Systematic bias from waveform modeling for binary black hole populations in next-generation gravitational wave detectors

Veome Kapil[,*] Luca Reali[,†] Roberto Cotesta, and Emanuele Berti[‡]

William H. Miller III Department of Physics and Astronomy, Johns Hopkins University,
3400 N. Charles Street, Baltimore, Maryland, 21218, USA

Next-generation gravitational wave detectors such as the Einstein Telescope and Cosmic Explorer will have increased sensitivity and observing volumes, enabling unprecedented precision in parameter estimation. However, this enhanced precision could also reveal systematic biases arising from waveform modeling, which may impact astrophysical inference. We investigate the extent of these biases over a year-long observing run with 10^5 simulated binary black hole sources using the linear signal approximation. To establish a conservative estimate, we sample binaries from a smoothed truncated power-law population model and compute systematic parameter biases between the `IMRPhenomXAS` and `IMRPhenomD` waveform models. For sources with signal-to-noise ratios above 100, we estimate statistically significant parameter biases in ~3%–20% of the events, depending on the parameter. We find that the average mismatch between waveform models required to achieve a bias of $\leq 1\sigma$ for 99% of detections with signal-to-noise ratios ≥ 100 should be $\mathcal{O}(10^{-5})$, or at least one order of magnitude better than current levels of waveform accuracy.

I. INTRODUCTION

Gravitational wave (GW) astronomy is revolutionizing our understanding of relativistic astrophysics through the direct detection of double compact object (DCO) mergers. The LIGO-Virgo-KAGRA GW detector network has already observed $\mathcal{O}(100)$ binary black hole (BBH) mergers [1], allowing us to chart the population of black holes in the local universe for the first time. The number of such observations is expected to increase dramatically with the introduction of next-generation (XG) GW detectors such as the Einstein Telescope (ET) [2] and Cosmic Explorer (CE) [3]. The increased sensitivities and observing volumes of these detectors will open the door to unprecedented science, ranging from studies of DCO formation channels and multimessenger astronomy to probes of physics beyond general relativity, nuclear physics and dark matter [4–9].

These applications will rely upon accurate measurements of the astrophysical parameters of DCOs, which in turn depend on the waveform models used to analyze the GW signals. Due to computational constraints, current waveform models are semianalytic approximations to two-body solutions in general relativity, which are then calibrated to numerical relativity (NR) simulations of DCO mergers. This gap between NR waveforms and semianalytic models can lead to systematic biases in the inferred binary parameters [10]. The problem of biased parameter inference is likely to be worse in regions of the parameter space where NR calibration points are sparse.

The limited signal-to-noise ratio (SNR) achievable in current detectors leads to large enough statistical uncertainties on binary parameters that systematic biases are generally not significant. At the sensitivity level of the third observing run of the LIGO-Virgo-KAGRA detectors (O3), the systematic biases from using inadequately calibrated waveform models are not particularly concerning (but see, e.g., [11,12] for cases in the O3 catalog where different waveform models produce inconsistent parameter posteriors). Future ground-based GW detectors, however, are expected to observe DCO coalescences with SNRs in excess of $\mathcal{O}(100)$ or even $\mathcal{O}(1000)$ [13,14]. The statistical uncertainties associated with such detections will be small enough that any systematic biases due to waveform errors could become significant.

Some of the existing literature estimates the impact of waveform errors on a handful of loud binaries. Studies by Owen *et al.* [15] and Read [16], among others, demonstrate the possibility of significant systematic biases due to waveform errors, even as early as O4. Pürrer and Haster [17] compute the expected parameter biases for a few high-SNR binaries in CE/ET to be in excess of 100σ. They show, albeit briefly, that waveform errors could also bias population model parameters. They further estimate that the mismatch between two given waveform models must be

*vkapil1@jhu.edu
†lreali1@jhu.edu
‡berti@jhu.edu

improved by ∼3 orders of magnitude to ensure accurate parameter estimation for the loudest binaries in XG detectors. Hu and Veitch [12] also conclude that a 3–4 order of magnitude improvement in waveform accuracy would be required at the SNRs expected in future ground-based detectors. Previous work has also shown that systematic bias will play a crucial role for the high-SNR massive black hole mergers observed by the Laser Interferometer Space Antenna (LISA) [10,18,19], and that parametrized tests of general relativity using high-SNR detections can produce biased results due to waveform modeling errors [20–23].

In this work, we expand and complement these studies by establishing a lower bound on the required waveform accuracy improvement for XG detectors. To this end, we consider the distribution of parameter biases across an astrophysical population of stellar mass BBHs. We first limit our focus to a pair of current waveform models from the same waveform family (IMRPhenomD [24,25] and IMRPhenomXAS [26]), conservatively quantifying the impact that current waveform errors will have on a realistic population of XG GW observations. As a second step, we synthetically tune the accuracy of our chosen waveforms to model future improvements in waveform calibration. Using waveforms interpolated between IMRPhenomD and IMRPhenomXAS, we estimate a minimum accuracy requirement for future waveform models to produce consistent parameter estimation across an astrophysically motivated BBH population.

The plan of the paper is as follows. In Sec. II we discuss how the Fisher information matrix formalism can be used to estimate waveform errors and systematic biases through the formalism developed by Cutler and Vallisneri [10], and we introduce our waveform interpolation scheme. In Sec. III we present our model of the astrophysical BBH population. In Sec. IV we compute waveform errors within the Cutler-Vallisneri formalism, and estimate the waveform accuracy requirement for these systematic errors to be subdominant with respect to statistical errors in XG detectors. In Sec. V we compare these results against full Bayesian parameter estimation calculations for selected binaries. In Sec. VI we discuss the implications of our work and possible directions for future research. In Appendix A we generalize the Cutler-Vallisneri calculation to a detector network, in Appendix B we illustrate the importance of waveform alignment in the calculation of the biases on luminosity distance, in Appendix C we show how the choice of a reference frequency parameter affects the behavior of frequency-domain waveforms and of the bias, and in Appendix D we show how the biases and statistical errors change as a function of the interpolation parameter between different waveforms.

II. STATISTICAL AND SYSTEMATIC ERRORS

In this section we discuss how to compute parameter errors and the Cutler-Vallisneri parameter bias within the Fisher information matrix formalism. We introduce the mismatch between waveforms, and we present a waveform interpolant that will be used in the rest of the paper to quantify how close two waveforms should be in order to achieve some desired maximum bias in the estimated parameters.

A. Fisher information matrix

The Fisher information matrix for a detected waveform $h(\vec{\theta})$ which depends on a set of parameters $\vec{\theta} = \{\theta_i\}$ is defined as

$$\Gamma_{ij} \equiv \left(\left. \frac{\partial h}{\partial \theta^i} \right| \frac{\partial h}{\partial \theta^j} \right), \qquad (1)$$

where the inner product between two waveforms in the frequency domain is defined as

$$(a|b) \equiv 4\mathrm{Re} \int_{f_{\min}}^{f_{\max}} \frac{\tilde{a}(f) \cdot \tilde{b}(f)^*}{S_n(f)} \, df, \qquad (2)$$

and $S_n(f)$ is the noise power spectral density for the given detector. For a network of detectors, the network Fisher matrix is given by

$$\Gamma_{\mathrm{Net}} = \Sigma_{\mathrm{D}=1}^{\mathrm{N_D}} \Gamma_{\mathrm{D}}, \qquad (3)$$

where the subscript D labels each detector in a network made up of $\mathrm{N_D}$ total detectors. In the linear signal approximation and under the assumption of Gaussian posteriors [27], the network's covariance matrix is

$$\Sigma_{\mathrm{Net}} = (\Gamma_{\mathrm{Net}})^{-1}. \qquad (4)$$

The variances and covariances for each parameter are given by the diagonal and off-diagonal elements of Σ_{Net}, respectively, such that

$$\sigma_{\theta_i} = \sqrt{\Sigma_{ii,\mathrm{Net}}}. \qquad (5)$$

We use the GWBENCH [28] code to compute Fisher information matrices and waveform derivatives for our simulated BBH population. The detector network consists of one 40 km Cosmic Explorer detector in Idaho, USA and one 20 km Cosmic Explorer detector in New South Wales, Australia, plus one Einstein Telescope detector at design sensitivity in Cascina, Italy (from now on, for brevity, a "2CE + ET" network). The specific locations and orientations of these detectors can be found in Table III of Ref. [28] under the labels C, S, and E, respectively. The power spectral densities (PSDs) adopted here correspond to the 4020ET network of Ref. [8].

We set the minimum frequency for the detector response to be 5 Hz, which is also the reference frequency used to generate the waveforms.

We compute the waveform derivatives and Fisher matrices for the following parameters: chirp mass $\mathcal{M}_c$, symmetric mass ratio η, components of the primary and secondary dimensionless spins along the orbital angular momentum $\chi_{1,z}$ and $\chi_{2,z}$, luminosity distance D_L, coalescence time t_c, coalescence phase ϕ_c, binary inclination ι, right ascension RA, declination DEC, and polarization angle ψ. These parameters encapsulate all the degrees of freedom in the waveform models considered for this study.

Although we compute derivatives with respect to all of these parameters, in this study we only consider biases on the intrinsic binary parameters and on the luminosity distance D_L. Our tests show that the posteriors for the intrinsic parameters are generally well approximated by Gaussian distributions, and so the Fisher formalism is better suited to these parameters. To incorporate physical bounds on the statistical errors (particularly for the spins) within the Fisher matrix formalism, we crudely impose prior bounds following Refs. [29–31], i.e., we rewrite the Fisher matrix for each binary as

$$\Gamma_{\text{bounded}} = \Gamma + \Gamma^{(0)}. \tag{6}$$

Here, Γ is the naïve output of the Fisher matrix computation, and $\Gamma^{(0)}$ is the Fisher matrix corresponding to the multivariate Gaussian prior imposed on the system. The covariance matrix for the system can then be written as

$$\Sigma_{\text{bounded}} = (\Gamma + \Gamma^{(0)})^{-1}. \tag{7}$$

For simplicity, we only impose priors on $\chi_{1,z}$ and $\chi_{2,z}$ such that their statistical uncertainties are bounded in the range $[-1, 1]$. We find that imposing priors on other parameters (such as $\mathcal{M}_c$ and η) does not significantly impact the covariance, since the naïve error estimates for those parameters are generally small compared to the prior range.

B. Cutler-Vallisneri bias

Consider the scenario where a DCO merger produces some true gravitational waveform h_{TR}, which we might detect with our network of ground-based GW detectors. Let us denote by h_{AP} some hypothetical state-of-the-art waveform model approximant used to infer the binary parameters. Due to imperfect calibration of the approximate waveform model, there will be regions of the parameter space $\vec{\theta}$ such that $h_{\text{TR}}(\vec{\theta}) \neq h_{\text{AP}}(\vec{\theta})$. The resulting systematic bias on the parameter θ^i is given by

$$\Delta\theta^i = \left(\Gamma_{\text{AP,Net}}^{-1}\right)^{ij} \Sigma_{\text{D}=1}^{\text{N}_\text{D}} \left(\partial_j(h_{\text{AP}})_\text{D}|(h_{\text{TR}})_\text{D} - (h_{\text{AP}})\text{D}\right), \tag{8}$$

where $\Gamma_{\text{AP,Net}}$ is the network Fisher matrix defined in Eq. (3) evaluated using the h_{AP} waveform, and $(h_{\text{AP}})_\text{D}$ and $(h_{\text{TR}})_\text{D}$ represent the responses of the approximate and true waveforms in detector D, respectively.

To our knowledge, this bias was first computed in Refs. [10,32]. The generalization to a detector network given in Eq. (8) above can be found in Appendix A. From now on, we will refer to $\Delta\theta^i$ as the Cutler-Vallisneri bias.

In practice, we do not have access to a perfect model for the "true" waveform h_{TR}. To emulate the effect of waveform modeling errors, we thus use two waveform models from the IMRPhenom family of phenomenological GW approximants. Specifically, we use IMRPhenomXAS [26] to represent the "true" waveform h_{TR}, and IMRPhenomD [24,25] as the imperfectly calibrated waveform model h_{AP}.

These are both frequency-domain models of the gravitational radiation from nonprecessing BBHs which include only the dominant (2, 2) spherical harmonic mode. IMRPhenomXAS is an update to IMRPhenomD with better calibration to reference waveforms with large mass ratios and unequal spins [26]. As a result, any differences between IMRPhenomXAS and IMRPhenomD should be largest for volumes in the parameter space where the fewest calibration waveforms exist. We use semianalytic waveform models, rather than NR waveforms, to represent the true as well as the approximate waveforms so that we can study waveform errors for a population of BBHs spanning a large volume of the parameter space. Since our goal is to estimate waveform accuracy at the order-of-magnitude level, we deem these waveform models sufficient for this study. We expect waveform errors to be larger between models from different families and with different physics, so our study represents a conservative estimate of the resulting biases. We leave more detailed comparisons with other waveform models to future work.

C. Waveform mismatch

The differences in waveform models can be quantified by the relative disagreement, or mismatch, between the waveforms at a given point in parameter space. The mismatch $\mathcal{M}$ between an approximate waveform h_{AP} and the true waveform h_{TR} can be computed as

$$\mathcal{M} = 1 - \max_{\phi_c, t_c, \psi} \frac{(h_{\text{AP}}(\vec{\theta}, \phi_c, t_c, \psi)|h_{\text{TR}})}{\sqrt{(h_{\text{AP}}|h_{\text{AP}})}\sqrt{(h_{\text{TR}}|h_{\text{TR}})}}, \tag{9}$$

where we maximize over the parameters that capture differences in waveform convention. A mismatch of zero corresponds to two exactly equivalent waveforms, while a mismatch of unity indicates maximally distinct waveforms in the frequency domain.

D. Interpolating waveform models

We are interested in not only quantifying systematic biases between current waveform models, but also estimating the requirements on future waveform models to mitigate these biases. To find these accuracy requirements in practice, we would like to generate waveforms with

arbitrary levels of calibration with respect to a reference waveform model. To accomplish this, we first decompose our true and approximate waveforms into an amplitude and a phase:

$$h_{\mathrm{TR}}(f) = A_{\mathrm{TR}}(f) \exp[i\phi_{\mathrm{TR}}(f)], \qquad (10)$$

$$h_{\mathrm{AP}}(f) = A_{\mathrm{AP}}(f) \exp[i\phi_{\mathrm{AP}}(f)]. \qquad (11)$$

We can now construct a new waveform

$$h'_{\mathrm{AP}}(f) = A'_{\mathrm{AP}}(f) \exp[i\phi'_{\mathrm{AP}}(f)], \qquad (12)$$

where the amplitude and phase are obtained by linearly interpolating between the two original waveforms as follows:

$$A'_{\mathrm{AP}} = \lambda A_{\mathrm{TR}} + (1 - \lambda)A_{\mathrm{AP}}, \qquad (13)$$

$$\phi'_{\mathrm{AP}} = \lambda\phi_{\mathrm{TR}} + (1 - \lambda)\phi_{\mathrm{AP}}. \qquad (14)$$

This approach allows us to place waveforms anywhere between two given waveform models, and thus to emulate future levels of waveform mismatch. By construction, the h'_{AP} waveform is expected to be in better agreement with h_{TR} than the h_{AP} waveform (see Appendix D). The parameter λ is best interpreted as a perturbation around $h_{\mathrm{TR}}(f)$, in the direction of $h_{\mathrm{AP}}(f)$, such that $|1 - \lambda| \ll 0$. We avoid larger values of λ, which may lead to unphysical parameter biasing artifacts.

III. ASTROPHYSICAL POPULATION OF BINARY BLACK HOLES

The accuracy of available waveform models can vary drastically as a function of the intrinsic parameters of the binary (such as masses and spins), depending on the availability and accuracy of NR calibration waveforms. One of our goals is to understand what fraction of the BBH systems observable with XG GW detectors would be strongly impacted by these biases. To answer this question, we will quantify the aggregate severity of waveform calibration biases over simulated binary populations compatible with current LIGO-Virgo-KAGRA observations. Our results are naturally sensitive to the chosen population model, so it will be worth expanding the present analysis using various parametric and nonparametric populations in the future. Following Ref. [5], we assume that our detector network will detect 10^5 BBH sources over a one year-long observing run. We sample the masses, spins, orientations, and sky locations of these BBHs from probability density functions consistent with our astrophysical assumptions. While the specific population model chosen for our study may not precisely represent the true distribution of BBHs in the Universe, it should provide a reasonable first estimate of the distribution of systematic biases for some of the most interesting binary parameters.

Masses. The mass distribution of our BBH population is chosen to follow the TRUNCATED model of Ref. [33], with some modifications at the low-mass end. For our population of 10^5 BBH systems, the primary mass m_1 is drawn from a modified truncated power-law distribution

$$p(m_1|\alpha, \eta, m_0, m_{\min}, m_{\max}) \propto m_1^\alpha \times h(m_1|m_0, \eta)$$
$$\times \mathcal{B}(m_1|m_{\min}, m_{\max}). \qquad (15)$$

The proportionality sign indicates that the final probability density function should be normalized, but we omit the full expression here for clarity. The first term of Eq. (15) represents the usual power law with exponent α. The next term is a high-pass Butterworth filter with roll-off mass $m_0 = 5M_\odot$ and sharpness parameter $\eta = 50$, given by

$$h(m_1|m_0, \eta) = \left[1 + \left(\frac{m_0}{m_1}\right)^\eta\right]^{-1}. \qquad (16)$$

This term is responsible for smoothing the lower end of the primary mass distribution so that it remains continuous. The last term is the boxcar function, used to enforce that m_1 is between the limits $m_{\min}$ and $m_{\max}$, and given by

$$\mathcal{B}(m_1|m_{\min}, m_{\max}) = \begin{cases} 1 & \text{if } m_1 \in [m_{\min}, m_{\max}], \\ 0 & \text{otherwise.} \end{cases} \qquad (17)$$

In our simulations, we fix $m_{\min} = 3M_\odot$ and $m_{\max} = 60M_\odot$. The Butterworth filter and power law ensure that the peak of the primary mass distribution is at $5M_\odot$. The secondary mass m_2 is then drawn uniformly between $m_{\min}$ and m_1.

When computing the waveform responses and derivatives, these masses are converted to chirp mass $\mathcal{M}_c = (m_1 m_2)^{3/5}(m_1 + m_2)^{-1/5}$ and symmetric mass ratio $\eta = (m_1 m_2)(m_1 + m_2)^{-2}$.

Due to the gravitational redshift, only the redshifted masses of a binary are observable via GWs. To properly emulate GW detections, the binary component masses sampled above must be converted into detector-frame redshifted masses such that $m_{\mathrm{det}} = m(1 + z)$, where the redshift for each binary is sampled from a suitable probability distribution. Similarly, we have $\mathcal{M}_{c,\mathrm{det}} = \mathcal{M}_c(1 + z)$.

Spins. The spin magnitude for each binary component χ_i is independently drawn from a beta distribution

$$p(\chi_i|a, b) \propto \chi_i^{a-1}(1 - \chi_i)^{b-1}. \qquad (18)$$

We choose $a = 2$ and $b = 7$, which generally reproduces the DEFAULT spin model from Ref. [33]. The spin orientations are isotropically distributed on the sphere, such that the polar angle θ obeys $p(\cos\theta) = U[-1, 1]$ and the

azimuthal angle ϕ is distributed as $p(\phi) = U[0, 2\pi]$. We limit our study to waveform models with aligned spins, so we consider only the $\chi_{i,z} = \chi_i \cos\theta$ component.

Sky position and orientation. The right ascension and declination of the binaries are sampled uniformly on the sphere:

$$p(\mathrm{RA}) \propto U[0, 2\pi], \tag{19}$$

$$p(\cos(\Theta)) \propto U[-1, 1], \tag{20}$$

where $\Theta = \mathrm{DEC} + \pi/2$. The inclination angle ι is sampled uniformly in $\cos\iota$, where an angle of 0 ($\pi/2$) represents a face-on (edge-on) binary. The polarization angle ψ is sampled uniformly between 0 and 2π.

Coalescence time and phase. The coalescence time, t_c, and phase, ϕ_c, are arbitrarily set to zero for all binaries in the population when evaluating h_{TR}. These parameters are extrinsic, and can be chosen freely for each waveform model such that the overlap with the GW signal is maximized [34–36]. We therefore evaluate h_{AP} at the values of t_c and ϕ_c that minimize the mismatch defined in Eq. (9), to allow for differences in conventions between the two waveform models. We became aware of the importance of aligning the waveforms when computing the bias formula from the authors of Ref. [34]. We therefore evaluate h_{AP} at the values of t_c and ϕ_c that minimize the mismatch defined in Eq. (9), to allow for differences in conventions between the two waveform models. In practice, aligning `IMRPhenomD` and `IMRPhenomXAS` in coalescence phase and time can lower the computed biases by several orders of magnitude, as first derived by Dhani *et al.* [34].

Redshift. The source redshifts are drawn from the Madau and Fragos [37] probability distribution, given by

$$\Psi(z) = 0.01 \frac{(1+z)^{2.6}}{1 + [(1+z)/3.2]^{6.2}} M_\odot \, \mathrm{yr}^{-1} \, \mathrm{Mpc}^{-3}. \tag{21}$$

The range of possible redshifts is chosen to be $z \in [0.02, 50]$ to match the population in Ref. [14]. Wherever necessary, the redshift is converted to D_{L} (and vice versa) using the Planck18 [38] cosmology. For simplicity, we assume that the star formation redshift distribution is the same as the merger redshift distribution, i.e., we do not explicitly model the delay times. The effect of this assumption will likely be to underestimate the significance of waveform biases, since a more careful modeling of delay times will cause BBHs to merge at lower redshifts, and thus produce smaller statistical errors on average.

In Fig. 1 we show the redshift distribution of the simulated observable population at various SNR thresholds. The chosen detector network can detect simulated BBHs with SNR of $\mathcal{O}(100)$ out to $z \sim 3$. The median SNR across the entire population is about 22, and the maximum
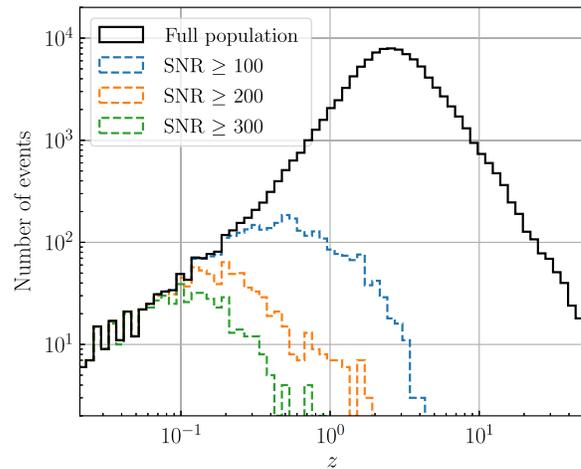
FIG. 1. Redshift distributions of the binaries observed at various SNR thresholds by a 2CE + ET network, as sampled from the Madau and Fragos [37] distribution.

SNR for a binary in our population is $\sim$4000. This SNR distribution is in reasonable agreement with recent forecasts of the capabilities of XG detectors [39–41].

IV. RESULTS

In our estimates of statistical and systematic parameter estimation errors, we will limit our attention to the loudest BBH sources, i.e., those with SNR ≥ 100. These systems (roughly 3000 binaries in our population model) are particularly valuable for astrophysical applications due to the small parameter uncertainties, and they are also most affected by any systematic biases between different waveform models. The event posteriors of high-SNR BBHs are also generally well approximated by Gaussian distributions, and so the Fisher information formalism used in this work is expected to be a good approximation for these events. For the same reason, we also expect the systematic bias computed under the linear signal approximation to be reliable for these binaries. Wherever necessary, we impose a limit in postprocessing to ensure that the biases do not exceed physical bounds.

A. Current waveform model accuracy

In Fig. 2 we show the distribution of Cutler-Vallisneri biases between `IMRPhenomXAS` and `IMRPhenomD`, $\Delta\theta$, normalized to the statistical uncertainties obtained from the Fisher analysis, σ_θ, for the following binary parameters: $m_{1,\mathrm{det}}$, $m_{2,\mathrm{det}}$, $\chi_{1,z}$, $\chi_{2,z}$, and D_{L}. Note that while both h_{TR} and h_{AP} enter the calculation of $\Delta\theta$ through the inner product of Eq. (8), the Fisher matrix is always computed only in terms of the recovered signal h_{AP}.

As a rule of thumb, we will say that a parameter is unbiased (biased) if $|\Delta\theta/\sigma_\theta| \leq 1$ ($|\Delta\theta/\sigma_\theta| > 1$, respectively). The differences between `IMRPhenomXAS` and `IMRPhenomD` can bias parameter values by at least 1σ
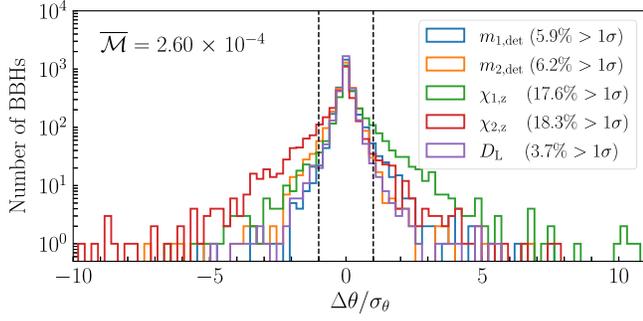
FIG. 2. Significance of waveform calibration biases between `IMRPhenomD` and `IMRPhenomXAS`, for various parameters, over the SNR ≥ 100 subpopulation. All the biases have been normalized by the statistical uncertainties, such that the dashed lines indicate the 1σ bias thresholds. The average mismatch between the true and approximate waveforms computed across this set of binaries is shown in the top left corner, and the percentages of events with $> 1\sigma$ bias are listed in the legend.

for a few percent of this high-SNR population, with spin being the most affected parameter.

Although not shown in this figure (but see Fig. 3), the largest biases can reach $\geq 20\sigma$ for the masses, and $\geq 50\sigma$ for the spins. Regardless, there are at least 2500 binaries in our population with SNR ≥ 100 and unbiased intrinsic parameters. The average mismatch between the two waveform models across the SNR ≥ 100 binaries is $\overline{\mathcal{M}} = 2.6 \times 10^{-4}$. Note that the parameters shown can be biased in either direction, depending on the specific configuration of the binary system. As such, the average values of the biases are consistent with zero for most parameters when we consider the entire population.

In Fig. 3 we show a few alternate visualizations of the magnitudes of the normalized biases $|\Delta\theta/\sigma_\theta|$ for various binary parameters.

The top panel of this plot displays a strong positive correlation between the SNR of a binary and the significance of the waveform biases. This is expected, as the statistical uncertainties scale as SNR^{-1}, while it can be shown that the systematic uncertainties from Eq. (8) do not depend on SNR.

The spin of a binary also seems to have some impact on the magnitude of the biases: the second panel from the top in Fig. 3 shows that large differences in spin magnitudes are weakly correlated with larger biases. Although this is not a particularly strong effect in the data, it aligns with the differences in calibration between `IMRPhenomD` and `IMRPhenomXAS`, with the former waveform model being calibrated primarily to equal-spin binary simulations [25], and the latter being calibrated to more unequal-spin simulations [26]. As the spin parameter space gets explored better in simulations, differences in calibration between different waveform models will likely become more apparent.
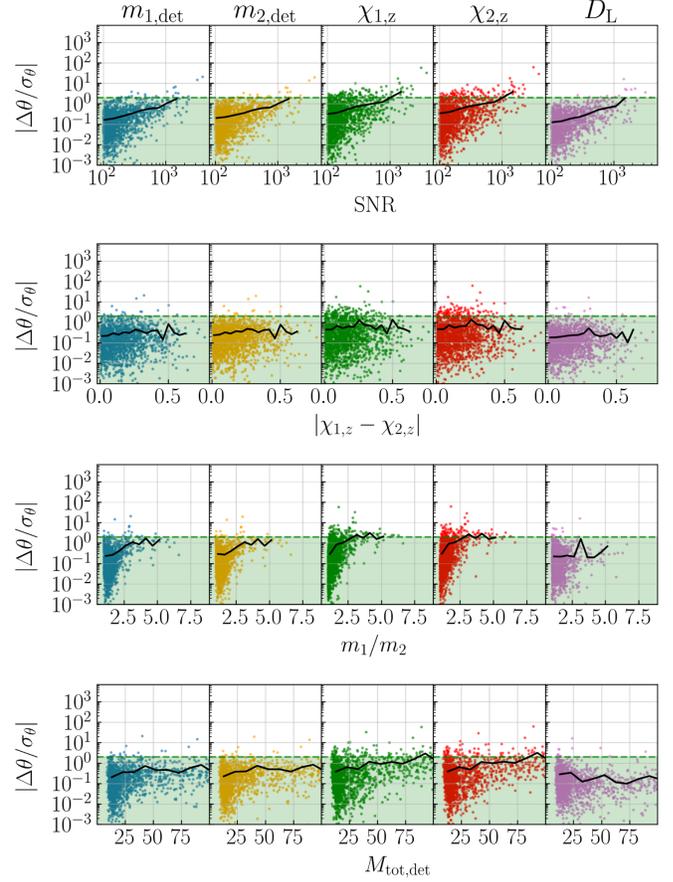
FIG. 3. The absolute value of Cutler-Vallisneri biases for various parameters vs SNR (top), absolute difference between spins (2nd from top), mass ratio (3rd from top), and detector-frame total mass (bottom). Each dot is one binary from the SNR ≥ 100 subpopulation, and the solid black line tracks the rolling average of the bias for each parameter. Biases are normalized to the statistical errors for each binary. The green shaded region represents biases below 1σ.

The significance of Cutler-Vallisneri biases for intrinsic parameters is also somewhat positively correlated to the binary mass ratio m_1/m_2 (third panel) and to the total detector-frame mass (bottom panel). Compared to `IMRPhenomD`, the `IMRPhenomXAS` waveform model is calibrated to additional waveforms with large mass ratios. Additionally, `IMRPhenomD` exhibits worse calibration to high-mass binaries than to low-mass binaries (see Fig. 15 of [26]), which may explain the trends in Fig. 3. Although not explored further in this study, other population models allowed by the Gravitational Wave Transient Catalog 3 (GWTC-3) [42] exhibit overdensities in the primary mass spectrum around $m_1 \approx 10 M_\odot$ or higher. Our assumed population is dominated by lower mass black holes, and hence, our study is conservative in terms of the assumed mass distribution. The median biases for the true observed BBH population will likely be larger than our estimates.

TABLE I.   Median magnitudes of the Cutler-Vallisneri biases, $\overline{|\Delta\theta|}$ (and their significance with respect to statistical errors, $\overline{|\Delta\theta/\sigma_\theta|}$), for subpopulations above various SNR thresholds in the 2CE + ET detector configuration. $\mathrm{SNR}_{1\sigma}$ denotes the approximate SNR threshold above which the median significance of the bias in each parameter is expected to exceed unity. For parameters that never exceed a median bias of 1σ in our population due to limited statistics, the $\mathrm{SNR}_{1\sigma}$ thresholds are calculated by extrapolation, and are marked with an asterisk (*).

| Parameter | Median absolute bias $\overline{|\Delta\theta|}$ ($\overline{|\Delta\theta/\sigma_\theta|}$) | | | |
	SNR $\geq$ 100	SNR $\geq$ 200	SNR $\geq$ 300	$\mathrm{SNR}_{1\sigma}$
$m_{1,\mathrm{det}}$	$0.24 M_\odot$ (0.1σ)	$0.25 M_\odot$ (0.2σ)	$0.26 M_\odot$ (0.4σ)	$\sim$500*
$m_{2,\mathrm{det}}$	$0.18 M_\odot$ (0.1σ)	$0.18 M_\odot$ (0.3σ)	$0.21 M_\odot$ (0.4σ)	$\sim$500*
m_1	$0.18 M_\odot$ (0.1σ)	$0.22 M_\odot$ (0.3σ)	$0.25 M_\odot$ (0.4σ)	$\sim$500*
m_2	$0.14 M_\odot$ (0.1σ)	$0.16 M_\odot$ (0.3σ)	$0.19 M_\odot$ (0.4σ)	$\sim$500*
$\chi_{1,z}$	0.07 (0.2σ)	0.12 (0.4σ)	0.14 (0.5σ)	$\sim$400
$\chi_{2,z}$	0.10 (0.2σ)	0.16 (0.4σ)	0.19 (0.6σ)	$\sim$400
D_L	12.7 Mpc (0.1σ)	3.4 Mpc (0.2σ)	2.2 Mpc (0.3σ)	$\sim$700*
z	1.9×10^{-3} (0.1σ)	6.1×10^{-4} (0.2σ)	4.4×10^{-3} (0.3σ)	$\sim$700*

In Table I, we show the behavior of the median magnitude of Cutler-Vallisneri biases for our population of BBHs for subpopulations with various SNR thresholds, with predictions for the threshold above which the median biases would overcome statistical uncertainties. For each parameter in the table, we present the median absolute values of Cutler-Vallisneri biases, $\overline{|\Delta\theta|}$, for three sets of SNR cutoffs. The SNR $\geq$ 100 threshold is the fiducial value adopted for Fig. 2.

The absolute parameter biases computed via the Cutler-Vallisneri formalism are independent of SNR, and so these values are largely constant across SNR thresholds. In parentheses, we indicate how significant the biases are, normalized by the statistical uncertainty in each of the parameters ($\overline{|\Delta\theta/\sigma_\theta|}$). Since σ_θ scales inversely with SNR, we find that the significance of the bias increases almost linearly with the SNR cutoff. The final column of Table I lists the SNR threshold, $\mathrm{SNR}_{1\sigma}$, above which the median bias is expected to exceed 1σ. Due to lack of sufficient statistics at higher SNR thresholds, we do not always find subpopulations with $\overline{|\Delta\theta/\sigma_\theta|} \geq 1$ for the listed parameters. In those instances, we compute $\mathrm{SNR}_{1\sigma}$ by extrapolating from subpopulations with lower SNR cutoffs. Based on our data, we find that parameter biases due to waveform errors could start becoming significant around SNR $\geq$ 400 for a typical binary from the specified population. In other words, we expect half of the detected population with SNR $\geq$ 400 to have at least one parameter that is significantly affected by waveform modeling errors.

B. Estimating future accuracy requirements

Waveform models with better calibration to NR, and subsequently smaller mutual mismatch across the parameter space, will be crucial to making optimal use of XG GW detections. A natural question to ask is: how much better do such waveform models need to be? One approach could be

to set a benchmark for the performance of these future waveforms, say, a waveform calibration effort that results in 99% of the expected detections to be biased by $\leq 1\sigma$ for a few chosen parameters. We can use the formalism described in Sec. II D to quantify the average mismatch for which this condition will be met.

In Fig. 4, we show the distribution of systematic biases between `IMRPhenomXAS` and various sets of interpolated waveforms with selected values of the interpolation parameter λ. The precise percentage by which each parameter is biased above the acceptable threshold depends on the specifics of the interpolated waveform, and is merely demonstrative. The general trend, however, is robust: improving the agreement between two waveforms models lowers the overall systematic bias between them. Note that biases in D_L have been omitted from the following discussion, since they are particularly sensitive to the interpolated waveforms and can be affected in unphysical ways. However, as we showed in the previous section, biases in the intrinsic parameters are typically larger than biases in D_L. As such, consideration of the intrinsic parameters should be sufficient for determining the strongest constraints on the waveform accuracy threshold. See Appendix D for an illustration of how the intrinsic parameters vary with λ for a single binary.

Performing this analysis for a finer grid of interpolated waveforms, we can develop a picture of how biased each parameter is for a range of average waveform mismatches. This result is shown in the left panel of Fig. 5. As the average mismatch between waveform models decreases, the percentage of events with SNR $\geq$ 100 and bias $\geq 1\sigma$ decreases for every parameter. In the right panel of Fig. 5, we show how this mismatch requirement varies over a range of SNR cutoffs, and not just the fiducial SNR $\geq$ 100 threshold. As one considers lower SNR binaries, the average mismatch requirement for keeping 99% of the binaries unbiased becomes less stringent across all parameters. On the other
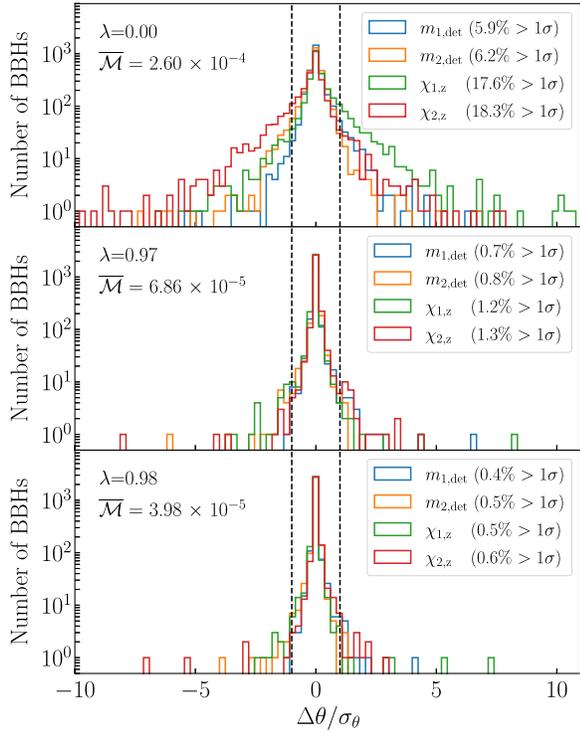
FIG. 4. Significance of waveform calibration biases for the SNR ≥ 100 subpopulation, between interpolated waveforms with varying levels of simulated accuracy to `IMRPhenomXAS`. The value of λ used to generate the approximate waveforms and the average mismatch $\overline{\mathcal{M}}$ across the chosen subpopulation is shown in the top left corner of each panel. The dashed black lines mark the 1σ bias threshold in either direction, and the percentage of binaries with bias $> 1\sigma$ is given in the legend for each parameter.

hand, the average mismatch required is much lower for subpopulations with higher SNRs.

Based on the SNR ≥ 100 results, the average mismatch between the two given waveform models would need to be $\sim 6 \times 10^{-5}$ for the intrinsic parameters of 99% of the

population to remain unbiased. The mismatch requirement can be as strict as 1×10^{-5} if we consider subpopulations with larger SNRs. Recall that the default average mismatch for the SNR ≥ 100 binaries was 2.6×10^{-4}. Therefore, the required mismatch is lower by nearly an order of magnitude.

This statement is sensitive to several assumptions, including the choice of population, the waveform models in question, and even the parameters for which the bias should be minimized. Also, we stress that this requirement applies to the *average* mismatch across our population, and not to individual binaries. Many low-SNR events in specific regions of the parameter space may require less demanding waveform accuracy, while louder binaries may require even better waveform calibration. However, the approach adopted here provides a reasonable and conservative order-of-magnitude estimate for the mismatch requirement across an expected population.

V. COMPARISON WITH FULL PARAMETER ESTIMATION

We now validate our Fisher results and Cutler-Vallisneri estimates by comparing them with full Bayesian parameter estimation runs using `BILBY` [43,44]. To mimic the biases from imperfect waveform modeling, we inject a GW signal with the `IMRPhenomXAS` model and recover it with the `IMRPhenomD` model.

In Fig. 6 we show a representative case belonging to the subpopulation of binaries with SNR ≥ 100. This particular BBH has a network SNR of ~ 140. We compare the parameter estimation results against Gaussian posteriors centered at the biased value $\theta_{i,\text{biased}} = \theta_i + \Delta\theta_i$, where θ_i are the injected parameters, $\Delta\theta_i$ is the Cutler-Vallisneri bias, and the covariance matrix is the inverse of the Fisher matrix. Solid (dashed) black lines refer to the case in which we do (or do not) impose Gaussian priors on the spin components following the simple recipe of Eq. (6).
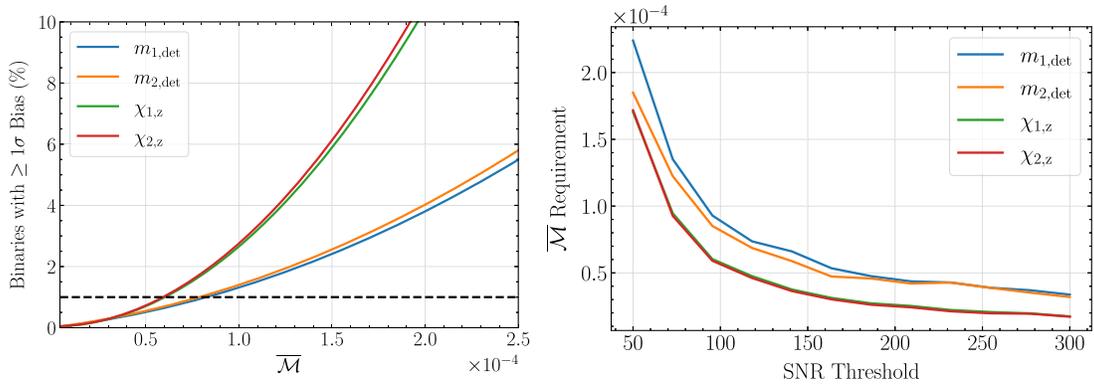
FIG. 5. Mismatch requirements to have unbiased parameters, computed using the interpolated waveform formalism. Left panel: percentage of binaries that are biased by more than 1σ as a function of the average mismatch, for the SNR ≥ 100 subpopulation. The dashed black line marks the threshold at which 1% of events are biased by at least 1σ, or equivalently, 99% of the binaries are biased by at most 1σ. Right panel: average mismatch requirement for at most 1% of the binaries to be biased by $\geq 1\sigma$, for various SNR thresholds.
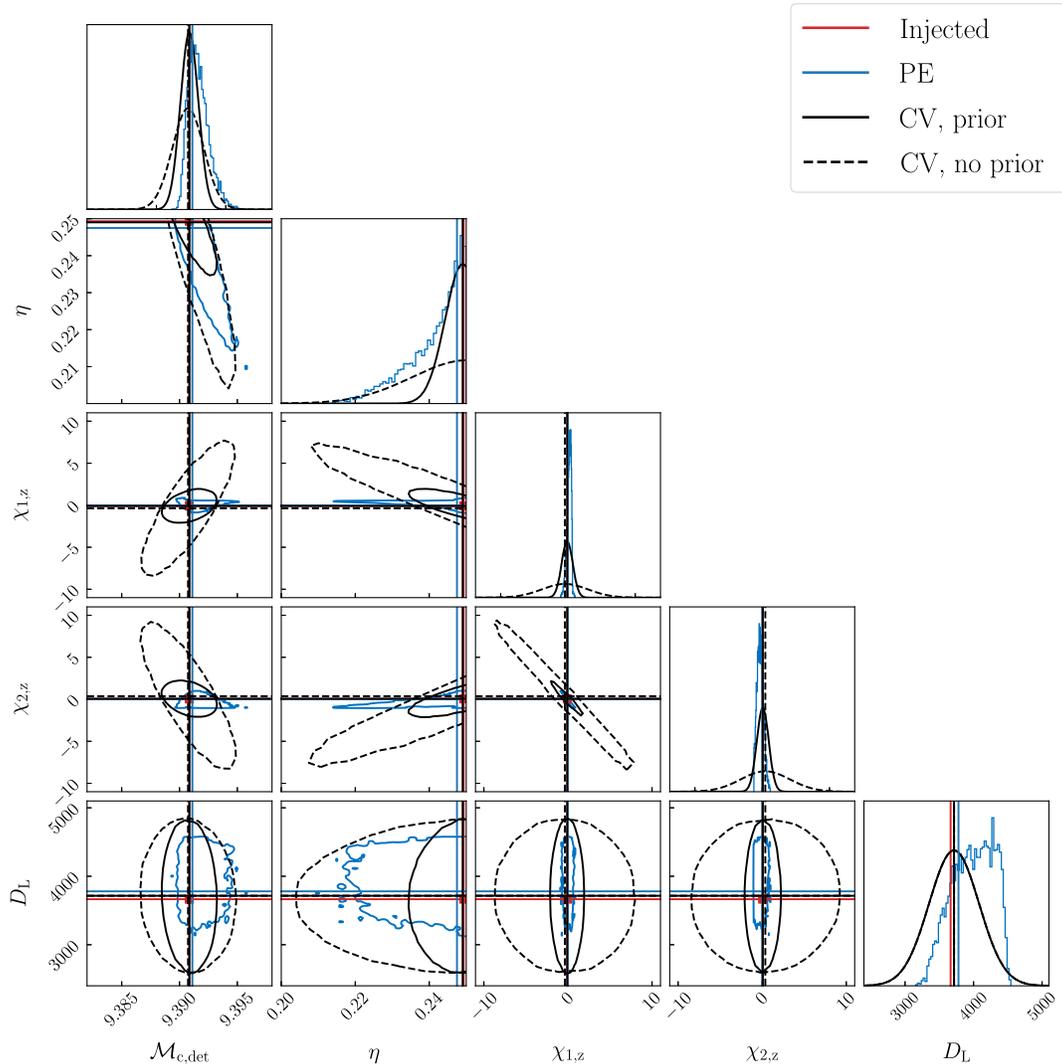
FIG. 6.   Posterior distributions for a representative BBH in our population. The blue curves results from a full Bayesian parameter estimation run. The black curves are estimated through the Cutler-Vallisneri formalism, with (solid lines) and without (dashed lines) imposing a Gaussian priors on the spin components. The 2D contours correspond to 90% confidence levels. The red lines correspond to the injected values. This particular BBH has the following true parameters: $\mathcal{M}_{c,\mathrm{det}} = 9.391 M_\odot, \eta = 0.249, \chi_{1,z} = -0.077, \chi_{2,z} = 0.047,$ $D_\mathrm{L} = 3665.092$ Mpc, $\iota = 0.631,$ RA $= 0.849,$ DEC $= 0.728, \psi = 5.696.$

We find that the biased estimates from Cutler-Vallisneri, $\theta_{i,\mathrm{biased}}$, for all parameters relevant to this study are close to the best-estimate parameters recovered with BILBY, whether we do or do not impose priors. The Fisher errors $\sigma_{i,\mathrm{stat}}$, instead, tend to be larger than the corresponding intervals in the 1D marginalized posteriors. For $\mathcal{M}_{c,\mathrm{det}}$, η, and D_L (or equivalently $m_{1,\mathrm{det}}$, $m_{2,\mathrm{det}}$, and z) the discrepancy between Fisher errors and 1σ intervals in the marginalized posteriors is within a factor of $\lesssim 2$. If we do not impose a prior on the spins, the Fisher errors on $\chi_{1,z}$ and $\chi_{2,z}$ (dashed black contours) are larger by more than an order of magnitude than the corresponding confidence intervals in the 1D marginalized posteriors from BILBY. When we impose a Gaussian prior (solid black contours), the disagreement decreases to a factor of $\lesssim 3$ for both spin components.

However, due to parameter degeneracies, imposing a prior on the spins also affects $\mathcal{M}_{c,\mathrm{det}}$ and η. For $\mathcal{M}_{c,\mathrm{det}}$, the Fisher error goes from being $\sim 50\%$ larger than the 1σ interval in the marginalized posteriors to being $\sim 15\%$ smaller. For η, the Fisher error goes from ~ 2 times larger to $\sim 40\%$ smaller.

The discrepancy in the errors might lead us to underestimate the statistical significance of the biases with the Cutler-Vallisneri formalism. To illustrate how this may happen, in Fig. 7 we compare Fisher estimates and parameter estimation results for the "pessimistic" case of a BBH for which the recovered parameters are significantly biased. The corresponding signal has a network SNR of ~ 350. For this BBH, the Cutler-Vallisneri biases and the Fisher matrix estimates of the errors on $\mathcal{M}_{c,\mathrm{det}}$, η, and D_L are in reasonable agreement with the best-estimate parameters
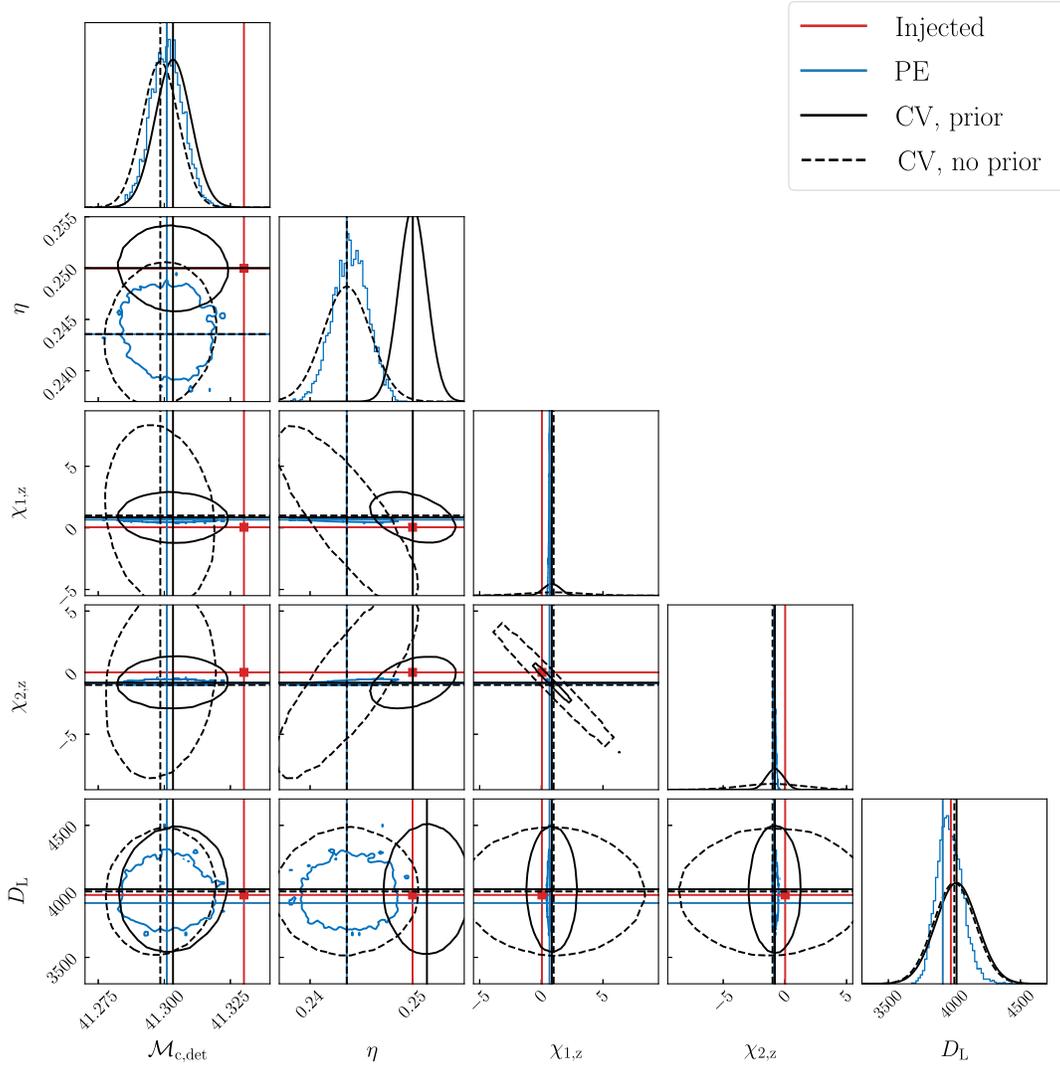
FIG. 7.   Same as Fig. 6, but now for a pessimistic case where the intrinsic parameters recovered from parameter estimation are strongly biased. The BBH has the following true parameters: $\mathcal{M}_{c,\text{det}} = 41.330 M_\odot, \eta = 0.250, \chi_{1,z} = 0.050, \chi_{2,z} = 0.024, D_\text{L} = 3972.875$ Mpc, $\iota = 0.733$, RA $= 2.680$, DEC $= -0.516$, $\psi = 0.868$.

from full parameter estimation without imposing any Gaussian prior. The Fisher errors on these parameters agree within $\lesssim 50\%$ with the 1σ intervals in the 1D marginalized posteriors. However, the Fisher matrix estimates of the errors on $\chi_{1,z}$ and $\chi_{2,z}$ are larger than the parameter estimation results by a factor ~ 40 without any prior, while the Cutler-Vallisneri biases would far exceed the physical range, thus we force them to take the boundary values $\theta_{\chi_{i,z},\text{biased}} \in [-1, 1]$. This is a clear example of why the spin priors are necessary. After imposing the prior, the Cutler-Vallisneri biases agree with the best estimate from parameter estimation within $\sim 30\%$ for both spin components, and the 1σ errors are now larger by a factor of ~ 10 instead of ~ 40. Still, the spin components recovered with full parameter estimation are both biased at $\sim 10\sigma$ with respect to the injected values, while they are barely biased (at $\sim 1\sigma$) when we use the Cutler-Vallisneri formalism.

Moreover, while the 1σ Fisher errors on the marginalized distributions for $\mathcal{M}_{c,\text{det}}$, η and D_L are only slightly affected by the imposition of a spin prior, the Cutler-Vallisneri bias on η changes significantly, now pushing this parameter to hit the physical boundary of 0.25.

In general, we find that the combination of Fisher and Cutler-Vallisneri formalism might perform poorly for bounded parameters when the estimated uncertainty exceeds the prior bound, as in Fig. 7. A more systematic comparison of the Cutler-Vallisneri formalism with full parameter estimation in the whole parameter space is clearly necessary, but it is beyond the scope of this exploratory study.

VI. DISCUSSION

Our estimates suggest that, in order for 99% of the intrinsic parameters of an astrophysical BBH population

with SNR $\geq$ 100 to remain unbiased by waveform errors, an order of magnitude improvement in the average mismatch is required. Our estimate is meant to be a lower bound, especially when considered alongside other work where the authors study individual binaries with SNRs of the order $\mathcal{O}(1000)$ [12,17]. We stress that the mismatch and waveform calibration requirements will ultimately depend on the intended application. Over a population of binaries, many systems will have insignificant biases in intrinsic parameters due to either lower SNRs or the existing calibration accuracy of specific waveform models. For the loudest binaries observed in XG detectors, the biases will be extremely significant, which may be particularly relevant for applications such as dark siren cosmology or tests of general relativity.

One relevant extension of this work would be to investigate how waveform systematics over a population could impact the inference of specific hyperparameters of a given population model. The accuracy requirements for ensuring unbiased population hyperparameters may differ from the results presented here in nontrivial ways.

Additionally, the waveform models used in this study both belong to the IMRPhenom family, and represent only a subset of the models that may be used in a parameter estimation campaign. Models with higher harmonics, spin precession, etc. could produce different mutual biases, and would likely require higher accuracy thresholds to ensure consistent parameter inference. The interpolated waveform formalism presented in this work (or generalizations thereof) could be especially useful in guiding the calibration efforts for these models. By varying waveform accuracy in specific parts of the parameter space, one could deduce where to iteratively add BBH simulations to most efficiently improve waveform accuracy with additional simulations. These applications are left for future studies.

APPENDIX A: WAVEFORM BIASES IN A DETECTOR NETWORK

In this appendix we generalize the calculation of systematic parameter biases due to waveform errors from Ref. [10] to a network of detectors. Let us assume that a GW data stream can be modeled as a combination of some true gravitational waveform and noise, such that

$$d(t) = h_{\mathrm{TR}}(t; \vec{\theta}_{\mathrm{tr}}) + n(t), \quad (A1)$$

where the signal is a function of time t, and $\vec{\theta}_{\mathrm{tr}}$ represents the set of true parameter values that we would like to infer. In practice, this is done using an approximate waveform model h_{AP}, such that

$$h_{\mathrm{TR}}(t; \vec{\theta}_{\mathrm{tr}}) \neq h_{\mathrm{AP}}(t; \vec{\theta}_{\mathrm{tr}}). \quad (A2)$$

Parameter estimation involves minimizing the following log-likelihood with respect to $\vec{\theta}$ [45]:

$$\log p(d|\vec{\theta}) = -\frac{1}{2}\big(d(t) - h_{\mathrm{AP}}(t; \vec{\theta})|d(t) - h_{\mathrm{AP}}(t; \vec{\theta})\big). \quad (A3)$$

In the presence of multiple detectors, the various antenna patterns and sky locations will give rise to detector-dependent waveforms, such that parameter estimation requires minimizing the following log-likelihood instead:

$$\log p(d|\vec{\theta}) = \Sigma_{\mathrm{D}=1}^{N_{\mathrm{D}}} -\frac{1}{2}\big(d_{\mathrm{D}}(t) - h_{\mathrm{AP,D}}(t; \vec{\theta})|d_{\mathrm{D}}(t) - h_{\mathrm{AP,D}}(t; \vec{\theta})\big). \quad (A4)$$

The best-fit parameters $\vec{\theta}_{\mathrm{bf}}$ will thus satisfy

$$\Sigma_{\mathrm{D}=1}^{N_{\mathrm{D}}}\big(\partial_i h_{\mathrm{AP,D}}(\vec{\theta}_{\mathrm{bf}})|d_{\mathrm{D}}(t) - h_{\mathrm{AP,D}}(\vec{\theta}_{\mathrm{bf}})\big) = 0, \quad (A5)$$

where here and below we omit the time dependence for brevity, and the index i represents the components of $\vec{\theta}_{\mathrm{bf}}$. Using Eq. (A1), we can write

$$d_{\mathrm{D}}(t) - h_{\mathrm{AP,D}}(\vec{\theta}_{\mathrm{bf}}) = n_{\mathrm{D}}(t) + h_{\mathrm{TR,D}}(\vec{\theta}_{\mathrm{tr}}) - h_{\mathrm{AP,D}}(\vec{\theta}_{\mathrm{tr}}) + h_{\mathrm{AP,D}}(\vec{\theta}_{\mathrm{tr}}) - h_{\mathrm{AP,D}}(\vec{\theta}_{\mathrm{bf}}), \quad (A6)$$

where we have added and subtracted a term $h_{\mathrm{AP,D}}(\vec{\theta}_{\mathrm{tr}})$ for convenience. We can think of $\vec{\theta}_{\mathrm{bf}}$ as a perturbation around the true parameters, such that

$$\vec{\theta}_{\mathrm{bf}} = \vec{\theta}_{\mathrm{tr}} + \Delta\vec{\theta}. \quad (A7)$$

If $\Delta\vec{\theta}$ is sufficiently small, the linear signal approximation gives us

$$h_{\rm AP,D}(\vec{\theta}_{\rm tr}) - h_{\rm AP,D}(\vec{\theta}_{\rm bf}) \approx \partial_i h_{\rm AP,D}(\vec{\theta}_{\rm bf})\Delta\theta^i. \quad \text{(A8)}$$

Therefore, we can rewrite Eq. (A6) as

$$d_{\rm D}(t) - h_{\rm AP,D}(\vec{\theta}_{\rm bf}) \approx n_{\rm D}(t) + \delta h_{\rm D}(\vec{\theta}_{\rm tr}) - \partial_i h_{\rm AP,D}(\vec{\theta}_{\rm bf})\Delta\theta^i, \quad \text{(A9)}$$

where $\delta h_{\rm D}(\vec{\theta}_{\rm tr}) \equiv h_{\rm TR,D}(\vec{\theta}_{\rm tr}) - h_{\rm AP,D}(\vec{\theta}_{\rm tr})$ is the waveform residual.

Substituting this expansion back into Eq. (A5) and redefining some indices, we obtain

$$\Sigma_{\rm D=1}^{N_{\rm D}}\big(\partial_i h_{\rm AP,D}(\vec{\theta}_{\rm bf})|n_{\rm D}(t) + \delta h_{\rm D}(\vec{\theta}_{\rm tr}) - \partial_j h_{\rm AP,D}(\vec{\theta}_{\rm bf})\Delta\theta^j\big) = 0. \quad \text{(A10)}$$

Expanding the inner product terms, we get

$$\Sigma_{\rm D=1}^{N_{\rm D}}\big(\partial_i h_{\rm AP,D}(\vec{\theta}_{\rm bf})|n_{\rm D}(t)\big) + \Sigma_{\rm D=1}^{N_{\rm D}}\big(\partial_i h_{\rm AP,D}(\vec{\theta}_{\rm bf})|\delta h_{\rm D}(\vec{\theta}_{\rm tr})\big)$$
$$= \Sigma_{\rm D=1}^{N_{\rm D}}\big(\partial_i h_{\rm AP,D}(\vec{\theta}_{\rm bf})|\partial_j h_{\rm AP,D}(\vec{\theta}_{\rm bf})\big)\Delta\theta^j \quad \text{(A11)}$$

The inner product on the right-hand side can be recognized as the network Fisher information matrix of $h_{\rm AP,D}$, such that

$$\Sigma_{\rm D=1}^{N_{\rm D}}\big(\partial_j h_{\rm AP,D}(\vec{\theta}_{\rm bf})|n_{\rm D}(t)\big) + \Sigma_{\rm D=1}^{N_{\rm D}}\big(\partial_j h_{\rm AP,D}(\vec{\theta}_{\rm bf})|\delta h_{\rm D}(\vec{\theta}_{\rm tr})\big)$$
$$= \Gamma_{ij,\rm Net}\Delta\theta^i, \quad \text{(A12)}$$

where we have used the symmetry of the Fisher matrix and switched the i, j indices for convenience. The overall bias on parameter θ^i can now be expressed as

$$\Delta\theta^i = \big(\Gamma_{\rm Net}^{-1}\big)^{ij}(\vec{\theta}_{\rm bf})\Big[\Sigma_{\rm D=1}^{N_{\rm D}}\big(\partial_j h_{\rm AP,D}(\vec{\theta}_{\rm bf})|n_{\rm D}(t)\big)$$
$$+ \Sigma_{\rm D=1}^{N_{\rm D}}\big(\partial_j h_{\rm AP,D}(\vec{\theta}_{\rm bf})|\delta h_{\rm D}(\vec{\theta}_{\rm tr})\big)\Big]. \quad \text{(A13)}$$

Each of the two terms on the right-hand side has a clear interpretation: the first term represents the error contribution from noise, while the second is the systematic bias due to waveform residuals. Isolating the systematic waveform error term, we can write

$$\Delta\theta^i_{\rm sys} \equiv \big(\Gamma_{\rm Net}^{-1}\big)^{ij}(\vec{\theta}_{\rm bf})\Sigma_{\rm D=1}^{N_{\rm D}}\big(\partial_j h_{\rm AP,D}(\vec{\theta}_{\rm bf})|\delta h_{\rm D}(\vec{\theta}_{\rm tr})\big). \quad \text{(A14)}$$

Since in practice we do not have access to $\vec{\theta}_{\rm tr}$, we can instead evaluate the waveform residual at $\vec{\theta}_{\rm bf}$ to a good approximation. This yields the final result:

$$\Delta\theta^i_{\rm sys} = \big(\Gamma_{\rm Net}^{-1}\big)^{ij}(\vec{\theta}_{\rm bf})\Sigma_{\rm D=1}^{N_{\rm D}}\big(\partial_j h_{\rm AP,D}(\vec{\theta}_{\rm bf})|h_{\rm TR,D}(\vec{\theta}_{\rm bf}) - h_{\rm AP,D}(\vec{\theta}_{\rm bf})\big). \quad \text{(A15)}$$

APPENDIX B: IMPACT OF WAVEFORM ALIGNMENT ON LUMINOSITY DISTANCE BIASES

Systematic errors due to different waveforms are of the form

$$\Delta\theta^i = \big(\Gamma_{\rm AP,Net}^{-1}\big)^{ij}\big(\partial_j(h_{\rm AP})|h_{\rm TR} - h_{\rm AP}\big). \quad \text{(B1)}$$

In the frequency domain, the overall amplitudes of $h_{\rm TR}$ and $h_{\rm AP}$ are independent of systematic shifts in phase and time. However, differences in the coalescence time and phase affect the relative phases of the two waveforms. Since the Cutler-Vallisneri bias formula depends on the difference $(h_{\rm TR} - h_{\rm AP})$, the biases are also affected by the relative phase between the two waveform models.

In Fig. 8 we illustrate the effect of changing one of these parameters, $\phi_{\rm c}$. As shown in the bottom panel, the difference between the two waveforms can depend strongly on their relative phase difference. Biases in luminosity distance have the primary effect of changing the relative amplitude of the template waveform, which can cancel out this effect. Therefore, biases in luminosity distance from the Cutler-Vallisneri formalism can be unphysically large if the waveforms are not aligned first.

As such, the alignment procedure alluded to in Sec. III has a dramatic impact on the bias in $D_{\rm L}$.
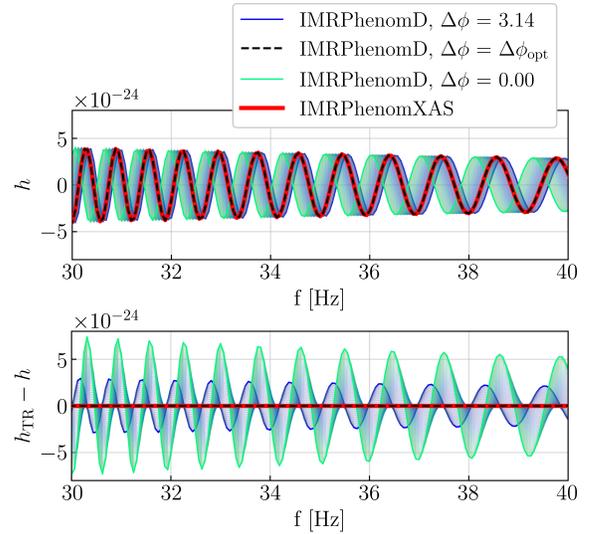
FIG. 8. Top: GW strain amplitude for a chosen binary when using various waveform models. The reference waveform generated using `IMRPhenomXAS` is shown in red. We also show different versions of an `IMRPhenomD` waveform with the same injected parameters and an additional phase shift in the range $[0, \pi]$ (curves ranging from green to blue). The waveform which minimizes the mismatch with respect to the reference waveform is plotted as a dashed black curve. Bottom: the same set of waveforms, with the y-axis now showing the difference between the given waveform and the `IMRPhenomXAS` reference waveform.
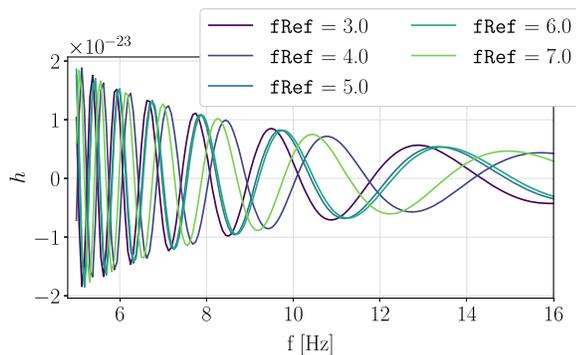
FIG. 9. Detector response of a given BBH from our population, for various choices of `fRef`. The detector is set to the 40 km CE in Idaho, and the BBH has the following parameters: $\mathcal{M}_{c,\mathrm{det}} = 121.181 M_\odot$, $\eta = 0.249$, $\chi_{1,z} = 0.026$, $\chi_{2,z} = 0.036$, $D_L = 16215.376$ Mpc, $\iota = 2.709$, RA $= 1.511$, DEC $= 0.188$, $\psi = 2.049$.

APPENDIX C: CHOICE OF REFERENCE FREQUENCY

The waveforms for this study are generated using the `LALSuite` library [46,47], which requires a parameter "`fRef`" to set the reference frequency at which the phase and orientation of a binary are defined. For this study we set `fRef` to 5 Hz, which is the minimum frequency chosen for the detector network. In Fig. 9 we show the effect of varying `fRef` while keeping all other parameters constant, for a specific binary.

Evidently, the choice of `fRef` can affect the behavior of a frequency domain waveform in much the same way as an additional phase factor, and in this sense it has a similar effect as modifying the coalescence phase parameter ϕ_c.

Although the waveform alignment process alluded to in Sec. III should generally account for variations in the waveform phase, there may still be some minor effects on the Cutler-Vallisneri bias as a result of different choices of `fRef`. The tests shown in Fig. 10 demonstrate that the

choice of `fRef` has a minor impact on the parameter biases. Using the biases at `fRef=5` Hz as a reference, changing the reference frequency only changes the normalized biases by a few parts in a hundred.

APPENDIX D: BIASES USING INTERPOLATED WAVEFORMS

Using the interpolation scheme described in Sec. II D, we synthetically produce waveforms that vary monotonically in mismatch between `IMRPhenomD` and `IMRPhenomXAS`. In Fig. 11 we illustrate the modulation of the statistical and systematic uncertainties from the Fisher and Cutler-Vallisneri formalism, by focusing on a single test binary. As expected, the magnitude of the biases as well as the statistical uncertainties of the intrinsic parameters can be seen changing from the value in `IMRPhenomD` ($\lambda = 0$) to the value in `IMRPhenomXAS` ($\lambda = 1$). This example shows that the interpolated waveform formalism can be trusted to produce qualitatively sensible results for the intrinsic parameters considered in Sec. IV B.

FIG. 10. Average deviations in parameter biases over a subset of 100 binaries with SNR ≥ 100 from our population, for various choices of `fRef`. The y-axis shows the deviation of the Cutler-Vallisneri biases with respect to their respective values at `fRef` = 5 Hz, normalized to the statistical errors for each binary. The dashed black line marks the location of `fRef` = 5 Hz.
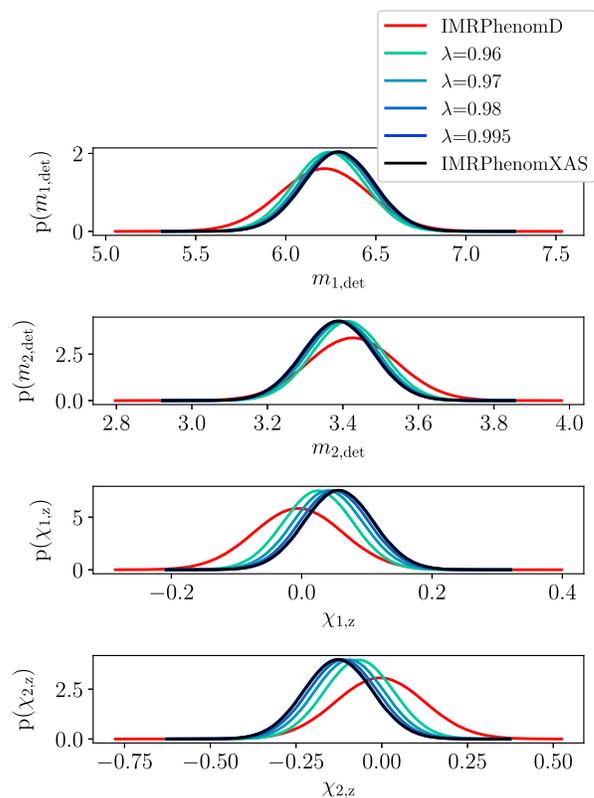
FIG. 11. Behavior of statistical and systematic errors for a single BBH event, for various interpolated waveforms defined by the parameter λ. The chosen BBH has the following parameters: $\mathcal{M}_{c,\mathrm{det}} = 3.982 M_\odot$, $\eta = 0.227$, $\chi_{1,z} = 0.056$, $\chi_{2,z} = -0.126$, $D_L = 131.168$ Mpc, $\iota = 2.162$, RA $= 5.575$, DEC $= -0.654$, $\psi = 2.663$.

[1] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), Phys. Rev. X **13**, 041039 (2023).

[2] M. Punturo *et al.*, Classical Quantum Gravity **27**, 194002 (2010).

[3] D. Reitze *et al.*, Bull. Am. Astron. Soc. **51**, 035 (2019).

[4] M. Maggiore *et al.*, J. Cosmol. Astropart. Phys. 03 (2020) 050.

[5] M. Evans *et al.*, arXiv:2109.09882.

[6] M. Branchesi *et al.*, J. Cosmol. Astropart. Phys. 07 (2023) 068.

[7] M. Evans *et al.*, arXiv:2306.13745.

[8] I. Gupta *et al.*, arXiv:2307.10421.

[9] A. Corsi *et al.*, arXiv:2402.13445.

[10] C. Cutler and M. Vallisneri, Phys. Rev. D **76**, 104018 (2007).

[11] H. Estellés *et al.*, Astrophys. J. **924**, 79 (2022).

[12] Q. Hu and J. Veitch, Phys. Rev. D **106**, 044042 (2022).

[13] B. P. Abbott *et al.* (LIGO Scientific Collaboration), Classical Quantum Gravity **34**, 044001 (2017).

[14] S. Borhanian and B. S. Sathyaprakash, arXiv:2202.11048.

[15] C. B. Owen, C.-J. Haster, S. Perkins, N. J. Cornish, and N. Yunes, Phys. Rev. D **108**, 044018 (2023).

[16] J. S. Read, Classical Quantum Gravity **40**, 135002 (2023).

[17] M. Pürrer and C.-J. Haster, Phys. Rev. Res. **2**, 023151 (2020).

[18] E. Berti, Classical Quantum Gravity **23**, S785 (2006).

[19] C. Pitte, Q. Baghi, S. Marsat, M. Besançon, and A. Petiteau, Phys. Rev. D **108**, 044053 (2023).

[20] C. J. Moore, E. Finch, R. Buscicchio, and D. Gerosa, iScience **24**, 102577 (2021).

[21] P. Saini, M. Favata, and K. G. Arun, Phys. Rev. D **106**, 084031 (2022).

[22] P. Saini, S. A. Bhat, M. Favata, and K. G. Arun, arXiv:2311.08033.

[23] Q. Hu and J. Veitch, Astrophys. J. **945**, 103 (2023).

[24] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. Jiménez Forteza, and A. Bohé, Phys. Rev. D **93**, 044006 (2016).

[25] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez Forteza, and A. Bohé, Phys. Rev. D **93**, 044007 (2016).

[26] G. Pratten, S. Husa, C. Garcia-Quiros, M. Colleoni, A. Ramos-Buades, H. Estelles, and R. Jaume, Phys. Rev. D **102**, 064001 (2020).

[27] M. Vallisneri, Phys. Rev. D **77**, 042001 (2008).

[28] S. Borhanian, Classical Quantum Gravity **38**, 175014 (2021).

[29] C. Cutler and E. E. Flanagan, Phys. Rev. D **49**, 2658 (1994).

[30] E. Poisson and C. M. Will, Phys. Rev. D **52**, 848 (1995).

[31] E. Berti, A. Buonanno, and C. M. Will, Phys. Rev. D **71**, 084025 (2005).

[32] E. E. Flanagan and S. A. Hughes, Phys. Rev. D **57**, 4566 (1998).

[33] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), arXiv:2111.03634.

[34] A. Dhani, S. H. Völkel, A. Buonanno, H. Estelles, J. Gair, H. P. Pfeiffer, L. Pompili, and A. Toubiana, arXiv:2404.05811.

[35] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, Phys. Rev. D **85**, 122006 (2012).

[36] A. Buonanno, B. Iyer, E. Ochsner, Y. Pan, and B. S. Sathyaprakash, Phys. Rev. D **80**, 084043 (2009).

[37] P. Madau and T. Fragos, Astrophys. J. **840**, 39 (2017).

[38] N. Aghanim *et al.* (Planck Collaboration), Astron. Astrophys. **641**, A6 (2020); **652**, C4(E) (2021).

[39] N. Muttoni, D. Laghi, N. Tamanini, S. Marsat, and D. Izquierdo-Villalba, Phys. Rev. D **108**, 043543 (2023).

[40] M. Pieroni, A. Ricciardone, and E. Barausse, Sci. Rep. **12**, 17940 (2022).

[41] F. Iacovelli, M. Mancarella, S. Foffa, and M. Maggiore, Astrophys. J. **941**, 208 (2022).

[42] R. Abbott *et al.* (KAGRA, Virgo, and LIGO Scientific Collaborations), Phys. Rev. X **13**, 011048 (2023).

[43] G. Ashton *et al.*, Astrophys. J. Suppl. Ser. **241**, 27 (2019).

[44] K. Krishna, A. Vijaykumar, A. Ganguly, C. Talbot, S. Biscoveanu, R. N. George, N. Williams, and A. Zimmerman, arXiv:2312.06009.

[45] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Classical Quantum Gravity **37**, 055002 (2020).

[46] LIGO Scientific, Virgo, and KAGRA Collaborations, LVK Algorithm Library—LALSuite, Free software (GPL) (2018).

[47] K. Wette, SoftwareX **12**, 100634 (2020).