Multidimensional hierarchical tests of general relativity with gravitational waves

Haowen Zhong,[1,*] Maximiliano Isi,[2,†] Katerina Chatziioannou,[3,4,‡] and Will M. Farr[2,5,§]

[1]*School of Physics and Astronomy, University of Minnesota, Minneapolis, Minnesota 55455, USA*
[2]*Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, New York 10010, USA*
[3]*Department of Physics, California Institute of Technology, Pasadena, California 91125, USA*
[4]*LIGO Laboratory, California Institute of Technology, Pasadena, California 91125, USA*
[5]*Department of Physics and Astronomy, Stony Brook University, Stony Brook, New York 11794, USA*

(Received 7 June 2024; accepted 2 August 2024; published 26 August 2024)

Tests of general relativity with gravitational waves typically introduce parameters for putative deviations and combine information from multiple events by characterizing the population distribution of these parameters through a hierarchical model. Although many tests include multiple such parameters, hierarchical tests have so far been unable to accommodate this multidimensionality, instead restricting to separate one-dimensional analyses and discarding information about parameter correlations. In this paper, we extend population tests of general relativity to handle an arbitrary number of dimensions. We demonstrate this framework on the two-dimensional inspiral-merger-ringdown consistency test, and derive new constraints from the latest LIGO-Virgo-KAGRA catalog, GWTC-3. We obtain joint constraints for the two parameters introduced by the classic formulation of this test, revealing their correlation structure both at the individual-event and population levels. We additionally propose a new four-dimensional formulation of the inspiral-merger-ringdown test that we show contains further information. As in past work, we find the GW190814 event to be an outlier; the 4D analysis yields further insights on how the low mass and spin of this event biases the population results. Without (with) this event, we find consistency with general relativity at the 60% (92%) credible level in the 2D formulation, and 76% (80%) for the 4D formulation. This multidimensional framework can be immediately applied to other tests of general relativity in any number of dimensions, including the parametrized post-Einsteinian tests and ringdown tests.

DOI: 10.1103/PhysRevD.110.044053

I. INTRODUCTION

In their first three observing runs (O1, O2, O3), LIGO [1] and Virgo [2], have detected gravitational waves (GWs) from over 90 compact binary coalescences (CBCs) [3–6]. These observations have not only opened up new frontiers for astrophysics and cosmology [7–14] but also bolstered support for general relativity (GR) [15–17]. Each individual GW detection furnishes a test of GR, leading our cumulative sensitivity to increase with the number of observations. The expanding catalog of events calls for robust statistical methods to combine these tests and produce constraints on deviations from GR from sets of detections [18–24]. Isi *et al.* [19], building upon work by Zimmerman *et al.* [18], proposed a hierarchical-inference framework [25–30] for this purpose. The framework enables a null test of GR that does not hinge on assumptions about the true theory of gravity or about how deviations manifest in different events.

Starting from measurements of parameters controlling deviations away from GR in individual events, the hierarchical framework characterizes the distribution of the true parameter across the population of events. Typically, parametrizations are constructed so that GR is recovered in the limit of vanishing deviation parameters; this translates to a population distribution that is a delta function at the origin if GR is correct, i.e., with the deviation vanishing for all events. Since O2, population distributions have been obtained for parametrized post-Einsteinian (ppE) deviations in the GW phase coefficients [31–34], ringdown analyses [35,36], and inspiral-merger-ringdown (IMR) consistency tests [37,38], among others [16,17,19,39]; recently, the framework has been extended to simultaneously model the GR deviations and the astrophysical properties of sources [22], as well as to factor in selection biases [23,24].

However, so far, results have been limited to one-dimensional tests of GR, which model a *single* deviation

*Contact author: zhong461@umn.edu
†Contact author: misi@flatironinstitute.org
‡Contact author: kchatziioannou@caltech.edu
§Contact author: wfarr@flatironinstitute.org

parameter at a time, even for tests that are inherently multidimensional. For example, ringdown tests introduce deviations in both the frequency and damping rate of one or multiple quasinormal modes, while the IMR consistency test introduces two parameters that quantify agreement of the remnant mass and spin as inferred independently from high versus low frequencies of the signal. In these cases, multidimensional posteriors are produced at the individual-event level, but then only the marginal distributions are considered when combining events. On the other hand, the ppE test is typically carried out by varying a single deviation coefficient at a time, in spite of the existence of multiple ppE coefficients that should, in principle, be measured jointly (as has been done only occasionally [40–43]).

Previous work has considered how a deviation in one parameter can manifest in multiple coefficients when each of them is measured independently rather than jointly, as is typically the case for the ppE test. In that case, the deviation is eventually detected by the one-dimensional hierarchical test of GR given enough observations (Fig. 2 in Ref. [19]); however, there will be little indication regarding the true combinations of coefficients that can explain the observed departure from GR, since individual-event measurements did not contain information about correlations across coefficients in the first place. Furthermore, there have been no studies of the case in which such correlation information exists at the individual-event level but it is ignored at the catalog level, as has been the case for the IMR and ringdown tests so far. Reducing a multidimensional test to a single dimension discards information about potential correlations between the deviation parameters (both at the single-event and population levels) and decreases its sensitivity.

In this paper, we generalize the hierarchical test of GR to handle an arbitrary number of deviation parameters simultaneously. This allows us to properly deal with likelihood correlations at the individual-event level (induced by the measurement process, e.g., through parameter degeneracies), as well as potential correlation structure appearing in the intrinsic distribution of GR deviations, were any of them to be detected. Correlations in the intrinsic distribution of GR deviations would be expected if the observed deviations were a function of binary parameters, like masses or spins—a common feature of several extensions to GR. We demonstrate an application in two and four dimensions on the IMR test and GWTC-3 data. The multidimensional analysis uncovers the structure of correlations between the test parameters, while confirming the data's consistency with GR with significance comparable to existing one-dimensional results.

The four-dimensional formulation also sheds further light on the role of the GW190814 event, an outlier for this test. With a remnant spin of $\chi_{\rm f} \approx 0.28$, this event is an outlier compared to the majority of the catalog that has $\chi_{\rm f} \approx 0.75$. Since the remnant spin is correlated with the remnant spin inferred from pre-merger and post-merger data, ignoring the former leads to a preference for a nonzero value in the variance of the latter. Such a variance would signal a GR deviation. The four-dimensional analysis gains access to this correlation and restores consistency with GR.

The organization of this paper is as follows. In Sec. II, we detail the hierarchical formalism for an arbitrary-dimensional parameter space. In Sec. III we summarize the two-dimensional IMR test and introduce an extended four-dimensional formulation, which we argue can better encompass the structure of the data. Sections IV and IV B present results for a two- and four-dimensional test respectively, and discuss the role of GW190814 in both. We conclude in Sec. V.

II. METHOD

We adopt a hierarchical framework following Isi *et al.* [19]. Consider N events and K beyond-GR parameters $\{\boldsymbol{\varphi}\} \equiv \{\varphi_1, \varphi_2, \ldots, \varphi_K\}$. Each individual event has a true underlying value $\{\hat{\boldsymbol{\varphi}}\}$; GR is recovered for $\hat{\boldsymbol{\varphi}} = 0$. We target the first two moments of the true distribution of $\{\hat{\boldsymbol{\varphi}}\}$ by modeling it as a K-dimensional Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is a vector of length K and Σ is a $K \times K$ covariance matrix. This is the K-dimensional generalization of the one-dimensional Gaussian of Refs. [16,17,19]. The goal is to determine the posterior distribution of the $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ hyperparameters, which consist of $\frac{1}{2}K(K+3)$ numbers: K components of $\boldsymbol{\mu}$, and $\frac{1}{2}K(K+1)$ unique components of $\boldsymbol{\Sigma}$, which is symmetric. GR is recovered in the limit that all means and variances (diagonal of $\boldsymbol{\Sigma}$) vanish.

Disregarding selection effects [23,24], the posterior distribution for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \{\boldsymbol{d}_i\}_{i=1}^N) \propto p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathcal{L}(\{\boldsymbol{d}_i\}_{i=1}^N | \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{1}$$

where $p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the (hyper)prior, $\mathcal{L}(\{\boldsymbol{d}_i\}_{i=1}^N | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the hierarchical likelihood, and $\boldsymbol{d}_i$ is the data for the ith GW event; the constant of proportionality normalizes the distribution. Selection effects can be accounted for by enhancing Eq. (1) with a detection efficiency factor following the usual procedure, e.g., [23,24].

A. Hyperpriors

We adopt separable (hyper)priors for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$: $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(\boldsymbol{\mu}) p(\boldsymbol{\Sigma})$. For the mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, we choose an uncorrelated zero-mean Gaussian with some characteristic scale $\varsigma_{\mu,k}$ for each k, i.e.,

$$p(\boldsymbol{\mu}) = \prod_{k=1}^{K} \mathcal{N}(0, \varsigma_{\mu,k}^2)[\mu_k]. \tag{2}$$

To avoid being overly restrictive, the prior scale $\varsigma_{\mu,k}$ should match or exceed the typical magnitude of the φ_k

measurements from individual events.[1] One could also replace this by a flat or Jeffreys prior, but Gaussian priors are computationally beneficial. See [20] (including Appendix A therein) for a discussion of the number of events required for the likelihood to inform the posterior as a function of prior scale.

To set the prior $p(\boldsymbol{\Sigma})$ for the covariance matrix $\boldsymbol{\Sigma}$, we first decompose the matrix itself as

$$\boldsymbol{\Sigma} = \boldsymbol{\sigma}^{\intercal}\mathcal{C}\boldsymbol{\sigma}, \tag{3}$$

where the vector $\boldsymbol{\sigma} = (\sigma_0, \ldots, \sigma_K)^{\intercal}$ encodes the intrinsic standard deviations of each parameter, and the matrix $\mathcal{C}_{kj} \coloneqq \Sigma_{kj}/\sqrt{\Sigma_{kk}\Sigma_{jj}}$ is the associated correlation matrix. While $\boldsymbol{\sigma}$ encodes the typical magnitude of each $\hat{\varphi}_k$, $\mathcal{C}_{ij}$ has unit-scale entries and reduces to the identity matrix if the parameters are uncorrelated; by construction, $\mathcal{C}$ is positive definite, with unit diagonal and $0 \leq |\mathcal{C}_{kj}| \leq 1$ for $k \neq j$. We set priors for the scale vector $\boldsymbol{\sigma}$ and correlation matrix $\mathcal{C}$ separately, so that $p(\boldsymbol{\Sigma}) = p(\boldsymbol{\sigma})p(\mathcal{C})$.

For the scale vector $\boldsymbol{\sigma}$ prior, we choose an uncorrelated (truncated) normal distribution as we did for $\boldsymbol{\mu}$, but now with a set of scales $\varsigma_{\sigma,k}$. In other words, we set

$$p(\boldsymbol{\sigma}) = \prod_{k=1}^{K} \mathcal{N}_{[0,\infty)}(0, \varsigma_{\sigma,k}^2)[\sigma_k], \tag{4}$$

with the $[0,\infty)$ subscript indicating the additional constraint that $\sigma_k \geq 0$ for all k, and $p(\boldsymbol{\sigma}) = 0$ otherwise. Here, again, the scale of the hyperprior $\varsigma_{\sigma,k}$ should match or exceed the expected scale of the $\hat{\varphi}_k$'s. As before, one could also replace this by a flat or Jeffreys prior.

For the correlation matrix, $\mathcal{C}$, we use the Lewandowsk-Kurowicka-Joe (LKJ) [44] distribution, which is a standard choice of prior for correlation matrices [45–48]. This is a probability density on the space of unit-diagonal, positive-definite correlation matrices; the density function can be defined as a power-law of the determinant, $|\mathcal{C}|$, such that

$$p(\mathcal{C}) = \mathrm{LKJCorr}(\mathcal{C}|\eta) \propto |\mathcal{C}|^{\eta-1}, \tag{5}$$

for some shape parameter $\eta > 0$. For any η, the LKJ prior always has the identity matrix (I) as the expected value, i.e., $\mathrm{E}[\mathcal{C}]_{\mathcal{C}\sim\mathrm{LKJ}} = I$, so that *on average* there will be no correlations imposed across different $\hat{\varphi}_k$'s. On the other hand, the choice of η controls the *spread* of the distribution, and thus the amount of support for off-diagonal elements of $\mathcal{C}$, with larger values of η more sharply favoring $\mathcal{C} = I$.

[1]As an implementation detail, we usually rescale all our parameters by a (potentially dimensionful) constant before sampling, bringing all coefficients to unit scale and allowing us to set $\varsigma_{\mu,k} = 1$. This can be beneficial for nonaffine samplers. We confirm that the prior does not affect subsequent results in Appendix C.

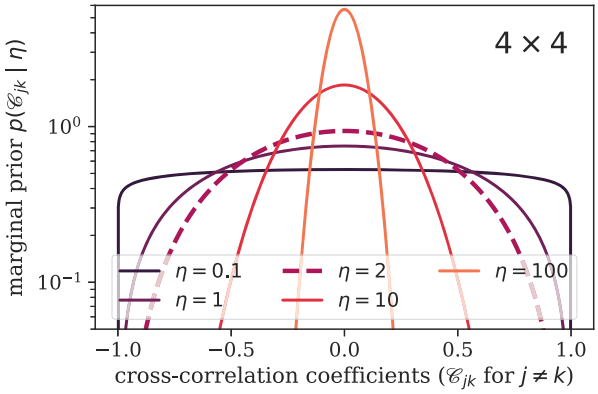

FIG. 1. Marginal prior on the off-diagonal components of the correlation matrix $\mathcal{C}$ corresponding to the LKJ prior of Eq. (5), assuming a 4×4 correlation matrix, i.e., $K = 4$ for different values of the shape parameter η. The marginal density follows a Beta distribution with shape parameters $\alpha = \beta = \eta - 1 + K/2$, such that larger values of η disfavor correlations more strongly. A dashed trace highlights our choice of $\eta = 2$ for the analyses in Secs. IV and IV B.

With this choice of prior on $\mathcal{C}$, each off-diagonal element $\mathcal{C}_{jk}$, $j \neq k$, will have a marginal prior given by a beta distribution, $B(\alpha,\beta)$, with shape parameters $\alpha = \beta = \eta - 1 + K/2$, for a $K \times K$ correlation matrix. Concretely, if $\alpha = \beta = 1$, then the density is uniform over correlation matrices; if $0 < \alpha = \beta < 1$, the prior probability density drops for the identity matrix; if $\alpha = \beta > 1$, the prior peaks at the identity, with increasing sharpness for larger η. For $K = 4$, as corresponds to our case below, we display this marginal prior for different choices of η in Fig. 1, and representations of correlation matrices drawn from Eq. (5) in Fig. 8 in Appendix A.

For concreteness, in our analyses below we choose a hyperprior $\eta = 2$ that favors a weak correlation between beyond-GR parameters $\hat{\boldsymbol{\varphi}}$ (dashed trace in Fig. 1); this choice is not fixed and could be adjusted based on the specific problem at hand. The full hyperprior is thus $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = p(\boldsymbol{\mu})p(\boldsymbol{\sigma})p(\mathcal{C})$, with factors given by Eqs. (2), (3), and (5).

B. Hierarchical likelihood

The hierarchical likelihood, $\mathcal{L}(\{\boldsymbol{d}_i\}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, is obtained from the likelihoods of individual events, $p(\{\boldsymbol{d}_i\}|\hat{\boldsymbol{\varphi}})$, as

$$\mathcal{L}(\{\boldsymbol{d}_i\}_{i=1}^{N}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int \mathrm{d}\hat{\boldsymbol{\varphi}}\, p(\{\boldsymbol{d}_i\}_{i=1}^{N}|\hat{\boldsymbol{\varphi}})p(\hat{\boldsymbol{\varphi}}|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{6}$$

By construction of the population model, we have that $\hat{\boldsymbol{\varphi}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ so the second factor in the integrand is a Gaussian that can be evaluate in closed form, i.e., $p(\hat{\boldsymbol{\varphi}}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})[\hat{\boldsymbol{\varphi}}]$. The first factor is the likelihood of observing the data given true values of the deviation parameters $\hat{\boldsymbol{\varphi}}$. Since each individual observation is independent, this separates into a product

$$p(\{\boldsymbol{d}_i\}_{i=1}^N|\hat{\boldsymbol{\varphi}}) = \prod_{i=1}^N p(\boldsymbol{d}_i|\hat{\boldsymbol{\varphi}}). \quad (7)$$

Each factor in this product is a K-dimensional likelihood obtained by applying the test of GR to a single event in isolation. Other parameters that may have been measured jointly with the $\hat{\varphi}_k$'s have been implicitly marginalized over, assuming some fixed prior; it is often more appropriate to simultaneously model those parameters with the $\hat{\varphi}_k$'s at the population level [22], as we revisit in Sec. III B below.

The individual-event likelihoods are typically estimated by reweighting posterior samples obtained under some fiducial sampling prior; then Eq. (6) can be estimated via a Monte Carlo sum [29,30]. To further increase computational efficiency, we instead leverage the fact that the population model is a Gaussian and represent each individual-event likelihood through a Gaussian mixture model (GMM). That is, we express the multidimensional single-event likelihood distribution for the ith event as a weighted sum of $N_{g,i}$ Gaussians, such that

$$p(\boldsymbol{d}_i|\hat{\boldsymbol{\varphi}}) \approx \sum_{j=1}^{N_{g,i}} w_j \mathcal{N}(\boldsymbol{\mu}_i^{(j)}, \boldsymbol{C}_i^{(j)})[\hat{\boldsymbol{\varphi}}]. \quad (8)$$

Similar GMM representations of individual event likelihoods have been used in the GW literature before [e.g., [49]]. We can then analytically evaluate Eq. (6) for each term in the GMM and the hierarchical log-likelihood becomes

$$\log \mathcal{L}(\{\boldsymbol{d}_i\}_{i=1}^N|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{i=1}^N \left[\log\left(\sum_{j=1}^{N_{g,i}} w_j \frac{1}{\sqrt{(2\pi)^K|\boldsymbol{\Sigma}+\boldsymbol{C}_i^{(j)}|}} \exp\left[-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_i^{(j)})^\intercal(\boldsymbol{\Sigma}+\boldsymbol{C}_i^{(j)})^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_i^{(j)})\right]\right)\right]. \quad (9)$$

We provide a detailed derivation of Eq. (9) in Appendix A.

Given this likelihood and the hyperprior discussed above, we sample the posterior for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ per Eq. (1). The amount of support for GR can be inferred by computing the probability for $\boldsymbol{\mu} = \boldsymbol{\sigma} = 0$; on the other hand, to the extent that there is support for $\boldsymbol{\sigma} > 0$, the posterior for the correlation coefficients $\mathcal{C}_{jk}$ $(j \neq k)$ will encode information about the nature of the deviation.

C. Population-marginalized distribution

The result of the multidimensional hierarchical analysis is fully encompassed by the hyperposterior on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Nevertheless, as is the case for the 1D case [16,19], it is sometimes useful to further compute a population-marginalized expectation for the deviation parameters, $\boldsymbol{\varphi}$. This K-dimensional distribution, also known as the *observed population predictive distribution*, represents our expectation for $\boldsymbol{\varphi}$ conditioned on the population properties inferred by the hierarchical analysis, marginalized over hyperparameters. In arbitrary dimensions, this is formally

$$p(\boldsymbol{\varphi}|\{\boldsymbol{d}_i\}_{i=1}^N) = \int \mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\Sigma} p(\boldsymbol{\varphi}|\boldsymbol{\mu},\boldsymbol{\Sigma}) p(\boldsymbol{\mu},\boldsymbol{\Sigma}|\{\boldsymbol{d}_i\}_{i=1}^N), \quad (10)$$

and can be easily estimated by taking a draw from $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ for each value of $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ in the hyperposterior. As in the 1D case, although convenient, this estimate has important limitations. First, the shape of this distribution is directly related to the assumed Gaussian ansatz, i.e., $\boldsymbol{\varphi} \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, and therefore should not be taken to generally represent the shape of the true underlying distribution of deviation parameters. Second, consistency with $\boldsymbol{\varphi} = 0$ is not a guarantee of consistency with GR, as this can be satisfied even if $\sigma_k > 0$. In spite of these limitations, the population expectation has been used to compare different catalog analyses in a succinct way [16,17,19], so we demonstrate it below.

In the remainder of this paper, we apply this framework to the IMR consistency test to demonstrate the advantages of the multidimensional hierarchical model in obtaining improved observational constraints. We also validate our implementation with simulated data in Appendix B and further sanity checks in Appendix C.

III. INSPIRAL-MERGER-RINGDOWN TEST

A. Traditional formulation

In this section, we provide an overview of the IMR test [50,51], which we will use to showcase our method. The basic idea is to split the GW data for each event into low- and high-frequency parts to obtain independent measurements of the source parameters, and then compare the two estimates for consistency. The cut is performed in the Fourier domain to leverage the fact that the likelihood is diagonal in this space and so the two measurements are statistically independent, even though this does not technically separate the inspiral and merger-ringdown regimes exactly [52,53]. The cutoff frequency, f_c^{IMR}, is chosen for each event based on the merger frequency estimated from an analysis of the entire signal [17,50,51].

Once a value of the cutoff has been chosen, the low $(f < f_c^{\mathrm{IMR}})$ and high $(f > f_c^{\mathrm{IMR}})$ frequency data are analyzed using a standard, Fourier-domain waveform model

based in GR, typically of the IMRPhenom family [54–60]. The resulting posteriors are used to estimate the (detector-frame) remnant mass $M_{\rm f}$ and remnant spin $\chi_{\rm f}$ for each event; the estimates from low-frequency data are labeled $(M_{\rm f}^{\rm pre},\chi_{\rm f}^{\rm pre})$, whereas those from high-frequency data are labeled $(M_{\rm f}^{\rm post},\chi_{\rm f}^{\rm post})$.

If GR is correct and the waveform is a good description of the data, we expect these two independent estimates of the remnant parameters to be in agreement. We quantify departures from this expectation through the fractional deviations

$$\delta M_{\rm f} \equiv 2\frac{M_{\rm f}^{\rm pre}-M_{\rm f}^{\rm post}}{M_{\rm f}^{\rm pre}+M_{\rm f}^{\rm post}}, \tag{11a}$$

$$\delta \chi_{\rm f} \equiv 2\frac{\chi_{\rm f}^{\rm pre}-\chi_{\rm f}^{\rm post}}{\chi_{\rm f}^{\rm pre}+\chi_{\rm f}^{\rm post}}, \tag{11b}$$

so that GR is recovered for $\delta M_{\rm f}=\delta\chi_{\rm f}=0$. Since $\delta M_{\rm f}$ and $\delta\chi_{\rm f}$ are not parameters we control in waveform models, we cannot directly obtain posterior estimates for these quantities. Instead, their joint posterior is estimated by computing Eq. (1) for independent draws from the $(M_{\rm f}^{\rm pre},\chi_{\rm f}^{\rm pre})$ and $(M_{\rm f}^{\rm post},\chi_{\rm f}^{\rm post})$ posterior [15,17]; the likelihood on $(\delta M_{\rm f},\delta\chi_{\rm f})$ is estimated by doing the same for the prior on $(M_{\rm f}^{\rm pre},\chi_{\rm f}^{\rm pre})$ and $(M_{\rm f}^{\rm post},\chi_{\rm f}^{\rm post})$, and then reweighting the posterior accordingly [16,17].

The result of this process is an estimate of the two-dimensional likelihood function for $(\delta M_{\rm f},\delta\chi_{\rm f})$ for each event. Although these objects contain information about potential correlations between the two parameters, previous catalog analyses consider only one parameter at a time, i.e., they infer the population distribution of $\delta M_{\rm f}$ and $\delta\chi_{\rm f}$ separately [16,17]. However, in doing so, they ignore potential correlations, with the drawbacks highlighted above. To remedy this, we preserve the two-dimensional likelihood information for each event and apply our multidimensional hierarchical formalism, as encompassed by Eq. (1).

B. Extended formulation

The previous subsection describes the IMR test as it has been formulated in the literature so far, yielding a two-dimensional parameter space $(\delta M_{\rm f},\delta\chi_{\rm f})$. However, we may go one step further by noting that, intrinsically, this is not a two-dimensional problem but a four-dimensional one: there are four basic quantities in this problem $(M_{\rm f}^{\rm pre},\chi_{\rm f}^{\rm pre},M_{\rm f}^{\rm post},\chi_{\rm f}^{\rm post})$, not two. By considering only the fractional differences of Eq. (11), we have disregarded half of the relevant parameters.

To take advantage of all the information inherent in the original test, we introduce two additional parameters, $\mathcal{M}$ and $\mathcal{X}$, defined as

$$\mathcal{M} \equiv \frac{M_{\rm f}^{\rm pre}+M_{\rm f}^{\rm post}}{2}, \tag{12a}$$

$$\mathcal{X} \equiv \frac{\chi_{\rm f}^{\rm pre}+\chi_{\rm f}^{\rm post}}{2}. \tag{12b}$$

With this extension, the parameter space spanned by $\{\delta M_{\rm f},\delta\chi_{\rm f},\mathcal{M},\mathcal{X}\}$ is equivalent to the initial parameter space spanned by $\{M_{\rm f}^{\rm pre},M_{\rm f}^{\rm post},\chi_{\rm f}^{\rm pre},\chi_{\rm f}^{\rm post}\}$ up to a coordinate transformation.

Restricting the hierarchical analysis to the $(\delta M_{\rm f},\delta\chi_{\rm f})$ subspace would only be appropriate if these quantities were fully decoupled from $\mathcal{M}$ and $\mathcal{X}$ at the individual-event level, i.e., if the single-event likelihoods displayed no correlations across the two subspaces. However, there is no reason *a priori* to expect this to be the case, and indeed this is not the case for existing events, see Fig. 4. If any degree of correlation is present across the two subspaces, ignoring $\mathcal{M}$ and $\mathcal{X}$ is equivalent to marginalizing over these quantities by assuming a fixed prior distribution, cf., Eq. (7). This distribution is determined by the prior chosen for $(M_{\rm f}^{\rm pre},\chi_{\rm f}^{\rm pre},M_{\rm f}^{\rm post},\chi_{\rm f}^{\rm post})$ when projected onto this subspace, and is not physically meaningful or guaranteed to match the observed data. As long as there are any correlations across $(\delta M_{\rm f},\delta\chi_{\rm f})$ and $(\mathcal{M},\mathcal{X})$, this will bias the catalog test of GR.

This situation is similar to that identified by Payne *et al.* [22], who noted that parameters controlling deviations from GR may couple to astrophysical parameters, like the black hole (BH) masses and spins. The solution in that case, as well as here, is to simultaneously model all relevant degrees of freedom hierarchically at the population level. In our case, this means that we not only model the two-dimensional $(\delta M_{\rm f},\delta\chi_{\rm f})$ subspace, but rather the full $(\delta M_{\rm f},\delta\chi_{\rm f},\mathcal{M},\mathcal{X})$ space, applying the framework in Sec. II in four dimensions.[2] Consistency with GR is still established for $\mu_k=\sigma_k=0$ in the $(\delta M_{\rm f},\delta\chi_{\rm f})$ subspace alone, after marginalizing over all other (nuisance) hyperparameters, including those controlling $\mathcal{M}$ and $\mathcal{X}$.

In the following, we present results for both the traditional (2D) and extended (4D) formulations of this test.

IV. ANALYSIS OF GWTC EVENTS

Here we apply our method to the events analyzed in Ref. [17] to obtain higher-dimensional IMR-test constraints on deviations from GR. Reference [17] considered 18 CBC signals and combined them using a one-dimensional framework applied to $\delta M_{\rm f}$ and $\delta\chi_{\rm f}$ separately. That analysis found preference for a nonzero variance in the $\delta\chi_{\rm f}$ population, i.e., low support for $\sigma=0$ in Fig. 5 of

[2]Future studies could consider alternative parametrizations for the $\mathcal{M}$ population to match the observed structure of compact-binary masses, e.g., a power law plus a Gaussian peak [12].

TABLE I. Medians and 90% credible intervals for all hyperparameters and from all analyses. The first column indicates the hyperparameter, and the following columns shows their recovered values in each analysis. The superscript ⋆ indicates that we ignore GW190814 in that particular analysis. The 1D results we quote here are from our own reanalyses on the GWTC-3 events which are consistent to the results reported in Ref. [17]. The remaining columns are $2d$ and $4d$ results with and without GW190814 respectively.

	Parameters considered in the analysis							
Hyperparameter	$\delta M_{\rm f}$	$\delta\chi_{\rm f}$	$\delta M_{\rm f}^{\star}$	$\delta\chi_{\rm f}^{\star}$	$\{\delta M_{\rm f}, \delta\chi_{\rm f}\}$	$\{\delta M_{\rm f}, \delta\chi_{\rm f}\}^{\star}$	$\{\delta M_{\rm f}, \delta\chi_{\rm f}, \mathcal{M}, \mathcal{X}\}$	$\{\delta M_{\rm f}, \delta\chi_{\rm f}, \mathcal{M}, \mathcal{X}\}^{\star}$
$\mu_{\delta M_{\rm f}}$	$0.04^{+0.08}_{-0.07}$	⋯	$0.05^{+0.08}_{-0.07}$	⋯	$0.00^{+0.07}_{-0.07}$	$0.02^{+0.07}_{-0.06}$	$-0.02^{+0.06}_{-0.06}$	$-0.02^{+0.07}_{-0.06}$
$\mu_{\delta\chi_{\rm f}}$	⋯	$-0.04^{+0.11}_{-0.11}$	⋯	$0.01^{+0.10}_{-0.10}$	$-0.09^{+0.10}_{-0.10}$	$-0.02^{+0.08}_{-0.08}$	$-0.11^{+0.09}_{-0.10}$	$-0.07^{+0.08}_{-0.07}$
$\mu_{\mathcal{M}}/M_{\odot}$	⋯	⋯	⋯	⋯	⋯	⋯	$74.04^{+7.96}_{-7.87}$	$77.67^{+7.21}_{-6.86}$
$\mu_{\mathcal{X}}$	⋯	⋯	⋯	⋯	⋯	⋯	$0.75^{+0.04}_{-0.05}$	$0.80^{+0.02}_{-0.03}$
$\sigma_{\delta M_{\rm f}}$	$0.04^{+0.09}_{-0.04}$	⋯	$0.04^{+0.09}_{-0.04}$	⋯	$0.05^{+0.09}_{-0.04}$	$0.03^{+0.06}_{-0.03}$	$0.04^{+0.07}_{-0.03}$	$0.02^{+0.05}_{-0.02}$
$\sigma_{\delta\chi_{\rm f}}$	⋯	$0.14^{+0.16}_{-0.12}$	⋯	$0.06^{+0.11}_{-0.06}$	$0.14^{+0.12}_{-0.10}$	$0.03^{+0.06}_{-0.03}$	$0.13^{+0.11}_{-0.11}$	$0.02^{+0.05}_{-0.02}$
$\sigma_{\mathcal{M}}/M_{\odot}$	⋯	⋯	⋯	⋯	⋯	⋯	$18.59^{+7.27}_{-4.61}$	$15.75^{+6.91}_{-4.25}$
$\sigma_{\mathcal{X}}$	⋯	⋯	⋯	⋯	⋯	⋯	$0.09^{+0.06}_{-0.04}$	$0.02^{+0.03}_{-0.02}$
$\rho_{\delta M_{\rm f}\delta\chi_{\rm f}}$	⋯	⋯	⋯	⋯	$0.43^{+0.49}_{-0.96}$	$0.15^{+0.67}_{-0.82}$	$0.25^{+0.53}_{-0.72}$	$0.15^{+0.58}_{-0.69}$
$\rho_{\delta M_{\rm f}\mathcal{M}}$	⋯	⋯	⋯	⋯	⋯	⋯	$0.20^{+0.52}_{-0.68}$	$0.14^{+0.55}_{-0.66}$
$\rho_{\delta M_{\rm f}\mathcal{X}}$	⋯	⋯	⋯	⋯	⋯	⋯	$0.07^{+0.60}_{-0.65}$	$0.02^{+0.62}_{-0.64}$
$\rho_{\delta\chi_{\rm f}\mathcal{M}}$	⋯	⋯	⋯	⋯	⋯	⋯	$0.46^{+0.34}_{-0.60}$	$0.11^{+0.57}_{-0.67}$
$\rho_{\delta\chi_{\rm f}\mathcal{X}}$	⋯	⋯	⋯	⋯	⋯	⋯	$0.57^{+0.30}_{-0.67}$	$0.00^{+0.63}_{-0.62}$

Ref. [17]. The spread in the $\delta\chi_{\rm f}$ distribution was found to be driven by GW190814 [61], which yields nonvanishing $\delta M_{\rm f}$ and $\delta\chi_{\rm f}$ measurements with high credibility (possibly because of the lack of a sufficiently loud merger-ringdown [16]); removing this event from the set restored consistency with GR [17].

We revisit those results with a multi-dimensional analysis of both the traditional (2D) and extended (4D) formulations of the IMR test outlined above, with and without the inclusion of GW190814. We make use of prior and posterior samples for individual events made available by the LIGO-Virgo-KAGRA collaborations [62]. Our hyperprior is as described in Sec. II, with $\eta = 2$ and $\varsigma_{\mu} = \varsigma_{\sigma} = 1$ for all parameters except $\mathcal{M}$, for which we set $\varsigma_{\mu} = \varsigma_{\sigma} = 100M_{\odot}$. We summarize our results with medians and 90% credible intervals for all hyperparameters from all analyses in Table I.

A. Traditional 2D formulation

We start by applying a two-dimensional version of the formalism to the traditional formulation of the test, in which only $\delta M_{\rm f}$ and $\delta\chi_{\rm f}$ are explicitly considered (Sec. III A). This analysis introduces five hyperparameters consisting of two population means ($\mu_{\delta M_{\rm f}}$, $\mu_{\delta\chi_{\rm f}}$), two standard deviations ($\sigma_{\delta M_{\rm f}}$, $\sigma_{\delta\chi_{\rm f}}$), and one correlation coefficient ($\rho_{\delta M_{\rm f}\delta\chi_{\rm f}}$). In the notation of Sec. II, the mean vector is $\boldsymbol{\mu} = (\mu_{\delta M_{\rm f}}, \mu_{\delta\chi_{\rm f}})$, the scale vector is $\boldsymbol{\sigma} = (\sigma_{\delta M_{\rm f}}, \sigma_{\delta\chi_{\rm f}})$, and the only off-diagonal component of the 2×2 correlation matrix is $\mathcal{C}_{12} = \mathcal{C}_{21} = \rho_{\delta M_{\rm f}\delta\chi_{\rm f}}$. Consistency with GR is represented by $\mu_{\delta M_{\rm f}} = \mu_{\delta\chi_{\rm f}} = \sigma_{\delta M_{\rm f}} = \sigma_{\delta\chi_{\rm f}} = 0$, irrespective of $\rho_{\delta M_{\rm f}\delta\chi_{\rm f}}$.

1. Including GW190814

We first show the result of the 2D analysis applied to all 18 events in our set, including GW190814. Figure 2 shows posteriors for all four hyperparameters in the collective analysis (blue). For comparison, we also display the result of 1D analyses that treat $\delta M_{\rm f}$ and $\delta\chi_{\rm f}$ separately (orange), as was done in Ref. [17].

As in Ref. [17], we find that including GW190814 in our sample leads to mild support for a deviation from GR through a nonvanishing $\sigma_{\delta\chi_{\rm f}}$ (fourth diagonal panel). This deviation is more apparent under the multidimensional formalism that models correlations between $\delta M_{\rm f}$ and $\delta\chi_{\rm f}$, which also results in a preference for $\mu_{\delta\chi_{\rm f}} < 0$ (third diagonal panel). Accordingly, the 2D analysis recovers GR at the 92% credible level, as opposed to 81% for the 1D $\delta\chi_{\rm f}$ analysis (64% for the 1D $\delta M_{\rm f}$ analysis).[3]

The reason for the difference between the 2D and 1D analyses stems from the fact that there is evidence for correlations between the two deviation parameters, $\rho_{\delta M_{\rm f}\delta\chi_{\rm f}} > 0$. This is encoded in the structure of the 2D individual-event measurements: the 1D analyses, unable to access information contained in the 2D individual-event

[3]A higher value for this credible level (quantile) corresponds to *less* support for GR, since it means that the posterior mass is distributed further away from the GR point. As elsewhere in the literature [16,17], we estimate it in practice as the fraction of posterior samples with probability density higher than the $\mu_{\delta M_{\rm f}} = \sigma_{\delta M_{\rm f}} = \mu_{\delta\chi_{\rm f}} = \sigma_{\delta\chi_{\rm f}} = 0$ point (marginalized over ρ); for the 1D analyses, this reduces to either the $\mu_{\delta M_{\rm f}} = \sigma_{\delta M_{\rm f}} = 0$ point or the $\mu_{\delta\chi_{\rm f}} = \sigma_{\delta\chi_{\rm f}} = 0$ point. We use 2D Gaussian kernel density estimation to approximate the posterior, which may slightly underestimate support for $\sigma = 0$ due to edge effects.

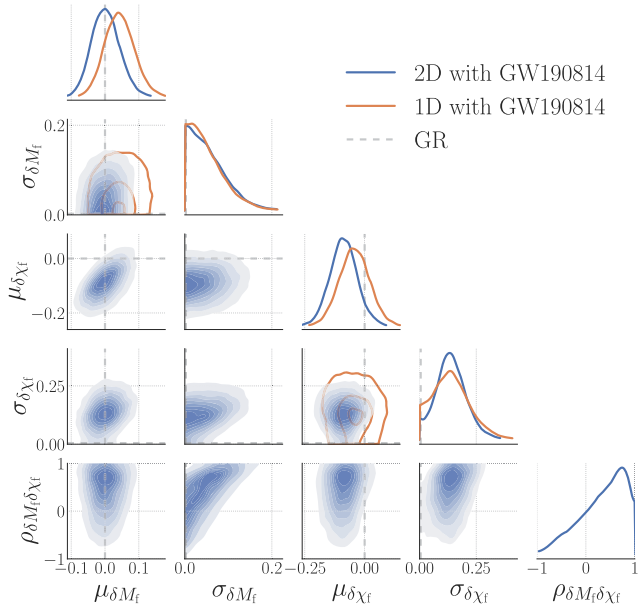

FIG. 2. Result of the 2D hierarchical analysis on δM_f and $\delta\chi_f$ joint measurements (Sec. III A) from 18 GWTC-3 events including GW190814 (blue), compared to 1D analyses of the same events looking at δM_f and $\delta\chi_f$ separately (orange). The parameters are the mean and spread of δM_f ($\mu_{\delta M_f}, \sigma_{\delta M_f}$), the mean and spread of $\delta\chi_f$ ($\mu_{\delta\chi_f}, \sigma_{\delta\chi_f}$), and the correlation between the two ($\rho_{\delta M_f\delta\chi_f}$); the 1D analyses only has access to marginals for δM_f and $\delta\chi_f$, so it can only measure their respective means and variances assuming no cross-correlation (orange subcorners). Blue contours enclose probability mass at increments of 10%, starting at 90% for the outermost contour; orange contours enclose 90%, 50%, and 10% of the probability mass. GR is recovered for $\mu_{\delta M_f} = \sigma_{\delta M_f} = \mu_{\delta\chi_f} = \sigma_{\delta\chi_f} = 0$ (gray dashed line). Inclusion of GW190814 leads to mild support for a deviation in $\delta\chi_f$.

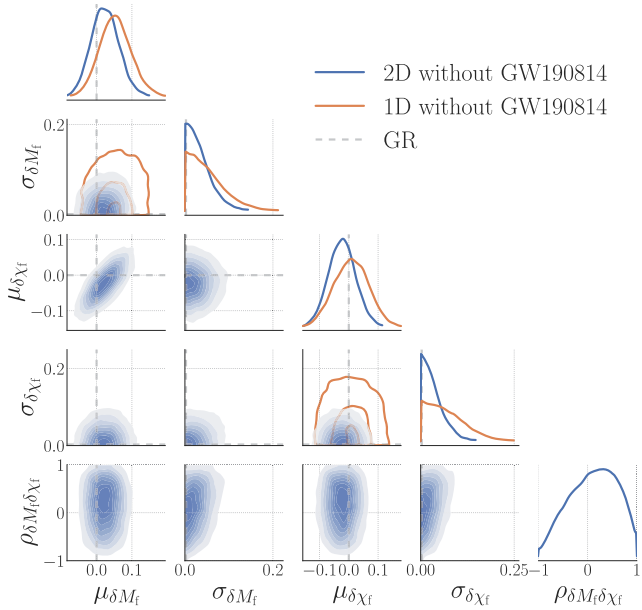

FIG. 3. Same as in Fig. 2 but for an analysis that excludes GW190814. Exclusion of GW190814 removes the support seen in Fig. 2 for $\sigma_{\delta\chi_f} > 0$ and $\mu_{\delta\chi_f} < 0$. The 2D analysis (blue) is able to ascertain consistency with GR with higher precision than the 1D analyses (orange).

likelihoods, cannot distinguish such correlations from statistical uncertainty in either δM_f or $\delta\chi_f$, leading to broader hyper-posteriors and, correspondingly, degraded confidence in a deviation from GR (fourth diagonal panel). As a result of neglecting correlations, the 1D analyses also infers a potential offset from $\mu_{\delta M_f} = 0$ (top left panel).

On the other hand, the 2D analysis is able to infer that there are correlations between the δM_f and $\delta\chi_f$ measurements at the individual-event level and that, typically, some linear combination of the two parameters is better measured than each parameter alone. The 2D measurement can pin down both the δM_f and $\delta\chi_f$ quantities simultaneously, thus inferring that there are actually no clear anomalies in the δM_f distribution (top left panel), but that there are indeed anomalies in $\delta\chi_f$ (third and fourth diagonal panels). Furthermore, it also directly reveals that there are likely correlations between the two parameters (bottom right panel), and that this interaction is the dominant cause of variance in δM_f (bottom row, second column). This can be gleaned from the structure of the 2D likelihoods for individual events, and in particular the large negative value of $\delta\chi_f$ for GW190814, see Fig. 3 in Ref. [16], or Fig. 4.

All information about 2D correlations is destroyed when we first marginalize either quantity, as we do for the 1D δM_f or $\delta\chi_f$ analyses. This highlights the power of the new method to better model deviations from GR in multidimensional tests: when there is a departure from the null hypothesis (as is indeed the case here due to GW190814), the 2D analysis is not only better able to pick that up, but also sheds light on the nature of the putative deviation.

2. Excluding GW190814

We repeat the analysis, but now excluding GW190814. Figure 3 shows the results, again comparing the 2D framework (blue) to the traditional 1D framework (orange), as we did in Fig. 2. The exclusion of GW190814 has done away with what support there was for $\mu_{\delta\chi_f} < 0$ or $\sigma_{\delta\chi_f} > 0$ in both the 2D and 1D analyses. However, only the 2D analysis also displays reduced support for $\mu_{\delta M_f} > 0$, as well as increased precision in the measurement of all parameters overall, i.e., tightening of blue versus orange distributions, as well as credible intervals in Table I. This leads to heightened credibility in GR: the 2D analysis recovers GR at the 60% credible level; the 1D analyses at 71% and 55% for δM_f and $\delta\chi_f$, respectively.

As discussed above, the 2D analysis is able to determine that the measurement process induces correlations in the joint δM_f and $\delta\chi_f$ likelihoods for individual events, so that

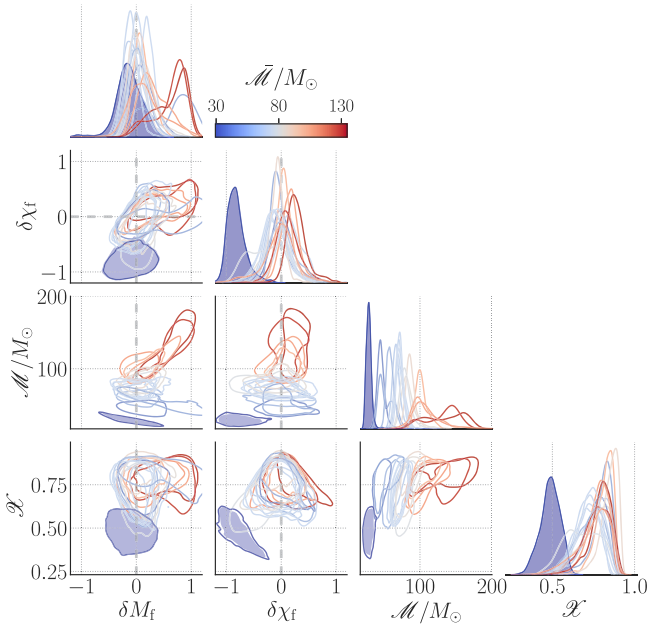

FIG. 4. Four-dimensional likelihoods for each of the 18 events considered in the hierarchical IMR test analysis of GWTC data. The parameters correspond to the extended version of the IMR consistency test defined in Sec. III B. Contours enclose 90% of the likelihood, colored by the inferred mean of the $\mathcal{M}$ parameter (labeled $\bar{\mathcal{M}}$), which is a proxy for the remnant mass; the null hypothesis requires $\delta M_f = \delta\chi_f = 0$ (dashed lines), irrespective of $\mathcal{M}$ and $\mathcal{X}$. The filled distribution highlights GW190814, an outlier in this population (Sec. IV B).

some linear combination of the two parameters is typically better measured than either quantity alone. This allows the 2D analysis to conclude that, in the absence of GW190814, the set of measurements is fully consistent with zero mean and no intrinsic spread in either deviation parameter ($\mu_{\delta M_f} = \mu_{\delta\chi_f} = \sigma_{\delta M_f} = \sigma_{\delta\chi_f} = 0$). It does so with better precision than the 1D analyses because it can disentangle contributions to the observed variance in δM_f or $\delta\chi_f$ that are due to the correlations in the measurement process rather than an intrinsic spread in the population of true parameters.

Even in the absence of a deviation from GR, the correlation structure in the δM_f and $\delta\chi_f$ joint likelihoods leaves an imprint in the 2D hierarchical posterior, manifesting as a correlation in the inferred joint distribution for $\mu_{\delta M_f}$ and $\mu_{\delta\chi_f}$ (second row, first column; also visible in Fig. 2). Irrespective of this correlation structure, there is no evidence for deviations from GR in the set without GW190814, and $\mu_{\delta M_f} = \mu_{\delta\chi_f} = 0$ is well supported. Since the posterior also favors vanishing variances ($\sigma_{\delta M_f} = \sigma_{\delta\chi_f} = 0$), there are no strong preferences for positive or negative correlations and the posterior for $\rho_{\delta M_f \delta\chi_f}$ resembles the prior (bottom right panel).

In summary, Fig. 3 again demonstrates the advantages of the 2D hierarchical framework, this time on a set of detections that show consistency with GR. Under these circumstances, the 2D analysis is able to achieve greater precision than the traditional 1D analysis by extracting information from the 2D likelihoods of individual events that is inaccessible to the 1D analyses. Figure 3 also confirms that GW190814 is the cause for the deviations from GR seen when analyzing the full GWTC-3 set, as was pointed out in Refs. [16,17].

B. Extended 4D formulation

We now turn to a hierarchical analysis of GWTC-3 over the full $4D$ parameter space comprised of $\{\delta M_f, \delta\chi_f, \mathcal{M}, \mathcal{X}\}$, as described in Sec. III B. Figure 4 shows the 4D individual-event likelihoods that make up the starting point for this analysis. Whereas the 2D analysis marginalized over some *ad hoc* implicit prior for the nuisance parameters average remnant mass $\mathcal{M}$ and average remnant spin $\mathcal{X}$, we here infer those populations simultaneously with δM_f and $\delta\chi_f$. This analysis thus introduces 14 hyperparameters consisting of four population means ($\mu_{\delta M_f}$, $\mu_{\delta\chi_f}$, $\mu_{\mathcal{M}}$, $\mu_{\mathcal{X}}$), four standard deviations ($\sigma_{\delta M_f}$, $\sigma_{\delta\chi_f}$, $\sigma_{\mathcal{M}}$, $\sigma_{\mathcal{X}}$), and six correlation coefficients ($\rho_{\delta M_f \delta\chi_f}$, $\rho_{\delta M_f \mathcal{M}}$, $\rho_{\delta M_f \mathcal{X}}$, $\rho_{\delta\chi_f \mathcal{M}}$, $\rho_{\delta\chi_f \mathcal{X}}$, $\rho_{\mathcal{M}\mathcal{X}}$). As before, consistency with GR is represented by $\mu_{\delta M_f} = \mu_{\delta\chi_f} = \sigma_{\delta M_f} = \sigma_{\delta\chi_f} = 0$, irrespective of the other parameters.

1. Including GW190814

We begin with the full set of 18 GWTC-3 events, including GW190814, and analyze it under the 4D hierarchical framework. Figure 5 displays a subspace of the posterior from the 4D analysis (green), excluding correlation coefficients for ease of display; the complete corner plot containing all 14 hyperparameters can be found in Fig. 11 in Appendix D. In addition to the 4D result, Fig. 5 also shows the 2D result obtained in Fig. 2 for reference (blue). Credible intervals for all hyperparameters can be found in Table I.

The subspace of the IMR test corresponds to the upper left corner of Fig. 5, showing the means and variances for the δM_f and $\delta\chi_f$ populations. In relation to the 2D result, there is a slight reduction in overall variance, i.e., shrinkage of green contours relative to blue, and the 4D analysis recovers the null at a higher credible level of 80%, as opposed to 92% for the 2D analysis. This suggests that there are correlations in the 4D likelihoods at the individual-event level, which was indeed the motivation for this extended analysis, see Sec. III B and Fig. 4.

The existence of correlations across the $(\delta M_f, \delta\chi_f)$ and $(\mathcal{M}, \mathcal{X})$ subspaces is apparent in Fig. 5. In particular, the 4D analysis identifies a clear correlation between the variances of $\delta\chi_f$ and $\mathcal{X}$, as can be seen from the $(\sigma_{\mathcal{X}}, \sigma_{\delta\chi_f})$ panel in Fig. 5 (bottom row, fourth column). Roughly speaking, there are two scenarios consistent with Fig. 5: either (1) there is larger variance in the average spin parameter $\mathcal{X}$ across the population of events ($\sigma_{\mathcal{X}} \gtrsim 0.1$)

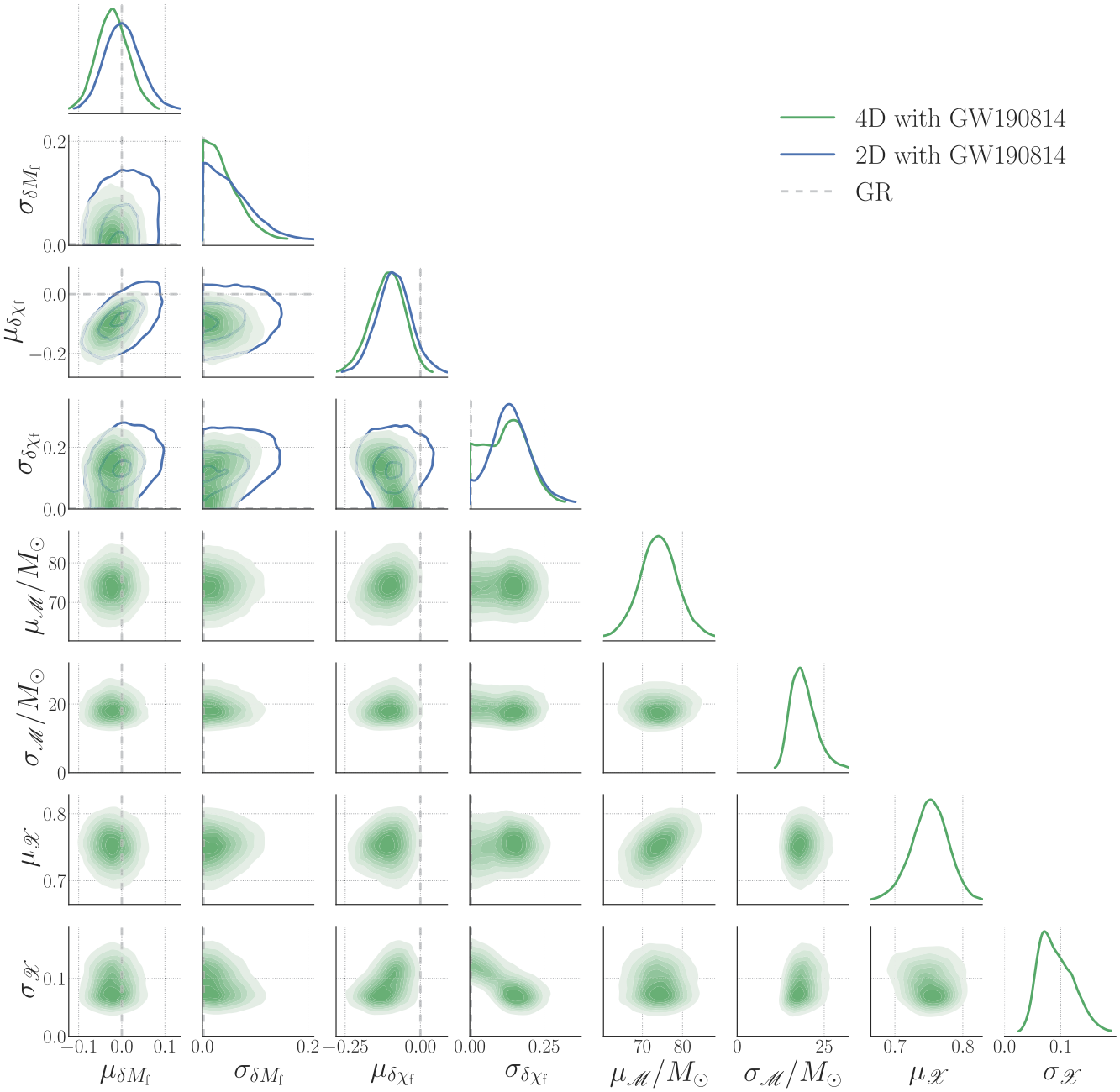

FIG. 5. Subspace of the posterior measurement obtained from the 4D hierarchical analysis (Sec. III B) of 18 GWTC-3 events including GW190814 (green), compared to the 2D analysis of the same events from Fig. 2 (blue). The parameters are the mean and spread of δM_f ($\mu_{\delta M_f}$, $\sigma_{\delta M_f}$), the mean and spread of $\mathcal{X}$ ($\mu_{\mathcal{X}}$, $\sigma_{\mathcal{X}}$), the mean and spread of $\delta\chi_f$ ($\mu_{\delta\chi_f}$, $\sigma_{\delta\chi_f}$), and the mean and spread of $\mathcal{M}$ ($\mu_{\mathcal{M}}$, $\sigma_{\mathcal{M}}$); we omit posterior for the cross correlation parameters, which we show in Fig. 11 of Appendix D. Green contours enclose probability mass at increments of 10%, starting at 90% for the outermost contour; blue contours enclose 90%, 50% and 10% of the probability mass. GR is recovered for $\mu_{\delta M_f} = \sigma_{\delta M_f} = \mu_{\delta\chi_f} = \sigma_{\delta\chi_f} = 0$ (gray dashed line). Inclusion of GW190814 leads to two largely distinct solutions per the 4D analysis: higher variance in $\mathcal{X}$ and no variance in $\delta\chi_f$, or a lower variance in $\mathcal{X}$ and a nonzero variance in $\delta\chi_f$—the latter of which corresponds to a deviation from GR (or other systematic).

and there is no variance in the $\delta\chi_f$ population or (2) the $\mathcal{X}$ population has standard deviation $\sigma_{\mathcal{X}} \lesssim 0.1$ and there is a markedly nonzero variance in $\delta\chi_f$, which would imply a violation from GR per this test. The first of these scenarios also corresponds to a mean $\delta\chi_f$ closer to zero (bottom row, third column), and a likely lower variance in δM_f (bottom row, second column).

The structure in the 4D result helps further elucidate the anomalies in $\delta\chi_f$ in the 2D and 1D analyses when including GW190814 (Fig. 2, as well as Refs. [16,17]). Unable to

directly access information about $\mathcal{X}$, the 2D analyses effectively average over the possible scenarios outlined above, with some implied weighting imposed by the sampling prior on $\mathcal{X}$ and $\mathcal{M}$. Accordingly, the result of the 2D analysis for $\sigma_{\delta\chi_f}$ does not correspond to either of the two modes exactly ($\sigma_{\delta\chi_f} = 0$ is disfavored but not excluded), although the second scenario appears to be upweighted.

We can understand the above observations by referring to the individual event likelihoods (Fig. 4). The measurement for GW190814 stands out in all dimensions, with the exception of δM_f. In particular, the structure of this likelihood shows clear correlations in the $(\mathcal{X}, \delta\chi_f)$ subspace: if $\mathcal{X}$ were to take on a value closer to the bulk of the population (i.e., $\mathcal{X} \approx 0.5$ or higher, closer to the population concentrated around $\mathcal{X} \approx 0.75$), then we must have $\delta\chi_f \approx -1$; on the other hand, if $\delta\chi_f$ were to be closer to zero, then we must have that $\mathcal{X}$ is much lower than the bulk of the population (i.e., $\mathcal{X} \approx 0.35$).

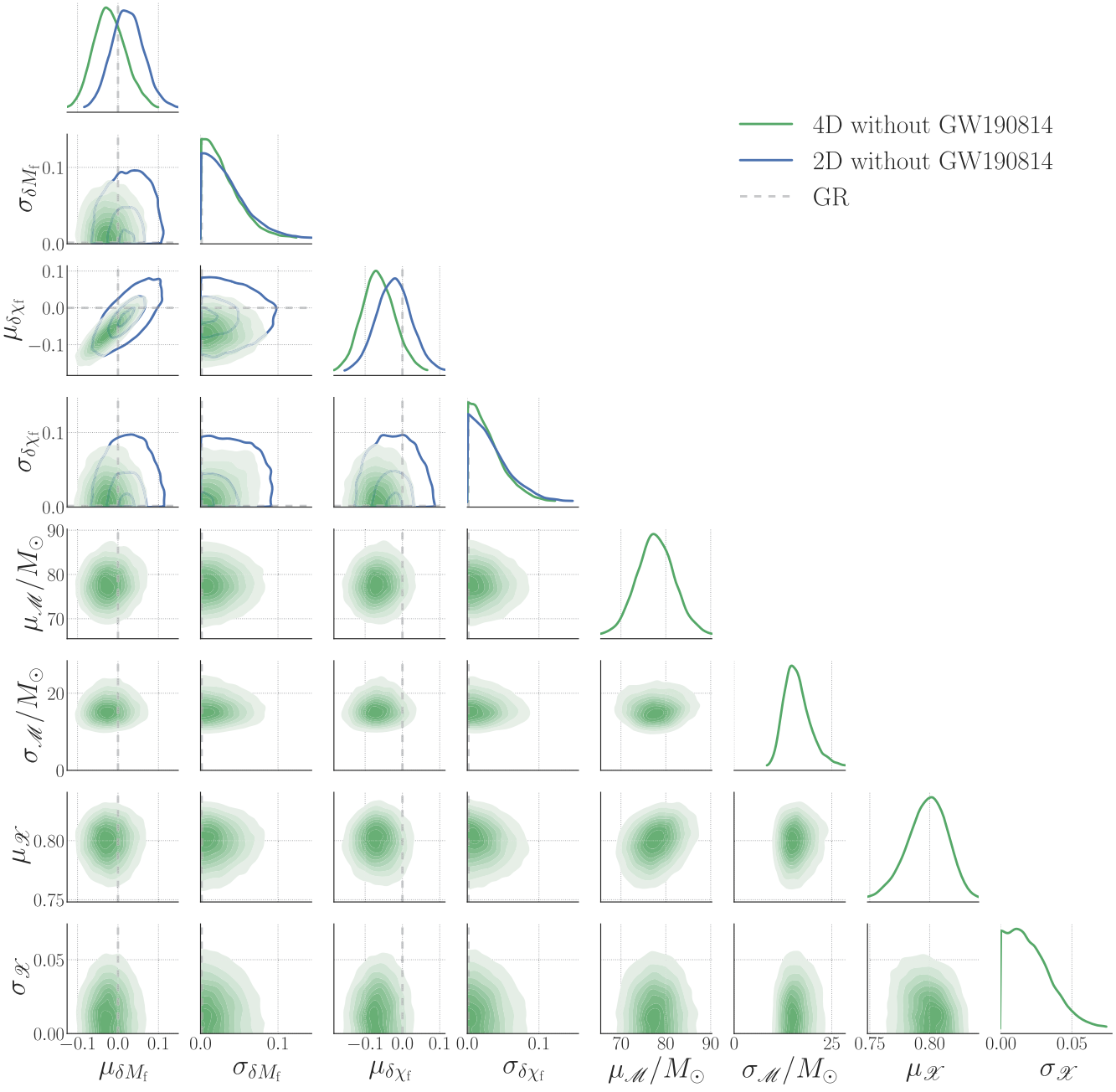

FIG. 6. Same as Fig. 5 but now excluding GW190814. We show both a subspace of the posterior from the 4D hierarchical analysis (green) and the 2D result from Fig. 3 for comparison (blue); the corresponding full 14D corner plot including correlation parameters for the 4D hierarchical analysis is shown in Fig. 11. Excluding GW190814 gets rid of the bimodalities in the posterior from the 4D analysis.

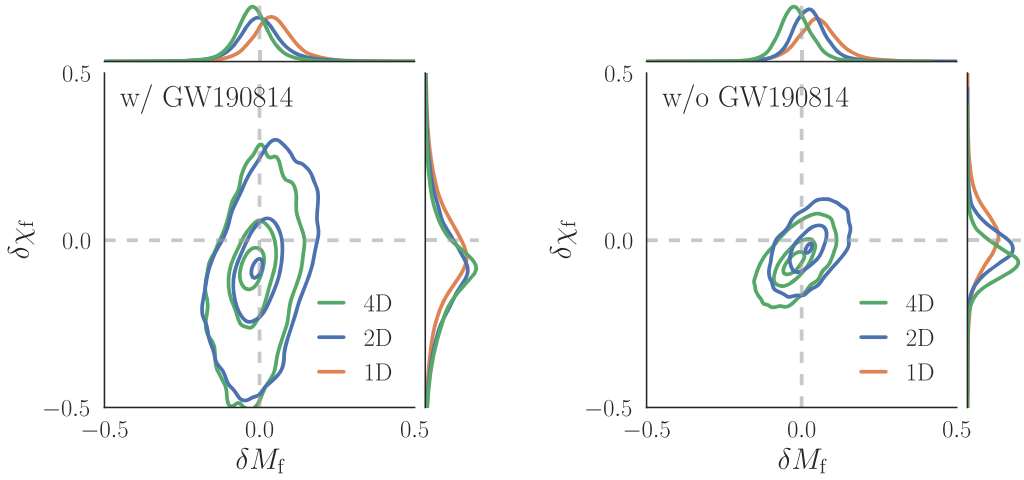

FIG. 7. Population-marginalized distribution for the IMR test parameters δM_f and $\delta\chi_f$, Eq. (10), as derived from all of the hierarchical analyses of GWTC-3 data presented in this paper: 4D (green), 2D (blue), and 1D (orange), with GW190814 (left) and without GW190814 (right). The results on the left correspond to the hyperposteriors in Figs. 2 and 5, while those on the right correspond to Figs. 3 and 6. Contours enclose 90%, 50%, and 10% of the probability mass; dashed lines mark the null expectation, $\delta M_f = \delta\chi_f = 0$. Excluding GW190814 leads to tighter population-marginalized distributions, especially for $\delta\chi_f$. See Table II for constraints derived from these distributions.

2. Excluding GW190814

We repeat the 4D hierarchical analysis, but now excluding GW190814 from our set of detections. Figure 6 shows the result (green), in full analogy to Fig. 5, but this time displaying the corresponding 2D analysis without GW190814 from Fig. 3 for reference (blue). The full posterior including correlation coefficients for this analysis is shown in Fig. 11 of Appendix D.

Without GW190814, the 4D hyperposterior is now unimodal, without outstanding interactions across the $(\delta M_f, \delta\chi_f)$ and $(\mathcal{M}, \mathcal{X})$ subspaces. Now $\sigma_{\delta\chi_f} = 0$ is preferred (fourth diagonal panel) and, although the $\mu_{\delta M_f}$ and $\mu_{\delta\chi_f}$ are slightly offset from zero (third row, first column), the overall posterior is broadly consistent with GR, with the null hypothesis recovered at 76% credibility. Further observations will be required to determine whether the slight shift in the means is simply due to statistical uncertainty or whether it represents a true systematic in the measurement or event selection process.

Although it does not directly factor into the test of GR, this analysis infers low or vanishing variance in the $\mathcal{X}$ parameter, with a typical mean value of $\mu_{\mathcal{X}} \approx 0.8$. In terms of mass, the 17 events in this set are inferred to have a mean of $\mu_{\mathcal{M}} \approx 80M_\odot$, with a spread of $\sigma_{\mathcal{M}} \approx 15M_\odot$, see Table I. With a remnant mass and spin of $M_f \approx 25M_\odot$ and $\chi_f \approx 0.28$ [61], GW190814 would be a clear outlier for this population.

C. Population-marginalized expectations

As described in Sec. II C, we may cast the hierarchical analysis result in a different light by computing the population-marginalized expectation (also known as the observed population predictive distribution) for δM_f and $\delta\chi_f$. Although these derived distributions contain less information than the hyperposteriors, we compute them to facilitate comparison to past work and to derive constraints in directly in the δM_f and $\delta\chi_f$ space. Figure 7 shows the joint population expectation for δM_f and $\delta\chi_f$ derived from all the hierarchical analyses of GWTC-3 data presented in this section: 4D (green), 2D (blue), and 1D (orange), both with (left) and without (right) GW190814. In all cases, these distributions are computed from a Monte Carlo estimate, as described below Eq. (10).

The population-marginalized distributions reveal some of the same features already described in the discussion of Figs. 2–6, but now directly in the space of δM_f and $\delta\chi_f$, rather than the hyperparameters. The most obvious feature is the increased precision achieved by excluding GW190814 in the sample set, regardless of the dimensionality of the analysis, and particularly with regards to $\delta\chi_f$. It is also notable that the population expectations derived

TABLE II. Population-marginalized constraints (median and 90% credible symmetric interval).[a]

Analysis	δM_f	$\delta\chi_f$	$\mathcal{M}/M_\odot$	$\mathcal{X}$
1D	$0.04^{+0.14}_{-0.12}$	$-0.04^{+0.28}_{-0.29}$	⋯	⋯
1D*	$0.05^{+0.13}_{-0.12}$	$0.01^{+0.17}_{-0.17}$	⋯	⋯
2D	$0.00^{+0.14}_{-0.12}$	$-0.09^{+0.28}_{-0.27}$	⋯	⋯
2D*	$0.02^{+0.10}_{-0.09}$	$-0.02^{+0.11}_{-0.10}$	⋯	⋯
4D	$-0.02^{+0.11}_{-0.10}$	$-0.10^{+0.23}_{-0.28}$	$73.80^{+32.28}_{-32.75}$	$0.75^{+0.16}_{-0.17}$
4D*	$-0.02^{+0.09}_{-0.09}$	$-0.07^{+0.10}_{-0.09}$	$77.78^{+28.12}_{-27.91}$	$0.80^{+0.05}_{-0.05}$

[a]Denotes the analyses *excluding* GW190814.

from the multidimensional analyses (4D and 2D) carry information about the correlations in the inferred population means $\mu_{\delta M_{\rm f}}$ and $\mu_{\delta\chi_{\rm f}}$, which here manifest as correlations between $\delta M_{\rm f}$ and $\delta\chi_{\rm f}$ themselves. These results also yield direct constraints on the $\delta M_{\rm f}$ and $\delta\chi_{\rm f}$ values. We report such constraints in Table II, including for $\mathcal{M}$ and $\mathcal{X}$ which are not shown in Fig. 7.

V. CONCLUSION

In this paper, we have generalized the hierarchical inference framework for testing GR with gravitational wave observations from a single deviation parameter to an arbitrary number of parameters. For tests that are formulated in terms of more than a single parameter, e.g., the ringdown and IMR tests, this generalization gains access to potential correlations between the test parameters both at the individual-event level, i.e., correlated likelihoods, and at the population level, i.e., correlated hyperparameters.

We applied the multidimensional framework to the IMR consistency test using GWTC-3 events. The IMR test divides a CBC signal into high- and low-frequency portions and estimates the remnant mass and spin independently from each. The test is parametrized via two deviation parameters, $\delta M_{\rm f}$ and $\delta\chi_{\rm f}$, and two parameters that encode the remnant mass and spin, $\mathcal{M}$ and $\mathcal{X}$. Formulations of the IMR test with reduced dimensionality, i.e., considering the population distribution of only $\delta M_{\rm f}$ and $\delta\chi_{\rm f}$ separately, have previously yielded mild evidence for a violation of GR or other systematics (cf., Fig. 2), attributed to the GW190814 event. Restoring the full four-dimensional formulation resolves this apparent deviation which is attributed to a correlation between $\mathcal{X}$ and $\delta\chi_{\rm f}$, cf., Fig. 4. This application emphasizes the need to expand the dimensionality of tests of GR to all relevant parameters in order to avoid potential systematics from improper assumptions, such as ignoring correlations.

The expanded dimensionality means that more parameters need to be included in the analysis models and selection terms. This analysis focuses on multivariate Gaussian population distributions for all hyperparameters. Although this model is reasonable for deviation parameters whose distribution cannot be motivated otherwise [19], extended formulations could explore more complex distributions for the remnant mass and spin parameters. This situation is akin to the analysis of Payne *et al.* [22] that extended the parametrized phase deviation test to include the BH masses and spins and made use of distributions such as power laws. Moreover, the nature of the IMR test (that hinges on events with informative post- *and* premerger data) makes estimating its selection effect particularly involved. We leave such extensions to future work with the expectation that their importance will increase as more events are detected and constraints are becoming more stringent.

ACKNOWLEDGMENTS

We thank Geraint Pratten for feedback on this manuscript. The authors are grateful for computational resources provided by the LIGO Laboratory under NSF Grants No. PHY-0757058 and No. PHY-0823459. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. H. Z. was supported by NSF Grant No. DGE-1922512. The Flatiron Institute is funded by the Simons Foundation. K. C. was supported by NSF Grant No. PHY-2110111. This paper carries LIGO document number LIGO-P2400214.

APPENDIX A: HYPERPRIOR AND LIKELIHOOD DERIVATION

In this appendix, we provide a visualization of the covariance matric hyperprior in Fig. 8, detail the derivation of Eq. (9), and provide further details on constructing the GMM for each event. We determine the number of Gaussians in the GMM, $N_{g,i}$, and their parameters $\boldsymbol{\mu}_i^{(j)}$ and $\boldsymbol{C}_i^{(j)}$ by optimizing the Bayesian information criterion (BIC) [63]

$$\mathrm{BIC} \coloneqq k\ln(n) - 2\ln(\hat{L}), \tag{A1}$$

where k is the number of model parameters, n is the sample size, and $\hat{L}$ is the maximized likelihood function value of the chosen model given the data. For each event, we employ a grid search [64,65] to identify the optimal $N_{g,i}$ that minimizes the BIC. Subsequently, we compute the best-fit $\boldsymbol{\mu}_i^{(j)}$ and $\boldsymbol{C}_i^{(j)}$ for each chosen $N_{g,i}$. Typically, N_g ranges from $\mathcal{O}(3\text{–}10)$.

To evaluate the integral in Eq. (6), we leverage the fact that the product of two Gaussians of arbitrary dimension can be refactored into the product of two different Gaussians as [66,67]

$$\mathcal{N}(x|\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)\mathcal{N}(x|\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) = \mathcal{C}\mathcal{N}(x|\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3), \tag{A2}$$

where x denotes data samples, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ $(i=1,2,3)$ are the mean vectors and covariance matrices of the corresponding multivariate Gaussians, and $\mathcal{C}$ is a normalization factor. Explicitly, $\mathcal{C}, \boldsymbol{\mu}_3$ and $\boldsymbol{\Sigma}_3$ are given by

$$\mathcal{C} = \mathcal{N}(\boldsymbol{\mu}_1|\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2), \tag{A3a}$$

$$\boldsymbol{\mu}_3 = (\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1})^{-1}(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1+\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2), \tag{A3b}$$

$$\boldsymbol{\Sigma}_3 = (\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1})^{-1}, \tag{A3c}$$

where the Gaussian represented by $\mathcal{C}$ becomes a factor independent of the data x and, in that sense, can be thought

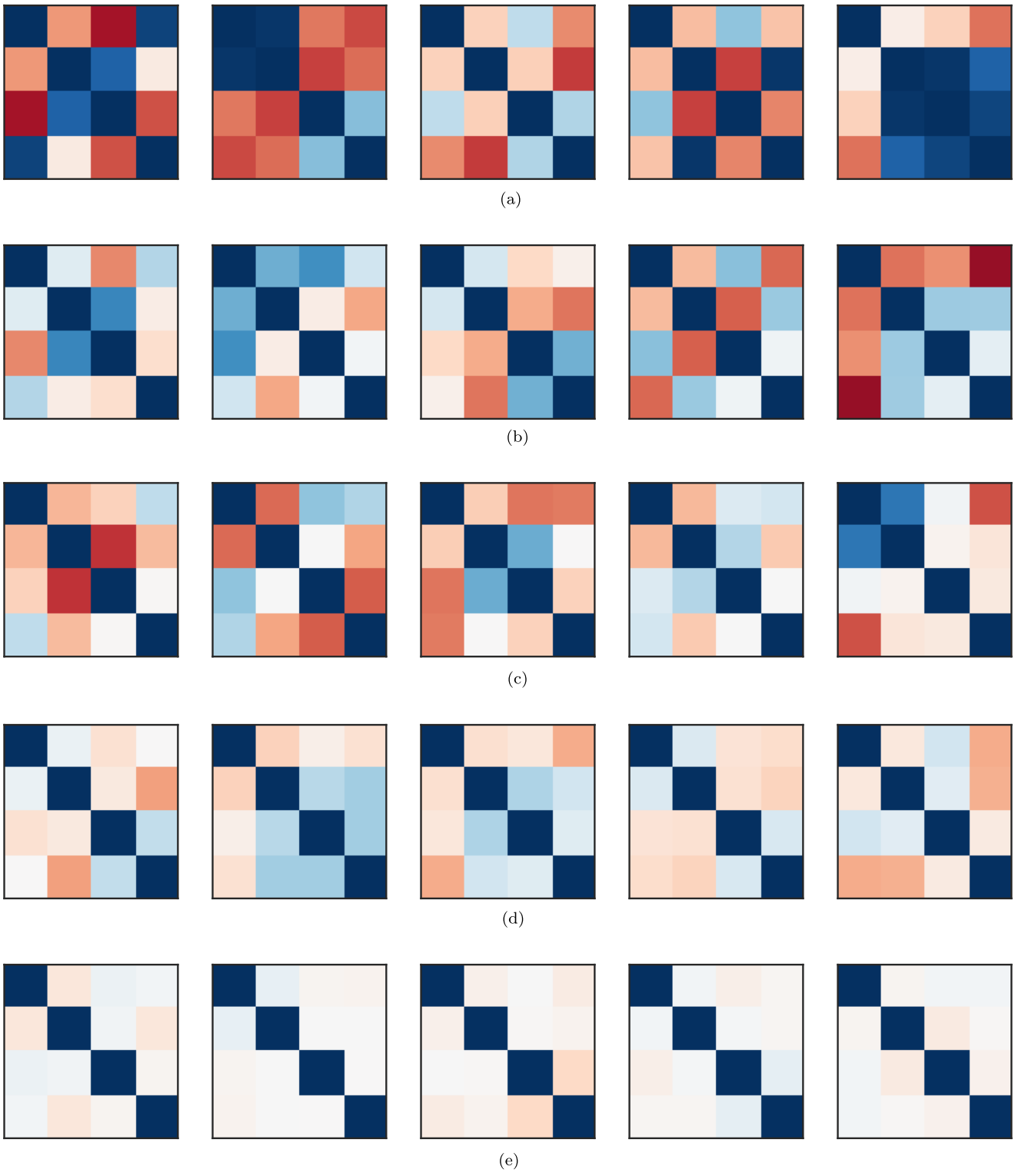

FIG. 8. Visualization of 4×4 correlation matrices $\mathcal{C}$ drawn from an LKJ prior for different values of the shape parameter η per Eq. (5). For each η (rows), we display five random draws (columns), with the color of each cell encoding the value of the corresponding $\mathcal{C}_{jk}$ entries: blues (reds) represents positive correlations $\mathcal{C}_{jk} > 0$ ($\mathcal{C}_{jk} < 0$), and the intensity of the color encodes the magnitude of the correlation such that dark blue represents full positive correlation ($\mathcal{C}_{jk} = 1$), white represents no correlation ($\mathcal{C}_{jk} = 0$), and dark red represents full anticorrelation ($\mathcal{C}_{jk} = -1$). The diagonal of $\mathcal{C}$ is always unity; the off-diagonal entries follow the marginal distributions shown in Fig. 1. Larger values of η favor the identity, i.e., lack of correlations, more strongly. (a) $\eta = 0.1$. (b) $\eta = 1$. (c) $\eta = 2$. (d) $\eta = 10$. (e) $\eta = 100$.

of as a normalizing constant. With the help of Eq. (A3a), plugging Eqs. (7) and (8) back into Eq. (6) yields Eq. (9).

APPENDIX B: SIMULATED DATA

To verify the implementation of the multi-dimensional hierarchical analysis, we consider two simulated data scenarios: (a) GR is correct, and (b) GR is violated. The hyperparameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \tag{B1}$$

For the GR is correct case, $\hat{\boldsymbol{\mu}} = \mathbf{0}$ and $\hat{\boldsymbol{\Sigma}} = \mathbf{0}$, while for the second case, we simulate deviations in both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. We simulate measurement uncertainty through a covariance matrix $\Sigma_{\rm obs}$, such that the simulated posterior samples, $\varphi_{\rm data}$, are drawn from

$$\begin{cases} \hat{\boldsymbol{\varphi}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ \boldsymbol{\varphi}_{\rm obs} \sim \mathcal{N}(\hat{\boldsymbol{\varphi}}, \boldsymbol{\Sigma}_{\rm obs}), \\ \boldsymbol{\varphi}_{\rm data} \sim \mathcal{N}(\boldsymbol{\varphi}_{\rm obs}, \boldsymbol{\Sigma}_{\rm obs}). \end{cases} \tag{B2}$$

To simulate the population, we set

$$\boldsymbol{\Sigma}_{\rm obs} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}. \tag{B3}$$

which corresponds to two strongly correlated beyond-GR parameters.

Results are show in Fig. 9. For the case where GR is correct (left), we simulate 20 events and 1000 likelihood samples for each event. The true values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are recovered at the 90% credible level. For the case where GR is incorrect (right), we simulate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as follows:

$$\hat{\boldsymbol{\mu}} = (1, 2), \qquad \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}. \tag{B4}$$

Compared to the case where GR is correct where $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}_{\rm obs}$ differ significantly, $\hat{\boldsymbol{\Sigma}}$ is now comparable to $\boldsymbol{\Sigma}_{\rm obs}$. We simulate 100 events and 1000 likelihood samples for each and again recover the true parameters to within the 90% credible level.

APPENDIX C: ANALYSIS SANITY CHECKS

In this appendix, we confirm that (a) the choice of hyperprior η does not affect the hyperposteriors, and (b) we recover the 1D results from the 2D analysis in the appropriate limit, i.e., for $\rho \to 0$ and no correlation between the individual-event $\delta M_{\rm f}$ and $\delta\chi_{\rm f}$ likelihoods. For the latter, we start with individual-event 2D samples denoted as $\{(\delta M_{{\rm f},i}, \delta\chi_{{\rm f},i})\}_{i=1}^{Ns}$. We then independently shuffle the sets $\{\delta M_{{\rm f},i}\}_{i=1}^{Ns}$ and $\{\delta\chi_{{\rm f},i}\}_{i=1}^{Ns}$ and create a new set of paired samples, $\{(\delta M_{{\rm f},i'}, \delta\chi_{{\rm f},i'})\}_{i'=1}^{Ns}$. This process removes any correlations between these two parameters in the individual-event likelihood.

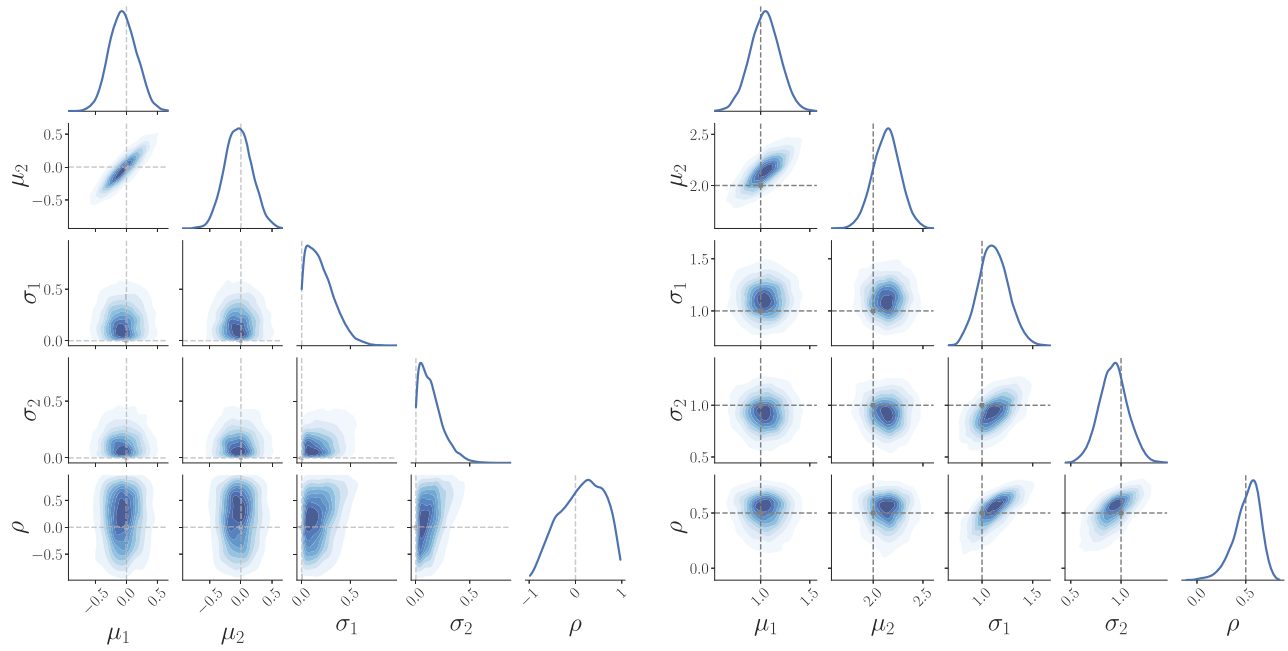

FIG. 9. Marginalized posterior distributions on the hyperparameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of mock beyond-GR parameters φ_1 and φ_2 assuming GR is correct (left) or incorrect (right). The gray dashed lines indicate the location of true values. The correct parameters are always recovered within the distributions.

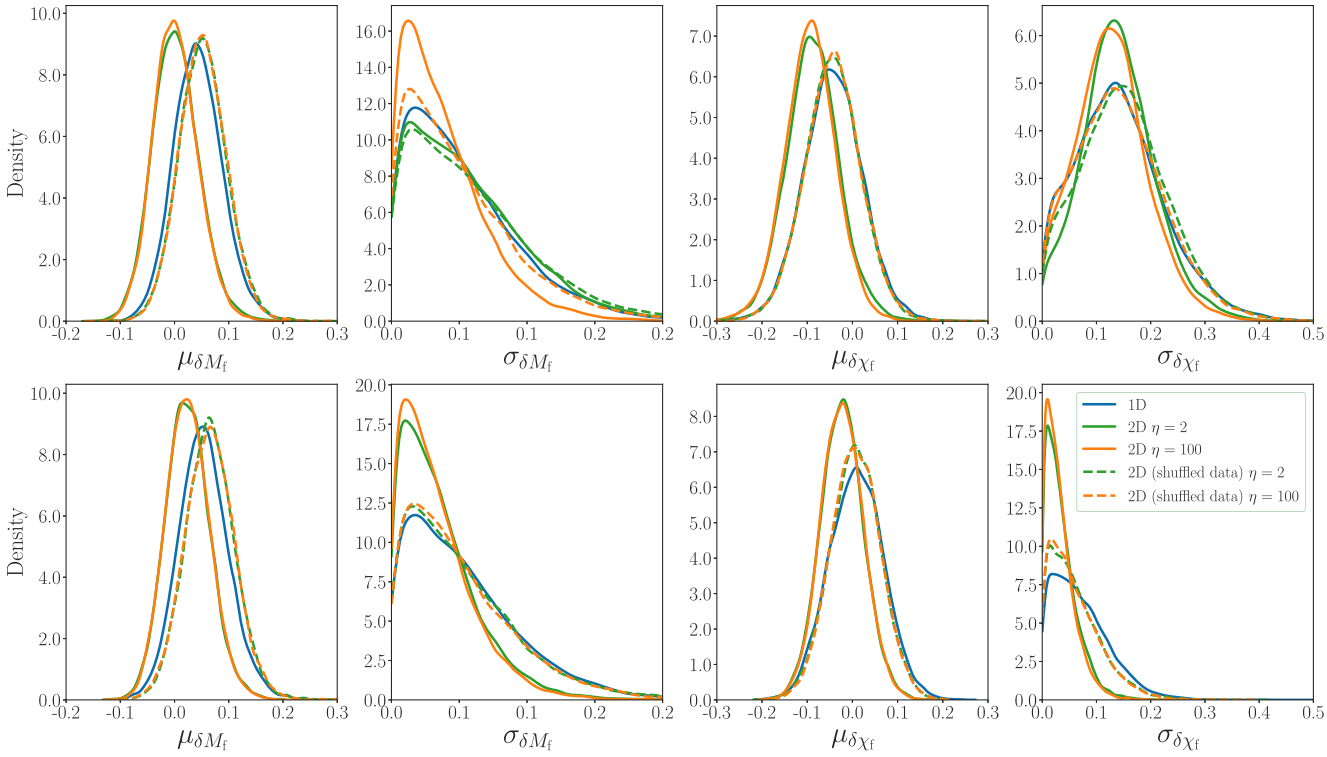

FIG. 10. Hyperparameter posterior distributions derived from GWTC IMR analyses with varying configurations with (left) and without (right) GW190814. The blue curves correspond to 1D analyses, while the orange and green curves show results from 2D analyses, with priors of $\eta = 2$ and $\eta = 100$ respectively. Dashed curves show 2D results from shuffled individual-event likelihoods (Appendix C). When the individual-event likelihoods are informative, hyperposteriors are not sensitive to the choice of η (solid orange versus green orange); 1D results can be recovered by erasing correlation information from the individual-event likelihoods through sample shuffling (dashed orange and green versus solid blue).

We repeat the hierarchical analysis and show results in Fig. 10. Each subplot shows hyperparamater posterior distributions from analyses with varying configurations. The blue curves show the results from 1D analyses, while other curves give results from 2D analyses. Orange and green curves correspond to the prior $\eta = 2$ and $\eta = 100$ cases, respectively. Dashed curves are results of 2D analyses on shuffled samples. Comparing the orange and green curves indicates that varying the prior η has minimal impact on the resulting posterior distributions. When the samples are shuffled, the 1D results and 2D results are identical.

APPENDIX D: FULL 4D RESULTS

In this appendix we show corner plots for all hyperparameters of the full 4D analysis with (purple) and without (light blue) GW190814 in Fig. 11.

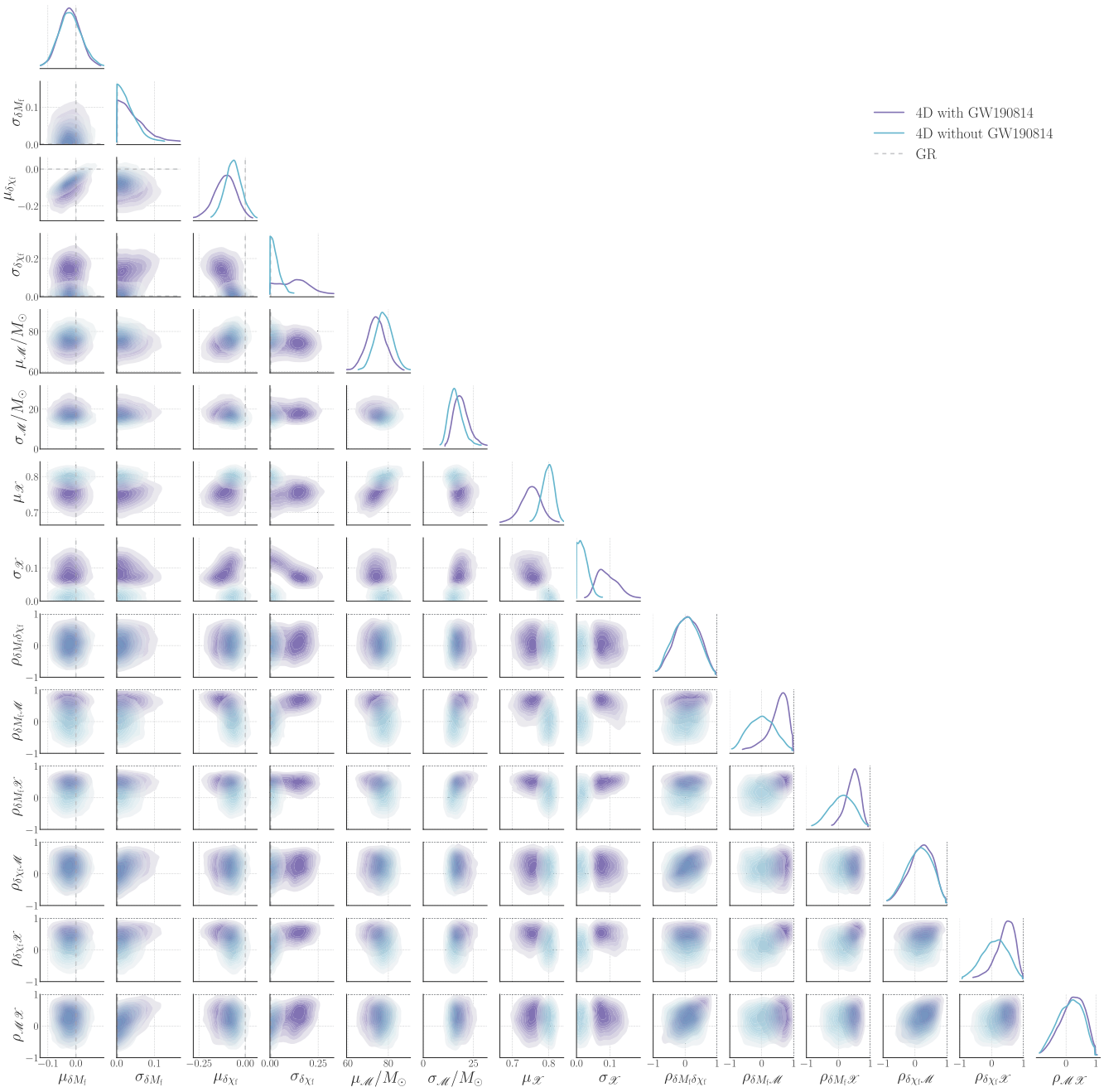

FIG. 11. Posterior distributions of all hyperparameters of the 4D IMR analysis of GWTC-3, including (purple) and excluding (light blue) GW190814. Contours enclose increments of 10% of probability mass starting at 90% for the outermost contour. The purple (light blue) distribution here is the same as the green distribution in Fig. 5 (Fig. 6). See Sec. IV B in the main text for a discussion of these results.

[1] J. Aasi *et al.* (LIGO Scientific Collaboration), Advanced LIGO, Classical Quantum Gravity **32**, 074001 (2015).

[2] F. Acernese *et al.* (Virgo Collaboration), Advanced Virgo: A second-generation interferometric gravitational wave detector, Classical Quantum Gravity **32**, 024001 (2015).

[3] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), GWTC-1: A gravitational-wave transient catalog of compact binary mergers observed by LIGO and Virgo during the first and second observing runs, Phys. Rev. X **9**, 031040 (2019).

[4] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), GWTC-2: Compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run, Phys. Rev. X **11**, 021053 (2021).
[5] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), GWTC-2.1: Deep extended catalog of compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run, Phys. Rev. D **109**, 022001 (2024).
[6] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run, Phys. Rev. X **13**, 041039 (2023).
[7] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Binary black hole population properties inferred from the first and second observing runs of Advanced LIGO and Advanced Virgo, Astrophys. J. Lett. **882**, L24 (2019).
[8] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), A gravitational-wave measurement of the Hubble constant following the second observing run of Advanced LIGO and Virgo, Astrophys. J. **909**, 218 (2021).
[9] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Population properties of compact objects from the second LIGO-Virgo gravitational-wave transient catalog, Astrophys. J. Lett. **913**, L7 (2021).
[10] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Search for gravitational waves associated with gamma-ray bursts detected by Fermi and Swift during the LIGO-Virgo Run O3a, Astrophys. J. **915**, 86 (2021).
[11] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Search for lensing signatures in the gravitational-wave observations from the first half of LIGO–Virgo's third observing run, Astrophys. J. **923**, 14 (2021).
[12] R. Abbott *et al.* (KAGRA, Virgo, and LIGO Scientific Collaborations), Population of merging compact binaries inferred using gravitational waves through GWTC-3, Phys. Rev. X **13**, 011048 (2023).
[13] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), Constraints on the cosmic expansion history from GWTC–3, Astrophys. J. **949**, 76 (2023).
[14] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), Search for gravitational waves associated with gamma-ray bursts detected by Fermi and Swift during the LIGO–Virgo Run O3b, Astrophys. J. **928**, 186 (2022).
[15] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Tests of general relativity with the binary black hole signals from the LIGO-Virgo catalog GWTC-1, Phys. Rev. D **100**, 104036 (2019).
[16] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Tests of general relativity with binary black holes from the second LIGO-Virgo gravitational-wave transient catalog, Phys. Rev. D **103**, 122002 (2021).
[17] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), Tests of general relativity with GWTC-3, arXiv:2112.06861.
[18] A. Zimmerman, C.-J. Haster, and K. Chatziioannou, On combining information from multiple gravitational wave sources, Phys. Rev. D **99**, 124044 (2019).
[19] M. Isi, K. Chatziioannou, and W. M. Farr, Hierarchical test of general relativity with gravitational waves, Phys. Rev. Lett. **123**, 121101 (2019).
[20] M. Isi, W. M. Farr, and K. Chatziioannou, Comparing Bayes factors and hierarchical inference for testing general relativity with gravitational waves, Phys. Rev. D **106**, 024048 (2022).
[21] C. Pacilio, D. Gerosa, and S. Bhagwat, Catalog variance of testing general relativity with gravitational-wave data, Phys. Rev. D **109**, L081302 (2024).
[22] E. Payne, M. Isi, K. Chatziioannou, and W. M. Farr, Fortifying gravitational-wave tests of general relativity against astrophysical assumptions, Phys. Rev. D **108**, 124060 (2023).
[23] R. Magee, M. Isi, E. Payne, K. Chatziioannou, W. M. Farr, G. Pratten, and S. Vitale, Impact of selection biases on tests of general relativity with gravitational-wave inspirals, Phys. Rev. D **109**, 023014 (2024).
[24] R. Essick and M. Fishbach, DAGnabbit! Ensuring consistency between noise and detection in hierarchical Bayesian inference, Astrophys. J. **962**, 169 (2024).
[25] W. James and C. Stein, Estimation with quadratic loss, in *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability* (University California Press, Berkeley, California, 1961), Vol. I, pp. 361–379.
[26] D. V. Lindley and A. F. M. Smith, Bayes estimates for the linear model, J. R. Stat. Soc. Ser. B **34**, 1 (1972).
[27] B. Efron and C. Morris, Stein's paradox in statistics, Sci. Am. **236**, No. 5, 119 (1977).
[28] D. B. Rubin, Estimation in parallel randomized experiments, J. Educ. Stat. **6**, 377 (1981).
[29] E. Thrane and C. Talbot, An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models, Publ. Astron. Soc. Aust. **36**, e010 (2019); **37**, e036(E) (2020).
[30] S. Vitale, D. Gerosa, W. Farr, and S. Taylor, Inferring the properties of a population of compact binaries in presence of selection effects, in *Handbook of Gravitational Wave Astronomy* (Springer, Singapore, 2022).
[31] N. Yunes and F. Pretorius, Fundamental theoretical bias in gravitational wave astrophysics and the parameterized post-Einsteinian framework, Phys. Rev. D **80**, 122003 (2009).
[32] T. G. F. Li, W. Del Pozzo, S. Vitale, C. Van Den Broeck, M. Agathos, J. Veitch, K. Grover, T. Sidery, R. Sturani, and A. Vecchio, Towards a generic test of the strong field dynamics of general relativity using compact binary coalescence, Phys. Rev. D **85**, 082003 (2012).
[33] M. Agathos, W. Del Pozzo, T. G. F. Li, C. Van Den Broeck, J. Veitch, and S. Vitale, TIGER: A data analysis pipeline for testing the strong-field dynamics of general relativity with gravitational wave signals from coalescing compact binaries, Phys. Rev. D **89**, 082001 (2014).
[34] A. K. Mehta, A. Buonanno, R. Cotesta, A. Ghosh, N. Sennett, and J. Steinhoff, Tests of general relativity with gravitational-wave observations using a flexible theory-independent method, Phys. Rev. D **107**, 044020 (2023).
[35] G. Carullo, W. Del Pozzo, and J. Veitch, Observational black hole spectroscopy: A time-domain multimode analysis of GW150914, Phys. Rev. D **99**, 123029 (2019); **100**, 089903(E) (2019).

[36] M. Isi, M. Giesler, W. M. Farr, M. A. Scheel, and S. A. Teukolsky, Testing the no-hair theorem with GW150914, Phys. Rev. Lett. **123**, 111102 (2019).
[37] A. Ghosh *et al.*, Testing general relativity using golden black-hole binaries, Phys. Rev. D **94**, 021101 (2016).
[38] A. Ghosh, N. K. Johnson-Mcdaniel, A. Ghosh, C. K. Mishra, P. Ajith, W. Del Pozzo, C. P. L. Berry, A. B. Nielsen, and L. London, Testing general relativity using gravitational wave signals from the inspiral, merger and ringdown of binary black holes, Classical Quantum Gravity **35**, 014002 (2018).
[39] T. C. K. Ng, M. Isi, K. W. K. Wong, and W. M. Farr, Constraining gravitational wave amplitude birefringence with GWTC-3, Phys. Rev. D **108**, 084068 (2023).
[40] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Tests of general relativity with GW150914, Phys. Rev. Lett. **116**, 221101 (2016); **121**, 129902(E) (2018).
[41] S. Perkins and N. Yunes, Are parametrized tests of general relativity with gravitational waves robust to unknown higher post-Newtonian order effects?, Phys. Rev. D **105**, 124047 (2022).
[42] A. Pai and K. G. Arun, Singular value decomposition in parametrised tests of post-Newtonian theory, Classical Quantum Gravity **30**, 025011 (2013).
[43] M. Saleem, S. Datta, K. G. Arun, and B. S. Sathyaprakash, Parametrized tests of post-Newtonian theory using principal component analysis, Phys. Rev. D **105**, 084062 (2022).
[44] D. Lewandowski, D. Kurowicka, and H. Joe, Generating random correlation matrices based on vines and extended onion method, J. Multivariate Anal. **100**, 1989 (2009).
[45] D. Akinc and M. Vandebroek, Bayesian estimation of mixed logit models: Selecting an appropriate prior for the covariance matrix, J. Choice Modell. **29**, 133 (2018).
[46] M. Lieu, W. M. Farr, M. Betancourt, G. P. Smith, M. Sereno, and I. G. McCarthy, Hierarchical inference of the relationship between concentration and mass in galaxy groups and clusters, Mon. Not. R. Astron. Soc. **468**, 4872 (2017).
[47] Y. Tao, K.-K. Phoon, H. Sun, and Y. Cai, Hierarchical Bayesian model for predicting small-strain stiffness of sand, Can. Geotech. J. **61**, 668 (2024).
[48] Y. Feng, K. Gao, A. Mignan, and J. Li, Improving local mean stress estimation using Bayesian hierarchical modelling, Int. J. Rock Mech. Min. Sci. Geomech. Abstr. **148**, 104924 (2021).
[49] J. Golomb and C. Talbot, Hierarchical inference of binary neutron star mass distribution and equation of state with gravitational waves, Astrophys. J. **926**, 79 (2022).
[50] A. Ghosh *et al.*, Testing general relativity using golden black-hole binaries, Phys. Rev. D **94**, 021101 (2016).
[51] A. Ghosh, N. K. Johnson-Mcdaniel, A. Ghosh, C. K. Mishra, P. Ajith, W. Del Pozzo, C. P. L. Berry, A. B. Nielsen, and L. London, Testing general relativity using gravitational wave signals from the inspiral, merger and ringdown of binary black holes, Classical Quantum Gravity **35**, 014002 (2018).
[52] M. Isi, W. M. Farr, M. Giesler, M. A. Scheel, and S. A. Teukolsky, Testing the black-hole area law with GW150914, Phys. Rev. Lett. **127**, 011103 (2021).
[53] M. Cabero, C. D. Capano, O. Fischer-Birnholtz, B. Krishnan, A. B. Nielsen, A. H. Nitz, and C. M. Biwer, Observational tests of the black hole area increase law, Phys. Rev. D **97**, 124069 (2018).
[54] P. Schmidt, M. Hannam, S. Husa, and P. Ajith, Tracking the precession of compact binaries from their gravitational-wave signal, Phys. Rev. D **84**, 024046 (2011).
[55] P. Schmidt, M. Hannam, and S. Husa, Towards models of gravitational waveforms from generic binaries: A simple approximate mapping between precessing and non-precessing inspiral signals, Phys. Rev. D **86**, 104063 (2012).
[56] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Simple model of complete precessing black-hole-binary gravitational waveforms, Phys. Rev. Lett. **113**, 151101 (2014).
[57] S. Khan, F. Ohme, K. Chatziioannou, and M. Hannam, Including higher order multipoles in gravitational-wave models for precessing binary black holes, Phys. Rev. D **101**, 024056 (2020).
[58] G. Pratten *et al.*, Computationally efficient models for the dominant and subdominant harmonic modes of precessing binary black holes, Phys. Rev. D **103**, 104056 (2021).
[59] C. García-Quirós, M. Colleoni, S. Husa, H. Estellés, G. Pratten, A. Ramos-Buades, M. Mateu-Lucena, and R. Jaume, Multimode frequency-domain model for the gravitational wave signal from nonprecessing black-hole binaries, Phys. Rev. D **102**, 064002 (2020).
[60] G. Pratten, S. Husa, C. Garcia-Quiros, M. Colleoni, A. Ramos-Buades, H. Estelles, and R. Jaume, Setting the cornerstone for a family of models for gravitational waves from compact binaries: The dominant harmonic for nonprecessing quasicircular black holes, Phys. Rev. D **102**, 064001 (2020).
[61] R. Abbott, T. D. Abbott, S. Abraham *et al.*, GW190814: Gravitational waves from the coalescence of a 23 solar mass black hole with a 2.6 solar mass compact object, Astrophys. J. Lett. **896**, L44 (2020).
[62] KAGRA, LIGO Scientific, and Virgo Collaborations, Data release for tests of general relativity with GWTC-3, 10.5281/zenodo.7007370 (2022).
[63] P. Stoica and Y. Selen, Model-order selection: A review of information criterion rules, IEEE Signal Process. Mag. **21**, 36 (2004).
[64] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, SCIKIT-LEARN: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011), http://jmlr.org/papers/v12/pedregosa11a.html.
[65] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, API design for machine learning software: Experiences from the scikit-learn project, in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning* (2013), pp. 108–122.
[66] P. A. Bromiley, Products and convolutions of Gaussian probability density functions (2013).
[67] D. W. Hogg, A. M. Price-Whelan, and B. Leistedt, Data analysis recipes: Products of multivariate Gaussians in bayesian inferences, arXiv:2005.14199.