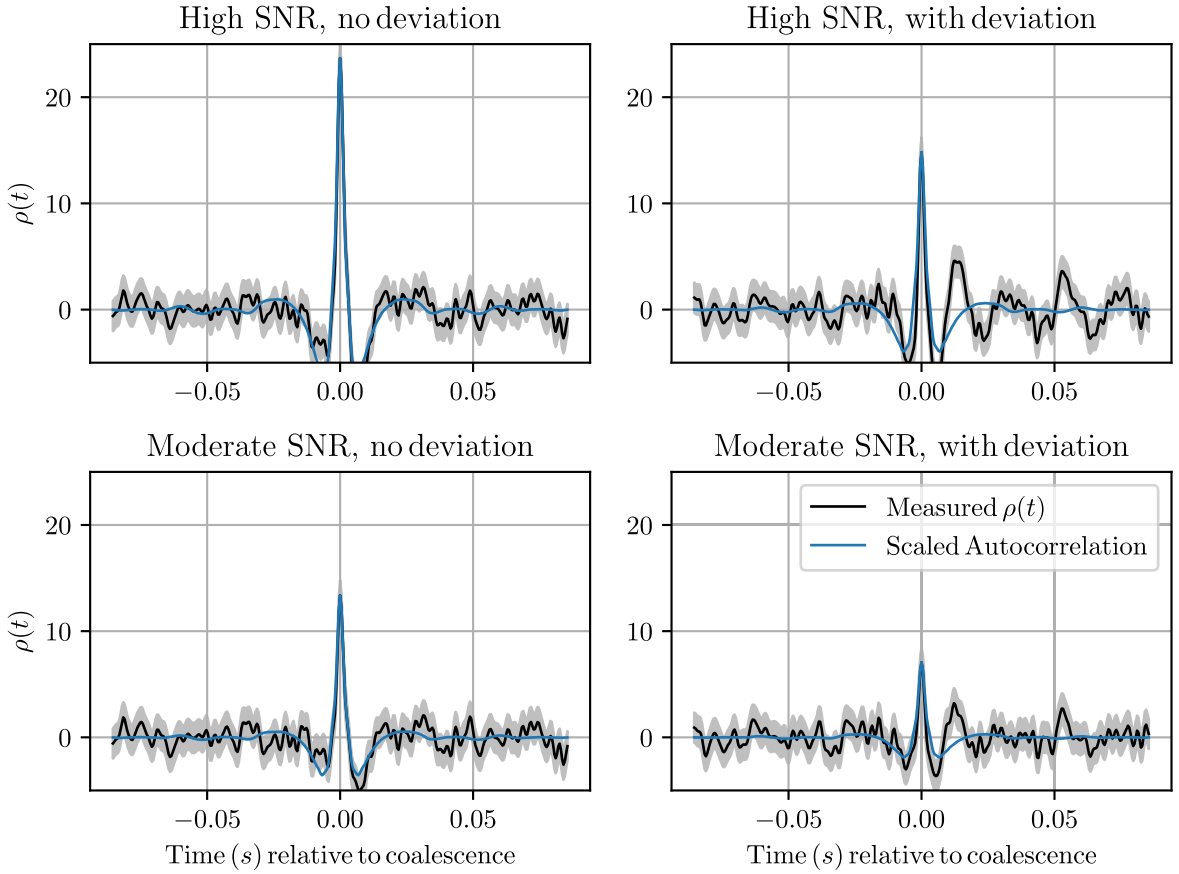

FIG. 2. The response of a single search template to a 30–30$M_\odot$ BBH without (left) and with (right) deviations to $\delta\varphi_{-2}$ for SNR ~ 24 (top) and ~ 15 (bottom) injections in Gaussian noise colored to O3 sensitivities. The injections that deviate from GR use $\delta\varphi_{-2} = -0.1$. The black line shows the measured SNR time series for a single template waveform, with the gray band denoting the 1σ measurement uncertainty. The beyond-GR phasing results in an SNR loss of $\sim$40% between the left and right columns. Additionally, there is a mismatch between the measured SNR time series and the SNR scaled autocorrelation that weakens the signal consistency test, ξ^2. Both effects lead to a reduction of our detection statistic $\bar{\rho}$, Eq. (1), and thus a loss in sensitivity.

detection statistic $\bar{\rho}$ for a characteristic BBH with redshifted masses 30–30$M_\odot$ in the detector frame. The two components of $\bar{\rho}$, ρ and ξ^2, are represented in these plots by, respectively, the peak of the SNR time series (black) and the integrated area between it and the scaled autocorrelation time series (blue). Mismatches between a signal and the template bank induced by a GR deviation will impact detection efficiency due to both a loss in the recovered SNR ρ (reduction in the peak height) and increase in the signal consistency check value ξ^2 (increased disagreement between blue and black curves).

Indeed, the beyond-GR deviation causes a reduction in the recovered SNR, seen through a reduced peak between the left and right panels of Fig. 2, thus directly affecting $\bar{\rho}$. Moreover, the introduction of beyond-GR effects creates secondary peaks in the SNR time series obtained from filtering with a GR waveform. The oscillations in SNR further reduce the signal consistency check, ξ^2—that is, the square difference between the measured SNR and the scaled autocorrelation, per Eq. (2). These oscillations become harder to discern from the Gaussian background with decreasing SNR, thus minimizing the effect of ξ^2 on the detectability of the signal. Figure 2 is helpful in understanding the interplay between ρ and ξ^2 in the presence of a deviation from GR. However, it is not sufficient to determine the degree of selection bias against beyond-GR signals, as it only shows the effect of a single injection relative to the corresponding GR template with the same parameters. In an actual search, we compare a beyond-GR injection against the entire bank, and the detection statistic is based on the best match.

To quantify the actual impact of GR deviations on the detection efficiency, we study the distribution of parameters of the signals that made it through our simplified detection pipeline, i.e., those that returned a value of $\bar{\rho} > 10$ when compared against *any* template in the GR bank. This amounts to measuring the detectable fraction:

$$\hat{\mathcal{E}}(\Lambda) = \int \mathrm{d}\theta\, p_{\mathrm{det}}(\theta)\pi(\theta|\Lambda), \tag{3}$$

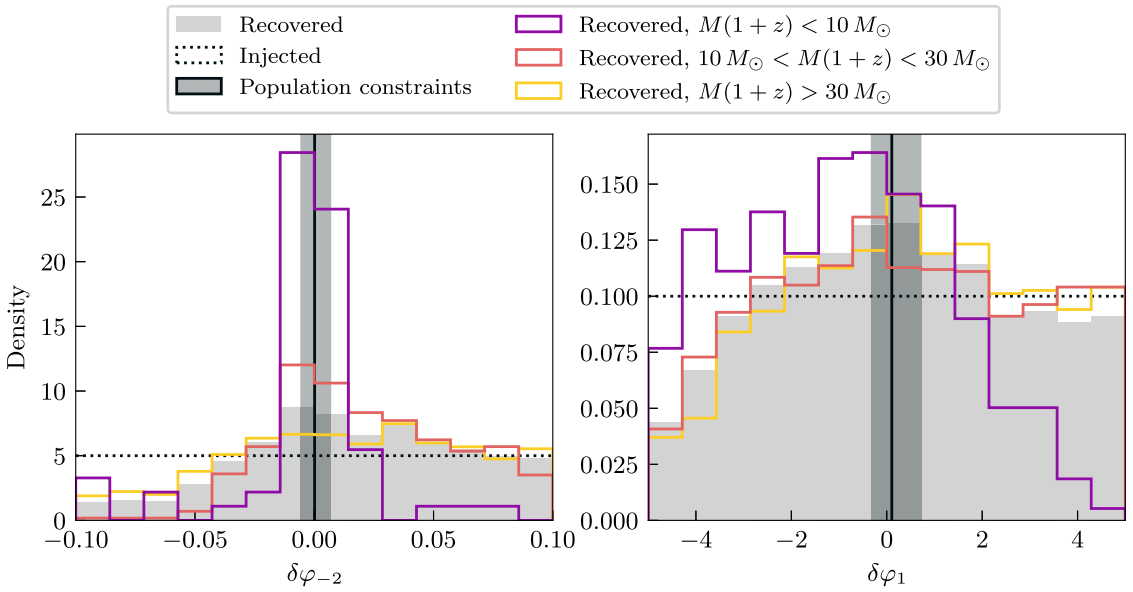

FIG. 3. Histograms of recovered injections with deviations from GR in the −1PN ($\delta\varphi_{-2}$, left) and 0.5PN ($\delta\varphi_1$, right) coefficients. Although the initial injection set was assigned deviations from a uniform distribution (dotted black), the pipeline selects against large negative values of the deviation parameters, as indicated by the dearth of detections in the leftmost bins (gray histograms). Besides the total set of injections, we show subdistributions corresponding to different injected mass bins in the detector frame (colored histograms). The distributions of recovered injections are largely flat over the span of values allowed by the analysis of the 12 events considered in Sec. IV (which are ∼4× broader than GWTC-3 constraints [7]; vertical gray band, median and 90% credible level), suggesting that the selection bias is not strong enough to affect the population constraints.

where Λ is the set of hyperparameters that describe the underlying population distribution, $\pi(\theta|\Lambda)$, and $p_{\rm det}(\theta)$ is the selection function that describes the probability of detecting a system with parameters θ. Figure 3 shows the marginal selection function, $p_{\rm det}(\delta\varphi)$, for the −1PN coefficient ($\delta\varphi_{-2}$, left) and the 0.5PN coefficient ($\delta\varphi_1$, right), over the whole mass space (gray) as well as subsections for different BBH mass bins (colors). For both parameters, the distribution of detected signals departs from the uniform intrinsic distribution that we injected (black): There is a dearth of detected signals with large negative values of the deviation parameters, indicating that such signals are selected against. This can be explained by the fact that a negative value for these parameters will shorten the inspiral, which in turn reduces the SNR of the signal. This effect is more pronounced for the −1PN coefficient, which is consistent with the intuition that this coefficient should have a larger impact on the GW phase than the 0.5PN coefficient over the duration of an inspiral because it is associated with a correction entering at a lower power of the frequency. The drop in detection efficiency is also sharper for lower masses, as expected given the scaling of the inspiral length with the BBH mass.

In spite of the drop in sensitivity observed at the edges of the histograms in Fig. 3, the recovered distributions are generally flat in the region that is allowed by the population constraints from GWTC-3 (gray band). Lower detector-frame masses demonstrate a larger gradient across these regions [e.g., $M(1+z) < 10M_\odot$; purple]. However, the observed events considered here do not reside in this region of the mass parameter space. Since there is no gradient in the region allowed by the observations, there is no preference for any particular value of the deviation parameter in the range still consistent with current data. This suggests that the selection bias is not strong enough to affect the population constraints, which are more sensitive to GR deviations than the detection pipeline. We confirm this below by repeating catalog analysis of GR deviations with and without the selection effects.

IV. UPDATED POPULATION ESTIMATES

We incorporate the selection function computed from Sec. III into population-level inference for inspiral tests of GR. By computing the astrophysical distribution of beyond-GR parameters, we can now make statements about the types of GR deviations consistent with an observed set of detections. In practice, computing the astrophysical distribution requires incorporating knowledge of the detection efficiency over parameter space to deconvolve the instrument's selection function from the set of observed measurements.

We evaluate the consistency of a set of observations with GR through a hierarchical analysis without imposing strong assumptions about the nature of the deviation across events. As a null test, we follow [7,24,25,48] in parametrizing the intrinsic distribution of individual-event values for some deviation parameter $\delta\phi$ as a Gaussian $\delta\phi \sim \mathcal{N}(\mu, \sigma)$. This

model targets the mean μ and variance σ^2 of GR deviations, regardless of the true shape of the underlying distribution. Beyond-GR parameters are typically defined to vanish in GR, so that the null hypothesis that GR is valid for all events predicts $\mu = \sigma = 0$. If GR is not correct, then the deviation parameters may take different (nonzero) values as a function of source parameters, resulting in nonvanishing μ or σ. We apply the approach in [48] to simultaneously model the distribution of astrophysical parameters.

Existing implementations of this hierarchical analysis characterize the set of *observed* events but do not inform about possible *intrinsic* deviation distributions that predict events with such large deviations that are undetectable. To factor this in, we use the result of Sec. III following the techniques used in the context of astrophysical inference to study the astrophysical distribution of within-GR parameters, such as masses and spins. The key additional step is to incorporate the detection efficiency into the hierarchical likelihood through a term that can be approximated as the Monte Carlo sum population weights over a set of m detected injections with parameters θ_k [72–74]:

$$\hat{\mathcal{E}}(\Lambda) = \frac{1}{M}\sum_{k}^{m}\frac{\pi(\theta_k|\Lambda)}{p(\theta_k|\mathrm{draw})}, \tag{4}$$

where M is the total number of drawn injections (out of which m were detected) and $p(\theta_k|\mathrm{draw})$ is the probability of drawing parameters θ_k from the population adopted in the injection campaign, with $\Lambda = \{\mu, \sigma\}$, in addition to the parameters describing the astrophysical population of GR quantities (like masses and spins). The hierarchical likelihood, $p(\{d\}|\Lambda)$, governing the inferred astrophysical population from N observations with dataset $\{d\}$ is

$$p(\{d\}|\Lambda) = \frac{1}{\hat{\mathcal{E}}(\Lambda)^N}\prod_{i}^{N}\int \mathrm{d}\theta_i\, p(d_i|\theta_i)\pi(\theta_i|\Lambda), \tag{5}$$

where $p(d_i|\theta_i)$ are the individual event likelihoods. The selection function influences the inferred hyperparameters through its inclusion in Eq. (5).

In order to include an injection in the "detected" sum of Eq. (4), besides GSTLAL's detection threshold of $\bar{\rho} > 10$ from Sec. II B, we additionally require that the measured SNR in the inspiral satisfy $\rho_{\mathrm{insp}} > 6$. The latter corresponds to the selection criterion for estimating the inspiral PN coefficients in [5–7]. In order to avoid computing the inspiral SNR for each injection in the set, we approximate the fraction of SNR in the inspiral as a linear function of the detector frame total mass as in [48].

In addition to hierarchically modeling the beyond-GR astrophysical distribution, we incorporate population models for the within-GR population distributions. Due to a lower number of recovered injections than the standard set of injections used in population studies [74], we only infer the primary mass and mass ratio distributions jointly with the beyond-GR population, using the models outlined in Ref. [48]. We fix the spin distribution to be uniform in spin magnitude and isotropic about all possible spin orientations; the redshift distribution is consistent with the *maximum a posteriori power law found in Ref. [18].*

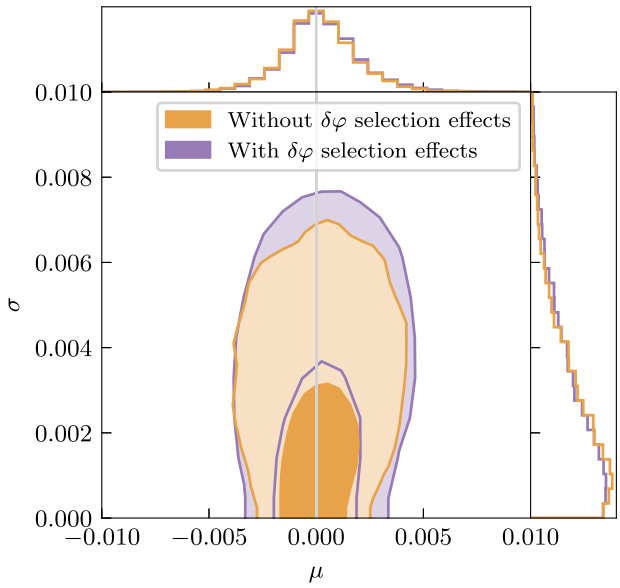

FIG. 4. Inference on the mean and standard deviation of the −1PN coefficient, $\delta\varphi_{-2}$. The orange contours show the result of the hierarchical analysis without accounting for selection effects, while the purple contours show the result when the selection function is included. The two results are consistent with each other, with the selection function widening the population only slightly. We find no difference in the coupling between μ and σ and the parameters controlling the mass distribution either (not shown).

With the setup described above, we repeat the hierarchical analysis in [6,7,24,48] applied to 12 events in O3a, to be consistent with times over which the selection function is estimated. A list of the included events can be found in Table I of Ref. [48]. Figure 4 shows the resulting inference on μ and σ for the −1PN coefficient, $\delta\varphi_{-2}$, compared to the result that does not account for selection biases in the beyond-GR parameters. Although this was the coefficient with the strongest detection bias as evaluated in the previous section (Fig. 3), this effect is very small, and the two results, with and without selection, are consistent with each other up to a slight widening of the population when selection is factored in. This is consistent with the expectation from Fig. 3, which suggested the impact of selection should be minimal in light of the accuracy of the constraint from parameter estimation. Figure 5 shows that this is the case for all coefficients, none of which show significant differences between the two results.

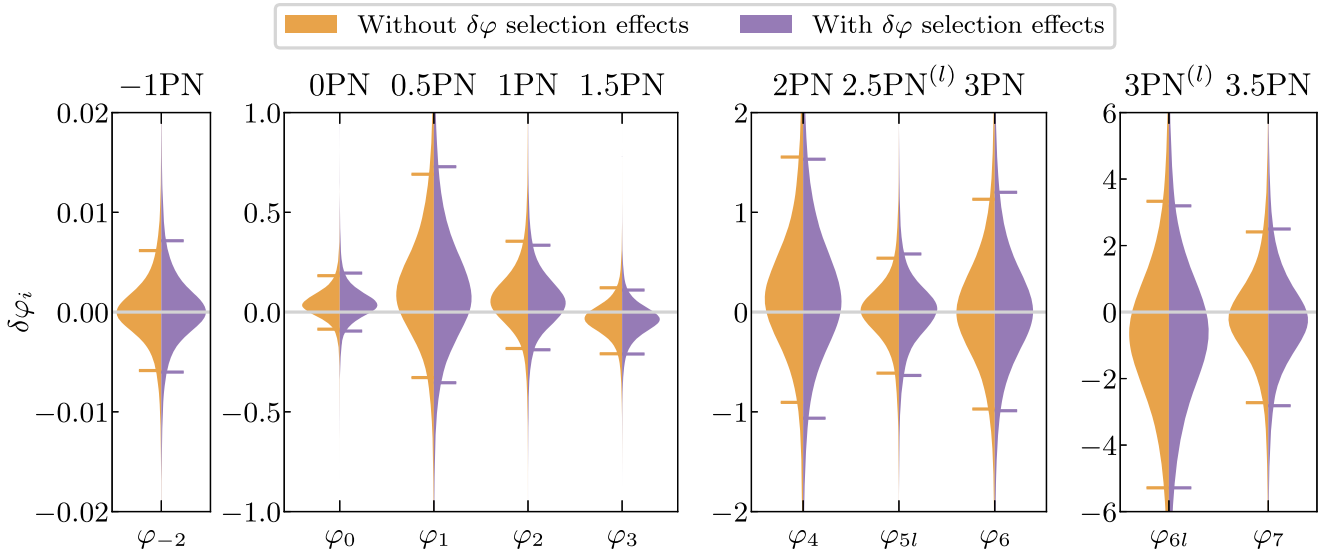

FIG. 5. Posterior predictive distributions (also known as the population-marginalized expectation) for deviations at all PN orders we consider, without (orange) and with (purple) selection effects factored in. No coefficient shows a significant impact when factoring in the selection: The $\delta\varphi_{-2}$ displays the strongest effect, with a slight broadening of the inferred distribution at the level of ∼10%.

V. CONCLUSIONS

In this study, we revisited tests of GR from the inspiral GW phase by accounting for the selection effect of templated searches against signals with GR deviations. We estimated the selection function by considering the performance of a simplified version of the GSTLAL search pipeline against simulated signals with beyond-GR effects affecting the PN evolution of a BBH inspiral. Since GSTLAL detects signals by comparing them to a template bank constructed with GR waveforms, its detection efficiency decreases under sufficiently large deviations from GR. However, we found that this threshold for deviations is less stringent than the precision of GWTC-3 constraints, suggesting that population inference on the inspiral deviation parameters is minimally affected by selection effects. In other words, existing constraints are already a very good approximation to the full astrophysical population of deviation parameters, apart from the possibility of a disconnected subpopulation of sources with very high deviations.

This finding can be understood by noting that the sensitivity of parameter estimation to deviations from GR scales inversely with the SNR of the signal, while the detection threshold imposed by the search pipelines is best represented as a hard SNR cutoff. A deviation $\delta\varphi$ that induces a mismatch $\mathcal{M}$ relative to the best-fitting GR template will result in an SNR loss of order $\rho \to \mathcal{M}\rho$; accordingly, the measurement precision in parameter estimation will scale as $\Delta(\delta\varphi) \sim 1/\rho$. For a given SNR, the mismatch tolerated by the search pipeline will be much higher than the sensitivity of the parameter estimation. Therefore, signals that incur an SNR penalty would still be detectable as long as they remain above the search's threshold; meanwhile, given a GR signal in the data, parameter estimation will constrain the magnitude of a deviation tightly around zero, with much better precision than would be directly associated with the pipeline's detection threshold.

In other words, the tolerance for detection is much larger than the tolerance for parameter estimation, and the latter is what determines the population constraints. Since the population of observed deviations is extremely narrow (a delta function at zero if GR is correct), the hierarchical measurement is minimally affected by selection effects, as we have shown in Fig. 5. This argument does not apply to other parameters, such as the BH masses, since their distribution is intrinsically broad.

Our main conclusion is that the deviation population is already narrower than the extent of the selection effects, and thus the latter do not impact the former. However, this assumes that deviations form a single, compact population whose mean and standard deviation we constrain. Since no observed events are inconsistent with GR, the inferred width of this population grows smaller as the catalog increases. We are therefore not considering, and thus not ruling out, disjoint populations with a subset of events that have extremely large (and potentially undetectable) deviations or a mass-dependent deviation population model. It remains conceivable that a subpopulation of signals with extremely high deviations could exist and remain hidden from GR-based pipelines, motivating dedicated searches [13–15]. However, that does not translate into selection biases for the components of the population that are already constrained by the existing catalog.

This distinction also suggests that there is no contradiction between our results and those of Refs. [13–15]: We both find appreciable selection effects for sufficiently large values of the deviation parameters; cf. Fig. 3. Our study, however, highlights that under the assumption of a single, unimodal population distribution of the deviation parameters, such large values of the deviation parameters are already ruled out.

As Essick and Fishbach [26] recently pointed out, the existence of prominent selection biases would complicate the interpretation of hierarchical constraints that do not factor in selection effects, as the inferred population would not be strictly representative of neither the true astrophysical distribution nor the observed distribution of parameters. However, in the absence of strong selection effects, hierarchical inference *without* a selection term remains a valid tool to constrain the population of beyond-GR parameters, as we have shown here for PN tests of the BBH inspiral. This, of course, may not be the case for other tests or implementations.

Our results are subject to a number of caveats, and selection effects might be stronger for different GR tests or population models. First, to mitigate computational costs, we have used an approximate ranking statistic that only incorporates information from a single detector. We impose a detection threshold of $\rho \geq \bar{\rho} \geq 10$ to maximize purity in accordance with the FAR threshold adopted in past GR tests [5–7]. We do not expect a full injection campaign utilizing the complete ranking statistic described in [47] would yield more precise results at this threshold and for the inspiral deviation test considered here. However, our results do not obviate the need for a full injection campaign for other tests of GR or other pipelines.

Besides the adopted threshold, the $\bar{\rho}$ ranking statistic differs from the full likelihood ratio also on the information it considers. The latter also includes information about the phase and time of the signal in different detectors. Though we do not expect those terms to be important for the inspiral deviation parameters we consider here, they could become important for other tests of GR, such as those considering propagation effects or the signal polarization. Quantifying selection effects for such tests would require a full multi-detector and likelihood ratio calculation.

We produce injected signals with GR deviations using standard infrastructure [28–30,51–53] and choose parameter ranges consistent with priors used in LIGO-Virgo-KAGRA publications. However, for some of these extreme values, the resulting waveform could become pathological [75] and may not represent a physically meaningful configuration [76]. Although this might affect the overall applicability and physical interpretation of the tests, it does not affect the interpretation of our results that relate to the selection effects of the tests as formulated. Reformulations of the inspiral tests to ensure the GW phase calculation remains in the convergent series expansion regime [76,77] would likely be affected by selection effects even less, as they restrict the allowed range of possible deviations.

Among the compact-binary pipelines, we restrict to a simplified version of GSTLAL. We expect the impact of this assumption to be small, as we only consider the most confidently detected BBHs with single detector SNRs $\gtrsim 10$, all of which are detectable by GSTLAL. If we decreased the SNR threshold, we might encounter events detected by other compact-binary pipelines, in which case we would need to quantify their selection effects. However, we expect that relaxing SNR or FAR thresholds should only make pipelines more tolerant to signals beyond GR.

Extending beyond matched-filter pipelines, we expect weakly modeled search methods [8,9] to surpass template-based ones for sufficiently large GR deviations. However, it is the case that both all events we consider here and all events that have been detected in general are detected significantly by at least one template-based search. Ultimately, the sensitivity of weakly modeled searches should also be quantified and taken into account, though some have started to explore the biases this would introduce [15].

As the sensitivity of GW detectors improves, so does the number and quality of detections, leading to increasing sensitivity to both subtle deviations from GR and systematics in our models. While here we have focused on tests of GR based on GW inspiral phases and single-Gaussian populations, exploring the effect of selection biases in other tests or under other population models will also become important. As both our detectors and techniques evolve, future studies need to evaluate this and other potential systematics.

ACKNOWLEDGMENTS

We thank Reed Essick for helpful discussions. We thank Leo Tsukada for discussions on GSTLAL. The Flatiron Institute is funded by the Simons Foundation. K. C. was supported by National Science Foundation (NSF) Grant No. PHY-2110111. G. P. gratefully acknowledges support from Royal Society University Research Fellowship No. URF\R1\221500 and No. RF\ERE\221015. LIGO was constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the National Science Foundation and operates under Cooperative Agreement No. PHY-1764464. This research has made use of data, software and/or web tools obtained from the Gravitational Wave Open Science Center, 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. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation.

The authors are grateful for computational resources provided by the LIGO Laboratory and supported by NSF Grants No. PHY-0757058 and No. PHY-0823459. This research has made use of data or software obtained from the Gravitational Wave Open Science Center, a service of LIGO Laboratory, the LIGO Scientific Collaboration, the Virgo Collaboration, and KAGRA. This paper carries LIGO Document No. LIGO-P2300381. The filtering was performed with the GSTLAL library [44–47], built on the LALSUITE software library [78]. Our hierarchical analysis utilizes NumPyro [79,80], JAX [81], ASTROPY [82–84], NumPy [85], and SciPy [86]. The plots shown in this work use MATPLOTLIB [87], seaborn [88], arViz [89], and corner [90].

[1] J. Aasi *et al.* (LIGO Scientific Collaboration), Advanced LIGO, Classical Quantum Gravity **32**, 074001 (2015).
[2] F. Acernese *et al.* (Virgo Collaboration), Advanced Virgo: A second-generation interferometric gravitational wave detector, Classical Quantum Gravity **32**, 024001 (2015).
[3] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Tests of general relativity with GW150914, Phys. Rev. Lett. **116**, 221101 (2016); **121**, 129902(E) (2018).
[4] N. Yunes, K. Yagi, and F. Pretorius, Theoretical physics implications of the binary black-hole mergers GW150914 and GW151226, Phys. Rev. D **94**, 084002 (2016).
[5] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Tests of general relativity with the binary black hole signals from the LIGO-Virgo catalog GWTC-1, Phys. Rev. D **100**, 104036 (2019).
[6] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Tests of general relativity with binary black holes from the second LIGO-Virgo gravitational-wave transient catalog, Phys. Rev. D **103**, 122002 (2021).
[7] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), Tests of general relativity with GWTC-3, arXiv:2112.06861.
[8] S. Klimenko *et al.*, Method for detection and reconstruction of gravitational wave transients with networks of advanced detectors, Phys. Rev. D **93**, 042004 (2016).
[9] N. J. Cornish, T. B. Littenberg, B. Bécsy, K. Chatziioannou, J. A. Clark, S. Ghonge, and M. Millhouse, BayesWave analysis pipeline in the era of gravitational wave observations, Phys. Rev. D **103**, 044006 (2021).
[10] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Observing gravitational-wave transient GW150914 with minimal assumptions, Phys. Rev. D **93**, 122004 (2016); **94**, 069903(A) (2016).
[11] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), All-sky search for short gravitational-wave bursts in the first Advanced LIGO run, Phys. Rev. D **95**, 042003 (2017).
[12] R. Abbott *et al.* (KAGRA, Virgo, and LIGO Scientific Collaborations), All-sky search for short gravitational-wave bursts in the third Advanced LIGO and Advanced Virgo run, Phys. Rev. D **104**, 122004 (2021).
[13] H. S. Chia and T. D. P. Edwards, Searching for general binary inspirals with gravitational waves, J. Cosmol. Astropart. Phys. 11 (2020) 033.
[14] H. S. Chia, T. D. P. Edwards, D. Wadekar, A. Zimmerman, S. Olsen, J. Roulet, T. Venumadhav, B. Zackay, and M. Zaldarriaga, In pursuit of love: First templated search for compact objects with large tidal deformabilities in the LIGO-Virgo data, arXiv:2306.00050.
[15] H. Narola, S. Roy, and A. S. Sengupta, Beyond general relativity: Designing a template-based search for exotic gravitational wave signals, Phys. Rev. D **107**, 024017 (2023).
[16] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Binary black hole population properties inferred from the first and second observing runs of Advanced LIGO and Advanced Virgo, Astrophys. J. Lett. **882**, L24 (2019).
[17] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Population properties of compact objects from the second LIGO-Virgo gravitational-wave transient catalog, Astrophys. J. Lett. **913**, L7 (2021).
[18] R. Abbott *et al.* (KAGRA, Virgo, and LIGO Scientific Collaborations), Population of merging compact binaries inferred using gravitational waves through GWTC-3, Phys. Rev. X **13**, 011048 (2023).
[19] W. James and C. Stein, Estimation with quadratic loss, in *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I (University California Press, Berkeley, California, 1961), pp. 361–379.
[20] D. V. Lindley and A. F. M. Smith, Bayes estimates for the linear model, J. R. Stat. Soc. Ser. B **34**, 1 (1972).
[21] B. Efron and C. Morris, Stein's paradox in statistics, Sci. Am. **236**, 119 (1977).
[22] D. B. Rubin, Estimation in parallel randomized experiments, J. Educ. Stat. **6**, 377 (1981).
[23] A. Zimmerman, C.-J. Haster, and K. Chatziioannou, On combining information from multiple gravitational wave sources, Phys. Rev. D **99**, 124044 (2019).
[24] M. Isi, K. Chatziioannou, and W. M. Farr, Hierarchical test of general relativity with gravitational waves, Phys. Rev. Lett. **123**, 121101 (2019).
[25] 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).
[26] R. Essick and M. Fishbach, DAGnabbit! Ensuring consistency between noise and detection in hierarchical Bayesian inference, arXiv:2310.02017.
[27] N. Yunes and F. Pretorius, Fundamental theoretical bias in gravitational wave astrophysics and the parameterized post-Einsteinian framework, Phys. Rev. D **80**, 122003 (2009).

[28] T. 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).

[29] T. 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: Further investigations, J. Phys. Conf. Ser. **363**, 012028 (2012).

[30] 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).

[31] N. Cornish, L. Sampson, N. Yunes, and F. Pretorius, Gravitational wave tests of general relativity with the parameterized post-Einsteinian framework, Phys. Rev. D **84**, 062003 (2011).

[32] T. 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).

[33] 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: Further investigations, J. Phys. Conf. Ser. **363**, 012028 (2012).

[34] L. Sampson, N. Cornish, and N. Yunes, Mismodeling in gravitational-wave astronomy: The trouble with templates, Phys. Rev. D **89**, 064037 (2014).

[35] L. Sampson, N. Cornish, and N. Yunes, Gravitational wave tests of strong field general relativity with binary inspirals: Realistic injections and optimal model selection, Phys. Rev. D **87**, 102001 (2013).

[36] 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).

[37] 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).

[38] A. Ghosh *et al.*, Testing general relativity using golden black-hole binaries, Phys. Rev. D **94**, 021101(R) (2016).

[39] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Binary black hole mergers in the first Advanced LIGO observing run, Phys. Rev. X **6**, 041015 (2016); **8**, 039903(E) (2018).

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

[41] 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, Classical Quantum Gravity **35**, 014002 (2018).

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

[43] T. Akutsu *et al.* (KAGRA Collaboration), Overview of KAGRA: Detector design and construction history, Prog. Theor. Exp. Phys. **2021**, 05A101 (2021).

[44] C. Messick *et al.*, Analysis framework for the prompt discovery of compact binary mergers in gravitational-wave data, Phys. Rev. D **95**, 042001 (2017).

[45] S. Sachdev *et al.*, The GSTLAL search analysis methods for compact binary mergers in Advanced LIGO's second and Advanced Virgo's first observing runs, arXiv:1901.08580.

[46] K. Cannon *et al.*, GSTLAL: A software framework for gravitational wave discovery, SoftwareX **14**, 100680 (2021).

[47] L. Tsukada *et al.*, Improved ranking statistics of the GSTLAL inspiral search for compact binary coalescences, Phys. Rev. D **108**, 043004 (2023).

[48] E. Payne, M. Isi, K. Chatziioannou, and W. M. Farr, Fortifying gravitational-wave tests of general relativity against astrophysical assumptions, arXiv:2309.04528.

[49] R. Abbott *et al.* (LIGO Scientific, Virgo and KAGRA Collaborations), GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run—O3 search sensitivity estimates, 10.5281/zenodo.5546676 (2021).

[50] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run, Phys. Rev. X **13**, 041039 (2023).

[51] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. J. Forteza, and A. Bohé, Frequency-domain gravitational waves from nonprecessing black-hole binaries. I. New numerical waveforms and anatomy of the signal, Phys. Rev. D **93**, 044006 (2016).

[52] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez Forteza, and A. Bohé, Frequency-domain gravitational waves from non-precessing black-hole binaries. II. A phenomenological model for the advanced detector era, Phys. Rev. D **93**, 044007 (2016).

[53] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Simple model of complete precessing black-hole-binary gravitational waveforms, Phys. Rev. Lett. **113**, 151101 (2014).

[54] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), Open data from the third observing run of LIGO, Virgo, KAGRA and GEO, Astrophys. J. Suppl. Ser. **267**, 29 (2023).

[55] C. Hanna *et al.*, Fast evaluation of multidetector consistency for real-time gravitational wave searches, Phys. Rev. D **101**, 022003 (2020).

[56] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, FINDCHIRP: An algorithm for detection of gravitational waves from inspiraling compact binaries, Phys. Rev. D **85**, 122006 (2012).

[57] B. Allen, A chi**2 time-frequency discriminator for gravitational wave detection, Phys. Rev. D **71**, 062001 (2005).

[58] T. Dal Canton *et al.*, Implementing a search for aligned-spin neutron star-black hole systems with advanced ground based gravitational wave detectors, Phys. Rev. D **90**, 082004 (2014).
[59] S. A. Usman *et al.*, The PyCBC search for gravitational waves from compact binary coalescence, Classical Quantum Gravity **33**, 215004 (2016).
[60] A. H. Nitz, T. Dent, T. Dal Canton, S. Fairhurst, and D. A. Brown, Detecting binary compact-object mergers with gravitational waves: Understanding and Improving the sensitivity of the PyCBC search, Astrophys. J. **849**, 118 (2017).
[61] T. Adams, D. Buskulic, V. Germain, G. M. Guidi, F. Marion, M. Montani, B. Mours, F. Piergiovanni, and G. Wang, Low-latency analysis pipeline for compact binary coalescences in the advanced gravitational wave detector era, Classical Quantum Gravity **33**, 175012 (2016).
[62] F. Aubin *et al.*, The MBTA pipeline for detecting compact binary coalescences in the third LIGO-Virgo observing run, Classical Quantum Gravity **38**, 095004 (2021).
[63] Q. Chu, Low-latency detection and localization of gravitational waves from compact binary coalescences, Ph.D. thesis, The University of Western Australia, 2017.
[64] T. Venumadhav, B. Zackay, J. Roulet, L. Dai, and M. Zaldarriaga, New search pipeline for compact binary mergers: Results for binary black holes in the first observing run of Advanced LIGO, Phys. Rev. D **100**, 023011 (2019).
[65] D. Mukherjee *et al.*, Template bank for spinning compact binary mergers in the second observation run of Advanced LIGO and the first observation run of Advanced Virgo, Phys. Rev. D **103**, 084047 (2021).
[66] S. Sakon *et al.*, Template bank for compact binary mergers in the fourth observing run of Advanced LIGO, Advanced Virgo, and KAGRA, arXiv:2211.16674.
[67] A. Bohé *et al.*, 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).
[68] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), 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).
[69] K. Cannon, C. Hanna, and J. Peoples, Likelihood-ratio ranking statistic for compact binary coalescence candidates with rate estimation, arXiv:1504.04632.
[70] J. Abadie *et al.* (LIGO Scientific and Virgo Collaborations), Search for gravitational waves from low mass compact binary coalescence in LIGO's sixth science run and Virgo's science runs 2 and 3, Phys. Rev. D **85**, 082002 (2012).
[71] S. Babak, R. Biswas, P. Brady, D. Brown, K. Cannon *et al.*, Searching for gravitational waves from binary coalescence, Phys. Rev. D **87**, 024033 (2013).
[72] W. M. Farr, Accuracy requirements for empirically-measured selection functions, Res. Not. AAS **3**, 66 (2019).
[73] S. Miller, T. A. Callister, and W. Farr, The low effective spin of binary black holes and implications for individual gravitational-wave events, Astrophys. J. **895**, 128 (2020).
[74] R. Essick and W. Farr, Precision requirements for Monte Carlo sums within hierarchical bayesian inference, arXiv:2204.00461.
[75] N. K. Johnson-McDaniel, A. Ghosh, S. Ghonge, M. Saleem, N. V. Krishnendu, and J. A. Clark, Investigating the relation between gravitational wave tests of general relativity, Phys. Rev. D **105**, 044020 (2022).
[76] 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).
[77] N. E. Wolfe, C. Talbot, and J. Golomb, Accelerating tests of general relativity with gravitational-wave signals using hybrid sampling, Phys. Rev. D **107**, 104056 (2023).
[78] LIGO Scientific and Virgo Collaborations, LALSUITE software (2018).
[79] D. Phan, N. Pradhan, and M. Jankowiak, Composable effects for flexible and accelerated probabilistic programming in NumPyro, arXiv:1912.11554.
[80] 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).
[81] 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).
[82] T. P. Robitaille *et al.* (Astropy Collaboration), Astropy: A community PYTHON package for astronomy, Astron. Astrophys. **558**, A33 (2013).
[83] A. M. Price-Whelan *et al.* (Astropy Collaboration) (Astropy Contributors), The Astropy Project: Building an open-science project and status of the v2.0 Core Package, Astron. J. **156**, 123 (2018).
[84] A. M. Price-Whelan *et al.* (Astropy Collaboration) (Astropy Project Contributors), The Astropy Project: Sustaining and growing a community-oriented open-source project and the latest major release (v5.0) of the core package, Astrophys. J. **935**, 167 (2022).
[85] C. R. Harris *et al.*, Array programming with NumPy, Nature (London) **585**, 357 (2020).
[86] P. Virtanen *et al.*, (SciPy 1.0 Contributors), SciPy 1.0: Fundamental Algorithms for Scientific Computing in PYTHON, Nat. Methods **17**, 261 (2020).
[87] J. D. Hunter, MATPLOTLIB: A 2d graphics environment, Comput. Sci. Eng. **9**, 90 (2007).
[88] M. L. Waskom, seaborn: Statistical data visualization, J. Open Source Softwaare **6**, 3021 (2021).
[89] R. Kumar, C. Carroll, A. Hartikainen, and O. Martin, Arviz a unified library for exploratory analysis of bayesian models in PYTHON, J. Open Source Softwaare **4**, 1143 (2019).
[90] D. Foreman-Mackey, corner.py: Scatterplot matrices in PYTHON, J. Open Source Softwaare **1**, 24 (2016).