Effect of ignoring eccentricity in testing general relativity with gravitational waves

Purnima Narayan, Nathan K. Johnson-McDaniel, and Anuradha Gupta
Department of Physics and Astronomy, The University of Mississippi, University, Mississippi 38677, USA

(Received 11 June 2023; accepted 9 August 2023; published 5 September 2023)

Detections of gravitational waves emitted from binary black hole coalescences allow us to probe the strong-field dynamics of general relativity (GR). One can compare the observed gravitational-wave signals with theoretical waveform models to constrain possible deviations from GR. Any physics that is not included in these waveform models might show up as apparent GR deviations. The waveform models used in current tests of GR describe binaries on quasicircular orbits, since most of the binaries detected by ground-based gravitational-wave detectors are expected to have negligible eccentricities. Thus, a signal from an eccentric binary in GR is likely to show up as a deviation from GR in the current implementation of these tests. We study the response of four standard tests of GR to eccentric binary black hole signals with the forecast O4 sensitivity of the LIGO-Virgo network. Specifically, we consider two parametrized tests (TIGER and FTI), the modified dispersion relation test, and the inspiral-merger-ringdown consistency test. To model eccentric signals, we use nonspinning numerical relativity simulations from the SXS catalog with three mass ratios (1, 2, 3), which we scale to a redshifted total mass of $80M_\odot$ and luminosity distance of 400 Mpc. For each of these mass ratios, we consider signals with eccentricities of $\sim$0.05 and $\sim$0.1 at 17 Hz. We find that signals with larger eccentricity lead to very significant false GR deviations in most tests while signals having smaller eccentricity lead to significant deviations in some tests. For the larger eccentricity cases, one would even get a deviation from GR with TIGER at $\sim$90% credibility at a distance of $\gtrsim$1.5 Gpc. Thus, it will be necessary to exclude the possibility of an eccentric binary in order to make any claim about detecting a deviation from GR.

DOI: 10.1103/PhysRevD.108.064003

I. INTRODUCTION

At present, general relativity (GR) is the most successful theory of gravity as it explains current astronomical observations and laboratory experiments [1,2]. GR has been rigorously tested over the years but no statistically significant deviation has been found yet when tested using solar system observations [1], binary pulsar observations [2–5], and gravitational-wave (GW) observations [6–10]. Testing GR with GWs from mergers of binary systems has a special significance since it allows us to probe gravity in the highly nonlinear and dynamical regime that is not probed by other tests. In these tests, one compares theoretical waveform models with the data collected by GW detectors such as LIGO [11] and Virgo [12]. Any disagreement between the models and the data may hint toward a possible deviation from GR (modulo any effects due to nonstationary and non-Gaussian noise in the data). Hence it is crucial to have waveform models that are as accurate as possible; i.e., they should include all known physics in GR and have systematic errors that are well below the statistical errors in the observations.

The current tests of GR carried out by the LIGO-Virgo-KAGRA Collaboration (LVK) that use waveform models that describe the entire signal are based on waveform models designed for coalescing binary black holes (BBHs) on quasicircular orbits and lack information on the eccentricity of the orbit. This is a reasonable choice since binaries formed through the isolated formation channel [13] get efficiently circularized by gravitational radiation [14,15] and hence are expected to have negligible eccentricities shortly before the merger when their GWs enter the frequency band of ground-based detectors. However, there are other pathways that can lead to a significant eccentricity at small binary separations. For instance, binary formation from primordial black holes (e.g., [16,17]), dynamical interactions in dense stellar environments such as galactic cores or globular clusters (e.g., [18–29]), active galactic nuclei (e.g., [30,31]), and the evolution of isolated triple systems (e.g., [32–34]). In these scenarios, the eccentricity could be as high as $\sim$1 at 10 Hz.

Various N-body simulations on the evolution of binaries in globular clusters (e.g., [22,27]) suggest that $\gtrsim$5% of binaries can have eccentricities > 0.1 when their GWs enter the advanced LIGO frequency window. This suggests that at least a fraction of binaries detected by LIGO-Virgo detectors will have non-negligible eccentricity. Recent analyses of data from GWTC-3 [35] events found evidence

of eccentricity [36–39] but are not able to distinguish between the effects of spin precession and eccentricity at present (see, e.g., [40]) since there are no waveform models including the merger-ringdown portion of the waveform that contain both effects. Moreover, it has been shown that inferred binary parameters will be biased if the detected binaries are on eccentric orbits [36,41,42], and nonnegligible residual eccentricity in the ground-based detector band can also mimic a significant deviation from GR [43,44].

In this paper, we study the effect of ignoring eccentricity while testing GR using some of the standard tests employed by the LVK. Specifically, we consider the test infrastructure for general relativity (TIGER) [45,46], flexible-theory-independent (FTI) [47], modified dispersion relation (MDR) [48], and inspiral-merger-ringdown (IMR) consistency [49,50] tests and check their response to simulated eccentric BBH GW signals in the LIGO-Virgo network at its forecast O4 sensitivity [51]. To simulate eccentric GW signals, we use numerical relativity (NR) waveforms from the simulating extreme spacetimes (SXS) catalog [52]. Our simulated observations have the following properties: all binaries are nonspinning, have three mass ratios ($q = 1$, 2, 3) and a redshifted total mass of $80M_\odot$, and are observed face-on at a luminosity distance of 400 Mpc. Moreover, for each mass ratio, we choose NR simulations (from [53]) where the binary's eccentricity is $\sim$0.05 and $\sim$0.1 at 17 Hz for our total mass of $80M_\odot$. For comparison, we also consider a quasicircular NR waveform for each mass ratio.

We found that, as expected, all quasicircular signals are consistent with GR at 90% credibility in all tests except for $q = 2$, 3 in the IMR consistency test. We find that the biases obtained in the IMR consistency can be attributed to the inclusion of higher modes in the analysis for the face-on signals we consider, though the same set of higher modes is used in both the simulated signal and the recovery waveform, and this bias is even present when the same waveforms are used for both the simulated signal and recovery. Ongoing studies [54] have found that these biases are only significant for binaries very close to face-on (or face-off). The signals with lower eccentricity show significant GR deviations in the TIGER and FTI tests for higher order post-Newtonian (PN) testing parameters while the higher-eccentricity signals show very significant GR deviations for almost all testing parameters. Both lower- and higher-eccentricity signals are found to be consistent with GR at 90% credibility for almost all testing parameters in MDR test. On the contrary, the higher-eccentricity signals show strong GR deviations in the IMR consistency test even in an analysis without higher modes. We also study the scaling of the posterior probability distributions of testing parameters with luminosity distance for a few cases. This is because increasing the distance leads to fainter signals which in turn lead to broader posteriors, so any GR deviation that is present might be lost in the statistical error. We found that we can still observe GR deviations at $\sim$90% credibility from eccentric signals placed at distances $\gtrsim$1.5 Gpc, $\gtrsim$1.2 Gpc, and $\gtrsim$0.5 Gpc in the TIGER, FTI, and MDR tests, respectively.

This paper is organized as follows: In Sec. II, we give the details of the four tests of GR we consider, and in Sec. III we give the specifics of our simulated observations. In Sec. IV, we discuss the results from our analysis, and we conclude in Sec. V. We use geometrized ($G = c = 1$) units throughout.

II. TESTS OF GR

The tests we consider are all based on waveform models for quasicircular BBHs in GR, viz., `IMRPhenomPv2` [55–57] (TIGER), `SEOBNRv4HM_ROM` [58,59] (FTI), and `IMRPhenomXPHM` [60] (IMR consistency and MDR). `IMRPhenomPv2` is a frequency-domain phenomenological model which only has the dominant $(l, m) = (2, \pm 2)$ modes in the coprecessing frame and a simple, single-spin model for precession. `SEOBNRv4HM_ROM` is a frequency-domain reduced-order model of a (time-domain) aligned-spin effective-one-body model that includes the $(2, \pm 1)$, $(3, \pm 3)$, $(4, \pm 4)$, and $(5, \pm 5)$ modes in addition to the dominant $(2, \pm 2)$ modes. `IMRPhenomXPHM` is a frequency-domain phenomenological model that improves the accuracy of `IMRPhenomPv2`, including two-spin precession and the $(2, \pm 1)$, $(3, \pm 3)$, $(3, \pm 2)$, and $(4, \pm 4)$ subdominant modes in the coprecessing frame. The latest LVK testing GR catalog paper [10] also uses `IMRPhenomXPHM` for the IMR consistency test and the version without higher modes (`IMRPhenomXP`) for the MDR test. The latest LVK testing GR catalog paper does not include the TIGER test, but the previous testing GR catalog paper [9] also uses `IMRPhenomPv2`. The LVK FTI analyses use `SEOBNRv4HM_ROM` (the version without higher modes) for most events, but the previous testing GR catalog paper [9] uses `SEOBNRv4HM_ROM` when applying FTI to signals with significant evidence for higher modes.

A. TIGER and FTI

The TIGER test [45,46] introduces parametrized deviations in the frequency-domain phase of the BBH signal. The version used in, e.g., [8,9] modifies the phase of the aligned-spin dominant mode `IMRPhenomD` waveform model [56], and then this modified phase is twisted up using GR spin precession (as in the unmodified `IMRPhenomPv2`) to obtain a modified version of the precessing `IMRPhenomPv2` waveform. There is a new `IMRPhenomXP`-based version of TIGER that was not ready for inclusion in [10] and is not yet publicly available. The parametrized deviations are introduced in the PN coefficients (φ_k, φ_{kl}) in the Fourier-domain inspiral phase

of the waveform (leaving off additive constants and phase and time shifts),

$$\Phi(f) = \frac{3}{128\eta v^5}\sum_{k=0}^{7}(\varphi_k v^k + 3\varphi_{kl} v^k \ln v) \tag{1}$$

(the factor of 3 in the log term is due to the definition of PN coefficients used in TIGER) as well as in the phenomenological coefficients in the late-inspiral and merger phases of the waveform (β_k and α_k). See Table I in [6] for a summary of the frequency dependence of these terms. In the frequency-domain phase expression, $\eta \coloneqq m_1 m_2/(m_1+m_2)^2$ is the symmetric mass ratio, where $m_{1,2}$ are the binary's individual masses, and $v \coloneqq (\pi M f)^{1/3}$, where $M \coloneqq m_1 + m_2$ is the binary's (redshifted) total mass and f is the GW frequency. Additionally, the logarithmic coefficients are only nonzero for $k \in \{5, 6\}$.

The `IMRPhenomD` phase is constructed to be C^1, so the changes to one of the lower-frequency portions of the phase affects the remainder of the phase due to the C^1 matching. For example, the deviations in the PN coefficients also affect the late inspiral and merger-ringdown portions of the signal. Denoting any of these coefficients by p_k, the deviation parameter $\delta\hat{p}_k$ is introduced by the replacement $p_k \to (1+\delta\hat{p}_k)p_k$, except for $\delta\hat{\varphi}_1$ which is zero in GR, so we just normalize by 0PN coefficient. Additionally, for the PN coefficients, the deviation parameter is normalized by the nonspinning portion of the coefficient, to prevent degeneracies in cases where the spins can cause a coefficient to vanish. The deviation parameters are all zero in GR.

While one expects all PN coefficients after a given order to be modified in an alternative gravity theory, we only vary one parameter at a time in our application of TIGER, as in the LVK catalog analyses [8,9]. We do this since one obtains uninformative results when allowing multiple parameters to vary simultaneously, as illustrated for GW150914 in [6], but one can still detect GR deviations that modify multiple PN coefficients (or other testing parameters) when varying a single one (even if this testing parameter is itself not modified), as illustrated in [46,61]. However, in the future, it will possible to constrain all PN coefficients at the same time with good accuracy using multiband observations of BBHs [62,63]. This is because degeneracies between parameters are removed when combining data from a ground-based detector such as Cosmic Explorer [64] and a space-based detector such as LISA [65], which improves the measurement of parameters. Principal component analysis is another method to perform multiparameter tests which is shown to be effective with observations from current [66,67] and future [68,69] GW detectors.

The FTI test [47] is similar to TIGER, except it only considers deviations in the PN coefficients and is applicable to any aligned-spin waveform model, though the current implementation of the higher-mode version is restricted to `SEOBNRv4HM_ROM`. Additionally, it tapers the deviations to zero above a given frequency instead of letting them affect the rest of the signal. FTI also normalizes the deviation parameter using the full PN coefficient, including the spin contributions, but we reweight the results to the TIGER convention as in the LVK analyses [8–10], for easy comparison.

B. Modified dispersion relation

The MDR test introduces a phenomenological dispersion relation, following [48], which gives a frequency-dependent propagation of GWs. Specifically, it considers

$$E^2 = p^2 + A_\alpha p^\alpha, \tag{2}$$

where E and p are the energy and momentum of the GWs, while A_α and α are phenomenological parameters that determine the strength of the GR deviation and the frequency dependence of the dispersion, respectively. For $\alpha = 0$ and $A_0 > 0$, this corresponds to the dispersion relation of a massive graviton. As discussed in [8], it is a good assumption to take the waveform close to the source to be that given by GR to a very good approximation, and the only modification to the waveform is due to the dispersive propagation. In general, this modification is an addition $\propto A_\alpha f^{\alpha-1}$ to the waveform's frequency-domain phase,[1] and the magnitude of the dephasing increases with distance (see, e.g., [8] for the specific expressions).[2] As in the LVK analyses (e.g., [10]), we consider $\alpha \in \{0, 0.5, 1.5, 2.5, 3, 3.5, 4\}$, where $\alpha = 2$ is omitted since there is no dispersion in this case. We also omit $\alpha = 1$ since the current implementation gives the logarithmic dephasing one gets using the particle velocity considered in [48], while the expression using the group velocity [71] (which it makes more sense to consider) gives a constant dephasing. Such a constant dephasing is detectable with waveforms including higher modes but is not implemented in the current implementation of the test in LALSuite [72]. Also as in the LVK analyses, we sample in an effective wavelength parameter (given in [8]) and consider the positive and negative A_α cases separately. We then combine together the results for the two different signs of A_α and reweight to a flat prior in A_α, as described in [8].

[1]While the MDR dephasing for $\alpha = 0$ has the same frequency dependance as the $\delta\hat{\varphi}_2$ TIGER/FTI and $\delta\hat{\alpha}_2$ TIGER testing parameters, the MDR dephasing affects the entire signal, while the TIGER and FTI dephasing is only restricted to certain frequencies.

[2]The exponent in Eq. (4) of [8] should be $1/(2-\alpha)$, as pointed out in [9]. Additionally, as in, e.g., [10], we use the TT + lowP + lensing + ext cosmological parameters from [70] in calculating the dephasing.

C. IMR consistency test

The IMR consistency test [49,50] checks the consistency of the low- and high-frequency portions of a BBH signal. The division between these portions of the signal is made at the median of the ($|m| = 2$) GW frequency of the innermost stable circular orbit (ISCO) of the final Kerr black hole [73] obtained from the GR analysis of the full signal. The LVK analysis uses a more involved procedure to obtain the cutoff frequency using the medians of the individual masses and spins. However, we found that this gives negligible differences (at most 1 Hz) compared to the more straightforward calculation we use. Thus, considering the dominant $(2, \pm 2)$ modes of the waveform, the low- and high-frequency portions of the signal correspond to the inspiral and postinspiral stages of the binary's coalescence.

The test assesses the consistency of the two portions of the signal by inferring the (redshifted) final mass M_f and spin χ_f from each portion, giving deviation parameters,

$$\frac{\Delta M_f}{\bar{M}_f} := 2\frac{M_f^{\mathrm{insp}} - M_f^{\mathrm{postinsp}}}{M_f^{\mathrm{insp}} + M_f^{\mathrm{postinsp}}}, \qquad \frac{\Delta \chi_f}{\bar{\chi}_f} := 2\frac{\chi_f^{\mathrm{insp}} - \chi_f^{\mathrm{postinsp}}}{\chi_f^{\mathrm{insp}} + \chi_f^{\mathrm{postinsp}}}, \tag{3}$$

where the "insp" and "postinsp" superscripts correspond to the low- and high-frequency portions of the signal. These deviation parameters should both be zero if the signal is consistent with the waveform model used in the analysis (which is a quasicircular BBH merger in GR in all current applications). The final mass and spin are computed as follows. One first performs parameter estimation analysis for each portion of the signal using a standard BBH waveform (in our case, `IMRPhenomXPHM`) parameterized by the binary's initial masses and spins. One then computes the final mass and spin using an average of fits to NR simulations [74–76].[3] As in [9,10], we reweight to a flat prior in the deviation parameters to obtain the final results.

III. SIMULATED OBSERVATIONS AND PARAMETER ESTIMATION SETUP

We consider simulated BBH observations in the LIGO-Virgo network with the forecast O4 sensitivity [51]—we use the more sensitive LIGO noise curve and do not include KAGRA since it is expected to be much less sensitive than LIGO and Virgo in O4 [78]. We also do not include noise in our simulated observations (i.e., taking the zero realization of Gaussian noise) in order to avoid biases due to specific noise realizations. We model the BBH waveforms using a selection of nonspinning NR simulations from the SXS catalog [52]. In particular, all the eccentric waveforms are ones used in [53], since that paper gives the eccentricities at a fixed dimensionless frequency obtained by comparison with a PN waveform. We consider cases with eccentricities around 0.05 and 0.1, for comparison. We do not consider the higher eccentricity simulations from that paper, since they are not long enough to include all the power starting at 20 Hz for our chosen total mass of $80M_\odot$. We also consider quasicircular waveforms with the same mass ratios, for comparison. We give the properties of all the simulations we consider in Table I. We use the $N = 2$ extrapolated waveforms and the highest resolution simulation available in the SXS catalog.

TABLE I. The SXS simulations we consider and their properties. All of these simulations have negligible spins. The eccentricities given are those from [53], which are quoted at a PN velocity squared of 0.075 (so an $|m| = 2$ GW frequency of ∼17 Hz for the $80M_\odot$ binaries we consider), except for the ones that give an upper bound of 10^{-4}. For these, we quote an upper bound that is greater than the eccentricities quoted in the SXS metadata. Those eccentricities come from the eccentricity reduction procedure, which is not designed to measure nonzero values of eccentricity.

ID	Mass ratio	Eccentricity
SXS:BBH:1155	1	$< 10^{-4}$
SXS:BBH:1355	1	0.053
SXS:BBH:1357	1	0.097
SXS:BBH:1222	2	$< 10^{-4}$
SXS:BBH:1364	2	0.044
SXS:BBH:1368	2	0.097
SXS:BBH:2265	3	$< 10^{-4}$
SXS:BBH:1371	3	0.055
SXS:BBH:1373	3	0.093

We choose a (redshifted) total mass of $80M_\odot$ so that the simulation is long enough that the binary's entire signal is in the detectors' sensitive band starting from a Fourier frequency of 20 Hz. We consider a face-on signal (inclination angle 0) so that only the $m = 2$ [spin-(−2)-weighted] spherical harmonic modes of the signal contribute, and thus we do not have to worry about the $m > 2$ modes, which would require significantly longer NR simulations in order for the signal to include all the power down to 20 Hz without a much higher total mass ($\sim 120M_\odot$ for the current simulations even if only including the $|m| = 3$ modes). For the other extrinsic parameters, we choose a luminosity distance of 400 Mpc, similar to GW150914 [79], and a randomly chosen sky location (right ascension, declination of 3.19, −0.14 rad), polarization angle (1.53 rad), and coalescence GPS time (1129708949). We also consider selected cases at a larger distance, for comparison, as discussed in Sec. IV E. In GW data analysis terminology we often refer to these simulated GW observations as *injections*, which we will henceforth use in the paper.

[3] We augment the aligned-spin final spin fits with the contribution from in-plane spins [77], but as in [8–10], we do not evolve the initial spins before applying the fits.

We choose the same (spin-weighted spherical harmonic) mode content in the injections as the (coprecessing frame) modes present in the waveform models used in the tests, to avoid any biases due to missing modes. Since only the $m = 2$ mode contributes in our face-on case, that means that we just include the (2, 2) mode in the injections to which we apply TIGER and FTI tests (using `IMRPhenomPv2` and `SEOBNRv4HM_ROM`, respectively) and also include the (3, 2) mode in the injections to which we apply the MDR and IMR consistency tests (using `IMRPhenomXPHM`). The (3, 2) mode makes a ∼10% correction to the (2, 2) mode's amplitude; in general the maximum amplitudes of the $(\ell, 2)$ modes scale roughly as $10^{2-\ell}$ with respect to the (2, 2) mode, so the (4, 2) and higher modes make a ∼1% correction. Thus, we do not expect significant differences between our results and the results one would obtain when applying these tests to a real eccentric signal with the same parameters as our signals. We obtain network signal-to-noise ratios (SNRs) for our $(2,2)+(3,2)$ [(2, 2) only] injections of about 120 (116), 106 (103), and 90 (87) for mass ratios $q = m_1/m_2 = 1$, 2, and 3, respectively. The SNR values given are those for the lower-eccentricity injections, rounded to the nearest integer. The SNRs of the quasicircular and higher-eccentricity injections differ by < 1.

We perform the parameter estimation using the implementation of nested sampling [80] in the LALInference code [81] in the LALSuite software library [72]. We use a lower frequency of 20 Hz and upper frequency of 512 Hz in analyzing the injections, except for the different upper and lower frequencies used in the inspiral and postinspiral analyses, respectively, for the IMR consistency test. We use the same priors on the GR parameters as in the LVK applications of these tests, which are the same as the GR parameter estimation analyses (see, e.g., the discussion in Appendix E of [35]), except with larger ranges in some cases to account for correlations with non-GR parameters and with a prior on the luminosity distance that is uniform in Euclidean volume, instead of the more complicated prior uniform in comoving frame merger rate used in the LVK GR parameter estimation analyses. Specifically, we use uniform priors in redshifted masses and spin magnitudes as well as isotropic priors in spin directions, binary orientation, and sky location. The priors on the non-GR parameters are all flat.

IV. RESULTS

We now discuss the results obtained when performing the four tests of GR described in Sec. II on the simulated eccentric signals discussed in Sec. III.

A. TIGER

We give the posterior probability distributions (henceforth posterior distributions or posteriors) of the TIGER testing parameters for all our injections in Fig. 1. As expected, results from quasicircular injections for all three mass ratios are consistent with GR at 90% credibility. For the lower-eccentricity injections, we find that GR is excluded at $\gtrsim$90% credibility in almost all cases. The higher-eccentricity injections all show strong deviations from GR, with GR excluded at $> 90\%$ credibility (often well above this) for all three mass ratios and all testing parameters.

We summarize the statistical level at which GR is excluded in Fig. 2, giving the equivalent Gaussian sigmas. However, we find that GR is excluded at such high credible levels in some cases that we cannot trust that the GR quantile is estimated accurately with the $\sim 10^4$ posterior samples we obtain. In order to estimate an appropriate lower bound in such cases, we drew 1.8×10^4 samples from a Gaussian and compared the analytically computed Gaussian sigma values with the ones obtained using the same kernel density estimator (KDE) calculation applied to the results of the tests of GR. We chose this number of samples to be similar to (and on the lower side of) the number of samples we obtain for many of our analyses. We also varied the mean and standard deviation of the Gaussian to produce different GR quantiles and to reproduce the rough properties of the posteriors we obtain for the testing parameters. We found that Gaussian sigma values above around 3σ had absolute errors (comparing the KDE and analytic results) of more than 0.1, so we quote a lower bound of 3σ on significances.

We find that GR is excluded at $> 3\sigma$ for all testing parameters for the $q = 1$ higher-eccentricity injections. GR is also excluded with $> 3\sigma$ for the $q = 2$, 3 higher-eccentricity injections with the exception of $\delta\hat{\varphi}_6$, $\delta\hat{\varphi}_7$, and $\delta\hat{\beta}_2$ for $q = 2$ and $\delta\hat{\beta}_2$ for $q = 3$, though in all of these cases GR is excluded at $> 2\sigma$ and close to 3σ in some cases. The lower-eccentricity $q = 1$ and $q = 2$ injections exclude GR at $< 3\sigma$ for all testing parameters with the exception of $\delta\hat{\beta}_3$, where it is excluded at $> 3\sigma$. For the $q = 3$ lower-eccentricity injection, GR is excluded at $> 3\sigma$ only for the $\delta\hat{\varphi}_{5l}$, $\delta\hat{\varphi}_6$, $\delta\hat{\varphi}_7$, and $\delta\hat{\beta}_2$ testing parameters.

In Fig. 1, we notice that the sign of the deviation from GR for a given mass ratio and eccentricity is different for different testing parameters. This is due to the PN coefficients and phenomenological parameters used in the normalization themselves having different signs. We also see that all the testing parameters are on the opposite sides of zero for the lower- and higher-eccentricity injections for $q = 1$. This can be attributed to these cases being well outside of the linear regime of the test's response to eccentricity, and these cases generally have significantly different sets of GR and non-GR parameters giving the best agreement with the observed signal for the two eccentricities. For instance, for $\delta\hat{\varphi}_0$ and $q = 1$, the chirp mass is biased to larger values for the smaller eccentricity and smaller values for the larger eccentricity, while for the

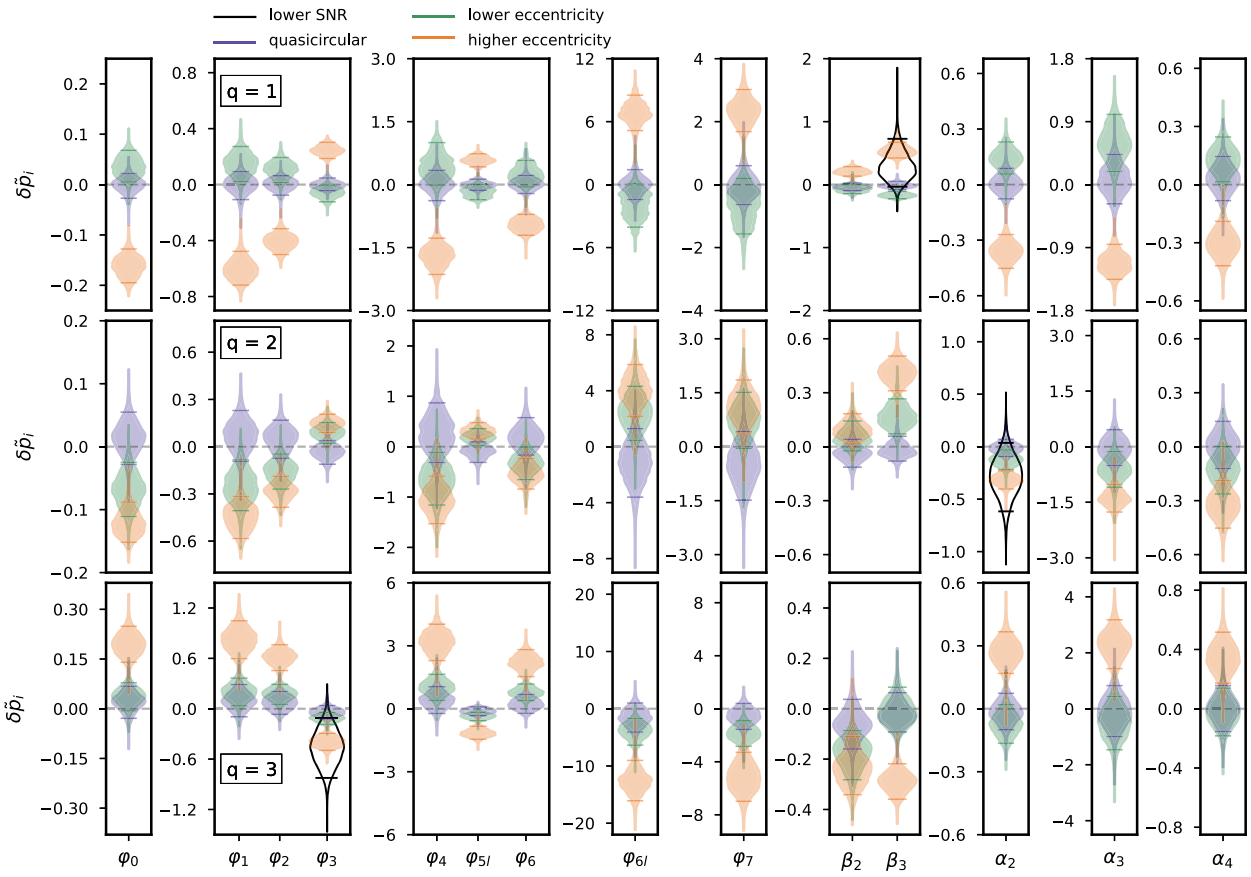

FIG. 1. The results of the TIGER test on the quasicircular, lower-eccentricity, and higher-eccentricity simulated injections of mass ratios 1, 2, and 3 in the top, middle and bottom panels, respectively. The posteriors of the testing parameters are presented as violin plots and the associated 90% credible intervals are labelled as horizontal bars. We mark the GR value of zero with dashed lines. We also show the results for the lower SNR injections, simulated by scaling the distance of selected higher-eccentricity runs for each mass ratio, as black unfilled violin plots. Details about the lower SNR cases are given in Sec. IV E.

same testing parameter and $q = 2$, the bias in the GR parameters generally increases monotonically from the smaller to the larger eccentricity, as does the value of the testing parameter.

Additionally, we find that the sign of a given testing parameter is different for different mass ratios. For the higher-eccentricity cases, there is generally a monotonic dependence of the value of the testing parameter on mass ratio, but for the lower-eccentricity cases, the signs of the PN coefficient testing parameters are the same for $q = 1$ and $q = 3$ and opposite those for $q = 2$. We investigate this difference in signs as follows: We first take the 3.5PN accurate quasicircular TaylorF2 nonspinning inspiral phase [82,83] and add the TIGER testing parameters at each PN order. We compare this phase with the eccentric PN inspiral phase from Moore *et al.* [84] which incorporates the effect of eccentricity to 3PN order (but is 3.5PN accurate in the quasicircular terms) and the leading-order (quadratic) terms in eccentricity. For each testing parameter, we obtain the value that minimizes the least-squares difference between the two phases with all the GR parameters fixed to the same values. However, we do not find any indication of a sign flip, and also find that this analysis returns values of the testing parameter of the order 10^{-3}, significantly smaller than what we find in the full analysis, suggesting that the merger-ringdown portion of the signal is quite important here. Nevertheless, we do find that the frequency derivative of the eccentric PN phase depends nonmonotonically on mass ratio (the ordering of the value by mass ratio is 1, 3, 2), showing that there is some nonmonotonicity present in the PN results.

We now compare our results with those from Saini *et al.* [43] which also studied the effect of ignoring eccentricity when performing the parametrized test of PN coefficients (though they do not consider the $\delta\hat{\varphi}_1$ testing parameter). The authors compare the expected value of the deviation parameters due to eccentricity using the formalism from [85], which is based on the Fisher matrix approach [86,87] that is used to obtain a prediction for the statistical errors in the deviation parameters. They only consider the inspiral portion of the signal and model the eccentric waveforms by adding the nonspinning eccentric contribution to the 3PN frequency domain phase from Moore *et al.* [84] to the aligned-spin 3.5PN GR phase [82,83,88,89], which is also

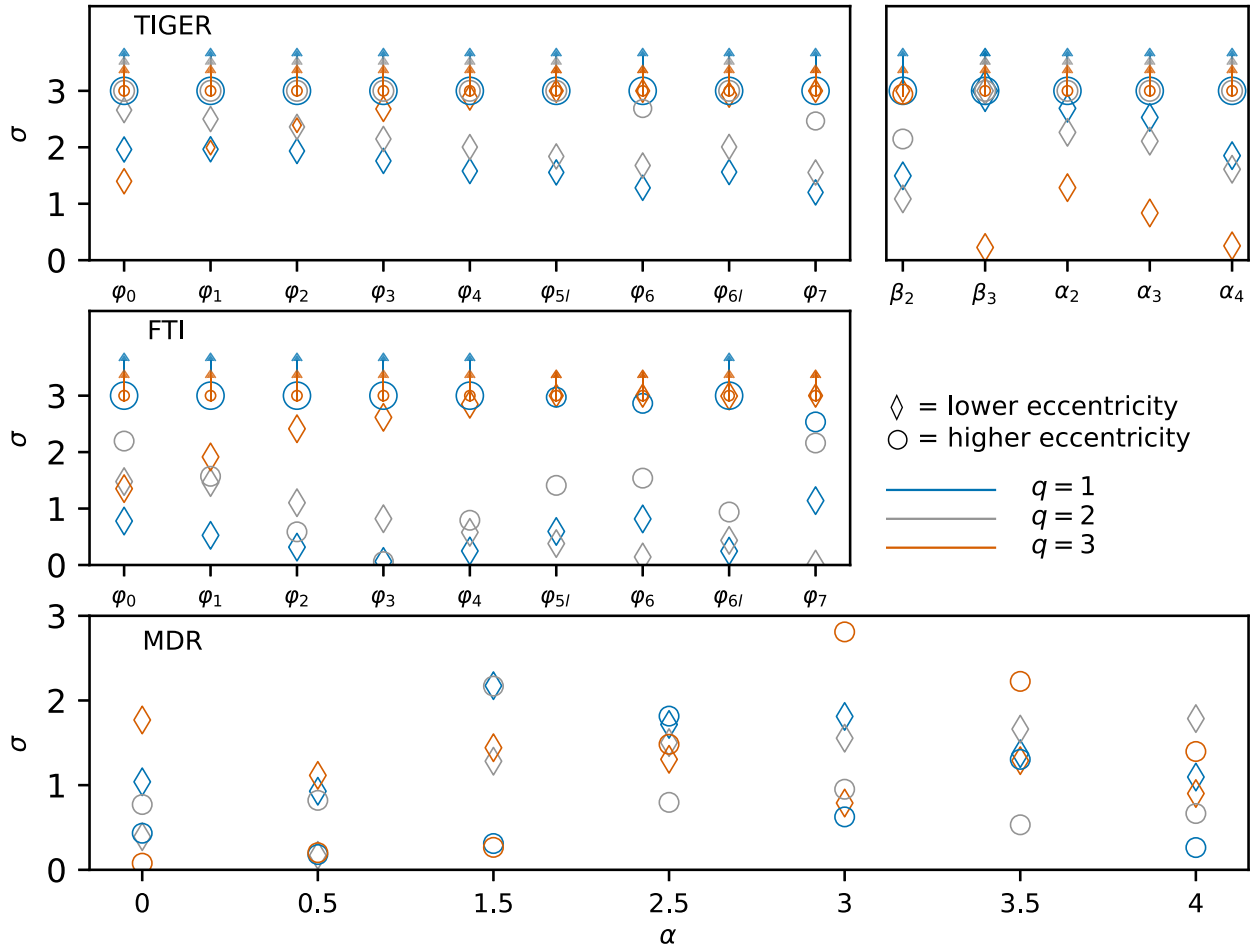

FIG. 2. The Gaussian sigma value at which GR is excluded for the eccentric injections with the TIGER, FTI, and MDR tests. A lower bound of 3σ is denoted using an upward arrow, above which the values cannot be stated with certainty from the order of 10^4 posterior samples in our analyses. The sizes of different markers and lengths of arrows have no significance and are varied to avoid overlaps with other data points as much as possible.

what they use to model the GR signals. Out of all the cases that Saini *et al.* considered, we make comparisons only for binaries with total redshifted masses of $72M_\odot$ and $99M_\odot$ (source-frame masses of $65M_\odot$ and $90M_\odot$) since they are the closest to the $80M_\odot$ used for our injections. Saini *et al.* only consider a mass ratio of 2, include aligned spins of 0.5 and 0.4, place the binaries at a distance of 500 Mpc, and specify the binaries' eccentricity at 10 Hz. Since we are using the eccentricity values from [53], which are given at the PN velocity squared of 0.075, corresponding to ∼17 Hz for our $80M_\odot$ binary, we use Eq. (4.17) in [84] to obtain an estimate of our lower and higher eccentricities at 10 Hz, giving 0.08 and 0.18, respectively, for the $q = 2$ case. We also have to convert the deviation parameters from the FTI convention (i.e., including the spinning terms in the PN coefficient used for the scaling) used by Saini *et al.* to the TIGER convention (i.e., only scaling by the nonspinning PN terms) that we use. We do this roughly using their injected parameters, which give scaling factors of 0.70, 0.75, 0.23, 0.029, and −1.2 (ratios of full to nonspinning PN coefficients) for $\delta\hat{\varphi}_3$, $\delta\hat{\varphi}_4$, $\delta\hat{\varphi}_{5l}$, $\delta\hat{\varphi}_6$, and $\delta\hat{\varphi}_7$. For the other cases, the PN coefficients have no dependence on the spin, so no rescaling is necessary.

Comparing the standard deviations of our posterior distributions of inspiral testing parameters for $q = 2$ to the statistical biases shown in Fig. 1 of Saini *et al.*, scaling their results to our SNRs, we found that the Saini *et al.* statistical errors are larger by a factor of ∼2–9 (smallest for $\delta\hat{\varphi}_3$ and largest for $\delta\hat{\varphi}_4$) with the exception of $\delta\hat{\varphi}_6$, $\delta\hat{\varphi}_{6l}$, and $\delta\hat{\varphi}_7$, where the Saini *et al.* statistical errors are smaller than our standard deviations by a factor of ∼3–5. (When making these comparisons here and below we always quote the smaller of the two differences between our results and the Saini *et al.* results for total masses of $72M_\odot$ and $99M_\odot$.) For these high-PN order coefficients, the smaller errors found by Saini *et al.* may be because they take the inspiral to extend up to the ISCO frequency of the final black hole ($Mf \simeq 0.06$ in their case), while the TIGER testing parameters are only applied up to the end of the `IMRPhenomD` inspiral phase ($Mf = 0.018$). Additionally, while we scale

the Saini *et al.* statistical errors to our SNR to make the results more comparable, there are many differences between our two analyses, so the significant differences we find are likely not unexpected. In particular, in addition to the differences in statistical methods and waveforms, Saini *et al.* also only consider a single LIGO detector with a slightly older noise curve and use a lower-frequency cutoff of 10 Hz, while we use 20 Hz.

We also compare the median of our $q = 2$ posteriors with the systematic biases shown in Fig. 1 of Saini *et al.* In general, we find that the agreement is better for the lower PN coefficients ($k \leq 4$). Specifically, for both eccentricities, we found that our median is contained within the Saini *et al.* range of systematic biases for the two total masses for $\delta\hat{\varphi}_0$ and $\delta\hat{\varphi}_4$. Additionally, for the larger eccentricity our median is only ∼5% smaller than the result for the larger total mass for $\delta\hat{\varphi}_2$ and ∼20% larger than the result for the smaller total mass for $\delta\hat{\varphi}_3$. For $\delta\hat{\varphi}_{5l}$, the Saini *et al.* systematic bias is ∼5 times (∼20%) larger than our median for the larger (smaller) eccentricity. For all the $\delta\hat{\varphi}_k$ not yet mentioned for a given eccentricity, the Saini *et al.* systematic bias is smaller than our median. For the larger eccentricity, the Saini *et al.* systematic bias is smaller by a factor of ∼10 for $\delta\hat{\varphi}_6$ and $\delta\hat{\varphi}_7$, while it is smaller by a factor of ∼20 for $\delta\hat{\varphi}_{6l}$. For the smaller eccentricity, the Saini *et al.* systematic bias is smaller by factors of ∼2 (for $\delta\hat{\varphi}_3$) to ∼60 (for $\delta\hat{\varphi}_{6l}$).

It is also useful to compare the statistical level at which GR is excluded in the two analyses. Here we can compare the ratio of the systematic bias to statistical error (scaled to our SNR) from Saini *et al.* with the Gaussian sigma equivalents to our GR quantiles (as given in Fig. 2). We find that both analyses agree that there is a significant GR deviation for the larger eccentricity and $\delta\hat{\varphi}_0$, $\delta\hat{\varphi}_2$, and $\delta\hat{\varphi}_4$. However, the Saini *et al.* systematic biases are smaller or comparable to their (SNR scaled) statistical errors for the other PN coefficients and the higher eccentricity, as well as for the smaller eccentricity, even though we find significant ($> 2\sigma$) GR deviations for most testing parameters in those cases. Nevertheless, the Saini *et al.* study finds that one can obtain a significant GR deviation for an eccentricity of ∼0.1 for the smaller total masses (particularly $15M_\odot$) for which their inspiral-only analysis is more reliable, so their overall conclusions are in agreement with those of our study.

Finally, we consider the biases in the GR parameters: As mentioned above for $\delta\hat{\varphi}_0$, we find that for $q = 1$, the chirp mass is biased in different directions for the smaller and larger eccentricity cases; we also find that the bias is in the opposite direction with the $\delta\hat{\alpha}_k$ deviation parameters than with the other deviation parameters. This bias primarily comes from a bias in the total mass, and we also find that there are biases in the total mass and chirp mass for the other mass ratios, though the biases are in the same direction for both eccentricity values for almost all testing parameters, and for $q = 3$, the total mass is consistent with the injected value for the PN deviation parameters. We also find biases in the effective spin [90,91], and find significant support for nonzero values of the effective precession spin parameter χ_p [55,92] for many cases, particularly for the higher eccentricities and $\delta\hat{\varphi}_k$ deviation parameters. There are even no samples near $\chi_p = 0$ for a few testing parameters in the higher-eccentricity cases. However, we do not find that the cases with larger support for precession have smaller GR deviations, as one might think would be the case (i.e., that the precession was absorbing some of the GR deviation). In fact, we usually finds both more support for precession and a larger GR deviation when increasing the eccentricity, though we generally do not find correlations between the testing parameter and χ_p. However, we find significant correlations between the testing parameter and $\chi_{\rm eff}$. We also find biases in the distance (both to larger and smaller values) and inclination angle, particularly for the larger eccentricity cases.

B. FTI

Similar to Fig. 1, Fig. 3 displays the posterior distributions of FTI testing parameters as violin plots. As we anticipate, all quasicircular injections are consistent with GR at 90% credibility. The $q = 1$ and 2 lower-eccentricity injections are both consistent with GR at 90% credibility. However, for the $q = 3$ lower-eccentricity injection, GR is excluded at $> 90\%$ credibility for all but $\delta\hat{\varphi}_0$, with the credible level at which GR is excluded increasing with increasing PN order. For the higher-eccentricity injections, the testing parameters show significant deviations from GR for $q = 1$ and $q = 3$ while moderate to no deviation for $q = 2$, which is consistent with GR at 90% credibility for most testing parameters. We refer to Fig. 2 again for the corresponding GR quantiles for all of our injections. We find that for the $q = 1$ higher-eccentricity injection, GR is excluded at $> 3\sigma$ for all testing parameters, except for $\delta\hat{\varphi}_{5l}$, $\delta\hat{\varphi}_6$, and $\delta\hat{\varphi}_7$. For the $q = 2$ higher-eccentricity injection, there is consistency with GR at $< 2\sigma$, except for the $\delta\hat{\varphi}_0$ and $\delta\hat{\varphi}_7$ testing parameters, where GR is excluded at slightly above 2σ.[4] For $q = 3$, we again find that GR is excluded at $> 3\sigma$ for all testing parameters for the higher-eccentricity injection and for $\delta\hat{\varphi}_{5l}$, $\delta\hat{\varphi}_6$ for the lower-eccentricity injection. We see the same general monotonic dependence of the value of the testing parameters with mass ratio seen for TIGER, but do not see the difference in signs for the lower- and higher-eccentricity cases seen for $q = 1$ with TIGER.

Comparing the standard deviations and medians of our FTI posterior distributions with the results from Saini *et al.*

[4]We checked that the significant difference between the TIGER and FTI results for $q = 2$ is not due to the inclusion of higher modes in the FTI analysis by applying FTI for $\delta\hat{\varphi}_3$ with just the dominant $l = |m| = 2$ modes to an injection containing just these modes and found that the posterior still peaks at 0.

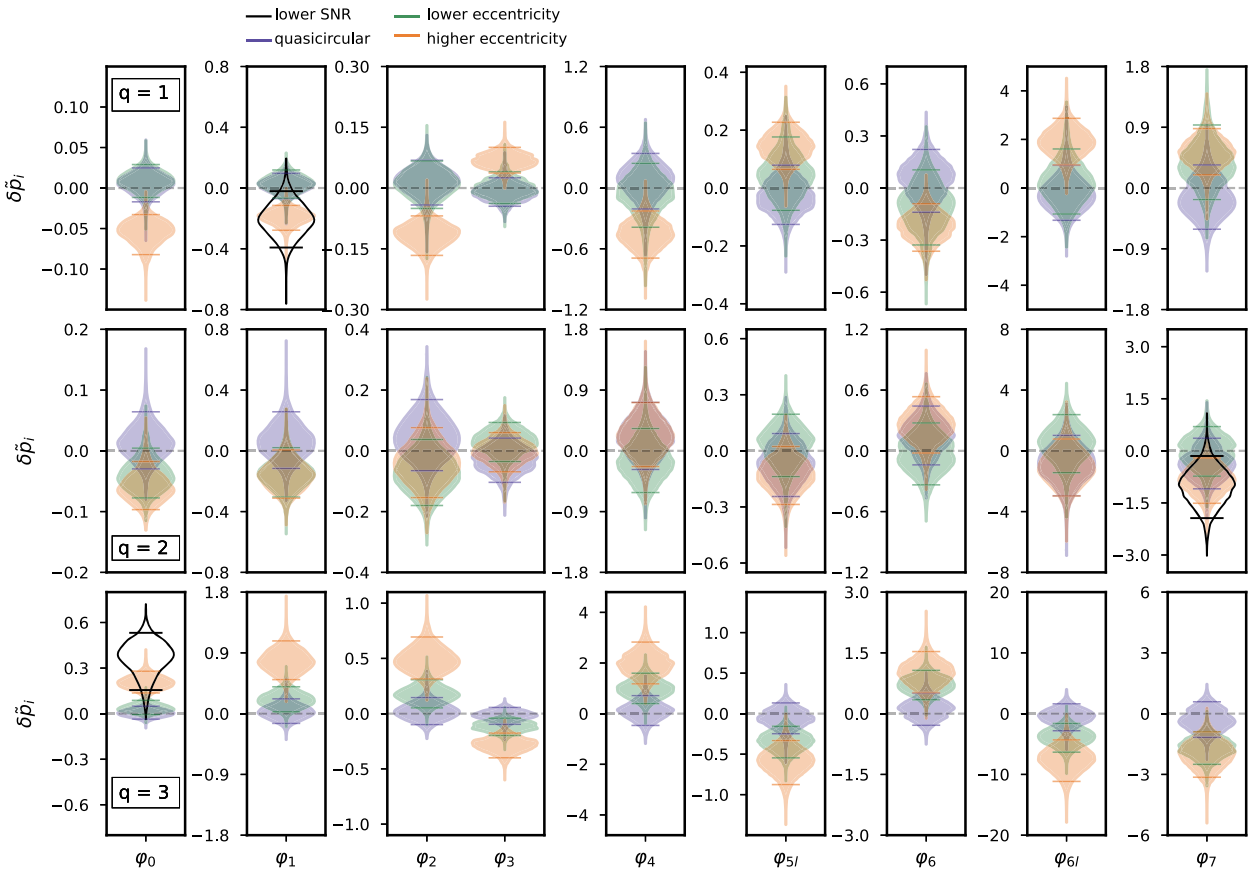

FIG. 3. Violin plots for FTI testing parameters. The color scheme and the layout of the subplots are similar to Fig. 1.

as we did for TIGER, we find the same relation between the statistical errors that we did with TIGER. The medians we obtain with FTI are all smaller in magnitude than those obtained with TIGER, so we find that for the smaller eccentricity they are contained between the values for the two (redshifted) total masses considered by Saini *et al.* ($72M_\odot$ and $99M_\odot$) for $\delta\hat{\varphi}_0$, $\delta\hat{\varphi}_2$, and $\delta\hat{\varphi}_4$. Our median is ∼2 times larger for $\delta\hat{\varphi}_3$ and $\delta\hat{\varphi}_6$, ∼30 times larger for $\delta\hat{\varphi}_{6l}$, and ∼6 (∼2) times smaller for $\delta\hat{\varphi}_{5l}$ ($\delta\hat{\varphi}_7$). For the larger eccentricity, only the median for $\delta\hat{\varphi}_0$ is between the two boundaries. For $\delta\hat{\varphi}_6$, $\delta\hat{\varphi}_{6l}$, and $\delta\hat{\varphi}_7$, our median is ∼5–8 times larger than the systematic bias obtained by Saini *et al.*, but in all the other cases our median is smaller than the Saini *et al.* systematic bias, only by a factor of ∼3 for $\delta\hat{\varphi}_4$, but up to a factor of ∼20 for $\delta\hat{\varphi}_3$. Since the FTI GR deviations are not as large as the TIGER ones for $q = 2$, there are not many cases as for TIGER where we find a significant GR deviation with FTI but Saini *et al.* find a systematic bias that is comparable to or smaller than the statistical error scaled to our SNR, though $\delta\hat{\varphi}_7$ larger eccentricity is a notable such case.

We find in general the same sorts of biases in the GR parameters as for TIGER, except mostly smaller (as is likely expected for the smaller parameter space of non-precessing systems considered here), and without any significant bias in the inclination angle. However, we do find that there are notable biases to larger total masses, more equal mass ratios, and larger distances for the $q = 3$ higher-eccentricity case.

C. MDR

Figure 4 illustrates the posteriors of the modified dispersion parameter $\tilde{A}_\alpha$ for different values of the modified dispersion relation exponent α. All the cases considered are consistent with GR at 90% credibility except for the $q = 2$, $\alpha = 1.5$; $q = 3$, $\alpha = 3$; and $q = 3$, $\alpha = 3.5$ higher-eccentricity and $q = 1$, $\alpha = 1.5$ lower-eccentricity cases where GR is excluded at $> 2\sigma$. However, GR is excluded at $< 3\sigma$ for all testing parameters as illustrated in Fig. 2. The lack of significant GR deviation in MDR test means that the modification to the waveform from the modified dispersion is mostly orthogonal to the modification introduced due to eccentricity (compared to the manifold of quasicircular waveforms). We find that in most cases, there is significant support for precession (usually well-constrained χ_p posteriors with no posterior samples near zero), except in a few lower-eccentricity $q = 2$ cases. These χ_p posteriors are generally considerably narrower than those for TIGER. There are also small biases in the effective spin (in both

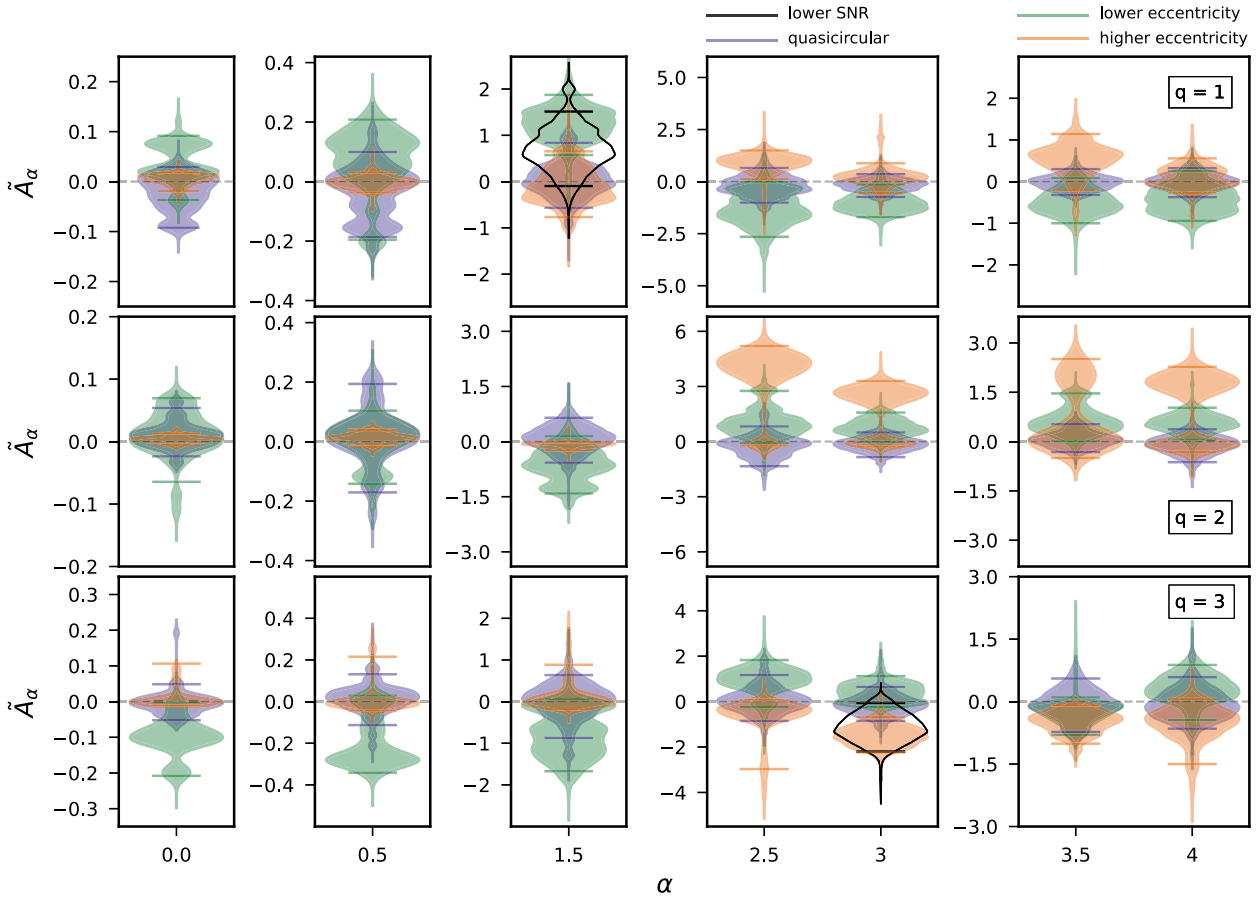

FIG. 4. Violin plots for the modified dispersion relation dimensionless parameter $\tilde{A}_\alpha \coloneqq 10^{43-12\alpha} A_\alpha/\mathrm{eV}^{2-\alpha}$ for different values of the modified dispersion relation exponent α, where the scaling is chosen to keep the plotted results of order unity. The color scheme and layout of subplots are similar to Fig. 1. The $q = 2$ lower-SNR result is omitted due to difficulty in obtaining reliable results.

directions) in some cases, as well as biases in the total mass to slightly larger values.

D. IMR consistency test

In Fig. 5 we show our results from the IMR consistency test.[5] We find a significant deviation from GR for all three mass ratios for the higher-eccentricity injections (GR quantiles of 99.9%, ∼100%, and 99.4% for $q = 1$, 2, and 3, respectively)[6] and also for the lower-eccentricity injections for $q = 2$, 3 (GR quantiles of 93% and 99.8%, respectively).[7] Surprisingly, we also find that $q = 2$, 3 quasicircular injections give noticeable GR deviations (GR quantiles of 96% and 99%, respectively). We find the same biases when applying the test to injections with the same parameters created using `IMRPhenomXPHM` instead of NR waveforms, so this is not due to waveform systematics. Thus, we suspect that this is due to the presence of higher modes in the recovery waveform model (i.e., `IMRPhenomXPHM`), since there are not yet extensive tests of the IMR consistency test using waveform models including higher order modes. However, while there are applications of the test to detected signals using `IMRPhenomXPHM` [10], the methods papers [49,50] do not consider waveform models with higher modes.

To verify this hypothesis, for these cases we applied the IMR consistency test to the (2, 2)-mode-only NR injections using `IMRPhenomXP`, which is the same as `IMRPhenomXPHM`, except it only has the $(2, \pm 2)$ modes in the coprecessing frame, as the recovery waveform model. We use the same cutoff frequencies as in the `IMRPhenomXPHM` analysis. We show the results from these analyses in Fig. 6. As expected, we now find that the quasicircular injections indeed agree with GR (as do the lower-eccentricity injections) while the higher-eccentricity injections still show a GR deviation (GR quantiles of

[5]The cutoff frequencies are $\{120, 119, 126\}$ Hz, $\{107, 107, 108\}$ Hz, and $\{95, 96, 94\}$ Hz for $q = 1$, 2, and 3, giving the values as {quasicircular, lower-eccentricity, higher-eccentricity}. These are obtained from a GR analysis of the signals using `IMRPhenomXPHM`.

[6]We have not checked the extent to which these large GR quantiles are reliable given the order of 10^4 posterior samples we have.

[7]Both inspiral and postinspiral analyses prefer equal masses, while the total mass is biased to lower values in the inspiral and to higher values in the postinspiral.

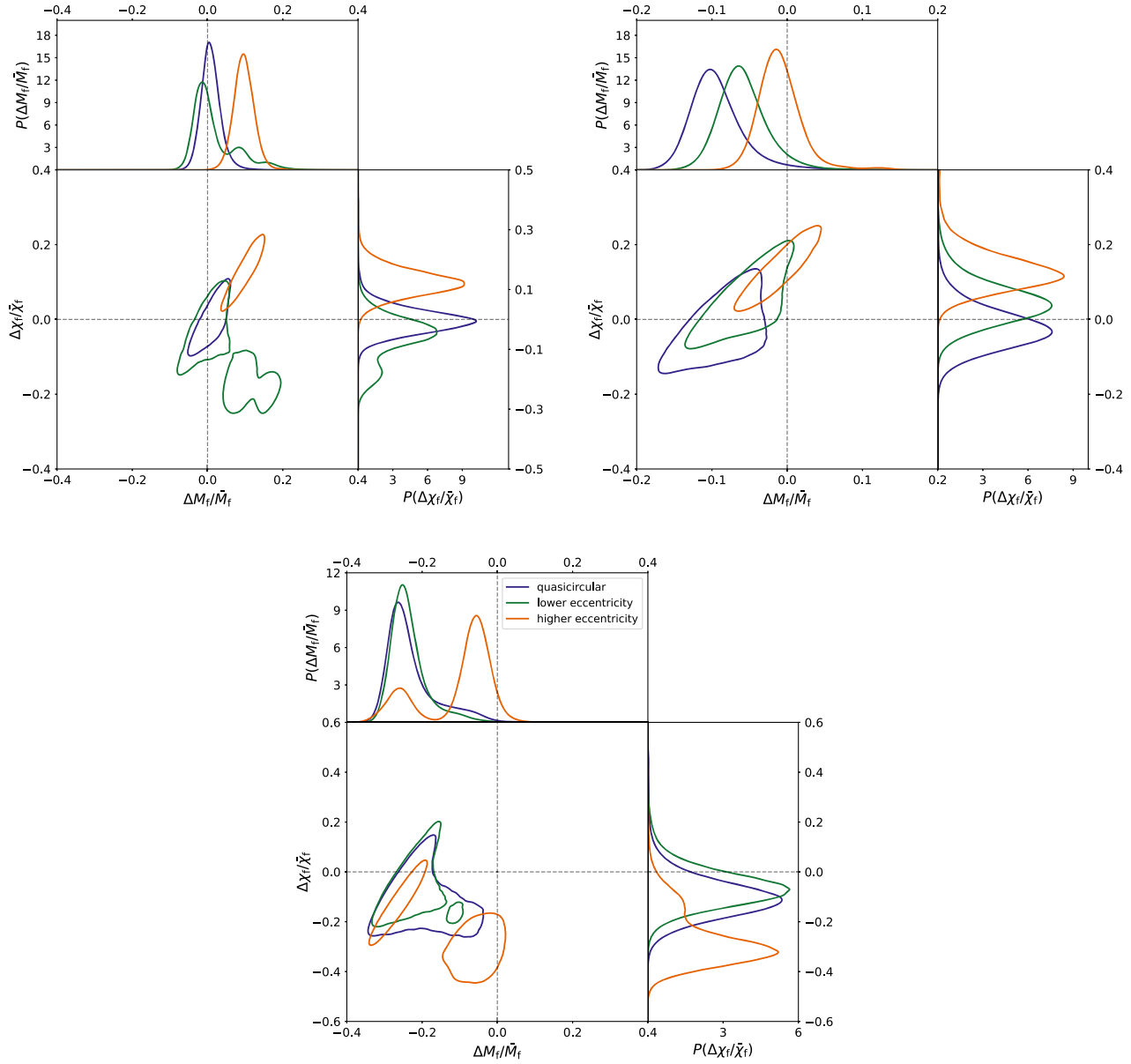

FIG. 5. The results from the IMR consistency test as the 90% credible regions of the joint posterior distributions of the recovered final mass and spin deviation parameters for quasicircular, lower-eccentricity, and higher-eccentricity injections for mass ratios 1 (top left), 2 (top right), and 3 (bottom). We also show the one-dimensional distributions for the marginalized deviation parameters. The color scheme is the same as in Fig. 1. Note that the range of the horizontal axes is smaller for the $q = 2$ plot than for the other two cases and the vertical axis range is larger for the $q = 3$ plot than for the other two cases.

~100% and 99% for $q = 2$ and 3, respectively). Therefore, we conclude that it is necessary to have a better understanding of the IMR consistency test when using waveforms with higher order modes. The studies necessary to obtain such a better understanding are currently underway, and they have already found that the bias is only significant when the binary is quite close to face-on (or face-off), though the bias does increase monotonically as the inclination varies from edge-on to face-on/face-off [54].

Bhat *et al.* [44] also studied the effect of missing eccentricity in the recovery waveform model when performing the IMR consistency test. They used the same Fisher matrix approach in this analysis as in the PN parameter analysis in Saini *et al.*, but here they use a full waveform model (the nonprecessing dominant mode `IMRPhenomD` model [56]), instead of just restricting to the PN inspiral waveform. They model the eccentric inspiral signal by adding the PN eccentric frequency domain phase contribution to the `IMRPhenomD` phasing. They also assume that the eccentricity has a negligible effect on the merger-ringdown part of the signal (as one expects will be the case for small eccentricities, since

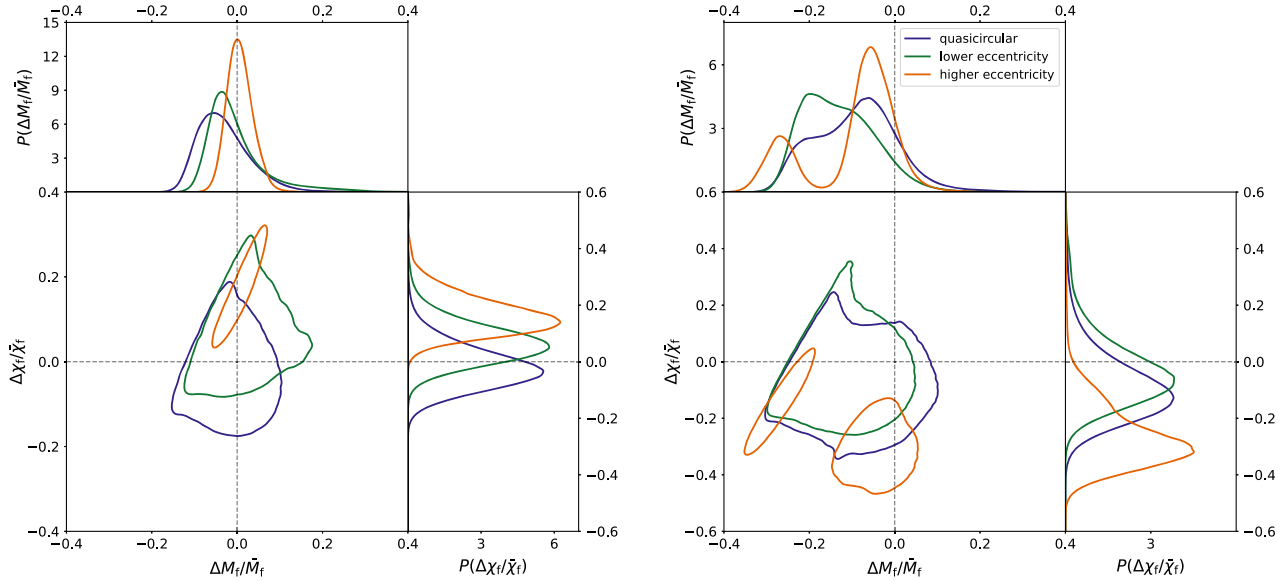

FIG. 6. The same as Fig. 5 except for just the $q = 2$, 3 cases and only including the $(2, \pm 2)$ (coprecessing frame) modes in the injections and recovery waveform model. Here the $q = 2$ plot now has the same horizontal axis range as the other plots (here and in Fig. 5) and the $q = 3$ plot has a larger range for the vertical axis.

eccentricity decreases during the inspiral), and ignored its effects on the mapping between the inspiral parameters and the final mass $M_{\rm f}$ and spin $\chi_{\rm f}$ of the merger remnant. Bhat *et al.* considered only a mass ratio of 2 (and aligned spins of 0.4 and 0.3) and specified the binary's eccentricity at 10 Hz. Thus, as described in Sec. IV A, we approximately obtain the eccentricity of our $q = 2$ injections at 10 Hz, giving $\sim$0.08 and $\sim$0.18 at 10 Hz.

We make a rough comparison between our $q = 2$ results for a (redshifted) total mass of $80M_\odot$ with the Bhat *et al.* results for redshifted total masses of $72M_\odot$ and $111M_\odot$ (source-frame masses of $65M_\odot$ and $100M_\odot$). While the $111M_\odot$ total mass is considerably further from our $80M_\odot$ total mass than the $72M_\odot$ mass is, we still consider it to bracket our total mass, particularly since the spins in the Bhat *et al.* signal make the dominant ($\ell = m = 2$, $n = 0$) quasinormal mode (QNM) frequency of the final black hole in the $111M_\odot$ case closer to the QNM frequency of the final black hole in our injections. Specifically, the QNM frequency in our $q = 2$ injections is 211 Hz (the same to this accuracy for all three simulations), while the QNM frequency for the $111M_\odot$ and $72M_\odot$ total mass Bhat *et al.* cases is 173 Hz and 267 Hz, respectively, so 38 Hz less and 56 Hz greater, respectively. We compute the QNM frequencies from the final masses and spins using the fit from [93]. For the Bhat *et al.* case, the final mass and spin are 0.950 times the total mass and 0.764, respectively (computed using the average of fits to NR results used in the IMR consistency test).

Since Bhat *et al.* just give ΔM_f and $\Delta\chi_f$, we divide these by the injected values of the final mass and spin (quoted above) for the purposes of this comparison. We compare our `IMRPhenomXP` results (see Fig. 6) and the results in Fig. 3 of Bhat *et al.*, which gives results for eccentricities of 0.08 and 0.15.[8] As in the comparison with Saini *et al.*, we scale their statistical errors to our SNRs (using the SNR of the entire signal, noting that the SNRs of the inspiral and postinspiral portions that Bhat *et al.* quote add in quadrature). We find that the scaled Bhat *et al.* statistical errors bracket the widths of the 68% credible intervals we obtain for the deviation parameters except for the larger eccentricity final mass, where our error is $\sim$30% smaller. Comparing our medians with the Bhat *et al.* systematic biases, we find that for the lower eccentricity, the median of our final mass posterior is about 2 times larger than the systematic bias found by Bhat *et al.*, but our final spin median is contained within the range for the two masses. For the larger eccentricity, our final mass median is $\sim$5 times smaller than the one from Bhat *et al.*, but only $\sim$30% smaller for the final spin, though this comparison underestimates the difference between the two results, since the Bhat *et al.* results with which we are comparing are for a somewhat smaller eccentricity (0.15 vs $\sim$0.18), and there is a relatively steep dependence of the systematic error on the eccentricity. As in the Saini *et al.* comparison, here we compare with the Bhat *et al.* total mass that gives the smaller differences (of the two total masses that bracket our total mass).

In general, we find that the Bhat *et al.* results qualitatively agree with ours. Specifically, we find that while the Bhat *et al.* statistical errors scaled to our SNR would give a

[8]The point for an eccentricity of 0.15 and a source-frame total mass of $65M_\odot$ is not included in the plotted region in Fig. 3 of Bhat *et al.*, but it has values of $\Delta M_f = 1.42M_\odot$, $\Delta\chi_f = 0.25$ [94].

significant GR deviation for the lower eccentricity and the lower total mass, the larger total mass would give consistency with GR in this case, in agreement with our results. Bhat *et al.* also find a significant GR deviation for an eccentricity of 0.15 for both total masses, in agreement with our result for an eccentricity of ∼0.18.

E. Checks of the scaling with SNR

As noted earlier, our injections with total mass $80M_\odot$ and luminosity distance 400 Mpc have network SNRs ∼ 90–120 at the forecast O4 sensitivity we are considering, which are on the higher side and way above the minimum SNRs of ∼10 one obtains for the high-significance signals considered in the LVK testing GR analyses [8–10]. There will be larger errors on testing parameters (broader posterior distributions) for lower-SNR signals, so apparent GR deviations due to eccentricity can be lost in the statistical error at smaller SNRs. In the high-SNR limit, the width of the posteriors scales like $1/\mathrm{SNR}$; we want to check if this scaling works well for our injections. We do not consider a low-SNR version of each of our injections (given in Table I) but make at least one for each of the higher-eccentricity cases, as well as for one lower-eccentricity case. We prepare our low-SNR injections as follows. For each injection and TIGER, FTI, and MDR testing parameter, we compute the luminosity distance at which we would expect to exclude GR at 90% credibility. We compute this by finding the scaling of the width of the 90% credible interval for which the edge of the scaled credible interval would just touch the GR value of zero, keeping the median of the posterior the same. We then apply this scaling to the injected luminosity distance, and for each test and mass ratio pick the injection and testing parameter that gives the largest luminosity distance for our low-SNR injection. The low-SNR injections then use this scaled luminosity distance while keeping the other binary parameters the same as for the high-SNR ones. We then apply the test to that injection just for the specific testing parameter used to find the scaling. For this study, we do not perform the IMR consistency test due to the significant systematic bias we find for that test due to the presence of higher order modes. The injected SNR in the low-SNR cases can be found by scaling the SNRs quoted in Sec. III by the appropriate distance scaling.

TIGER: For the TIGER test, it is the higher-eccentricity case that gives the largest GR quantiles for all three mass ratios. For $q = 1$, 2, and 3, we obtained the largest scaling factors for $\delta\hat{\beta}_3$, $\delta\hat{\alpha}_2$, and $\delta\hat{\varphi}_3$, giving 4.9, 3.7, and 3.8, respectively. The results are shown as black unfilled violins in Fig. 1. As expected, these posterior distributions are all broader than their high-SNR counterparts, and the 90% bound is very close to the GR value in the $q = 1$, 2 cases, though the median is also notably shifted to lower values in the $q = 1$ case, with a smaller shift in the $q = 2$ case. For $q = 3$, GR is still excluded at 2.2σ due to a shift in the median away from zero. These shifts in the median are due to degeneracies between the testing parameter and the chirp mass and between the chirp mass and distance. The distance prior we have chosen favors larger distances and thus larger chirp masses, while larger chirp masses are correlated with smaller values for these three testing parameters. Since a lower SNR allows for a larger mismatch of the model waveform with the injection, the degeneracy makes the testing parameter posterior peak at smaller values. As discussed for FTI in [47], the degeneracy between the testing parameter and chirp mass is most prominent for $\delta\hat{\varphi}_0$ and other low-PN-order parameters, but it is also present for other testing parameters.

FTI: For FTI test, it is again the higher-eccentricity case that gives the largest GR quantiles for all three mass ratios. For $q = 1$, 2, and 3, we obtained the largest scaling factors for $\delta\hat{\varphi}_3$, $\delta\hat{\varphi}_7$, and $\delta\hat{\varphi}_0$, giving 2.5, 1.3, and 3.0, respectively. The results are shown as black unfilled violins in Fig. 3. We observe that the scaling works very well for the $q = 1$ case, with the 90% bound almost exactly at the GR value, and fairly well for the $q = 2$ case, though GR is still excluded at 1.9σ. However, while the posterior for $q = 3$ is broadened, as expected, there is also a significant shift away from zero in the median, and GR is still excluded beyond 3σ, as it is in the high-SNR case. This is due to the significant degeneracy between $\delta\hat{\varphi}_0$ and chirp mass mentioned above, where larger $\delta\hat{\varphi}_0$ values are correlated with larger chirp masses, which are correlated with the larger distances favored by our distance prior.

MDR: For the MDR test, we find that the $q = 1$ lower-eccentricity $\alpha = 1.5$, $q = 2$ higher-eccentricity $\alpha = 1.5$, and $q = 3$ higher-eccentricity $\alpha = 3$ cases give the largest scaling factors, specifically 1.9, 1.3, and 1.9, respectively. The results for $q = 1$ and $q = 3$ are shown as black unfilled violins in Fig. 4. We do not plot the $q = 2$ results, since we were unable to obtain reliable results in this case, finding that the posterior peaks at significantly larger values of A_α than the A_α values that give the largest likelihood (the maximum likelihood is ∼4 orders of magnitude larger than the likelihood values near the peak of the posterior), while there is only a factor of $\lesssim 2$ difference in the prior values. We get the expected broadening of posteriors for $q = 1$ and $q = 3$, and also see shifts of the median to smaller values of A_α, which we expect, since a given A_α causes a larger dephasing on the waveform at a larger distance. To check that the shifts in the median are indeed the expected ones, we compare the posteriors on the dephasing in the low- and high-SNR cases. However, we find that the dephasing posterior peaks at zero for both eccentricities, it has a secondary peak at nonzero values in the high-SNR cases, making it difficult to draw any conclusions.

V. SUMMARY AND CONCLUSIONS

The waveform models employed in the LVK's current tests of GR do not account for the effects of eccentricity and

are only applicable to BBHs on quasicircular orbits. This has not been a serious issue since the binaries formed through isolated formation channels [13] are expected to have negligible eccentricity by the time their signals enter the sensitive band of ground-based GW detectors. However, there are many other formation pathways (such as dynamical formation, e.g., [22,24,27]) that can lead to non-negligible eccentricities in the frequency band of ground-based GW detectors for a small fraction of detected signals. Thus, as we detect more and more GW signals as current GW detectors improve in sensitivity [51], it is anticipated that a fraction of these binaries may be eccentric in nature. The mismatch between an eccentric GW signal in the data and the quasicircular waveform model used in the tests of GR can lead to a false GR deviation (see [43,44]). In this paper, we investigate the effect of ignoring eccentricity when performing some of the LVK's standard tests of GR on realistic eccentric BBH signals.

Specifically, we consider the TIGER [45,46], FTI [47], MDR [48], and IMR consistency [49,50] tests and check their response to simulated eccentric BBH GW signals in the LIGO-Virgo network at the forecast O4 sensitivity [51]. Our eccentric GW signals are modeled using NR waveforms from the SXS catalog [52]; all waveforms are nonspinning, and we consider three mass ratios ($q = 1, 2, 3$). We inject the signals with a total mass of $80M_\odot$ and at a luminosity distance of 400 Mpc, giving SNRs of about 120, 105, and 90 for the three mass ratios. We choose the SXS simulations such that the binary's eccentricity is $\sim$0.05 and $\sim$0.1 at 17 Hz. For each mass ratio, we also consider a quasicircular SXS waveform, to compare our results with eccentric cases. We inject these NR waveforms into zero noise.

As expected, all quasicircular injections are consistent with GR at 90% credibility when subjected to the TIGER, FTI, and MDR tests. However, for the IMR consistency test, the $q = 2$ and 3 quasicircular injections show a significant GR deviation (GR excluded at $> 2\sigma$). We find that this is attributable to the use of higher modes in the recovery waveform model (i.e., `IMRPhenomXPHM`). In particular, when we keep only the (2, 2) mode in the quasicircular $q = 2$, 3 injections and use `IMRPhenomXP` (which does not contain higher modes) to perform the IMR consistency test, no GR deviation is found.

For the TIGER test, we found that the lower-eccentricity injections are consistent with GR at $< 3\sigma$ except for the higher-PN-order parameters for $q = 3$ and $\delta\hat{\beta}_3$ for $q = 1, 2$ where they exclude GR at $> 3\sigma$. We found very significant GR deviations ($> 3\sigma$ in almost all cases) with TIGER for the higher-eccentricity injections. For the FTI test, we found the lower-eccentricity injections to be consistent with GR at 2σ, except for the higher-PN-order parameters in the $q = 3$ case, where these are $> 3\sigma$ deviations. Higher-eccentricity injections show large GR deviations in many cases, though not as large as in the TIGER analysis. In the MDR test, both lower- and higher-eccentricity injections are found to be consistent with GR at 3σ, with only three cases where GR is excluded at $> 2\sigma$. Further, the IMR consistency test with higher mode analysis reports strong GR deviations ($> 2.7\sigma$) for both lower- and higher-eccentricity injections, except for the lower-eccentricity case for $q = 1$, which is consistent with GR at 90% credibility. However, the analysis without higher modes for the $q = 2, 3$ cases finds that GR is excluded at 90% credibility (indeed $> 2.5\sigma$) only for the higher-eccentricity cases.

We also checked the scaling of our results with distance for the TIGER, FTI, and MDR results in a few cases that we expected to still give a significant GR deviation at much larger distances. Here we found that one will still exclude GR at the $\sim$90% credible level for at least one testing parameter at a distance of $\sim$2 Gpc ($\sim$1.5 Gpc) for TIGER and the $q = 1$ ($q = 2$, 3) higher-eccentricity injections; at distances of $\sim$1, 0.5, and 1.2 Gpc for FTI and the $q = 1$, 2, and 3 higher-eccentricity injections; and at distances of $\sim$0.7 Gpc ($\sim$0.5 Gpc) for MDR and the $q = 1$ lower-eccentricity and $q = 3$ higher-eccentricity ($q = 2$ higher-eccentricity) injections. We found that the $q = 1$ lower-eccentricity injection gives a larger GR deviation with MDR than the higher-eccentricity injection does.

The results we obtained in this paper suggest that one will obtain strong GR deviations when applying standard current LVK tests of GR to GW signals from binaries with non-negligible eccentricity ($\sim$0.05–0.1 at 17 Hz). While we have only considered a small portion of binary parameter space, in particular just one total mass, the Fisher matrix results in [43,44] suggest that this is a fairly general conclusion. Therefore, the possibility of the signal being from an eccentric binary needs to be ruled out before one can make any claims of GR violation.

Ruling out an eccentric binary as a possible cause of an apparent GR violation will require analysis of the signal using waveforms for eccentric BBHs in GR, and likely the implementation of tests of GR (those used in this paper or others) using eccentric GR waveform models as a baseline. It will be necessary to use waveforms including both eccentricity and precession in such analyses. While there are not yet any full IMR models for BBHs with both eccentricity and precession, there is a PN inspiral model with both these effects [95], as well as full IMR effective-one-body models for eccentric BBHs with aligned spins [96,97] and an NR surrogate model for nonspinning eccentric BBHs [98]. Thus, the prospects for having waveform models for precessing eccentric BBHs in the near future seem good, though even when these are available, it will be necessary to perform careful studies to determine the extent to which one can distinguish various possible GR deviations from the effects of eccentricity.

Additionally, there are several other physical effects that are missing in the current waveform models employed by the LVK's tests of GR, e.g., gravitational lensing and environmental effects, and could be important.

For instance, [99] showed that strong lensing can modify the lensed GW signal in such a way that it can become inconsistent with unlensed GR GW signal and [100] showed that this can indeed lead to biases in estimating parameters of lensed signal if the recovery waveform model does not account for lensing effects. Furthermore, even if the magnitude of the environmental effects are expected to be small [101] for ground-based detectors [102,103], there could be a possibility of detecting a GW signal with the effect of a third body at the forecast O4 sensitivity (see, e.g., [104]). Therefore, it will be interesting to see if and how such effects can potentially mimic a GR violation at the sensitivities that can be expected in the near future [51].

ACKNOWLEDGMENTS

We thank K. G. Arun and Pankaj Saini for useful comments, Mukesh Kumar Singh for sharing the results of his investigations into the IMR consistency test bias, and Archisman Ghosh for the code used to create the injections. We also thank all the LIGO-Virgo-KAGRA testing GR group members who implemented these tests in publicly available code. N. K. J.-M. is supported by NSF Grant No. AST-2205920. A. G. is supported in part by NSF Grants No. PHY-2308887 and No. AST-2205920. The authors are grateful for computational resources provided by the LIGO Laboratory and the Leonard E Parker Center for Gravitation, Cosmology and Astrophysics at the University of Wisconsin-Milwaukee and supported by National Science Foundation Grants No. PHY-0757058, No. PHY-0823459, No. PHY-1700765, and No. PHY-1626190. This study used the software packages LALSuite [72], Matplotlib [105], NumPy [106], PESummary [107], Positive [93], SciPy [108], and Seaborn [109]. This is LIGO Document No. P2300161.

[1] C. M. Will, Living Rev. Relativity **17**, 4 (2014).
[2] N. Wex, Testing relativistic gravity with radio pulsars, in *Frontiers in Relativistic Celestial Mechanics, Volume 2: Applications and Experiments*, edited by S. M. Kopeikin (De Gruyter, Berlin, 2014).
[3] J. M. Weisberg and J. H. Taylor, in ASP Conf. Ser. (2005), Vol. **328**, pp. 25–31 http://aspbooks.org/custom/publications/paper/328-0025.html.
[4] G. Voisin, I. Cognard, P. C. C. Freire, N. Wex, L. Guillemot, G. Desvignes, M. Kramer, and G. Theureau, Astron. Astrophys. **638**, A24 (2020).
[5] M. Kramer *et al.*, Phys. Rev. X **11**, 041050 (2021).
[6] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **116**, 221101 (2016).
[7] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **123**, 011102 (2019).
[8] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. D **100**, 104036 (2019).
[9] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. D **103**, 122002 (2021).
[10] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), arXiv:2112.06861.
[11] J. Aasi *et al.* (LIGO Scientific Collaboration), Classical Quantum Gravity **32**, 074001 (2015).
[12] F. Acernese *et al.* (Virgo Collaboration), Classical Quantum Gravity **32**, 024001 (2015).
[13] M. Mapelli, Formation channels of single and binary stellar-mass black holes, in *Handbook of Gravitational Wave Astronomy*, edited by C. Bambi, S. Katsanevas, and K. D. Kokkotas (Springer Singapore, Singapore, 2021).
[14] P. C. Peters, Phys. Rev. **136**, B1224 (1964).
[15] A. Tucker and C. M. Will, Phys. Rev. D **104**, 104023 (2021).
[16] I. Cholis, E. D. Kovetz, Y. Ali-Haïmoud, S. Bird, M. Kamionkowski, J. B. Muñoz, and A. Raccanelli, Phys. Rev. D **94**, 084013 (2016).
[17] Y.-F. Wang and A. H. Nitz, Astrophys. J. **912**, 53 (2021).
[18] L. Wen, Astrophys. J. **598**, 419 (2003).
[19] R. M. O'Leary, B. Kocsis, and A. Loeb, Mon. Not. R. Astron. Soc. **395**, 2127 (2009).
[20] F. Antonini, S. Chatterjee, C. L. Rodriguez, M. Morscher, B. Pattabiraman, V. Kalogera, and F. A. Rasio, Astrophys. J. **816**, 65 (2016).
[21] J. Samsing and E. Ramirez-Ruiz, Astrophys. J. Lett. **840**, L14 (2017).
[22] J. Samsing, Phys. Rev. D **97**, 103014 (2018).
[23] L. Gondán, B. Kocsis, P. Raffai, and Z. Frei, Astrophys. J. **860**, 5 (2018).
[24] C. L. Rodriguez, P. Amaro-Seoane, S. Chatterjee, and F. A. Rasio, Phys. Rev. Lett. **120**, 151101 (2018).
[25] J. Samsing, A. Askar, and M. Giersz, Astrophys. J. **855**, 124 (2018).
[26] M. Zevin, J. Samsing, C. Rodriguez, C.-J. Haster, and E. Ramirez-Ruiz, Astrophys. J. **871**, 91 (2019).
[27] C. L. Rodriguez, P. Amaro-Seoane, S. Chatterjee, K. Kremer, F. A. Rasio, J. Samsing, C. S. Ye, and M. Zevin, Phys. Rev. D **98**, 123005 (2018).
[28] L. Gondán and B. Kocsis, Mon. Not. R. Astron. Soc. **506**, 1665 (2021).
[29] M. Dall'Amico, M. Mapelli, S. Torniamenti, and M. Arca Sedda, arXiv:2303.07421.
[30] J. Samsing, I. Bartos, D. J. D'Orazio, Z. Haiman, B. Kocsis, N. W. C. Leigh, B. Liu, M. E. Pessah, and H. Tagawa, Nature (London) **603**, 237 (2022).
[31] H. Tagawa, B. Kocsis, Z. Haiman, I. Bartos, K. Omukai, and J. Samsing, Astrophys. J. Lett. **907**, L20 (2021).
[32] F. Antonini, N. Murray, and S. Mikkola, Astrophys. J. **781**, 45 (2014).
[33] J. M. Antognini, B. J. Shappee, T. A. Thompson, and P. Amaro-Seoane, Mon. Not. R. Astron. Soc. **439**, 1079 (2014).

[34] F. Antonini, S. Toonen, and A. S. Hamers, Astrophys. J. **841**, 77 (2017).
[35] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), arXiv:2111.03606.
[36] I. M. Romero-Shaw, P. D. Lasky, E. Thrane, and J. C. Bustillo, Astrophys. J. Lett. **903**, L5 (2020).
[37] V. Gayathri, J. Healy, J. Lange, B. O'Brien, M. Szczepanczyk, I. Bartos, M. Campanelli, S. Klimenko, C. O. Lousto, and R. O'Shaughnessy, Nat. Astron. **6**, 344 (2022).
[38] I. M. Romero-Shaw, P. D. Lasky, and E. Thrane, Astrophys. J. Lett. **921**, L31 (2021).
[39] I. M. Romero-Shaw, P. D. Lasky, and E. Thrane, Astrophys. J. **940**, 171 (2022).
[40] I. M. Romero-Shaw, D. Gerosa, and N. Loutrel, Mon. Not. R. Astron. Soc. **519**, 5352 (2023).
[41] E. O'Shea and P. Kumar, arXiv:2107.07981.
[42] M. Favata, C. Kim, K. G. Arun, J. Kim, and H. W. Lee, Phys. Rev. D **105**, 023003 (2022).
[43] P. Saini, M. Favata, and K. G. Arun, Phys. Rev. D **106**, 084031 (2022).
[44] S. A. Bhat, P. Saini, M. Favata, and K. G. Arun, Phys. Rev. D **107**, 024009 (2023).
[45] M. Agathos, W. Del Pozzo, T. G. F. Li, C. Van Den Broeck, J. Veitch, and S. Vitale, Phys. Rev. D **89**, 082001 (2014).
[46] J. Meidam *et al.*, Phys. Rev. D **97**, 044033 (2018).
[47] A. K. Mehta, A. Buonanno, R. Cotesta, A. Ghosh, N. Sennett, and J. Steinhoff, Phys. Rev. D **107**, 044020 (2023).
[48] S. Mirshekari, N. Yunes, and C. M. Will, Phys. Rev. D **85**, 024041 (2012).
[49] 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, Phys. Rev. D **94**, 021101(R) (2016).
[50] A. Ghosh, N. K. Johnson-McDaniel, A. Ghosh, C. K. Mishra, P. Ajith, W. Del Pozzo, C. P. L. Berry, A. B. Nielsen, and L. London, Classical Quantum Gravity **35**, 014002 (2018).
[51] B. P. Abbott *et al.* (KAGRA, LIGO Scientific, and Virgo Collaborations), Living Rev. Relativity **23**, 3 (2020), noise curves available from https://dcc.ligo.org/LIGO-T2000012/public.
[52] M. Boyle *et al.*, Classical Quantum Gravity **36**, 195006 (2019).
[53] I. Hinder, L. E. Kidder, and H. P. Pfeiffer, Phys. Rev. D **98**, 044015 (2018).
[54] M. K. Singh (private communication).
[55] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Phys. Rev. Lett. **113**, 151101 (2014).
[56] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez Forteza, and A. Bohé, Phys. Rev. D **93**, 044007 (2016).
[57] A. Bohé, M. Hannam, S. Husa, F. Ohme, M. Pürrer, and P. Schmidt, PhenomPv2—Technical notes for LAL implementation, Technical Report No. LIGO-T1500602, LIGO Project, 2016, https://dcc.ligo.org/LIGO-T1500602/public.
[58] R. Cotesta, A. Buonanno, A. Bohé, A. Taracchini, I. Hinder, and S. Ossokine, Phys. Rev. D **98**, 084028 (2018).
[59] R. Cotesta, S. Marsat, and M. Pürrer, Phys. Rev. D **101**, 124040 (2020).
[60] G. Pratten *et al.*, Phys. Rev. D **103**, 104056 (2021).
[61] N. K. Johnson-McDaniel, A. Ghosh, S. Ghonge, M. Saleem, N. V. Krishnendu, and J. A. Clark, Phys. Rev. D **105**, 044020 (2022).
[62] A. Gupta, S. Datta, S. Kastha, S. Borhanian, K. G. Arun, and B. S. Sathyaprakash, Phys. Rev. Lett. **125**, 201101 (2020).
[63] S. Datta, A. Gupta, S. Kastha, K. G. Arun, and B. S. Sathyaprakash, Phys. Rev. D **103**, 024036 (2021).
[64] M. Evans *et al.*, arXiv:2109.09882.
[65] P. Amaro-Seoane *et al.*, arXiv:1702.00786.
[66] A. A. Shoom, P. K. Gupta, B. Krishnan, A. B. Nielsen, and C. D. Capano, Gen. Relativ. Gravit. **55**, 55 (2023).
[67] M. Saleem, S. Datta, K. G. Arun, and B. S. Sathyaprakash, Phys. Rev. D **105**, 084062 (2022).
[68] S. Datta, M. Saleem, K. G. Arun, and B. S. Sathyaprakash, arXiv:2208.07757.
[69] S. Datta, arXiv:2303.04399.
[70] P. A. R. Ade *et al.* (Planck Collaboration), Astron. Astrophys. **594**, A13 (2016).
[71] J. M. Ezquiaga, W. Hu, M. Lagos, M.-X. Lin, and F. Xu, J. Cosmol. Astropart. Phys. 08 (2022) 016.
[72] LVK Algorithm Library Suite (LALSuite), 10.7935/GT1W-FZ16.
[73] J. M. Bardeen, W. H. Press, and S. A. Teukolsky, Astrophys. J. **178**, 347 (1972).
[74] F. Hofmann, E. Barausse, and L. Rezzolla, Astrophys. J. Lett. **825**, L19 (2016).
[75] J. Healy and C. O. Lousto, Phys. Rev. D **95**, 024037 (2017).
[76] X. Jiménez-Forteza, D. Keitel, S. Husa, M. Hannam, S. Khan, and M. Pürrer, Phys. Rev. D **95**, 064024 (2017).
[77] N. K. Johnson-McDaniel, A. Gupta, P. Ajith, D. Keitel, O. Birnholtz, F. Ohme, and S. Husa, Determining the final spin of a binary black hole system including in-plane spins: Method and checks of accuracy, Technical Report No. LIGO-T1600168, LIGO Project, 2016, https://dcc.ligo.org/LIGO-T1600168/public/main.
[78] Gravitational wave detector observing timeline, https://dcc.ligo.org/G2002127-v19/public.
[79] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **116**, 061102 (2016).
[80] J. Skilling, AIP Conf. Proc. **735**, 395 (2004).
[81] J. Veitch *et al.*, Phys. Rev. D **91**, 042003 (2015).
[82] L. Blanchet, T. Damour, G. Esposito-Farese, and B. R. Iyer, Phys. Rev. Lett. **93**, 091101 (2004).
[83] A. Buonanno, B. R. Iyer, E. Ochsner, Y. Pan, and B. S. Sathyaprakash, Phys. Rev. D **80**, 084043 (2009).
[84] B. Moore, M. Favata, K. G. Arun, and C. K. Mishra, Phys. Rev. D **93**, 124061 (2016).
[85] C. Cutler and M. Vallisneri, Phys. Rev. D **76**, 104018 (2007).
[86] C. Cutler and É. É. Flanagan, Phys. Rev. D **49**, 2658 (1994).
[87] E. Poisson and C. M. Will, Phys. Rev. D **52**, 848 (1995).
[88] K. G. Arun, A. Buonanno, G. Faye, and E. Ochsner, Phys. Rev. D **79**, 104023 (2009); **84**, 049901(E) (2011).
[89] C. K. Mishra, A. Kela, K. G. Arun, and G. Faye, Phys. Rev. D **93**, 084054 (2016).
[90] É. Racine, Phys. Rev. D **78**, 044021 (2008).

[91] L. Santamaría, F. Ohme, P. Ajith, B. Brügmann, N. Dorband, M. Hannam, S. Husa, P. Mösta, D. Pollney, C. Reisswig, E. L. Robinson, J. Seiler, and B. Krishnan, Phys. Rev. D **82**, 064016 (2010).
[92] P. Schmidt, F. Ohme, and M. Hannam, Phys. Rev. D **91**, 024043 (2015).
[93] L. London and E. Fauchon-Jones, Classical Quantum Gravity **36**, 235015 (2019).
[94] P. Saini (private communication).
[95] A. Klein, arXiv:2106.10291.
[96] A. Nagar, A. Bonino, and P. Rettegno, Phys. Rev. D **103**, 104021 (2021).
[97] A. Ramos-Buades, A. Buonanno, M. Khalil, and S. Ossokine, Phys. Rev. D **105**, 044035 (2022).
[98] T. Islam, V. Varma, J. Lodman, S. E. Field, G. Khanna, M. A. Scheel, H. P. Pfeiffer, D. Gerosa, and L. E. Kidder, Phys. Rev. D **103**, 064022 (2021).
[99] J. M. Ezquiaga, D. E. Holz, W. Hu, M. Lagos, and R. M. Wald, Phys. Rev. D **103**, 064047 (2021).
[100] A. Vijaykumar, A. K. Mehta, and A. Ganguly, arXiv:2202.06334.
[101] V. Cardoso and A. Maselli, Astron. Astrophys. **644**, A147 (2020).
[102] E. Barausse, V. Cardoso, and P. Pani, Phys. Rev. D **89**, 104059 (2014).
[103] C. Bonvin, C. Caprini, R. Sturani, and N. Tamanini, Phys. Rev. D **95**, 044029 (2017).
[104] A. Vijaykumar, A. Tiwari, S. J. Kapadia, K. G. Arun, and P. Ajith, arXiv:2302.09651.
[105] J. D. Hunter, Comput. Sci. Eng. **9**, 90 (2007).
[106] C. R. Harris *et al.*, Nature (London) **585**, 357 (2020).
[107] C. Hoy and V. Raymond, SoftwareX **15**, 100765 (2021).
[108] P. Virtanen *et al.*, Nat. Methods **17**, 261 (2020).
[109] M. Waskom, J. Open Source Software **6**, 3021 (2021).