

A meta inspiral-merger-ringdown consistency test of general relativity with gravitational wave signals from compact binaries

Sakshi Satish Madekar

Department of Physical Sciences, Indian Institute of Science Education & Research (IISER)
Mohali, Sector 81 SAS Nagar, Manauli PO 140306, Punjab, India

Nathan K Johnson-McDaniel

Department of Physics and Astronomy, University of Mississippi, University, Mississippi
38677, USA

Anuradha Gupta

Department of Physics and Astronomy, University of Mississippi, University, Mississippi
38677, USA

Abhirup Ghosh

Trade Republic Bank GmbH, Brunnenstraße 19-21, 10119 Berlin, Germany

Abstract. The observation of gravitational waves from compact binary coalescences is a promising tool to test the validity of general relativity (GR) in a highly dynamical strong-field regime. There are now a variety of tests of GR performed on the observed compact binary signals. In this paper, we propose a new test of GR that compares the results of these individual tests. This meta inspiral-merger-ringdown consistency test (IMRCT) involves inferring the final mass and spin of the remnant black hole obtained from the analyses of two different tests of GR and checking for consistency. If there is a deviation from GR, we expect that different tests of GR will recover different values for the final mass and spin, in general. We check the performance of the meta IMRCT using a standard set of null tests used in various gravitational-wave analyses: the original IMRCT, parameterized phasing tests (TIGER and FTI) and the modified dispersion test. However, the meta IMRCT is applicable to any tests of GR that infer the initial masses and spins or the final mass and spin, including ones that are applied to binary neutron star or neutron star–black hole signals. We apply the meta IMRCT to simulated quasi-circular GR and non-GR binary black hole (BBH) signals as well as to eccentric BBH signals in GR (analyzed with quasicircular waveforms). We find that the meta IMRCT gives consistency with GR for the quasi-circular GR signals and picks up a deviation from GR in the other cases, as do other tests. In some cases, the meta IMRCT finds a significant GR deviation for a given pair of tests (and specific testing parameters) while the individual tests do not, showing that it is more sensitive than the individual tests to certain types of deviations. In addition, we also apply this test to a few selected real compact binary signals and find them consistent with GR.

1. Introduction

Gravitational wave observations now regularly provide us with direct information about the highly dynamical strong-field mergers of binaries of black holes and neutron stars. The most

recent completed observing run (O3) has brought the total number of detection candidates to around 100 [1], and the ongoing observing run (O4, with some initial results released in [2]) will considerably increase this number. There is now a considerable variety of tests of general relativity (GR) applied to the most confident detections—see [3–6] for work by the LIGO-Virgo(-KAGRA) (LVK) collaboration ‡ and [7] for an overview of some of the other tests being applied or proposed. However, all the tests currently applied are null tests of some type—none of them is testing a specific alternative theory. Moreover, it is not clear to what extent they are sensitive to various possible deviations from GR. There are now some initial studies of the relation between a selection of these tests, looking at how they respond to various phenomenological non-GR signals [7] or signals beyond the standard quasicircular binary hypothesis used in the application of the tests [8]. However, there is still much to be done to understand the extent to which these tests are complementary or redundant.

Here we propose a new method for comparing tests of GR, which itself provides yet another test of GR. Specifically, we generalize the inspiral-merger-ringdown (IMR) consistency test for binary black hole (BBH) signals introduced in [9, 10] (which checks the consistency of the low- and high-frequency portions of the signal) to compare the posterior probability distributions for the mass and spin of the final black hole from two different tests of GR. As noticed in [7], these distributions are often quite different when different tests of GR (including different versions of the same test, e.g. parameterized tests varying different post-Newtonian parameters) are applied to the same non-GR signal, or when comparing the results of a test of GR and a standard GR analysis of the signal. In these analyses, the final mass and spin are calculated by applying fits to BBH numerical relativity simulations that give the final mass and spin in terms of the binary’s initial masses and spins. Thus, since the binary’s initial masses and spins are biased in different ways for different analyses when attempting to reproduce a non-GR signal, the final mass and spin are also in general biased differently in different analyses. This is why this method can detect deviations from GR (or, more generally, deviations from the signal model used in the analysis, which is a quasicircular binary in all LVK tests of GR).

The standard IMR consistency test (IMRCT) is restricted to binaries in a certain mass range where there is sufficient signal-to-noise (SNR) in both portions of the signal (see, e.g. [6]), due to its analysis of the low- and high-frequency portions of the signal independently. This generalized test has no such restrictions and can even be applied to binary neutron star or neutron star–black hole binary signals: One can even apply the same BBH fits in this case when applying the test, since the fits just provide a convenient way of mapping the initial mass and spin parameter space to a two-dimensional space. In fact, one does not have to apply the test to the final mass and spin, but could choose other GR parameters to compare. However, we will not explore this here. The generalized test is also computationally quite efficient, once one has the results from the individual tests, since it does not require any further stochastic sampling or evaluation of waveforms.

We illustrate the method, which we refer to as the *meta IMR consistency test* (or meta IMRCT), using the results from the tests of GR (and standard GR parameter estimation analysis) applied to quasicircular GR and non-GR signals from [7] and quasicircular and eccentric numerical relativity signals from [8] (here only analyzed with quasicircular waveform models). Specifically, we consider the TIGER [11] and FTI [12] parameterized tests, the modified dispersion relation test [4, 13], and the standard IMRCT [9, 10]. Like the individual tests, the meta IMRCT finds quasicircular GR signals in [7, 8] consistent with

‡ We refer to the LV(K) collaboration for cases where the KAGRA collaboration only contributed to some of the results being referenced.

GR but reports GR violation for non-GR and eccentric signals. We also apply the meta IMRCT to the compact binary signals GW170817 [14], GW190412 [15], GW190521 [16], GW190814 [17] and GW200225_060421 [1], using the results from the LV(K) collaboration's application of the aforementioned tests to these events. The meta IMRCT finds consistency with GR in these cases, as do the individual tests.

The paper is structured as follows: We introduce the existing tests of GR we consider and the Bayesian inference they are based on in section 2, and then provide the specifics of the meta IMRCT in section 3. We then give an overview of the simulated observations used to assess the performance of the tests in section 4 and the results of the meta IMRCT in section 5. Finally, we conclude in section 6. We use geometrised units with $G = c = 1$ throughout.

2. Tests of GR

This section provides an overview of standard GR parameter inference and the tests of GR that the meta IMRCT builds on. The original IMRCT [9, 10], the TIGER [11] and FTI [12] parameterised tests and the modified dispersion relation test [4, 13], are now a staple of testing-GR analyses by the LVK and others. They have thus been described in detail in previous publications, e.g. [4–7]. However, given the importance of the details of these analysis in motivating the meta IMRCT, we provide a brief review here again.

2.1. Bayesian parameter estimation

Before we review the tests, it is worth reviewing our parameter inference framework, which is Bayesian, as is the case for most of gravitational-wave data analysis. Data from a gravitational-wave detector is a noisy time series $\mathbf{d}(t)$ which occasionally contains a gravitational-wave signal $\mathbf{h}(t, \boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ denotes the parameters describing the gravitational-wave signal. The data can thus be written as

$$\mathbf{d}(t) = \mathbf{h}(t, \boldsymbol{\lambda}) + \mathbf{n}(t), \quad (1)$$

where $\mathbf{n}(t)$, the noise in detector data, is assumed to be Gaussian and characterised by a (frequency-domain) power spectral density $S_n(f)$. The gravitational-wave signal is described by a hypothesis H , which can either be GR or an alternate theory of gravity. Even within GR, waveforms of signals can be built with varying assumptions, and hence our hypothesis is the gravitational waveform used in the data analysis. This waveform is usually characterised by a multi-dimensional parameter set $\boldsymbol{\lambda}$, which includes a GR parameter set $\boldsymbol{\lambda}_{\text{GR}}$ and, in the case of a non-GR waveform, any additional non-GR parameter(s) $\boldsymbol{\lambda}_{\text{nGR}}$ that describe it.

Given a hypothesis H and any other information I , a posterior probability distribution of $\boldsymbol{\lambda}$, $P(\boldsymbol{\lambda}|\mathbf{d}, I)$, can be inferred using the Bayes Theorem:

$$P(\boldsymbol{\lambda}|\mathbf{d}, I) \propto p(\boldsymbol{\lambda})\mathcal{L}(\mathbf{d}|\boldsymbol{\lambda}, I), \quad (2)$$

where $p(\boldsymbol{\lambda})$ is the prior probability distribution and $\mathcal{L}(\mathbf{d}|\boldsymbol{\lambda}, I)$ is the likelihood function, i.e. the probability of the data being accurately described by a certain parameter choice. Assuming Gaussian noise, $\mathbf{n}(t)$, we have

$$\mathcal{L}(\mathbf{d}|\boldsymbol{\lambda}, I) \propto \exp \left[-\frac{1}{2} \langle \mathbf{d}(t) - \mathbf{h}(t, \boldsymbol{\lambda}) | \mathbf{d}(t) - \mathbf{h}(t, \boldsymbol{\lambda}) \rangle \right], \quad (3)$$

where $\langle \cdot | \cdot \rangle$ denotes the noise-weighted inner product defined as

$$\langle a | b \rangle = 2 \int_{f_{\text{low}}}^{f_{\text{high}}} \frac{\tilde{a}^*(f)\tilde{b}(f) + \tilde{a}(f)\tilde{b}^*(f)}{S_n(f)} df. \quad (4)$$

Here $\tilde{a}(f)$, $\tilde{b}(f)$ represent the Fourier transform of $a(t)$, $b(t)$ and f_{low} , f_{high} indicate the frequency cut-offs used in analysing the data which usually coincides with the bandwidth of the gravitational-wave detector's sensitivity. In practice, $P(\boldsymbol{\lambda}|\mathbf{d}, I)$ is sampled over the parameter space stochastically using Metropolis–Hastings Markov-Chain Monte Carlo [18, 19] or nested sampling [20]. Considering the results that we use later, Refs. [7, 8, 14] used the LALInference code [21], a part of LVK's LALSuite software library [22] while Refs. [1, 5, 6, 15–17] used both LALInference and Bilby [23] to construct the posteriors, $P(\boldsymbol{\lambda}|\mathbf{d}, I)$.

2.2. Inference using the GR model

As mentioned above, we need a hypothesis or waveform model to describe our gravitational-wave signal. For want of a credible alternative, our dominant hypothesis is GR, and any investigation into beyond-GR effects starts with setting a GR baseline.

There are three unique phases of a BBH merger: an initial *inspiral* where the two black holes slowly coalesce through the back-reaction of gravitational-wave emission, the actual process of *merger* whose onset is indicated by the formation of the apparent horizon, and the post-merger *ringdown* where the newly formed remnant object settles down to a stable state through the emission of (a superposition of) exponentially damped sinusoidal gravitational waves, the quasi-normal-mode spectrum.

In GR, the entire gravitational-wave signal from a quasi-circular BBH merger is completely described by the masses and spins of the initial black holes, $(m_1, m_2, \chi_1, \chi_2)$, its position in the sky (α, δ) and orientation (ι, ψ) , luminosity distance d_L and two reference quantities for phase and time (ϕ_c, t_c) . Together, they form the 15 dimensional parameter space $\boldsymbol{\lambda}_{\text{GR}} = \{m_1, m_2, \chi_1, \chi_2, \alpha, \delta, d_L, \iota, \psi, \phi_c, t_c\}$. Setting the GR baseline therefore involves performing parameter inference of $\boldsymbol{\lambda}_{\text{GR}}$ with a GR waveform model. As mentioned earlier, these analyses can use a variety of different waveform models with varying assumptions. In this paper we use results obtained with IMRPhenomPv2 [24–26] for the GW150914-like and GW170608-like simulated observations and IMRPhenomXPHM [27] for the quasicircular and eccentric simulated observations using numerical relativity waveforms. Both of these models describe the waveforms from precessing binary black holes, with IMRPhenomXPHM including two-spin precession effects and higher-order modes. We also use results obtained with the slightly earlier precessing model with higher modes, IMRPhenomPv3HM [28] for some real events, and with the precessing binary neutron star model IMRPhenomPv2_NRTidal [29, 30] for GW170817.

In our study, we are particularly interested in the mass M_f and spin χ_f of the remnant black hole. Analytical fits to numerical relativity simulations can translate our estimates of the initial masses and spins $(m_1, m_2, \chi_1, \chi_2)$ into the estimates of the final mass and spin (M_f, χ_f) . The different waveform models incorporate different fits to predict the final mass and spin. We instead use a single set of fits in this analysis (the same used by the LVK to infer the final mass and spin in, e.g. [1]) to avoid any biases due to different final mass and spin fits. Specifically, we use an average of fits to numerical relativity simulations [31–33] where the aligned-spin final spin fits are augmented with the contribution from in-plane spin components [34] for precessing waveform models. However, unlike the application in [1], we do not evolve the initial spins before applying the fits; this is the same procedure used in the computation of the final mass and spin for the standard IMRCT in the LVK test of GR analyses [4–6]. Since these fits are just used to provide a physically motivated two-dimensional parameter space in which to compare the results, the small decrease in accuracy of the fits due to not evolving the spins is not a concern.

The absence of a credible alternative to GR also means that most strong-field tests of gravity are essentially tests of GR, where we measure and constrain possible deviations from GR predictions. Such tests are called null tests, GR being the null hypothesis. Three of the most prominent null tests of GR using gravitational-wave observations of BBH mergers are the IMRCT, parameterised tests of the gravitational waveform and the test of the modified dispersion relation, as discussed below.

2.3. IMR Consistency Test

The three phases of a BBH merger, as described above, are conventionally pictured in the time domain (as in figure 2 of [35]). However, under the stationary phase approximation, the early (late) portions of the time-domain signal corresponds, to a high degree of accuracy, to the low- (high-) frequency portion of the Fourier-domain signal, as illustrated in [10]. We thus refer to these portions of the Fourier-domain signal as the inspiral and the merger-ringdown portions and refer to the dividing frequency as f_{cut} . [The merger-ringdown portion is referred to as the post-inspiral portion in the LV(K) testing GR catalogue papers [4–6].]

The IMRCT infers the mass and spin of the remnant black hole, (M_f, χ_f) from the low- and high-frequency portions of the signal, and compares these estimates with each other. If the entire gravitational-wave signal is accurately described by our GR hypothesis (and specifically by the family of waveforms used to carry out the test), then the low- and high-frequency inferences should correspond to the same underlying signal, or in terms of the remnant black hole, consistent measurements of (M_f, χ_f) . In order to quantify possible deviations between the two measurements, we define fractional deviations in the measurement of these quantities as:

$$\frac{\Delta M_f}{\bar{M}_f} = 2 \frac{M_f^{\text{I}} - M_f^{\text{MR}}}{M_f^{\text{I}} + M_f^{\text{MR}}}, \quad (5a)$$

$$\frac{\Delta \chi_f}{\bar{\chi}_f} = 2 \frac{\chi_f^{\text{I}} - \chi_f^{\text{MR}}}{\chi_f^{\text{I}} + \chi_f^{\text{MR}}}, \quad (5b)$$

where the “I” and “MR” superscripts refer to the inspiral and merger-ringdown portions of the signal. If GR is correct, the posteriors on $(\Delta M_f/\bar{M}_f, \Delta \chi_f/\bar{\chi}_f)$ will be consistent with $(0, 0)$.

The IMRCT takes f_{cut} to be the $|m| = 2$ gravitational-wave frequency of the innermost stable circular orbit (ISCO) of the remnant Kerr black hole [36] computed from GR analysis of the full signal. The LVK analyses [4–6] first compute the medians of the individual masses and spherical coordinate components of the spins from the GR analysis of the full signal and then use these values to compute a value for the final mass and spin and thence the ISCO frequency. The analysis in [8] computes the median of the ISCO frequency of the remnant Kerr black hole inferred from the GR analysis, but also checks that for the cases they consider the LVK method would give a negligible difference in f_{cut} .

The IMRCT can be performed using any GR waveform model. The analyses considered here use the IMRPhenomPv2 and IMRPhenomXPHM models.

2.4. Parameterised tests

The parameterized tests are designed to identify a GR violation by allowing deviations in the frequency-domain phasing coefficients of GR waveform models. The frequency-domain phase of the waveform in the early inspiral can be expressed analytically in a closed form

using the stationary phase approximation as

$$\Phi(f) = \frac{3}{128\eta v^5} \sum_{k=0}^7 (\varphi_k v^k + 3\varphi_{kl} v^k \ln v), \quad (6)$$

where φ_k and φ_{kl} are coefficients in the post-Newtonian (PN) expansion, and this expression extends to 3.5PN, the highest order known completely at the time the tests in question and the GR waveform models they are based on were finalized. All of the φ_k in the sum are nonzero, in general, except for φ_1 , while the φ_{kl} coefficients of the logarithmic terms are only nonzero for $k \in \{5, 6\}$ up to the PN order considered here. In the above expression, $\eta = m_1 m_2 / (m_1 + m_2)^2$ denotes the symmetric mass ratio and $v = (\pi M f)^{1/3}$, with $M = m_1 + m_2$ being the binary's total redshifted mass and f being the gravitational-wave frequency. The PN expressions are not sufficient to describe the late-inspiral and merger-ringdown stages of the signal, which are instead described in different ways in different GR waveform models. For the IMRPhenomD model [25] relevant here, they are described by analytical expressions using the phenomenological coefficients β_k and α_k that are fit to numerical relativity simulations. If p_k collectively denotes all of the inspiral and merger-ringdown phasing coefficients $\varphi_k, \varphi_{kl}, \beta_k$ and α_k , then deviations from GR at all stages can be introduced by inserting the dimensionless fractional deformation parameters $\delta\hat{p}_k$ into the phase such that $p_k \rightarrow (1 + \delta\hat{p}_k)p_k$. If the observed gravitational-wave signal is consistent with GR and more specifically described by the GR waveform family used to perform the analysis, all $\delta\hat{p}_k$ should to be consistent with zero.

As mentioned earlier there are two varieties of the parametrized test. The first one is the Test Infrastructure for General Relativity (TIGER) approach [11, 37] in which the deviations are introduced only in the non-spinning part of the PN coefficients φ_k, φ_{kl} , as well as in the phenomenological β_k and α_k coefficients. The restriction to modifying the non-spinning part of the PN coefficients prevents situations where the spin terms cancel the nonspinning terms in the PN coefficient, making the parameterisation singular. The LVK analyses so far have used a TIGER version that introduces the deviations in the phase of aligned-spin dominant mode waveform model IMRPhenomD [25] and its tidal extension IMRPhenomD_NRTidal [29, 30]. This modified IMRPhenomD (IMRPhenomD_NRTidal) phase is then twisted up to produce a modified precessing IMRPhenomPv2 (IMRPhenomPv2_NRTidal) waveform model using GR spin precession equations. The second variety is the Flexible-Theory-Independent (FTI) method [12] that considers deviations in only the inspiral PN coefficients (including the spin contributions) at low frequencies and tapers those deviations to zero at a specific cutoff frequency, unlike TIGER that allows inspiral deviations to affect the post-inspiral regime through the C^1 matching used to construct the IMRPhenomD model. The FTI test can be implemented based on any aligned-spin GR waveform model, but the version used in the LVK analyses [3–6] is based on the effective-one-body reduced-order model SEOBv4_ROM [38] and its higher-order mode version [39, 40] or tidal version SEOBv4_ROM_NRTidal [29, 30]. The FTI results are reweighted to the TIGER convention, as in the LVK analyses [3–6], so that the results of the two analyses are directly comparable.

2.5. Test of the modified dispersion relation

The modified dispersion relation (MDR) test of gravitational waves constrains dispersive propagation of gravitational waves, specifically considering a phenomenological modification of the standard dispersion relation. Gravitational waves in GR propagate non-dispersively as

described by the standard dispersion relation $E^2 = p^2$, where E and p denote the energy and momentum of the wave. The dispersion relation

$$E^2 = p^2 + A_\alpha p^\alpha \quad (7)$$

introduced in [13] includes the leading predictions of a number of non-GR theories, as discussed in that paper and [4]. Here A_α and α are phenomenological parameters, where A_α determines how much the phase of the gravitational-wave signal deviates from the predictions of GR and α determines the frequency dependence of the frequency-domain dephasing, which is $\propto f^{\alpha-1}$ (with a logarithmic expression for $\alpha = 1$ in some implementations); see the explicit expressions in [4, 13]. Here one assumes that the deviation from GR only comes from the propagation of the waves and not from its generation, i.e. the signal close to the source is still consistent with GR to a very good approximation, which is likely to be the case for at least some of the alternative theories that predict modified propagation, notably massive graviton theories, as discussed in [4].

In GR, $A_\alpha = 0$ for all α , but in a beyond GR theory they can take nonzero values. For example, in a massive graviton theory $\alpha = 0$ and $A_0 > 0$. The LVK analyses [3–6] consider $\alpha \in \{0, 0.5, 1, 1.5, 2.5, 3, 3.5, 4\}$ ($\alpha = 2$ is omitted since it corresponds to no dispersion of the wave), and we consider these values or a subset here: The analyses of the GR and non-GR waveforms from [7] just considers $\alpha = 0$ and the analysis of numerical relativity waveforms in [8] omits $\alpha = 1$ §. As described in [4], the analysis is carried out by sampling over an effective wavelength parameter, considering positive and negative values of A_α separately. The results from two signs of A_α are then combined and reweighted to a flat prior in A_α to obtain the final posterior on A_α for different values of α .

3. Meta IMR consistency test

The meta IMRCT is a straightforward extension of the standard IMRCT described in section 2.3. Instead of using the final mass and spin inferred from the low- and high-frequency portions of the analysis of a signal, one instead uses the final mass and spin inferred using two different tests of GR, or a test of GR and the standard GR analysis of the signal (henceforth referred to as GR parameter estimation [PE]). Specifically, one has

$$\frac{\Delta M_f}{\bar{M}_f} = 2 \frac{M_f^T - M_f^{T'}}{M_f^T + M_f^{T'}}, \quad (8a)$$

$$\frac{\Delta \chi_f}{\bar{\chi}_f} = 2 \frac{\chi_f^T - \chi_f^{T'}}{\chi_f^T + \chi_f^{T'}}, \quad (8b)$$

where T and T' are two separate analyses.

In this application, we break down each test of GR into its individual runs and apply the meta IMRCT to these. Specifically, we consider the inspiral and merger-ringdown analyses used in the standard IMRCT separately. Thus, the standard IMRCT is a special case of the meta IMRCT, though we exclude the standard IMRCT from our meta IMRCT results and just consider pairs of inspiral or merger-ringdown analyses with other tests or GR PE. For the

§ The analyses in [8] omit $\alpha = 1$ since the implementation of the test they use gives the logarithmic frequency dependence corresponding to using the particle velocity of the wave, while the group velocity that is physically preferable gives a constant dephasing (which still leads to a non-GR signal in the presence of higher modes). We still consider the $\alpha = 1$ cases in the application of the meta IMRCT to real events, since it is still a fine phenomenological test of GR in this case, even though it loses the motivation of coming from the dispersion relation.

TIGER, FTI, and modified dispersion analyses, we consider pairs involving all the different testing parameters separately, both with other testing parameters from the same test as well as all other tests.

As mentioned previously, one expects the meta IMRCT to be sensitive to deviations from GR (or more generally, from the waveform model used to perform the analysis), given the results in [7], which show that the different tests (or GR PE) have different biases in the final mass and spin when applied to non-GR signals (see specifically figures 10–12 in that paper). This is as one would expect: When one performs a test of GR or GR PE on a signal that is not described by the underlying GR waveform model used to perform the analysis, one will obtain different posteriors on the individual masses and spins depending on the deviations from GR allowed in the analysis. These differences in the individual masses and spins will, in general, translate to different final masses and spins.

As is done for the standard IMRCT, we reweight the posterior on the deviation parameters $(\Delta M_f/\bar{M}_f, \Delta \chi_f/\bar{\chi}_f)$ to a flat prior by dividing by the distribution of the deviation parameters induced by the prior distributions for the individual analyses. To have a single number to measure the consistency of the meta IMRCT result with GR, we compute the GR quantile Q_{GR} , which is the fraction of the reweighted posterior probability contained within the isodensity contour that intersects the GR value of $(0, 0)$, or

$$Q_{\text{GR}} = \int_{P(\delta_M, \delta_\chi) > P(0,0)} P(\delta_M, \delta_\chi) d\delta_M d\delta_\chi, \quad (9)$$

where P is the reweighted posterior probability density and $\delta_M := \Delta M_f/\bar{M}_f$, $\delta_\chi := \Delta \chi_f/\bar{\chi}_f$. Thus, small GR quantiles correspond to agreement with GR, while GR quantiles close to 100% correspond to significant GR deviations. We also sometimes quote the Gaussian σ equivalent to the GR quantile, i.e. the number of σ s away from the mean at which the probability contained within a symmetric interval with edges that distance from the mean is the same as the GR quantile, or symbolically, x , where

$$Q_{\text{GR}} = \int_{-x}^x \mathcal{N}(\bar{x}) d\bar{x}, \quad (10)$$

with \mathcal{N} the standard normal distribution.

One subtle point in interpreting the results of the meta IMRCT is its dependence on two prior probability distributions. (This also affects the standard IMRCT, but there it is customary to choose the same prior distributions for both the inspiral and merger-ringdown analyses.) In particular, if the priors on the masses and/or spins for the individual analyses are different, this can lead to a prior on the IMRCT deviation parameters that peaks away from the GR value of zero for one or both parameters. For instance, the prior on the deviation parameters peaks well away from the origin for several pairs of GW190412 and GW190814 analyses, due to different choices for the mass prior ranges, as discussed in the next section.

Finally, since one considers $N(N-1)/2$ pairs when applying the meta IMRCT to N tests of GR, one has to worry about the look elsewhere effect (see, e.g. [41]). We do not attempt to correct for this effect here with a trials factor, but just note that we will find that in the cases where the meta IMRCT recovers a GR deviation, it does so for a nonnegligible fraction of the pairs considered, not just one or two.

4. Simulated observations

We consider two sets of simulated observations: GR and non-GR simulated observations like GW150914 and GW170608 from [7] and quasicircular and eccentric simulated BBH

observations (using numerical relativity simulations) from [8]. We use the results of parameter estimation with GR waveforms (GR PE) and applications of the tests of GR described in section 2 obtained in these previous papers. All of these cases do not include noise in the simulated observations (i.e. consider the zero realization of the noise) to avoid biases from a specific nonzero noise realization.

4.1. GW150914-like and GW170608-like GR and non-GR cases

We first detail the simulated observations from [7], which are designed to emulate the GW150914 [42] and GW170608 [43] events, as fiducial high- and low-mass BBHs. These use both GR and non-GR waveforms, considering both large and small deviations in the GW150914-like case, and only small deviations in the GW170608-like case. The GR waveforms considered are the IHES effective-one-body (EOB) model [44] and the IMRPhenomD phenomenological model [25], and just include the dominant $\ell = |m| = 2$ modes of the waveform. The IHES EOB model is used to obtain a non-GR waveform by modifying its energy and angular momentum flux by multiplying the modes of the waveform that first enter these at 2PN by a factor $\sqrt{a_2}$. Thus, the inspiral phase is also modified starting at 2PN. Additionally, the final mass and spin are set self-consistently to satisfy energy and angular momentum balance, so the merger-ringdown portion of the signal also differs from that in GR for the same initial masses (and spins). The IMRPhenomD model is the basis for a waveform with dispersive propagation from a massive graviton [45] with graviton mass m_g , leading to a dephasing of the entire waveform in the frequency domain $\propto m_g^2/f$ (as in the MDR test with $\alpha = 0$, as discussed above). It is also the basis for waveforms with the TIGER and FTI modifications described in section 2, here considering the modification to the 2PN phasing coefficient.

The (redshifted) total masses of all these simulated observations are $72.2M_\odot$ and $19.9M_\odot$ for the GW150914-like and GW170608-like cases. The mass ratio is 1.22 for the EOB waveforms and 1 for the others (due to a typo in generating the waveforms). The IHES EOB model does not allow for spin, so the spins are set to zero in these simulated observations (and both GW150914 and GW170608 are consistent with being nonspinning). These simulated observations have SNRs of ~ 54 (GW150914-like) and ~ 21 (GW170608-like) in the O3-like LIGO-Virgo sensitivities [46] considered, except for the modified EOB GW150914-like case, where the SNR is as small as 40 (with the larger GR deviation).

The GR PE is carried out using the IMRPhenomPv2 waveform model, which also forms the basis for all the tests applied, except for FTI, where we use SEOBNRv4_ROM as in most applications to LIGO-Virgo data [4–6]. The applications of the TIGER and FTI tests to all these cases just consider the 2PN deviation parameter, as well as the 1PN deviation parameter in a few cases, and the $\delta\alpha_2$ and $\delta\beta_2$ post-inspiral parameters for TIGER. The MDR test just considers the $\alpha = 0$ case that agrees with the massive graviton dephasing.

The results use here include the corrected GW170608-like MDR simulated observations: The published results used an GW170608-like MDR simulated observation with no massive graviton dephasing, due to a typo in the command used to create it. An erratum for [7] is in preparation.

4.2. Eccentric cases

We now detail the simulated observations from [8], which use numerical relativity simulations of nonspinning BBHs from the Simulating eXtreme Spacetimes collaboration [47], both quasicircular and with eccentricities of ~ 0.05 and ~ 0.1 , at ~ 17 Hz for the total mass of

$80M_{\odot}$ used for these cases (see [48] for information about the eccentric simulations, including how the eccentricity is estimated). For each eccentricity, there are three mass ratios, q : 1, 2, and 3. These simulated observations are created to be face-on, so only the $m = 2$ modes contribute, and the ℓ modes considered are then matched to the test being carried out. Thus, they are just the dominant $\ell = m = 2$ modes for TIGER and FTI, where the analyses use the IMRPhenomPv2 and SEOBNRv4HM_ROM waveform models, respectively, but include the $\ell = 3$, $m = 2$ modes for MDR, which uses the IMRPhenomXPHM waveform model (and IMRCT, which also uses IMRPhenomXPHM, but we do not use those results here, as described below). The SNRs are ~ 120 , ~ 105 and ~ 90 for mass ratios of 1, 2 and 3, respectively, with the forecast O4 sensitivities [46] considered. We do not consider the IMRCT results in these cases, due to the bias found in [8] for even the quasicircular cases, which is attributable to the face-on inclination angle and the inclusion of higher modes.

5. Results

We now apply the meta IMRCT to the simulated observations discussed in the previous section, using the results of the tests given in section 2 along with GR PE. Here we consider the positive and negative A_{α} results from the MDR test separately. Additionally, as mentioned above, we exclude the (I, MR) pair that just reproduces the standard IMRCT.

Table 1. Summary of meta IMRCT on GW150914-like simulated observations (simul. obs.). Here $>$ represents for larger GR deviation and $<$ represents for smaller GR deviation. We show the specific testing parameters or other PE results used. We apply the meta IMRCT to all these pairs of results. The GR quantile and Gaussian σ columns give the median and 90% interval for the distribution of these quantities for all the pairs of tests considered. The GR quantile is two dimensional and 0 denotes agreement with GR. The quantity $N^{100\%}$ denotes the number of pairs where the GR quantile from the meta IMRCT and the maximum GR quantile from the individual tests are both 100%. The quantities $N_{\text{miMRCT}}^{\text{grtr},90\%}$ and $N_{\text{miMRCT}}^{\text{grtr}}$ denote the number of pairs where the meta IMRCT gives a larger GR quantile than the individual tests do, where the first one also restricts to the pairs where the meta IMRCT GR quantile is $\geq 90\%$. The quantity N_{tot} gives the total number of pairs.

Simul. Obs.	GR PE	IMRCT	TIGER	FTI	MDR	GR Quantile (%)	Gaussian σ	$N^{100\%}/N_{\text{miMRCT}}^{\text{grtr},90\%}/N_{\text{miMRCT}}^{\text{grtr}}/N_{\text{tot}}$
EOB, GR	✓	I, MR	$\delta\varphi_2, \delta\varphi_4, \delta\alpha_2, \delta\beta_2$	$\delta\varphi_2, \delta\varphi_4$	\tilde{A}_0	$2.8_{-1.8}^{+63.1}$	$0.0_{-0.0}^{+0.9}$	0/0/12/54
Phenom, GR	✓	I, MR	$\delta\varphi_2, \delta\varphi_4, \delta\alpha_2, \delta\beta_2$	$\delta\varphi_2, \delta\varphi_4$	\tilde{A}_0	$3.7_{-3.1}^{+57.8}$	$0.0_{-0.0}^{+0.8}$	0/0/7/54
modified EOB, $>$	✓	I, MR	$\delta\varphi_2, \delta\varphi_4, \delta\alpha_2, \delta\beta_2$	$\delta\varphi_2, \delta\varphi_4$	\tilde{A}_0	$99.6_{-89.0}^{+0.4}$	$2.9_{-2.8}^{+0.1}$	25/2/2/54
modified EOB, $<$	✓	I, MR	$\delta\varphi_4, \delta\alpha_2, \delta\beta_2$	$\delta\varphi_4$	\tilde{A}_0	$12.4_{-10.6}^{+58.8}$	$0.2_{-0.1}^{+0.9}$	0/0/5/35
MDR, $>$	✓	I, MR	$\delta\varphi_2, \delta\varphi_4, \delta\alpha_2, \delta\beta_2$	$\delta\varphi_2, \delta\varphi_4$	\tilde{A}_0	$71.5_{-68.7}^{+27.0}$	$1.1_{-1.0}^{+1.4}$	0/1/1/54
MDR, $<$	✓	I, MR	$\delta\varphi_4, \delta\alpha_2, \delta\beta_2$	$\delta\varphi_4$	\tilde{A}_0	$5.1_{-3.8}^{+66.9}$	$0.1_{-0.0}^{+1.0}$	0/0/3/35
TIGER, $>$	✓	I, MR	$\delta\varphi_4, \delta\alpha_2, \delta\beta_2$	$\delta\varphi_4$	\tilde{A}_0	$99.6_{-98.4}^{+0.3}$	$3.0_{-3.0}^{+0.1}$	17/0/0/35
TIGER, $<$	✓	I, MR	$\delta\varphi_4, \delta\alpha_2, \delta\beta_2$	$\delta\varphi_4$	\tilde{A}_0	$85.9_{-84.4}^{+14.1}$	$1.5_{-1.4}^{+1.5}$	4/0/0/35
FTI, $>$	✓	I, MR	$\delta\varphi_4, \delta\alpha_2, \delta\beta_2$	$\delta\varphi_4$	\tilde{A}_0	$100_{-42.5}^{+0.0}$	$3.0_{-2.1}^{+0.0}$	30/0/0/35
FTI, $<$	✓	I, MR	$\delta\varphi_4, \delta\alpha_2, \delta\beta_2$	$\delta\varphi_4$	\tilde{A}_0	$90.5_{-89.0}^{+9.4}$	$1.7_{-1.6}^{+1.3}$	10/1/1/35

5.1. GW150914-like cases

We summarize the meta IMRCT results for the GW150914-like GR and non-GR simulated observations in table 1 as well as in figures 1 (GR and larger GR deviations) and 2 (smaller GR deviations) \parallel . This table (and similar ones later) has a format where the different simulated observations are in separate rows, while the columns give the results of the different tests used in the application of the meta IMRCT as well as summaries of the meta IMRCT results, including various comparisons of the meta IMRCT GR quantiles with those of the individual tests, discussed below.

The meta IMRCT indeed finds that the GR simulated observations are consistent with GR, and recovers deviations from GR for many pairs for the larger GR deviations. In particular, the median GR quantile is $\geq 99.6\%$ for all the larger GR deviations except for the MDR simulated observation (though there are some pairs with considerably smaller GR quantiles for all simulated observations), while the median GR quantiles for the GR simulated observations are $< 4\%$. Indeed, the maximum GR quantiles for the GR simulated observations are only 68% and 64%, respectively, for the EOB and Phenom versions, coming from the $(\delta\varphi_2^{\text{TIGER}}, \delta\alpha_2)$ and $(\delta\varphi_2^{\text{TIGER}}, \delta\beta_2)$ pairs, respectively.

For a few of the simulated observations with larger GR deviations, the meta IMRCT gives a GR quantile $\geq 90\%$ and this GR quantile is larger than that from either of the individual tests (this is the number $N_{\text{mIMRCT}}^{\text{grtr}, 90\%}$ in the table) ∇ . Here, to make this comparison, we have converted the 1d GR quantiles Q_{1d} for the individual tests to the equivalent of the 2d GR quantile, given by $|2Q_{1d} - 1|$ (so that 0 indicates agreement with GR). Moreover, we round (2d or equivalent) GR quantiles above 99.7% to 100%. This lower bound arises because larger GR quantiles cannot be stated with certainty from the $\sim 10^4$ posterior samples we have, as discussed in [8]. We thus also quote in the tables (as $N^{100\%}$) the number of pairs for which the meta IMRCT and at least one of the individual tests both give a GR quantile of 100%. For the modified EOB simulated observation with the larger GR deviation, the pairs of $(\delta\varphi_2^{\text{TIGER}}, \delta\alpha_2)$ and $(\delta\varphi_2^{\text{TIGER}}, \delta\beta_2)$ gave GR quantiles of 100% and 99.7%, respectively, which are greater than the corresponding individual GR quantiles, which are 99.6%, 95.6% and 27.8% for $\delta\varphi_2^{\text{TIGER}}$, $\delta\alpha_2$ and $\delta\beta_2$, respectively. We also find that there are 25 pairs (close to half of the total of 54 pairs) where the meta IMRCT GR quantile and the larger of the two individual GR quantiles both round to 100%. In such cases a significantly larger number of samples would be necessary to determine whether the meta IMRCT identifies the GR deviation with higher significance than the individual tests do.

For the MDR simulated observation with the larger GR deviation, the meta IMRCT only gives a median GR quantile of 71.5%, but there is still one pair for which the meta IMRCT gives a larger GR quantile than individual tests: $(\delta\varphi_2^{\text{FTI}}, 1)$ gives a meta IMRCT GR quantile of 91.6%, while the individual GR quantiles are 84.4% and 90.1% for $\delta\varphi_2^{\text{FTI}}$ and the standard IMRCT, respectively. For the TIGER and FTI simulated observations with the larger GR deviation, there are no cases where the meta IMRCT gives a larger GR quantile than the individual tests do, though there are many cases where the meta IMRCT and at least one of the individual tests both identify these very significant GR deviations with a GR quantile of 100%. Indeed, this is the case for 17 (49%) of the pairs for the TIGER simulated observation and 30 (86%) of the pairs for the FTI simulated observation.

\parallel The sign of the deviation parameters depends on the ordering of the pair. We choose the following ordering for these plots and the similar ones below: GR PE, I, MR, $\delta\varphi_2^{\text{TIGER}}$, $\delta\varphi_4^{\text{TIGER}}$, $\delta\alpha_2$, $\delta\beta_2$, $\delta\varphi_2^{\text{FTI}}$, $\delta\varphi_4^{\text{FTI}}$, $A_0^{<0}$, $A_0^{>0}$, where not all of these are present for a given simulated observation.

∇ There are additional cases where the meta IMRCT gives a larger GR quantile than do the individual tests, but this GR quantile is $< 90\%$, and thus there is not a significant GR deviation identified. We thus give such results in the tables with $N_{\text{mIMRCT}}^{\text{grtr}}$ but do not comment on them in detail.

For the GW150914-like cases with smaller GR deviations, the meta IMRCT does not recover any significant deviation from GR in the modified EOB or massive graviton simulated observations, giving maximum GR quantiles of 75.6%, though the largest GR quantile from the individual tests in those cases is only 90.7%, for the MDR analysis of the modified EOB simulated observation. However, the meta IMRCT recovers a slightly larger GR quantile than the larger GR quantile for the individual tests for the $(\delta\beta_2, \delta\alpha_2)$ pair and the FTI simulated observation, where the pair gives a GR quantile of 97.81%, while the individual tests give GR quantiles of 97.77% and 12.69%, respectively. There are also 10 pairs for which the meta IMRCT GR quantile and the larger of the two individual GR quantiles both round to 100% for the FTI simulated observation, as well as 4 such pairs for the TIGER simulated observation, though there are no pairs for which the meta IMRCT gives a larger GR quantile.

5.2. GW170608-like cases

Table 2. The analogue of table 1 for the GW170608-like simulated observations.

Simul. Obs.	GR PE	IMRCT	TIGER	FTI	MDR	GR Quantile (%)	Gaussian σ	$N^{100\%}/N_{\text{mIMRCT}}^{\text{gtr},90\%}/N_{\text{mIMRCT}}^{\text{gtr}}/N_{\text{tot}}$
EOB, GR	✓	×	$\delta\varphi_4, \delta\alpha_2, \delta\beta_2$	$\delta\varphi_4$	\bar{A}_0	$16_{-15.4}^{+10.6}$	$0.2_{-0.2}^{+0.1}$	0/0/3/21
Phenom, GR	✓	×	$\delta\varphi_4, \delta\alpha_2, \delta\beta_2$	$\delta\varphi_4$	\bar{A}_0	$17.9_{-17.2}^{+10.9}$	$0.2_{-0.2}^{+0.1}$	0/0/10/21
modified EOB	✓	×	$\delta\varphi_4, \delta\alpha_2, \delta\beta_2$	$\delta\varphi_4$	\bar{A}_0	$27.2_{-26.1}^{+18.6}$	$0.3_{-0.3}^{+0.3}$	0/0/6/21
MDR	✓	×	$\delta\varphi_4, \delta\alpha_2, \delta\beta_2$	$\delta\varphi_4$	\bar{A}_0	$63.7_{-62.8}^{+36.3}$	$0.9_{-0.9}^{+2.1}$	5/0/4/21
TIGER	✓	×	$\delta\varphi_4, \delta\alpha_2, \delta\beta_2$	$\delta\varphi_4$	\bar{A}_0	$72.4_{-28.1}^{+17.3}$	$0.1_{-0.5}^{+0.5}$	0/0/3/21
FTI	✓	×	$\delta\varphi_4, \delta\alpha_2, \delta\beta_2$	$\delta\varphi_4$	\bar{A}_0	$55.7_{-55.5}^{+24.6}$	$0.8_{-0.8}^{+0.5}$	0/0/6/21

The GW170608-like simulated observations are for GR signals and non-GR signals with smaller GR deviations. We show the results for these simulated observations in table 2 and figure 3. The meta IMRCT finds that the GR signals are indeed consistent with GR, with a largest GR quantile of 30.4% [for the $(\delta\varphi_4^{\text{TIGER}}, \delta\alpha_2)$ pair and the Phenom simulated observation]. For the modified EOB simulated observation, none of the individual tests find a significant GR deviation (largest GR quantile of 71.7% for the MDR analysis), and the meta IMRCT finds even smaller GR quantiles [the largest is 49.4% for the $(\delta\varphi_4^{\text{TIGER}}, \delta\alpha_2)$ pair]. However, for the MDR simulated observation, the MDR test finds a significant GR deviation, with a GR quantile of 100%, though the maximum GR quantile for the other analyses is only 52.5% for $\delta\beta_2$. The meta IMRCT obtains GR quantiles of 100% for all pairs involving $A_0^{>0}$ (the sign of A_0 given in the injection), except for $(\delta\varphi_4^{\text{FTI}}, A_0^{>0})$, where the GR quantile is 99.1%. The largest GR quantile for the pairs not involving $A_0^{>0}$ is 89.3% for the (GR PE, $\delta\alpha_2$) pair; the $\delta\alpha_2$ GR quantile is 20.7%.

For the TIGER simulated observation, the $\delta\varphi_4^{\text{TIGER}}$, $\delta\varphi_4^{\text{FTI}}$ and $\delta\beta_2$ analyses find GR quantiles $> 90\%$, though the meta IMRCT only gives a GR quantile $> 90\%$ for the $(\delta\varphi_4^{\text{TIGER}}, \delta\alpha_2)$ pair, where the individual GR quantiles are 96.6% and 34.6%, respectively. For the FTI simulated observation, the maximum GR quantile for the individual tests is 87.9% for $\delta\varphi_4^{\text{FTI}}$. The maximum GR quantile from the meta IMRCT is 87.0% for the $(\delta\varphi_4^{\text{TIGER}}, \delta\alpha_2)$ pair, where the individual GR quantiles are 79.9% and 29.7%, respectively.

Table 3. Summary of meta IMRCT applied on numerical relativity quasicircular and eccentric simulated observations. Here q is the mass ratio and e is the eccentricity at ~ 17 Hz. The TIGER, FTI and MDR tests are performed with all deviation parameters. The other columns have the same meanings as in table 1.

Simulation	q	e	GR PE	IMRCT	TIGER	FTI	MDR	GR Quantile (%)	Gaussian σ	$N^{100\%}/N_{\text{mIMRCT}}^{\text{gr},90\%}/N_{\text{mIMRCT}}^{\text{gr}}/N_{\text{tot}}$
SXS:BBH:1155	1	0.0	✓	×	All	All	All	$25.4_{-14.5}^{+34.4}$	$0.3_{-0.2}^{+0.5}$	0/0/240/703
SXS:BBH:1355	1	0.05	✓	×	All	All	All	$37.0_{-32.5}^{+48.9}$	$0.5_{-0.4}^{+1.0}$	0/0/2/703
SXS:BBH:1357	1	0.10	✓	×	All	All	All	$99.2_{-86.7}^{+0.8}$	$2.6_{-2.5}^{+0.4}$	297/15/18/703
SXS:BBH:1222	2	0.0	✓	×	All	All	All	$6.1_{-2.6}^{+20.7}$	$0.1_{-0.0}^{+0.3}$	0/0/31/703
SXS:BBH:1364	2	0.05	✓	×	All	All	All	$46.3_{-41.7}^{+45.3}$	$0.6_{-0.6}^{+1.1}$	0/0/41/703
SXS:BBH:1368	2	0.10	✓	×	All	All	All	$87.9_{-77.6}^{+12.1}$	$1.6_{-1.4}^{+1.4}$	43/107/126/703
SXS:BBH:2265	3	0.0	✓	×	All	All	All	$4.9_{-2.6}^{+16.5}$	$0.1_{-0.0}^{+0.2}$	0/0/33/703
SXS:BBH:1371	3	0.06	✓	×	All	All	All	$44.5_{-42.2}^{+55.5}$	$0.6_{-0.6}^{+2.4}$	112/224/225/703
SXS:BBH:1373	3	0.09	✓	×	All	All	All	$96.9_{-88.2}^{+3.1}$	$2.2_{-2.0}^{+0.8}$	282/0/11/703

5.3. Quasicircular and eccentric numerical relativity waveforms

We show the results obtained by performing meta IMRCT on the TIGER, FTI, and MDR test results applied on eccentric and quasicircular simulated signals from [8] in table 3 and figures 4, 5 and 6⁺. The meta IMRCT finds that all quasicircular cases are consistent with GR at the 90% credible level, with largest GR quantiles of 86.4%, 40.3% and 37.4% for $q = 1, 2$ and 3, respectively. This result is as expected since the tests use waveform models that describe binaries on quasicircular orbits, though the significantly larger maximum and median GR quantiles for the $q = 1$ simulated observation are unexpected: One expects that the waveform models used are all quite accurate for nonspinning quasicircular binaries with $q \leq 3$, and the GR PE finds only the expected small biases from the true values for the quasicircular simulated observations. The largest GR quantiles for $q = 1$ are for pairs with $A_{0.5}^{<0}$, and the 46 largest GR quantiles (with values $\geq 55.6\%$) are for pairs with either $A_{0.5}^{<0}$ or $A_{0.5}^{>0}$. Both the $A_{0.5}$ analyses recover precessing binaries, and the posterior distributions of final masses and spins extend to larger values than the distribution from GR PE, particularly for the $A_{0.5}^{<0}$ analysis. However, the $A_{0.5}^{<0}$ analysis finds a notably smaller maximum likelihood than GR PE or the $A_{0.5}^{>0}$ analysis, so it may have been somewhat undersampled. This shows how the meta IMRCT can identify possible problems with analyses. However, when excluding the pairs with $A_{0.5}^{\geq 0}$, one still obtains a median GR quantile of 22.6%, considerably larger than the medians for the other quasicircular cases, though when excluding all the pairs with MDR, one obtains a median and maximum GR quantile of 17.0% and 43.6%, and when also excluding the pairs with FTI, 11.5% and 40.2%, showing that these are the analyses that give the largest bias away from GR in the meta IMRCT.

⁺ The ordering of the pairs in these figures is the same as in the previous figures, except that since we now have the full set of testing parameters, each set of testing parameters (i.e. the $\{\delta\varphi_k\}, \{\delta\alpha_k\}, \{\delta\beta_k\}$ and $\{A_{\alpha}^{\geq 0}\}$) is ordered by increasing values of its index. Explicitly, for the MDR parameters, we have $A_0^{<0}, A_0^{>0}, A_{0.5}^{<0}, A_{0.5}^{>0}, \dots$

Also as expected, the median and maximum meta IMRCT GR quantiles increase with eccentricity (the maxima are 99.5%, 95.4% and 100% for $q = 1, 2$ and 3 with the smaller eccentricity and 100% for all three mass ratios with the larger eccentricity), with the median value excluding GR at $> 2\sigma$ Gaussian equivalent for the $q = 1$ and 3 larger eccentricity simulated observations. The meta IMRCT also gives a GR quantile that is larger than 90% and larger than that from either of the individual tests for the $q = 1$ and 2 simulated observations with the larger eccentricities, and the $q = 3$ simulated observation with the smaller eccentricity. For the $q = 1$ simulated observation with the larger eccentricity, the 15 meta IMRCT GR quantiles in such cases range from 98.9% to 100% and all of the pairs involve $\delta\varphi_k^{\text{FTI}}$, $k \in \{6, 7\}$ paired with either GR PE or MDR. The largest difference between the meta IMRCT GR quantile and the largest individual GR quantile in these cases is 1.1 percentage points for the $(\delta\varphi_7^{\text{FTI}}, A_{1.5}^{<0})$ and $(\delta\varphi_7^{\text{FTI}}, A_{2.5}^{>0})$, both of which have a meta IMRCT GR quantile of 100%; the $\delta\varphi_7^{\text{FTI}}$, $A_{1.5}^{<0}$ and $A_{2.5}^{>0}$ GR quantiles are 98.8%, 24.4% and 93.1%, respectively. There are also 297 pairs (42% of the total) where the meta IMRCT gives a GR quantile of 100% and at least one of the individual tests does, as well.

For the $q = 2$ simulated observation with the larger eccentricity, the 107 cases (15% of the total) with meta IMRCT GR quantile $> 90\%$ and greater than the individual GR quantiles are mostly FTI–MDR pairs, as well as some GR PE–FTA pairs and the $\delta\varphi_7^{\text{TIGER}}$ and $\delta\beta_2$ pairs with $A_{2.5}^{>0}$. The largest meta IMRCT GR quantiles in this case are 100%, and they can be significantly larger than the individual GR quantiles, in particular being greater by 42.6 percentage points in 9 cases with $A_{2.5}^{>0}$ in the pair, including its pair with $A_{2.5}^{<0}$. The largest difference between the meta IMRCT GR quantile and the maximum individual GR quantile is for the pair (GR PE, $\delta\varphi_3^{\text{FTI}}$), where meta IMRCT gives a GR quantile of 90.4% and $\delta\varphi_3^{\text{FTI}}$ gives a GR quantile of 4.8%.

For the $q = 3$ simulated observation with the larger eccentricity, there are 282 pairs (40% of the total) with a meta IMRCT GR quantile of 100% (where at least one of the individual tests does, as well) and 391 pairs (56% of the total) with a meta IMRCT GR quantile $> 90\%$, though there are no cases where the meta IMRCT gives a larger GR quantile than the individual tests do. The pairs that give a GR quantile of 100% are mostly pairs of MDR with another analysis, but also involve pairs of TIGER and FTI with GR PE, and TIGER and FTI inspiral with $\delta\alpha_k$, as well as $(\delta\beta_3, \delta\alpha_2)$.

However, for the $q = 3$ simulated observation with the smaller eccentricity, there are 224 cases (32% of the total) with meta IMRCT GR quantile $> 90\%$ and greater than the individual GR quantiles. All of these have a meta IMRCT GR quantile of 100% and consist of pairs of GR PE, TIGER, or FTI with MDR. The largest difference between the meta IMRCT GR quantile and the largest individual GR quantile is 43.0 percentage points for the 6 pairs of GR PE, $\delta\beta_3$ or $\delta\alpha_4$ with $A_3^{\geq 0}$, where the meta IMRCT gives a GR quantile of 100%, while $\delta\beta_3$, $\delta\alpha_4$ and A_3 have GR quantiles of 17.9%, 20.1% and 57.0%, respectively. There are also 112 cases where the meta IMRCT and at least one of the individual tests both give a GR quantile of 100%.

For the $q = 1$ simulated observation with the smaller eccentricity, the largest meta IMRCT GR quantile is 99.5%, for the $(\delta\beta_3, \delta\alpha_3)$ pair, where the individual GR quantiles are 100% and 98.9%, respectively. For the $q = 2$ simulated observation with the smaller eccentricity, the largest meta IMRCT GR quantile is 95.4%, for the $(\delta\varphi_1^{\text{TIGER}}, A_3^{>0})$ pair, where the individual GR quantiles are 98.8% and 88.0%, respectively.

Table 4. Summary of meta IMRCT on real events, similar to table 1.

Event	GR PE	IMRCT	TIGER	FTI	MDR	GR Quantile (%)	Gaussian σ	$N^{100\%}/N_{\text{miMRCT}}^{\text{grtr},90\%}/N_{\text{miMRCT}}^{\text{grtr}}/N_{\text{tot}}$
GW170817	✓	×	All	All	×	$4.2_{-3.3}^{+14.3}$	$0.0_{-0.0}^{+0.2}$	0/0/0/210
GW190412	✓	I, MR	All	All	All	$14.3_{-10.9}^{+38.2}$	$0.2_{-0.1}^{+0.5}$	0/0/38/945
GW190521	✓	×	All	×	×	$4.2_{-1.2}^{+4.7}$	$0.0_{-0.0}^{+0.0}$	0/0/0/15
GW190814	✓	×	All	All	All	$6.2_{-5.8}^{+40.0}$	$0.1_{-0.1}^{+0.5}$	0/0/19/861
GW200225_060421	✓	I, MR	×	All	All	$59.1_{-52.0}^{+14.1}$	$0.8_{-0.7}^{+0.3}$	0/0/2/405

5.4. Real events

We now consider five real events: four exceptional events and one event that showed problematic results in the LVK O3b testing GR analyses [6]. The exceptional events are the likely binary neutron star GW170817 [14] (see also [49]); the unequal-mass binary GW190412 [15]; the high-mass event GW190521 [16, 50]; and the very unequal-mass event GW190814 [17], whose secondary is likely a light black hole, but could also be a heavy neutron star or something more exotic. See also [51] for the LVK’s latest analyses of the latter three signals. The problematic event is GW200225_060421 [1], which was an outlier in the LVK MDR results and had the smallest p -value in the residuals test (0.05) of all the O3 events [6]. It also had a larger significance in the minimally modelled coherent WaveBurst search [52] than in the templated searches that use GR compact binary waveforms [1].

We apply the meta IMRCT to these events using the LVC/LVK GR PE results and their testing GR results for the tests we considered in previous sections (also including the -1PN test, with testing parameter $\delta\varphi_{-2}$, for both FTI and TIGER) from [3, 5, 6], using the data from [53–55]. The waveform models employed for GR parameter estimation of these events is carried out using IMRPhenomPv2_NRTidal (GW170817), IMRPhenomPv3HM (GW190412, GW190521 and GW190814) and IMRPhenomXPHM (GW200225_060421). The tests we consider are based on these waveforms except for FTI, which uses SEOBNRv4HM_ROM, and MDR, which uses IMRPhenomPv2 for GW190412, GW190521 and GW190814. The TIGER and FTI tests for GW170817 use IMRPhenomPv2_NRTidal and SEOBNRv4_ROM_NRTidal, respectively. We show the meta IMRCT results in table 4 and figure 7, finding consistency with GR, as in the LVK results for the non-problematic events*. Specifically, the maximum GR quantile we obtain is 89.4 for the $(\delta\varphi_{-2}^{\text{FTI}}, \delta\alpha_4)$ pair and GW190412; the specifics of the maxima for each event are given in table 5. While the meta IMRCT does not find any GR quantiles $> 90\%$ for the problematic event GW200225_060421, the median GR quantile of 59.1% is considerably larger than for the other real events or GR simulated observations. We also found that the individual FTI GR quantiles for this event are $> 90\%$ for the 1.5PN $(\delta\varphi_3)$ and higher testing parameters, indeed as large as 94.4% for the 2.5PN log $(\delta\varphi_{5l})$ testing parameter, though this is not remarked on in the LVK testing GR paper, which only presents upper bounds for individual events.

Due to the specific mass priors used in the different tests for GW190412, GW190814 and GW200225_060421, the priors on the deviation parameters peak well away from the GR value

* While GW170817 is likely a binary neutron star, and the tests of GR are applied using waveforms that include tidal effects [30], in our application of the meta IMRCT we use the BBH final mass and spin fits, so we only consider the consistency of the initial masses and spins, not the tidal deformabilities.

Table 5. The details of the meta IMRCT pairs that give the largest GR quantile for the real events.

Event	Max. mIMRCT GR Quantile (%)	Pair	Indiv. GR quantiles (%)
GW170817	45.1	$(\delta\varphi_0^{\text{TIGER}}, \delta\varphi_4^{\text{TIGER}})$	81.1, 82.2
GW190412	89.4	$(\delta\varphi_{-2}^{\text{FTI}}, \delta\alpha_4)$	65.0, 0.5
GW190521	11.4	$(\delta\alpha_2, \delta\alpha_4)$	10.3, 49.5
GW190814	77.2	$(\delta\varphi_{-2}^{\text{TIGER}}, \delta\beta_3)$	94.5, 78.1
GW200225_060421	79.0	$(\delta\varphi_6^{\text{FTI}}, A_{1.5}^{<0})$	94.3, 95.9

for some pairs (distances of > 0.4 from the GR value). The cases that peak the furthest away from the GR value are any of the pairs of GR PE, FTI or MDR with the IMR consistency inspiral or merger-ringdown analyses for GW190412 and with TIGER for GW190814. In all these cases, the peak of the posterior is around $(\Delta M_f/\bar{M}_f, \Delta\chi_f/\bar{\chi}_f) = (1.3, 0)$, with the ordering of the pairs we use, except for the GR PE pair, where it has the opposite sign. Nevertheless, in all cases the posteriors still peak around the GR value. We find that for GW190814 the significant shift of the prior away from the region preferred by the posterior causes some artefacts when reweighting certain pairs to a flat prior in $(\Delta M_f/\bar{M}_f, \Delta\chi_f/\bar{\chi}_f)$. We are able to mitigate the artefacts by imposing a threshold on the probability density of 8×10^{-4} of the maximum, which only changes the GR quantiles of the pairs without artefacts by less than 1.6 percentage points. The largest GR quantile difference are for pairs with FTI. The largest difference for non-FTI pairs is 0.8 percentage points for the $(\delta\varphi_{-2}^{\text{TIGER}}, \delta\beta_2)$ pair.

For GW190412, we find that for all the aforementioned pairs where the prior peaks well away from the GR value the reweighted posteriors are all centred around zero, with no secondary peaks, while the unweighted posteriors have a secondary peak around the peak of the prior. However, for the pairs of the TIGER analyses with the merger-ringdown analysis, where the prior peaks around $(\Delta M_f/\bar{M}_f, \Delta\chi_f/\bar{\chi}_f) = (0.8, 0)$, the unweighted posterior still peaks at about the same location as for the FTI and merger-ringdown pairs, so one obtains the secondary peaks in the reweighted posterior seen in figure 7.

6. Conclusions

Gravitational waves from coalescing compact binaries are now routinely used to test GR in highly dynamical, non-linear gravity regions of space-time. There are now quite a few tests that have been applied to inspect the consistency of the LVK observations with GR predictions but no deviations from GR have been reported so far. In this paper, we proposed a new test of GR, called the meta IMRCT, that checks the consistency between the results from two different tests of GR analyses (or a GR analysis) on a given compact binary coalescence. Specifically, the meta IMRCT checks the consistency of the inference of the final mass and spin of the remnant black hole from two different tests of GR analyses or the standard GR analysis of the signal.

The meta IMRCT is a generalized version of the standard IMRCT (proposed in [9, 10]) that instead checks the consistency between the final mass and spin of the remnant estimated from independently analyzing the low- and high-frequency portions of the signal. Unlike the standard IMRCT, which is restricted to binaries whose masses are neither too low nor too high, so that there is sufficient SNR in both the inspiral and merger-ringdown portions of the signal (see, e.g. [6]), the meta IMRCT can be applied to BBHs of any total mass and SNR. It can even be applied to binaries involving neutron stars. Here one can use the BBH final mass and spin fits since these fits are just used to conveniently map the initial mass and spin parameter space to a two-dimensional space. However, it would also be possible to use fits for the final mass and spin of binary neutron star or neutron star–black hole binaries (e.g. the ones in [56] and [57]; the latter are just for neutron star–black hole binaries, but are more accurate, though the accuracy is not a great concern here, since these fits are just used to provide a convenient, physically inspired dimensionality reduction). The meta IMRCT is also computationally inexpensive to perform, without any additional stochastic sampling or evaluation of waveforms, once one has obtained the individual test results.

To assess the performance of the meta IMRCT we applied it to a variety of simulated GR and non-GR signals from quasicircular BBHs from [7] and numerical relativity GR signals from quasicircular and eccentric BBHs from [8] (see section 4 for details). Specifically, we utilize the tests of GR and standard GR parameter estimation analyses (all using quasicircular waveform models) performed on these simulated signals in [7, 8] and compare the final mass and spin estimates from these in pairs. The tests of GR that are performed in [7, 8] are the two parameterized tests, TIGER and FTI, the modified dispersion relation test, and the standard IMRCT (see section 2 for detail). We apply the meta IMRCT to all the different testing parameters for TIGER, FTI, and the modified dispersion relation test and consider the inspiral and merger-ringdown analyses of the standard IMRCT separately. (Of course, the meta IMRCT reproduces the standard IMRCT results when applied to the inspiral and merger-ringdown analyses.)

We found that the meta IMRCT indeed finds all simulated GR quasicircular signals (Phenom, EOB, or numerical relativity) are consistent with GR, as expected since the waveform models used in all the individual tests are quasicircular. The meta IMRCT reports high GR quantiles (meaning large GR deviations) for signals with larger GR deviations, larger than the individual tests' GR quantiles in some cases. For simulations with smaller GR deviation, the meta IMRCT does not report high GR quantiles, similar to the individual test analyses. For numerical relativity eccentric simulations the meta IMRCT GR quantiles increase with eccentricity, as expected, and give significantly large GR quantities ($\sim 100\%$) for larger eccentricity simulations. For one eccentric case, the meta IMRCT finds a GR deviation for several pairs where the individual tests to which it is being applied do not find a significant deviation, and finds a GR quantile $> 90\%$ and larger than the individual GR quantiles for almost a third of the pairs, so this larger GR deviation would not be significantly reduced by application of a trials factor.

We also applied the meta IMRCT to real compact binary signals GW170817, GW190412, GW190521, GW190814 and GW200225.060421 using the results from the LV(K) collaboration's application of the tests of GR (we considered only TIGER, FTI, MDR and IMRCT). The meta IMRCT finds all these events consistent with GR, giving GR quantiles below 90%.

In the future, it will be interesting to explore applying the meta IMRCT using other GR parameters (such as different combinations of the masses and spins or even extrinsic parameters like distance or inclination angle), or larger numbers of parameters, to see what gives the best performance. It will also be interesting to study the performance of the meta

IMRCT on simulated binary neutron star and neutron star–black hole binaries, in particular seeing whether including the tidal parameters improves the meta IMRCT.

Acknowledgments

We thank Muhammed Saleem, N. V. Krishnendu, James Clark and Purnima Narayan for their work on producing the results from [7, 8] that we use and Apratim Ganguly for useful comments on the paper. NKJ-M is supported by NSF grant AST-2205920. AG is supported in part by NSF grants AST-2205920 and PHY-2308887.

This research has made use of data obtained from the Gravitational Wave Open Science Center (<https://www.gwosc.org/>), a service of the LIGO Laboratory, the LIGO Scientific Collaboration and the Virgo Collaboration. LIGO Laboratory and Advanced LIGO are funded by the United States National Science Foundation (NSF) as well as the Science and Technology Facilities Council (STFC) of the United Kingdom, the Max-Planck-Society (MPS), and the State of Niedersachsen/Germany for support of the construction of Advanced LIGO and construction and operation of the GEO600 detector. Additional support for Advanced LIGO was provided by the Australian Research Council. Virgo is funded, through the European Gravitational Observatory (EGO), by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale di Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by institutions from Belgium, Germany, Greece, Hungary, Ireland, Japan, Monaco, Poland, Portugal, Spain. The construction and operation of KAGRA are funded by Ministry of Education, Culture, Sports, Science and Technology (MEXT), and Japan Society for the Promotion of Science (JSPS), National Research Foundation (NRF) and Ministry of Science and ICT (MSIT) in Korea, Academia Sinica (AS) and the Ministry of Science and Technology (MoST) in Taiwan.

This study used the software packages LALSuite [22], Matplotlib [58], NumPy [59], PESummary [60], and SciPy [61]. This is LIGO document P2400174.

- [1] Abbott R *et al.* (LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration) 2023 *Phys. Rev. X* **13** 041039 (*Preprint* 2111.03606)
- [2] Abac A G *et al.* (LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration) 2024 (*Preprint* 2404.04248)
- [3] Abbott B P *et al.* (LIGO Scientific Collaboration and Virgo Collaboration) 2019 *Phys. Rev. Lett.* **123** 011102 (*Preprint* 1811.00364)
- [4] Abbott B P *et al.* (LIGO Scientific Collaboration and Virgo Collaboration) 2019 *Phys. Rev. D* **100** 104036 (*Preprint* 1903.04467)
- [5] Abbott R *et al.* (LIGO Scientific Collaboration and Virgo Collaboration) 2021 *Phys. Rev. D* **103** 122002 (*Preprint* 2010.14529)
- [6] Abbott R *et al.* (LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration) 2021 (*Preprint* 2112.06861)
- [7] Johnson-McDaniel N K, Ghosh A, Ghonge S, Saleem M, Krishnendu N V and Clark J A 2022 *Phys. Rev. D* **105** 044020 (*Preprint* 2109.06988)
- [8] Narayan P, Johnson-McDaniel N K and Gupta A 2023 *Phys. Rev. D* **108** 064003 (*Preprint* 2306.04068)
- [9] Ghosh A, Ghosh A, Johnson-McDaniel N K, Mishra C K, Ajith P, Del Pozzo W, Nichols D A, Chen Y, Nielsen A B, Berry C P L and London L 2016 *Phys. Rev. D* **94** 021101(R) (*Preprint* 1602.02453)
- [10] Ghosh A, Johnson-McDaniel N K, Ghosh A, Mishra C K, Ajith P, Del Pozzo W, Berry C P L, Nielsen A B and London L 2018 *Class. Quantum Grav.* **35** 014002 (*Preprint* 1704.06784)
- [11] Meidam J *et al.* 2018 *Phys. Rev. D* **97** 044033 (*Preprint* 1712.08772)
- [12] Mehta A K, Buonanno A, Cotesta R, Ghosh A, Sennett N and Steinhoff J 2023 *Phys. Rev. D* **107** 044020 (*Preprint* 2203.13937)
- [13] Mirshekari S, Yunes N and Will C M 2012 *Phys. Rev. D* **85** 024041 (*Preprint* 1110.2720)
- [14] Abbott B P *et al.* (LIGO Scientific Collaboration and Virgo Collaboration) 2017 *Phys. Rev. Lett.* **119** 161101 (*Preprint* 1710.05832)
- [15] Abbott R *et al.* (LIGO Scientific Collaboration and Virgo Collaboration) 2020 *Phys. Rev. D* **102** 043015 (*Preprint* 2004.08342)

- [16] Abbott R *et al.* (LIGO Scientific Collaboration and Virgo Collaboration) 2020 *Phys. Rev. Lett.* **125** 101102 (Preprint 2009.01075)
- [17] Abbott R *et al.* (LIGO Scientific Collaboration and Virgo Collaboration) 2020 *Astrophys. J. Lett.* **896** L44 (Preprint 2006.12611)
- [18] Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1953 *J. Chem. Phys.* **21** 1087–1092
- [19] Hastings W K 1970 *Biometrika* **57** 97–109
- [20] Skilling J 2004 *AIP Conf. Proc.* **735** 395
- [21] Veitch J *et al.* 2015 *Phys. Rev. D* **91** 042003 (Preprint 1409.7215)
- [22] LIGO Scientific Collaboration, Virgo Collaboration and KAGRA Collaboration 2024 LVK Algorithm Library - LALSuite Free software (GPL) URL <https://doi.org/10.7935/GT1W-FZ16>
- [23] Ashton G *et al.* 2019 *Astrophys. J. Suppl.* **241** 27 (Preprint 1811.02042)
- [24] Hannam M, Schmidt P, Bohé A, Haegel L, Husa S, Ohme F, Pratten G and Pürrer M 2014 *Phys. Rev. Lett.* **113** 151101 (Preprint 1308.3271)
- [25] Khan S, Husa S, Hannam M, Ohme F, Pürrer M, Jiménez Forteza X and Bohé A 2016 *Phys. Rev. D* **93** 044007 (Preprint 1508.07253)
- [26] Bohé A, Hannam M, Husa S, Ohme F, Pürrer M and Schmidt P 2016 PhenomPv2 - Technical Notes for LAL Implementation Tech. Rep. LIGO-T1500602 LIGO Project URL <https://dcc.ligo.org/LIGO-T1500602/public>
- [27] Pratten G *et al.* 2021 *Phys. Rev. D* **103** 104056 (Preprint 2004.06503)
- [28] Khan S, Ohme F, Chatziioannou K and Hannam M 2020 *Phys. Rev. D* **101** 024056 (Preprint 1911.06050)
- [29] Dietrich T, Bernuzzi S and Tichy W 2017 *Phys. Rev. D* **96** 121501(R) (Preprint 1706.02969)
- [30] Dietrich T *et al.* 2019 *Phys. Rev. D* **99** 024029 (Preprint 1804.02235)
- [31] Hofmann F, Barausse E and Rezzolla L 2016 *Astrophys. J. Lett.* **825** L19 (Preprint 1605.01938)
- [32] Healy J and Lousto C O 2017 *Phys. Rev. D* **95** 024037 (Preprint 1610.09713)
- [33] Jiménez-Forteza X, Keitel D, Husa S, Hannam M, Khan S and Pürrer M 2017 *Phys. Rev. D* **95** 064024 (Preprint 1611.00332)
- [34] Johnson-McDaniel N K, Gupta A, Ajith P, Keitel D, Birnholtz O, Ohme F and Husa S 2016 Determining the final spin of a binary black hole system including in-plane spins: Method and checks of accuracy Tech. Rep. LIGO-T1600168 LIGO Project <https://dcc.ligo.org/LIGO-T1600168/public/main> URL <https://dcc.ligo.org/T1600168/public>
- [35] Abbott B P *et al.* (LIGO Scientific, Virgo) 2016 *Phys. Rev. Lett.* **116** 221101 [Erratum: Phys.Rev.Lett. 121, 129902 (2018)] (Preprint 1602.03841)
- [36] Bardeen J M, Press W H and Teukolsky S A 1972 *Astrophys. J.* **178** 347
- [37] Agathos M, Del Pozzo W, Li T G F, Van Den Broeck C, Veitch J and Vitale S 2014 *Phys. Rev. D* **89** 082001 (Preprint 1311.0420)
- [38] Bohé A *et al.* 2017 *Phys. Rev. D* **95** 044028 (Preprint 1611.03703)
- [39] Cotesta R, Buonanno A, Bohé A, Taracchini A, Hinder I and Ossokine S 2018 *Phys. Rev. D* **98** 084028 (Preprint 1803.10701)
- [40] Cotesta R, Marsat S and Pürrer M 2020 *Phys. Rev. D* **101** 124040 (Preprint 2003.12079)
- [41] Bayer A E and Seljak U 2020 *J. Cosmol. Astropart. Phys.* **10** 009 (Preprint 2007.13821)
- [42] Abbott B P *et al.* (LIGO Scientific Collaboration and Virgo Collaboration) 2016 *Phys. Rev. Lett.* **116** 061102 (Preprint 1602.03837)
- [43] Abbott B P *et al.* (LIGO Scientific Collaboration and Virgo Collaboration) 2017 *Astrophys. J. Lett.* **851** L35 (Preprint 1711.05578)
- [44] Damour T, Nagar A and Bernuzzi S 2013 *Phys. Rev. D* **87** 084035 (Preprint 1212.4357)
- [45] Will C M 1998 *Phys. Rev. D* **57** 2061 (Preprint gr-qc/9709011)
- [46] Abbott B P *et al.* (KAGRA Collaboration, LIGO Scientific Collaboration, and Virgo Collaboration) 2020 *Living Rev. Relativity* **23** 3 (Preprint 1304.0670)
- [47] Boyle M *et al.* 2019 *Classical Quantum Gravity* **36** 195006 (Preprint 1904.04831)
- [48] Hinder I, Kidder L E and Pfeiffer H P 2018 *Phys. Rev. D* **98** 044015 (Preprint 1709.02007)
- [49] Abbott B P *et al.* (LIGO Scientific Collaboration and Virgo Collaboration) 2019 *Phys. Rev. X* **9** 031040 (Preprint 1811.12907)
- [50] Abbott R *et al.* (LIGO Scientific Collaboration and Virgo Collaboration) 2020 *Astrophys. J. Lett.* **900** L13 (Preprint 2009.01190)
- [51] Abbott R *et al.* (LIGO Scientific Collaboration and Virgo Collaboration) 2024 *Phys. Rev. D* **109** 022001 (Preprint 2108.01045)
- [52] Klimentenko S *et al.* 2016 *Phys. Rev. D* **93** 042004 (Preprint 1511.05999)
- [53] LVC GW170817 testing GR data release, <https://dcc.ligo.org/LIGO-P1800059/public>
- [54] LVC O3a testing GR data release, <https://doi.org/10.7935/903s-gx73>
- [55] LVK O3b testing GR data release, <https://doi.org/10.5281/zenodo.7007369>
- [56] Deng C M 2020 *Mon. Not. R. Astron. Soc.* **497** 643 (Preprint 2007.02744)

- [57] Zappa F, Bernuzzi S, Pannarale F, Mapelli M and Giacobbo N 2019 *Phys. Rev. Lett.* **123** 041102 (*Preprint* 1903.11622)
- [58] Hunter J D 2007 *Comput. Sci. Eng.* **9** 90–95
- [59] Harris C R *et al.* 2020 *Nature (London)* **585** 357–362 (*Preprint* 2006.10256)
- [60] Hoy C and Raymond V 2021 *SoftwareX* **15** 100765 (*Preprint* 2006.06639)
- [61] Virtanen P *et al.* 2020 *Nat. Methods* **17** 261 (*Preprint* 1907.10121)

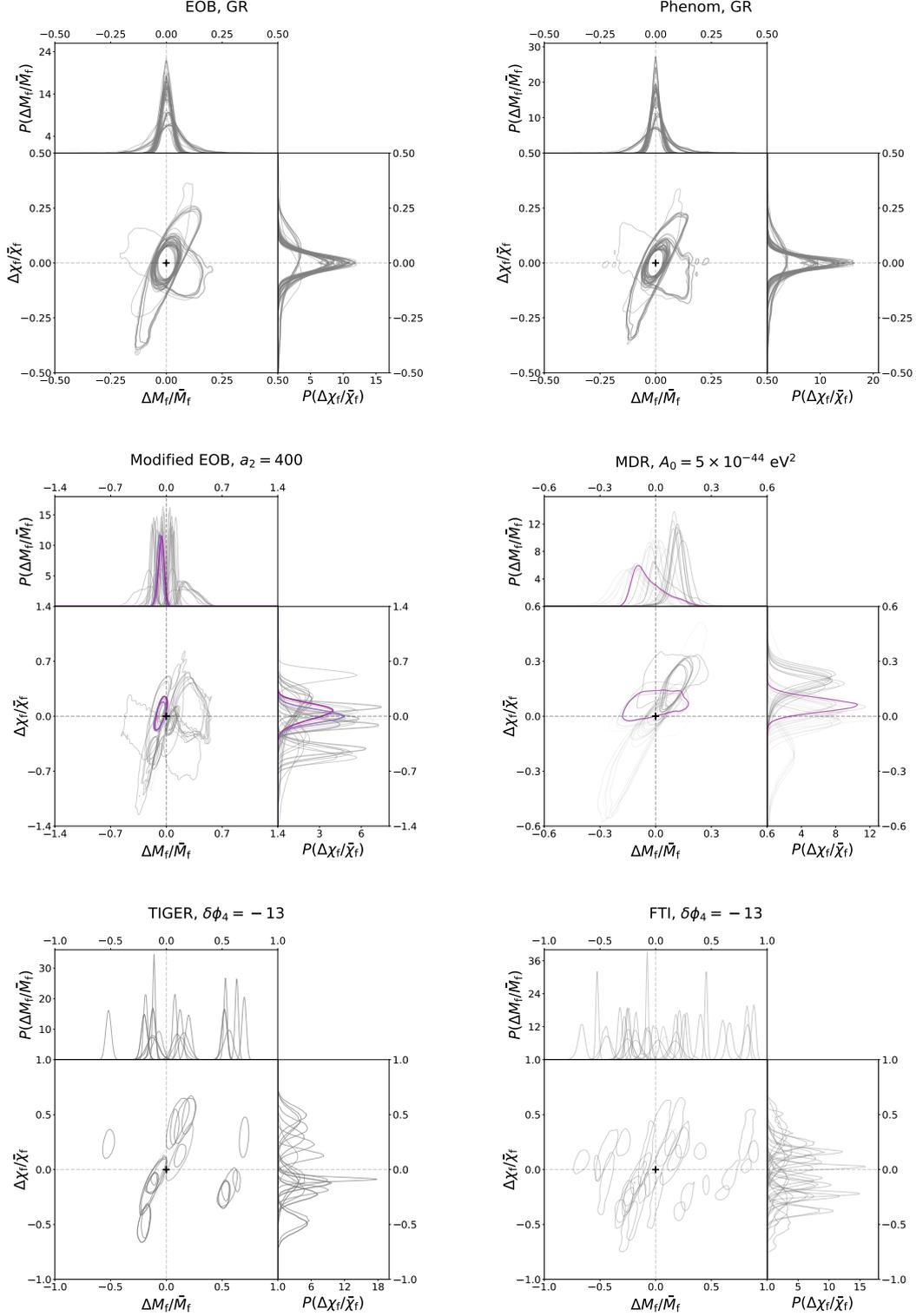


Figure 1. The results of meta IMRCT on the GW150914-like GR simulated observations and those with larger GR deviation presented as 2d contour plots of the posteriors on the deviation parameters. The side panels show the 1d histograms of the marginalized posteriors of $\Delta M_t/\bar{M}_t$ and $\Delta \chi_t/\bar{\chi}_t$. The gray contour delineates pairs of individual tests with color gradient varying based on the value of GR quantile (where darker gray denotes larger GR quantile and lighter gray denotes smaller GR quantile). We highlight in color the three cases with the largest meta IMRCT GR quantile where this is larger than the GR quantile from either of the tests of GR in the pair (or the single test of GR for pairs with GR PE) and also $> 90\%$. Specifically, we assign magenta color to the largest GR quantile, purple to the second-largest, and blue to the third-largest.

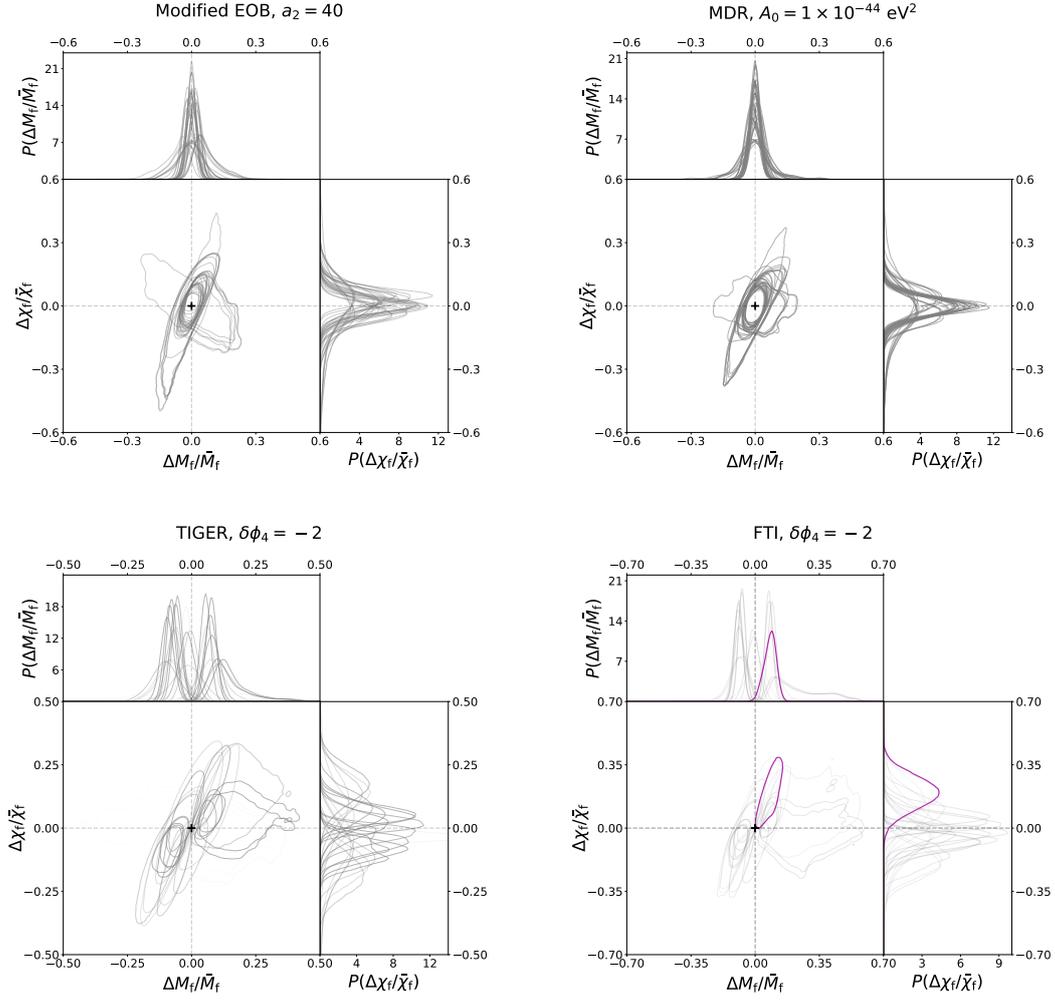


Figure 2. Similar to figure 1, but for the GW150914-like smaller GR deviation simulated observations.

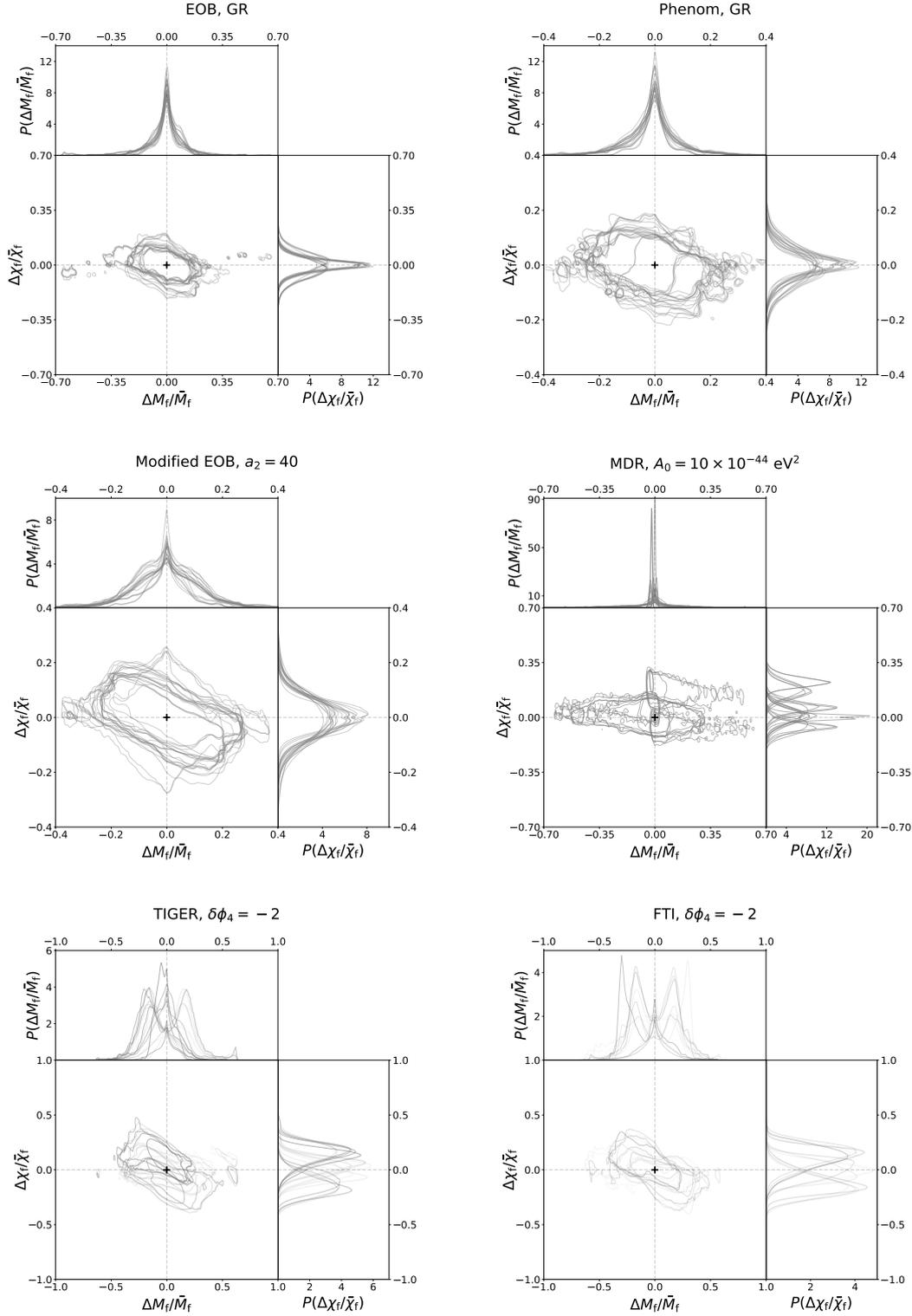


Figure 3. Similar to figure 1 for the GW170608-like smaller GR deviation simulated observations.

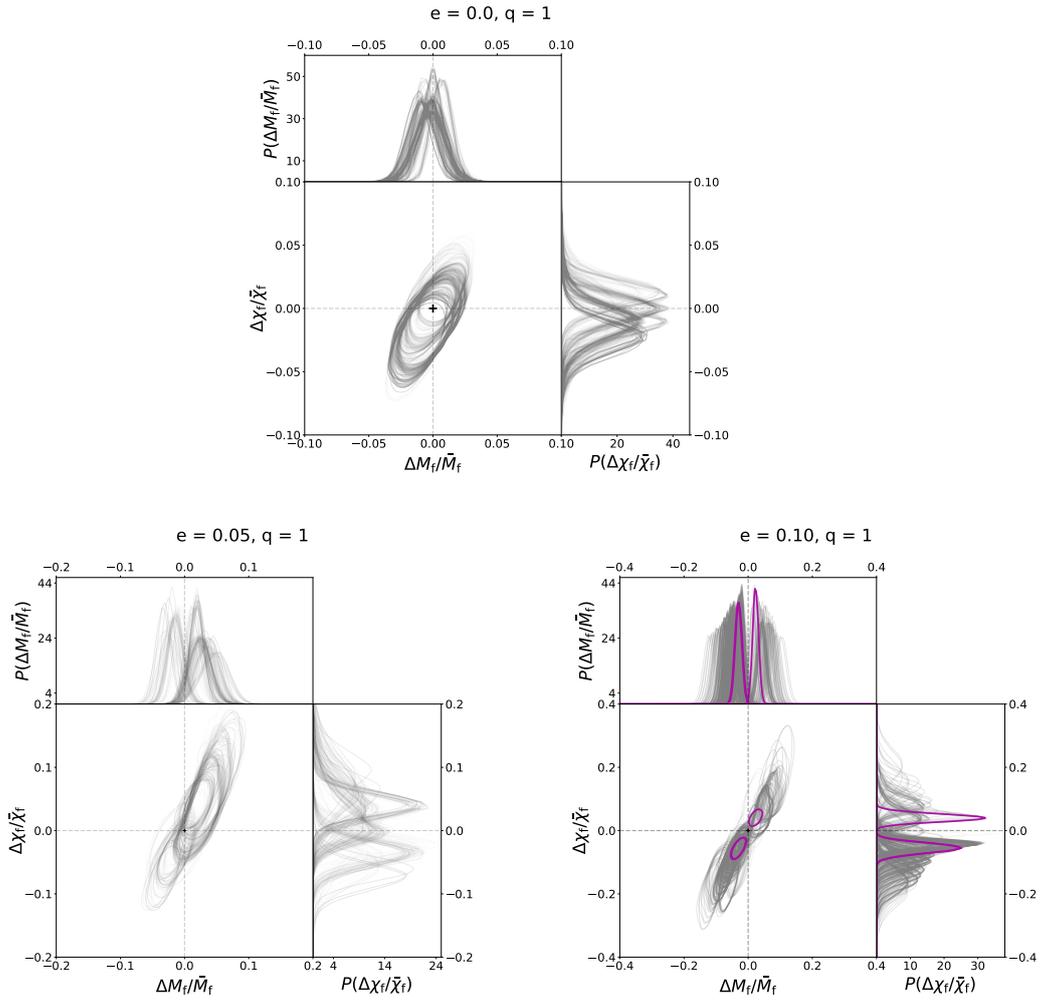


Figure 4. The results of meta IMRCT on the numerical relativity quasicircular and eccentric simulated observations with $q = 1$ presented as 2d contour plots of the posteriors on the deviation parameters. These are plotted the same way as in figure 1.

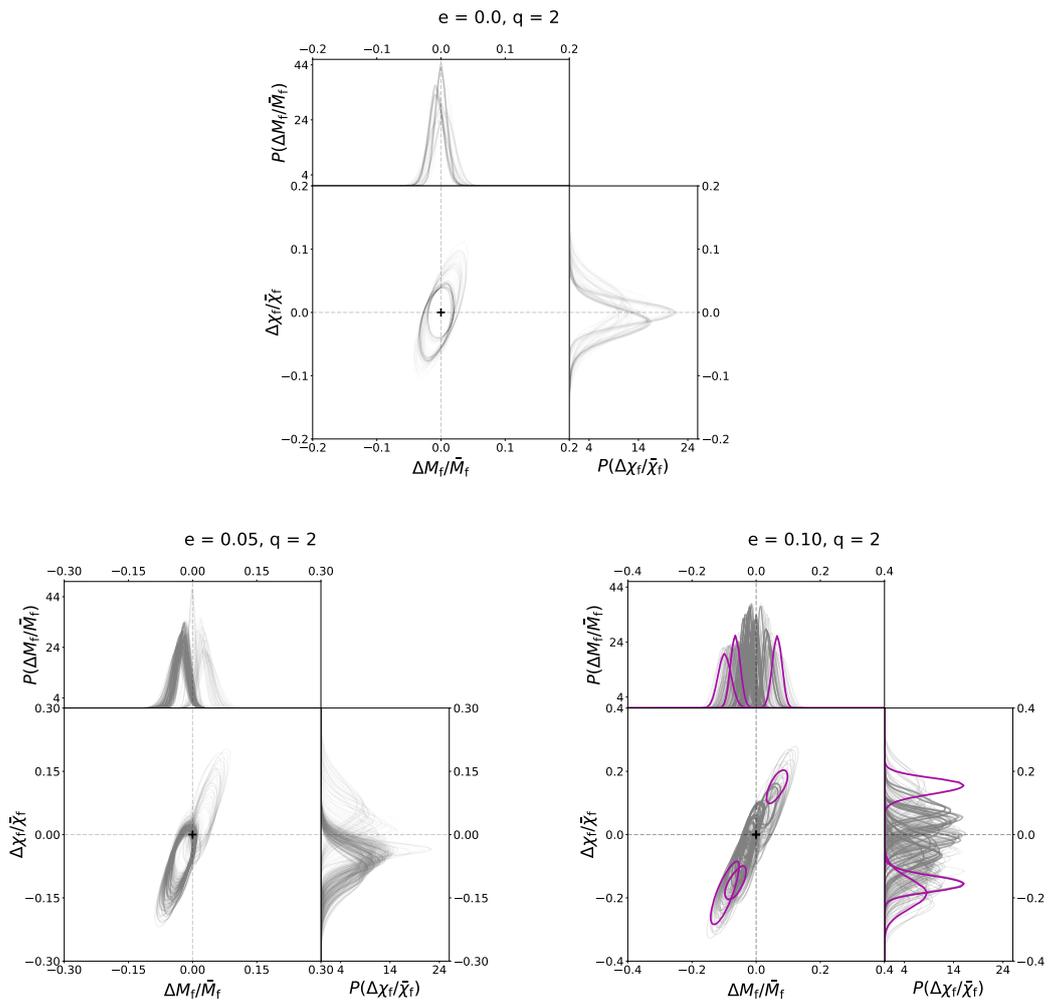


Figure 5. Similar to figure 4 for the $q = 2$ numerical relativity quasicircular and eccentric simulated observations.

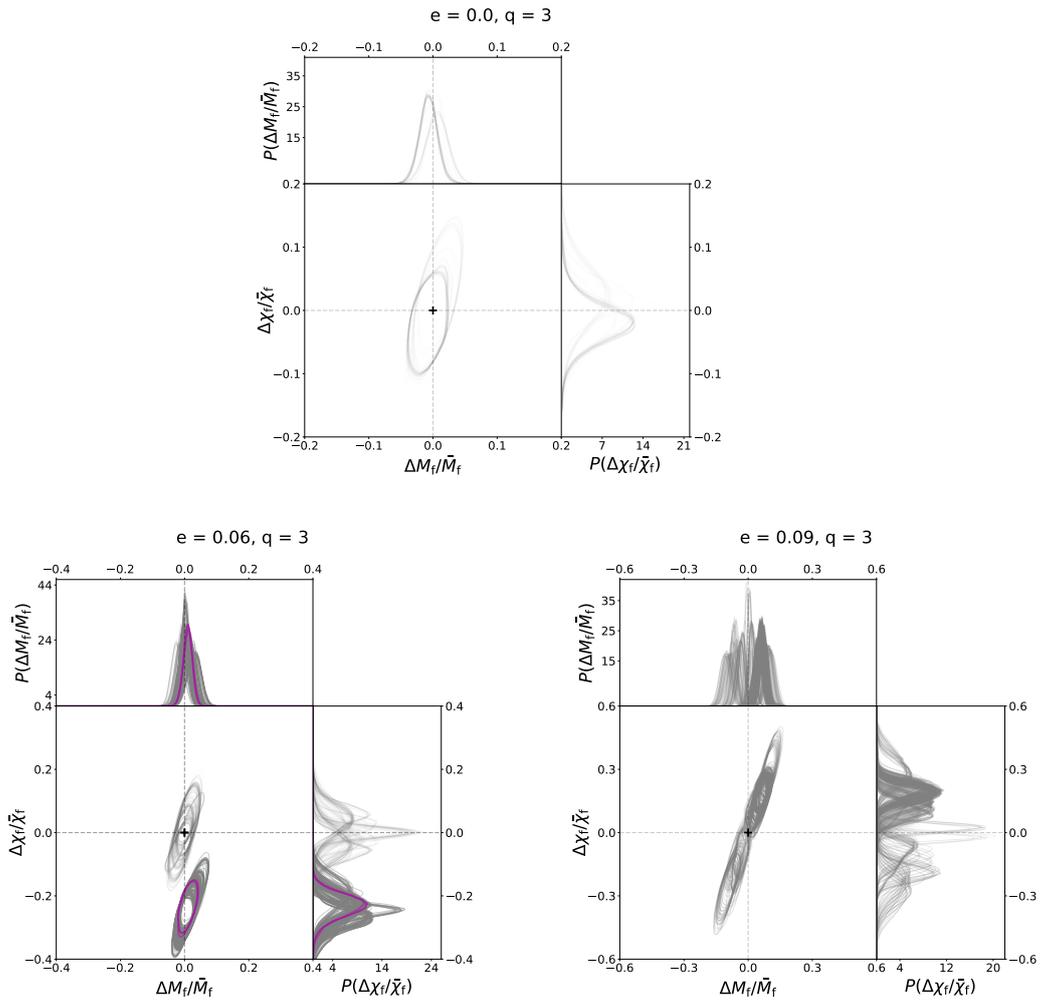


Figure 6. Similar to figure 4 for the $q = 3$ numerical relativity quasicircular and eccentric simulated observations.

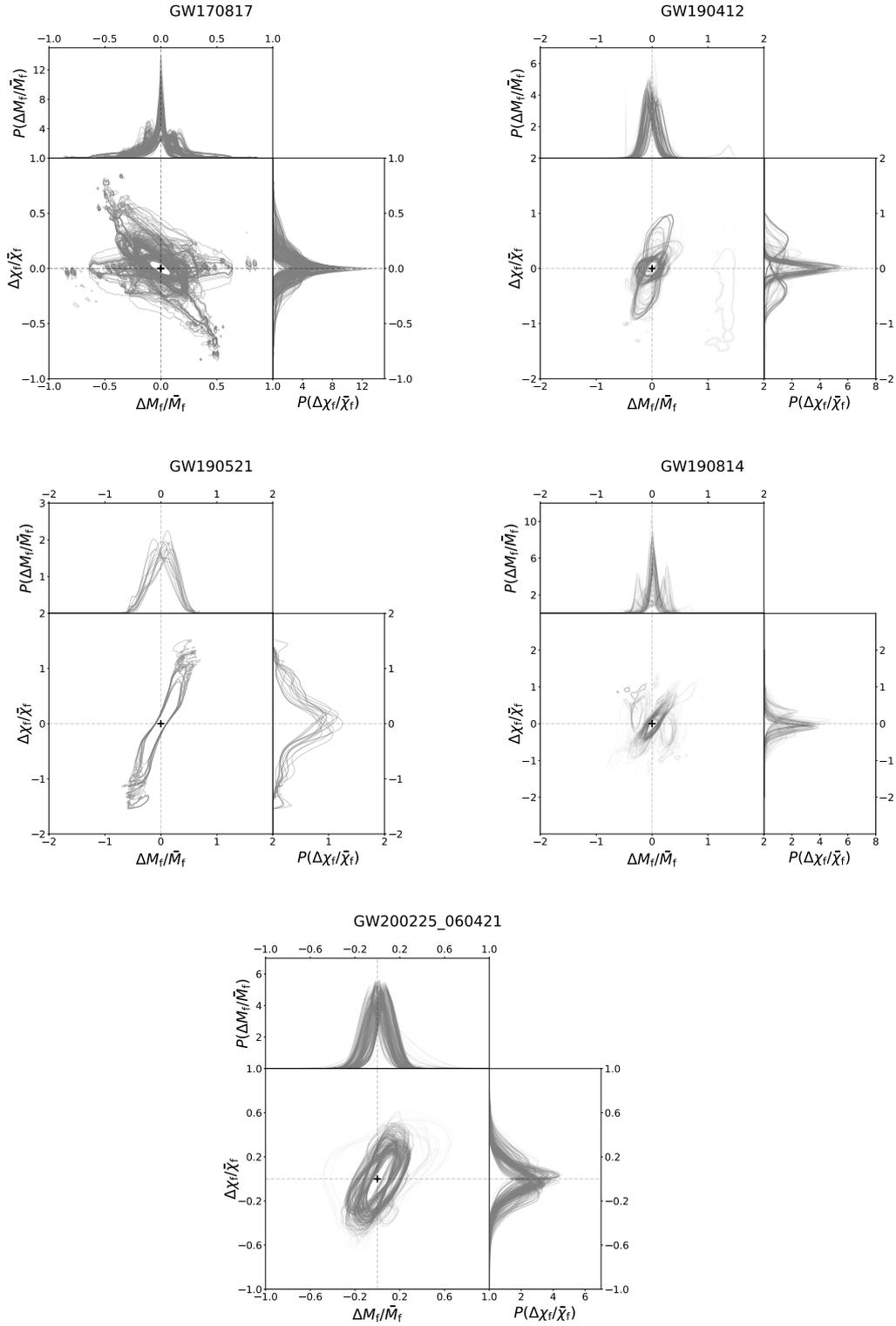


Figure 7. Result of meta IMRCT on real events presented as 2d contour plots and plotted in the same way as in figure 1.