Systematic Biases in Estimating the Properties of Black Holes Due to Inaccurate Gravitational-Wave Models

Arnab Dhani,[1,*] Sebastian H. Völkel,[1] Alessandra Buonanno,[1,2] Hector Estelles,[1] Jonathan Gair,[1] Harald P. Pfeiffer,[1] Lorenzo Pompili,[1] and Alexandre Toubiana[1]

[1]*Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Am Mühlenberg 1, Potsdam 14476, Germany*

[2]*Department of Physics, University of Maryland, College Park, Maryland 20742, USA*

(Received 6 May 2024; accepted 1 July 2025; published 8 August 2025)

Gravitational-wave (GW) observations of binary black-hole (BBH) coalescences are expected to address outstanding questions in astrophysics, cosmology, and fundamental physics. Inference of BBH parameters relies on waveform models, and realizing the full discovery potential of upcoming LIGO-Virgo-KAGRA observing runs and new ground-based facilities (such as the Einstein Telescope and Cosmic Explorer) hinges on the accuracy of these waveform models. Using linear-signal approximation methods and Bayesian analysis, we start to assess our readiness for what lies ahead using two state-of-the-art quasicircular, spin-precessing models: SEOBNRv5PHM and IMRPhenomXPHM. We find that systematic biases increase with the spin of the BH, with parameter biases being approximately 6 to 8 times likelier, if the primary-spin magnitude exceeds 0.5 compared to when it is less than 0.5. Additionally, we ascertain that current waveforms can accurately recover the distribution of masses in the LVK astrophysical population but not spins. Upon exploring the broader parameter space of BHs, we find that systematic biases increase with detector-frame total mass, binary asymmetry, and spin precession, with a majority of such binaries incurring parameter biases, extending up to redshifts around 3 in future detectors. Furthermore, we examine three "golden" events characterized by mass ratios of approximately 6 to 10, significant spin magnitudes ($0.6-0.9$), and high precession, evaluating how systematic biases may affect their scientific outcomes. Our findings reveal that current waveforms fail to enable the unbiased measurement of the Hubble-Lemaître parameter and sky localization from loud signals, even for current detectors. Moreover, highly asymmetric systems within the lower BH mass gap exhibit biased measurements of the secondary-companion mass, which impacts the physics of both neutron stars and formation channels. Similarly, we deduce that the primary mass of massive binaries ($> 60M_{\odot}$) will also be biased, affecting supernova physics. Future progress in analytical calculations and numerical-relativity simulations, crucial for calibrating the models, must target regions of the parameter space with significant biases to develop more accurate models. Only then can precision GW astronomy fulfill the promise it holds.

DOI: 10.1103/5pks-qz6b

Subject Areas: Astrophysics, Cosmology, Gravitation

I. INTRODUCTION

Almost a decade ago, the first observation of a gravitational wave (GW) from the coalescence of two black holes (BHs) marked an important milestone in the history of GW astronomy [1]. Since then, the LIGO-Virgo-KAGRA (LVK) Collaboration [2–4] has detected more than 90 compact binary mergers [5–7], and independent research groups [8–13] have discovered additional events. Thus, GWs have become a novel tool to explore the Universe. The observed signals have been used to measure the mass and spin distributions of BHs and neutron stars (NSs), their formation channels, and the coevolution of their properties with that of the Universe [14,15]. Binary neutron star (BNS) mergers have improved the bounds on the nuclear equation of state and the maximum allowed mass of a NS [16–18]. The mass distributions have been employed to constrain the observed lower and predicted upper mass gaps and other features in the mass spectrum. In conjunction with the electromagnetic (EM) counterparts observed for GW170817, or together with available galaxy catalogs, they have also been used to constrain

*Contact author: arnab.dhani@aei.mpg.de

the Hubble-Lemaître parameter (H_0) [19,20]. In addition, GW measurements have probed general relativity (GR) as the fundamental theory of gravity [21–23].

Improvements in the sensitivity of current GW detectors and proposed next-generation (XG) ground-based observatories like the Einstein Telescope (ET) and Cosmic Explorer (CE) [24–26] will significantly increase the observational volume and, with it, the number of GW sources. For instance, a network of XG detectors will observe *every* stellar-origin BBH merger and most BNS mergers across the observable Universe [27]. A number of studies have explored in detail the extent to which various science objectives can be accomplished [27–32]. With $\mathcal{O}(100)$ observations by the LVK Collaboration, and $\mathcal{O}(10^5)$ promised detections with XG detectors, meaningful inferences on the properties of the astrophysical distribution of BBHs will constrain more and more the underlying distribution of main sequence stars and their evolution. EM observations in our Galaxy indicate that stellar-origin BHs have masses above $5M_\odot$. However, these observations may be biased by properties that are unique to our Galaxy. Similarly, the pair-instability supernova (PISN) process is expected to suppress BH formation in the mass range of around 50–120$M_\odot$ [33,34]. Confident detections of BBHs in these mass ranges would pose challenges to stellar-evolution models, as well as constrain the $^{12}\mathrm{C}(\alpha,\gamma)^{16}\mathrm{O}$ reaction rate that drives the PISN process [35]. Gravitational-wave astronomy can also determine the cosmological evolution of the Universe. In particular, by combining many GW signals, it will contribute to resolving the H_0 tension and provide new constraints on structure formation. On the other hand, loud individual events carry a lot of information too. Individual golden BBHs can also resolve the Hubble-Lemaître tension [36]. Finally, precision tests of GR can be derived from high SNR observations [37–43]. All of these scientific objectives are vulnerable to false positives arising from waveform inaccuracies.

The source properties are estimated from the GWs via Bayesian inference using waveform models predicted by GR. Since there is no complete, closed-form analytic solution for the gravitational waveform of a compact-binary coalescence (CBC), various approximate and numerical methods have been developed to describe the GW signal faithfully. The effective-one-body (EOB) waveforms [44–58] combine and resum several perturbative results, such as post-Newtonian (PN), post-Minkowskian (PM), and gravitational self-force information for the conservative and dissipative dynamics, with physically motivated *Ansätze* for the merger and BH perturbation theory for the ringdown. They are made highly accurate through calibration to numerical relativity (NR) simulations [59–61]. The inspiral-merger-ringdown (IMR) phenomenological (IMRPhenom) models [62–67] are constructed in two steps. First, one stitches together in a time domain an inspiral EOB waveform with an IMR NR waveform and Fourier transforms it. Then, one fits the latter to a frequency-domain closed-form expression based on the PN stationary-phase approximation for the inspiral and plunge, and a physically motivated phenomenological *Ansatz* for the merger and ringdown. Where NR simulations are not available, EOB waveforms are used to calibrate the model. NR simulations give the most accurate representation of a GW signal, although they are still limited by numerical truncation errors [68–70], imperfect outer boundary conditions [71–73], and issues with GW extraction and extrapolation [74,75]. Moreover, NR simulations are not available in the entire parameter space and are limited in length due to their high computational cost. NR surrogate models (NRSur) [76–79] are constructed by directly interpolating NR waveforms, where available.

Thanks to advancement in GW modeling since the discovery of GW150914 [80], waveform models have been sufficiently accurate to analyze most signals in the LVK GW Transient Catalogs (GWTC) [80,81]. In Ref. [82], the authors used the absolute value of the difference between waveform models to quantify the accuracy of a given pair of models, finding that a few high signal-to-noise-ratio (SNR) events in the GWTC-3 and GWTC-2.1 fail their criterion. They also find that parameter estimation of such events shows greater inconsistencies. A reanalysis of the GWTC-3 by Ref. [83] finds that the NRSur7dq4 model recovers noticeably different parameters compared to LVK analyses using IMRPhenomXPHM and SEOBNRv4PHM waveform models for around 20% of the events where the NRSur7dq4 model can be used [84]. A hypermodel approach to identify waveform systematics has also been carried out on the 13 heaviest GW events from the GWTC-3. In this approach, waveform models are treated as parameters and directly sampled over, yielding a direct probability for each waveform model. The authors do not find any waveform model to be preferred except for three events that are marred by data quality issues [85]. Recently, there have also been efforts to marginalize over waveform modeling uncertainties [86,87]. Other studies have found that even relatively low SNR events could be affected by systematic biases if they lie in a region of the parameter space where calibration with NR is sparse, which would include binaries that are asymmetric or eccentric and/or have large spin magnitudes and precessing orbits [51,82].

With increasing detector sensitivity and number of detections, the median SNR of the observed population of binaries, as well as the likelihood of detecting a binary from a region of the parameter space where waveform inaccuracies are greater, will increase. While statistical uncertainties decrease with increasing SNR, systematic biases are independent of the signal power. Several studies have explored the validity of waveform models for the parameter estimation of quasicircular binary black-hole (BBH) mergers in upgraded and XG detectors, mainly

focusing on the biases for individual events [81,82,88,89], with Ref. [81] also showing the inferred distribution of the primary mass to be biased. While the negligence of subdominant modes can significantly bias the parameter estimation of individual events [90,91], a recent study indicated that such biases do not affect the inference of the LVK-like astrophysical distribution of BBHs [92]. Other studies have focused on waveform systematics in the presence of eccentricity [93,94], matter effects, and spin precession [95–99]. Recent studies have also explored the effect of truncation errors in NR simulations employing finite-differencing methods and concluded that current simulations are not accurate enough for highly asymmetric binaries and binaries whose orbits are inclined with respect to the line of sight [100,101]. However, state-of-the-art waveform models, such as SEOBNRv5PHM [50–53] and IMRPhenomXPHM [65,102,103], are calibrated to the Simulating-eXtreme-Spacetimes (SXS) Collaboration waveforms, which employ spectral methods, and the effect of truncation errors on these waveforms has not been explored systematically. An indistinguishability criterion [104] has also been used as an easy-to-compute metric to determine the accuracy requirements of waveforms [81,105]. However, this measure has been found to be very conservative. Reference [106] proposed a correction to improve the reliability of the measure.

We illustrate the effect of waveform mismodeling in Fig. 1, using a BBH with parameters given in Table I. [107] We show the multipolar, spin-precessing GW strains in the LIGO-Livingston detector from the SEOBNRv5PHM waveform model as the signal (black curve) and the IMRPhenomXPHM waveform model as the template (brown and green curves). For the green curve, we fix the polarization angle and time at coalescence, by maximizing its overlap against the SEOBNRv5PHM signal. We employ the LIGO-Virgo detectors assuming the sensitivity of the upcoming fifth observing (*O5*) run. If the green GW strain faithfully represented the signal (black), they would perfectly match throughout the coalescence. However, this is not the case; the amplitude modulations are different during the long inspiral and, in particular, during the late inspiral, merger, and ringdown. Furthermore, while the signal and the template phases match during the early inspiral, there is significant dephasing near the late inspiral, merger, and ringdown. In Fig. 1, we also show the IMRPhenomXPHM template (brown curve) evaluated at the maximum-likelihood parameters (obtained through a Bayesian analysis). It has a much better match to the SEOBNRv5PHM signal even during the late inspiral, merger, and ringdown. This best match is obtained at the expense of introducing a bias in the parameters; notably, the total mass, mass ratio, and the spin-precession parameter are biased by about 3%, 6%, and 13%, respectively. The brown curve also has an associated brown band representing the measurement errors at a 90% credible interval in the GW parameters, but it is barely visible to the naked eye, illustrating that this uncertainty, which represents the estimated statistical uncertainty from instrumental noise, is much smaller than the waveform difference between the signal and the template evaluated at the best-fit parameters. As previously stated, this inconsistency manifests itself as biased parameter estimation, which could affect the various science objectives.

In this work, we start to quantify the systematic biases that can be expected in future observing runs with current facilities and XG detectors using the SEOBNRv5PHM and IMRPhenomXPHM waveform models, which are employed for parameter-estimation studies of BBHs by the LVK Collaboration. Both models are valid for quasicircular binaries and incorporate subdominant spherical harmonics and spin-precession effects. While it would be ideal to quantify the biases of each of these models against the true GR signal, it is infeasible to do so everywhere in the parameter space since NR waveforms are not available.

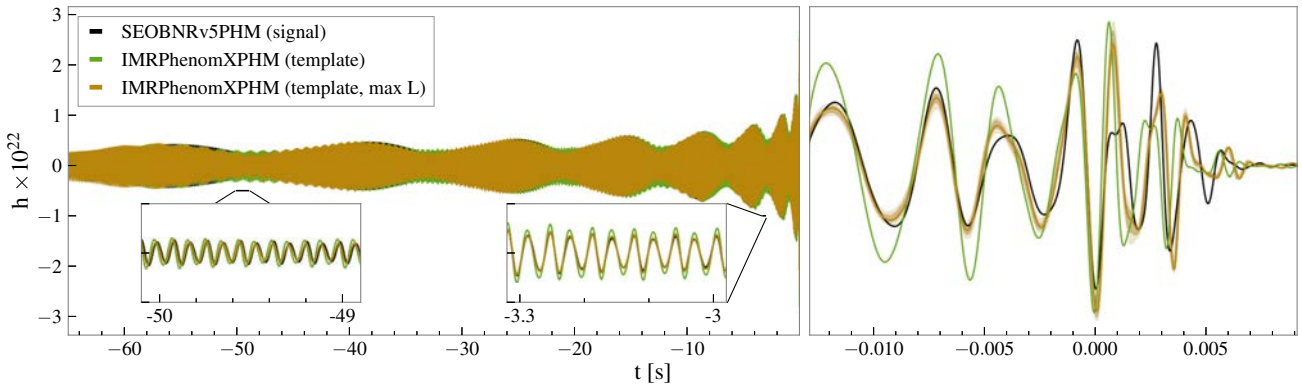

FIG. 1. GW strains for a BBH system with parameters given in Table I (binary 1) at the LIGO-Livingston detector of the *O5* network. The black curve is the injected signal SEOBNRv5PHM; the green curve is the template IMRPhenomXPHM, evaluated for the injection parameters, and time shifted and global-phase rotated to maximize their overlap with the signal; the brown curve is the template IMRPhenomXPHM evaluated at the maximum likelihood values obtained using a Bayesian analysis. The reference for the time axis, $t = 0$, is taken to be the peak of the GW multipole h_{22} of the signal.

We leave to a future study the use of NR waveforms as synthetic signals where available. Throughout this paper, we instead generate GW signals using SEOBNRv5PHM, considering these to represent the true signal, and analyze them using IMRPhenomXPHM. However, there is a drawback to this approach. If both the waveform models deviate in a similar way from the true GR signal, the present analysis would predict small biases even when the true bias is large. This outcome is especially true since the two waveform models are not completely independent. IMRPhenomXPHM uses the SEOBNR waveforms (although, from a previous version, i.e., SEOBNRv4) for calibration in parts of the parameter space where there is a dearth of NR simulations—precisely, the regions where systematic biases are expected to be more common. In this sense, our analysis is a conservative assessment of the prevalence of systematic biases.

To quantify the systematics of the aforementioned waveform models in a wide range of applications, we utilize Bayesian analysis as well as the linear-signal approximation (LSA). The former is the most reliable tool to obtain the posterior distribution for a GW signal, but it is computationally expensive. The latter allows for computational efficiency but approximates the predictions for the posterior properties, including systematic biases, and should become a good approximation only at large SNR [108–112]. We use the LSA to study biases for BBH populations, and a wider parameter space, which is not feasible with conventional Bayesian methods. We consider three detector networks comprised of the current LIGO-Virgo network at design sensitivity (*O5*), a planned network where the current LIGO detectors are upgraded to improved sensitivity (*A#*), and a XG network comprised of two CE and an ET. The BBH populations we consider follow the LVK-like distributions where the binary masses are distributed as determined by LVK while the spins are assumed to be isotropically oriented and distributed uniformly in magnitude. We use this approach to allow for a wider range of spins. The binaries extend up to a redshift of 3 following the Madau-Dickinson star formation rate (SFR) [113]. Next, we embark on a parameter-exploration study where we consider large redshifted total masses of $200M_\odot$, asymmetric systems with inverse mass ratios going up to 30, and highly spin-precessing systems. We study these as-yet unobserved regions of the parameter space in anticipation of future observations. We also consider three distinct prototypes of BBH mergers, which hold great potential for various science objectives but are nontrivial to model due to precession or large mass ratios. The details of these three golden binary systems can be found in Table I.

The paper is organized as follows. In Sec. II, we introduce the main characteristics of the GW signal and its parameters, the waveform models that we use, and the detector networks in which signals are simulated. In Sec. III, we describe our methodology comprised of Bayesian analysis and LSA. We point out the importance of having consistent parameter definitions across waveform models and the impact on the systematic bias, where we show a comparison of a Bayesian analysis with the estimates from LSA. We also discuss the limitations of the LSA for parameter estimation (notably, the Fisher information matrix, or FIM) and biases. The study of systematic biases in the LVK-like BBH population and a hierarchical Bayesian inference on parameter distributions, reweighted to the LVK population, is reported in Sec. IV. A much broader study across the binary parameter space with particular focus on massive, highly asymmetric, and spin-precessing binaries is reported in Sec. V. A ramification on the different science applications for GWs can be found in Sec. VI, where we study selected GW events or golden binaries. The discussion and conclusion can be found in Sec. VII. In Appendix A, we illustrate the effect of nonuniform-parameter definitions across waveform models on the estimates of the systematic bias through a toy model. In Appendix B, we discuss the effect of different harmonics of the EOB model starting at different frequencies. In Appendix C, we discuss the effect of the starting frequency of the analysis on parameter estimation and systematic biases. In Appendix D, we provide a complementary plot to Fig. 6 by reporting the dependence of the ratio of systematic bias to statistical error as a function of the SNR. In Appendix E, we show the effect of the SNR threshold on the distribution of the population parameters. In Appendix F, we report the bias horizon for the χ_1 parameter of the exploratory binaries of Sec. V.

II. GRAVITATIONAL-WAVE PARAMETERS, MODELS, AND DETECTORS

A. Gravitational-wave parameters

We are interested in estimating the properties of quasi-circular, spin-precessing BBHs observed with current and future ground-based detector networks. The GW strain emitted by such binaries is characterized by 15 parameters. The parameters intrinsic to the source are the component masses m_i [114] and the dimensionless spin vectors $\boldsymbol{\chi}_i = \boldsymbol{S}_i/m_i^2$ $(i = 1, 2)$. The position of the binary is described by its luminosity distance D_L and the coordinates on the plane of the sky, (α, δ). The orientation of the binary is described by the polar angle ι and the azimuthal angle φ to the observer in the source frame [115] at the reference frequency $f_{\rm ref}$, which we set to $f_{\rm ref} = 20$ Hz throughout this paper. Finally, the relative contribution of the two gravitational polarizations, $h_+(t)$ and $h_\times(t)$, is described by the polarization angle ψ, while the reference for the time is given by the coalescence time t_c. With these definitions, the GW strain can be expressed as

$$h(t) = h_+(t; m_{z,i}, \boldsymbol{\chi}_{1,2}, D_L, \iota, \varphi, t_c)F_+(\alpha, \delta, \psi) + h_\times(t; m_{z,i}, \boldsymbol{\chi}_{1,2}, D_L, \iota, \varphi, t_c)F_\times(\alpha, \delta, \psi), \quad (1)$$

where $F_{+,\times}(\alpha,\delta,\psi)$ are the antenna pattern functions [108,116]. The detector- and source-frame masses, $m_{z,i}$ and m_i, respectively, are related by $m_{z,i} = m_i(1+z)$, with z being the redshift of the source. A superscript on any mass parameter indicates that it is the detector frame while its absence indicates that it is the source frame. The parameters D_L and z are related for a given cosmological model, which we take to be the one from Planck18 [117]. The two GW polarizations can be decomposed in the basis of -2 spin-weighted spherical harmonics ${}_{-2}Y_{lm}$ as

$$h_+(t) - ih_\times(t) = \sum_{l=2}^{\infty}\sum_{m=-l}^{+l} {}_{-2}Y_{lm}(\iota,\varphi)h_{lm}(t), \tag{2}$$

where $h_{lm}(t)$ are the GW multipoles and $\varphi = \pi/2 - \phi_{\rm ref}$.

It is often helpful to express the GW signal in terms of parameters that are combinations of the component masses and spins, either because they appear in such combinations in PN expressions or because they are conserved up to certain PN orders. In particular, the chirp mass $\mathcal{M}_c$ and the symmetric mass ratio ν are defined by $\mathcal{M}_c = (m_1 m_2)^{3/5}/M^{1/5}$ and $\nu = (m_1 m_2)/M^2$, respectively, where $M = m_1 + m_2$ is the total mass. The effective spin $\chi_{\rm eff}$ [47,118,119] and spin-precession χ_p [120] parameters are given by

$$\chi_{\rm eff} = \frac{m_1\chi_{1z} + m_2\chi_{2z}}{m_1 + m_2}, \tag{3a}$$

$$\chi_p = \frac{1}{B_1 m_1^2}\max\left(B_1 m_1^2\chi_{1,\perp}, B_2 m_2^2\chi_{2,\perp}\right), \tag{3b}$$

where $B_{1,2} = 2 + 3m_{2,1}/m_{1,2}$, and $\chi_{i\perp}$ and χ_{iz} are the magnitudes of the projection of $\boldsymbol{\chi}_i$ onto the orbital plane and perpendicular to it, respectively. While alternative definitions of χ_p have been proposed [121,122], the GW–Bayesian-analysis package Bilby [123,124], which we use to analyze simulated signals, employs Eq. (3b).

When transforming to a spherical coordinate system with the z axis perpendicular to the instantaneous orbital angular momentum, the tilts of the two spin vectors with respect to the z axis, θ_i, are given by

$$\cos\theta_i = \frac{\chi_{iz}}{\chi_i}, \tag{4}$$

where $\chi_i \equiv |\boldsymbol{\chi}_i|$ are the magnitudes of the dimensionless spin vectors. The relative angle between them in the orbital plane is parametrized by $\phi_{12} = \phi_1 - \phi_2$, where $\phi_{1,2}$ are the azimuthal angles of the two spin vectors in the spherical coordinates. Finally, the direction of the total angular momentum $\boldsymbol{J}$ in the plane perpendicular to the orbital angular momentum $\boldsymbol{L}$ at some reference time is given by the parameter $\phi_{\rm JL}$. Since $\boldsymbol{J} = \boldsymbol{L} + \boldsymbol{S}_1 + \boldsymbol{S}_2$, $\phi_{\rm JL}$ also defines the direction of the total spin vector in the orbital plane. The total angular momentum also defines the angle $\theta_{\rm JN}$, which gives the orientation of the total angular momentum vector relative to the line of sight, $\boldsymbol{N}$, of the observer. The angle $\theta_{\rm JN}$ can be expressed in terms of the inclination angle ι at the reference frequency $f_{\rm ref}$, e.g., through Eq. (C9) of Pratten *et al.* [65]. In summary, the waveform depends on the following 15 parameters:

$$\boldsymbol{\vartheta} = \{\mathcal{M}_c, \nu, \boldsymbol{\chi}_i, \cos\theta_i, \phi_{\rm JL}, \phi_{12}, \alpha, \delta, D_L, t_c, \theta_{\rm JN}, \psi, \phi_{\rm ref}\}. \tag{5}$$

In the following, we use bold font, like $\boldsymbol{\vartheta}$, to describe a set of parameters and regular font, like ϑ, to describe a particular parameter in the set.

B. Waveform models

We consider two state-of-the-art, quasicircular, spin-precessing waveform models incorporating subdominant spherical harmonics—SEOBNRv5PHM and IMRPhenomXPHM. The GW modes $(l,m) \neq (2,2)$ are important both for detection [125–127], where their noninclusion leads to a loss of signal power for asymmetric binaries and inclined orbits, and parameter estimation, where these modes can break degeneracies between various parameters and improve the measurement accuracy [90,91,128,129].

The SEOBNRv5PHM waveforms contain the spherical harmonics $(l,|m|) = (2,2), (2,1), (3,3), (3,2)$, and $(4,4)$, $(4,3)$, $(5,5)$ in the coprecessing frame. However, in this paper, we do not include the $(l,m) = (5,5)$ mode. The IMRPhenomXPHM waveforms include the $(l,|m|) = (2,2)$, $(2,1), (3,3), (3,2), (4,4)$ modes in the coprecessing frame [130]. The coprecessing frame is a noninertial frame that tracks the instantaneous motion of the orbital plane, in which the GW radiation resembles that of an aligned-spin binary [132–136]. Both waveform models can be used for a wide range of mass ratios, as well as BH-spin magnitudes up to the maximal values. However, only the aligned-spin sectors of both waveform models were calibrated to NR simulations, and their accuracy has been assessed only in regions of parameter space where NR is available.

In this work, we consider a signal generated using the SEOBNRv5PHM model to be the true GW signal and analyze it using IMRPhenomXPHM as the template model. For the Bayesian analyses of this paper, we use this approach because IMRPhenomXPHM is quicker to evaluate due to it being a frequency-domain model, while SEOBNRv5PHM is a time-domain model and thus slower. Furthermore, the computational efficiency of IMRPhenomXPHM can be improved by utilizing the multibanding approach [137] while no such analogous methods exist for time-domain models. For the Fisher information matrix analysis discussed later, we find instabilities in the numerical derivatives of the SEOBNRv5PHM waveform with respect to the GW parameters, for some regions of the parameter space, and hence restrict ourselves to computing derivatives of the

IMRPhenomXPHM model. We expect to address this issue in the future.

C. Detector networks

The current detectors are expected to achieve design sensitivity in the next few years, during the O5 run, and continue operating till the end of the decade [138]. It is anticipated that the detectors would undergo major upgrades thereafter and operate until next-generation detectors come online or even in tandem with them. Since plans for future detector networks have not yet been finalized, a number of studies have explored the capabilities of different combinations of detector configurations to understand what the optimal design is for various science goals [27–30,139]. In this work, three GW detector networks—consisting of the current detectors at design and upgraded sensitivity, and proposed future detectors—are considered to emulate a highly probable observing scenario for the coming decades. [140] These networks are enumerated below:

(i) *O5* network: This network is comprised of the advanced LIGO detectors located at Hanford and Livingston, and the advanced Virgo detector operating at design sensitivities A+ and V+, respectively.

(ii) *A#* network: In this configuration, the LIGO detectors operate at upgraded A# sensitivity while the Virgo detector continues to operate at design sensitivity V+.

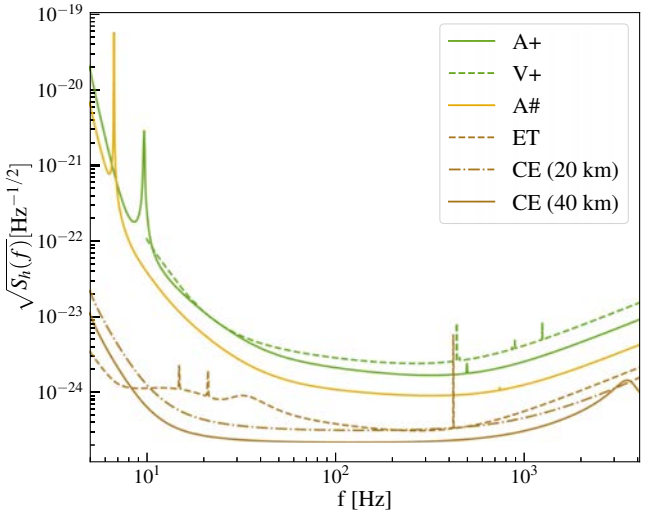

FIG. 2. Amplitude–spectral-density curves of the various detectors used in this paper (see Ref. [140]). The curves labeled by A+ and V+ denote the design sensitivity of the LIGO and Virgo detectors, respectively, which form part of the network of O5, while A# refers to the LIGO detectors at upgraded sensitivity. The next-generation observatories are the ET and CE, with the baseline network for the latter consisting of a 20-km and a 40-km detector.

(iii) *XG* network: This network comprises three proposed XG observatories consisting of the baseline 40-km and 20-km CE in the United States, and an ET in Europe.

The power spectral density (PSD) of the individual detectors is shown in Fig. 2.

III. STATISTICAL METHODS

This section describes the data-analysis methods used in this paper. In Sec. III A, we start by describing the Bayesian framework for analyzing GW signals and lay out the different choices of priors and frequency bands used for the different networks. Thereafter, in Sec. III B, we introduce the LSA for the likelihood and recount the FIM [108] method for estimating measurement errors. Following that, in Sec. III C, we elucidate the computation of biases (or systematic errors) under the LSA [104,109] and emphasize the importance of minimizing the mismatch between (i.e., aligning) the signal and template for reliable estimates of the bias. As an example, we compare the posterior distributions for a chosen binary system, as obtained from a full Bayesian analysis with the estimates from the LSA. Specifically, we point out the differences if the bias is computed without aligning the two waveforms. Finally, in Sec. III D, we discuss the hierarchical Bayesian method, which we employ to understand the impact of biases on the inference of the properties of the BBH population.

A. Bayesian analysis

The posterior probability distribution on the parameters of the waveform model, $\boldsymbol{\vartheta}$, given the observational data d and the hypothesis (model description) $\mathcal{H}$, is obtained using Bayes' theorem,

$$p(\boldsymbol{\vartheta}|d,\mathcal{H}) = \frac{p(d|\boldsymbol{\vartheta},\mathcal{H})p(\boldsymbol{\vartheta}|\mathcal{H})}{p(d|\mathcal{H})}, \tag{6}$$

where $p(\boldsymbol{\vartheta}|\mathcal{H})$ is the prior probability distribution, $p(d|\boldsymbol{\vartheta},\mathcal{H})$ is the likelihood function, and $p(d|\mathcal{H})$ is the evidence of the hypothesis $\mathcal{H}$. If one is interested solely in parameter estimation, and not in model selection, the latter serves as a normalization constant and can be discarded. For a detector with stationary Gaussian noise, the likelihood function for the data given the parameters $\boldsymbol{\vartheta}$ is defined as

$$\ln p(d|\boldsymbol{\vartheta}) = -\frac{1}{2}\langle d - h(\boldsymbol{\vartheta})|d - h(\boldsymbol{\vartheta})\rangle, \tag{7}$$

where we define the noise-weighted inner product as

$$\langle h_1|h_2\rangle = 4\mathrm{Re}\left[\int_{f_{\mathrm{low}}}^{f_{\mathrm{high}}} \frac{h_1(f)\times h_2^*(f)}{S_h(f)}\,df\right], \tag{8}$$

with $S_h(f)$ being the noise PSD, and $f_{\rm low}$ and $f_{\rm high}$ are the minimum and maximum frequencies in the detectors' bandwidth. This inner product also defines the optimal, matched-filtering SNR in a detector, ρ_n, by

$$\rho_n^2 = \langle h(\boldsymbol{\vartheta})|h(\boldsymbol{\vartheta})\rangle. \tag{9}$$

The total SNR is $\rho^2 = \sum_{n=1}^{N} \rho_n^2$, where N is the number of detectors in the network. We note that all our injections are noiseless, which corresponds to averaging over multiple noise realizations.

While the current detectors' sensitivity is limited to a minimum frequency of 20 Hz, at design sensitivity and with further upgrades, they are expected to reach a low-frequency sensitivity of 10 Hz. Meanwhile, XG observatories are aiming to further this improvement to 5 Hz. Therefore, the minimum frequency for the *O5* and *A#* networks is assumed to be $f_{\rm low}= 10$ Hz, while for *XG* detectors, it is taken as $f_{\rm low}= 5$ Hz. On the other hand, the maximum frequency is kept the same for all three networks at $f_{\rm high}= 1024$ Hz. This feature does not limit the analysis whatsoever since all the BBH systems considered in Sec. VI merge at much lower frequencies.

As we mentioned earlier, the signal is generated using the SEOBNRv5PHM model with the same starting frequency as the analysis—$f_{\rm low}= 10$ Hz for *O5* and *A#*; $f_{\rm low}= 5$ Hz for *XG*. Since SEOBNRv5PHM is a time-domain waveform model, this $f_{\rm low}$ refers to the starting frequency of the $(l, m) = (2, 2)$ mode. Subdominant harmonics with $m' \neq 2$ start at higher frequencies given by $f_{lo}^{m'} = m' f_{lo}/2$. For instance, in *O5* and *A#* networks, the $m' = 3$ modes start at 15 Hz while the $m' = 4$ modes start at 20 Hz. In Appendix B, we show that this choice does not affect our results because of the minimal additional information contained in the missing frequencies compared to the rest of the signal.

To simulate and analyze the GW signals in Sec. VI, we use the publicly available Bilby package [123,124], which incorporates the nested sampler DYNESTY [141], interfaced through the Bilby-pipe wrapper. Initially, a 14-dimensional GW parameter space is sampled using the DYNESTY sampler with a distance-marginalized likelihood. The full posterior probabilities are then reconstructed using semi-analytic methods [142,143].

All the detectors used in this study have an L-shaped interferometer configuration except the ET, which is proposed to have a triangular configuration. However, the Bilby-pipe wrapper is limited to L-shaped interferometer configurations. Consequently, the ET telescope is assumed to be L-shaped in Sec. VI. Our conclusions remain unaffected as the interferometer's shape has no significant impact on the science cases discussed here [30].

We make standard choices for the priors for all the parameters [144]. The priors for the component masses are taken to be uniform, and the spins are assumed to be isotropic in direction and uniform in magnitude. For the distance, we choose the prior proportional to d_L^2, corresponding to a uniform in comoving volume distribution at low redshift. We assume that the binary's position in the sky and the inclination of its orbit in the coprecessing frame are random. Therefore, we assign uniform priors on α, $\cos\delta$, and $\cos\theta_{\rm JN}$ across their domains. The other extrinsic parameters—namely, the polarization angle, coalescence time, and coalescence phase—are also taken to be uniform in their respective ranges.

B. Linear-signal approximation for measurement errors, systematic biases, and alignment

1. Measurement errors

The evaluation of the posterior probability distribution, as described in the previous section, is computationally expensive, which makes the estimation of the measurement accuracies and systematic biases for a large number of sources computationally prohibitive using the Bayesian method. An inexpensive approximate method is the LSA, which we now briefly introduce.

To estimate the parameter-estimation errors, the waveform model is expanded to linear order in the parameters around the maximum likelihood (best-fit) values $\boldsymbol{\vartheta}_{\rm bf}$. This process results in a Gaussian likelihood distribution whose covariance C_{ij} is given by the inverse of the FIM, $C_{ij} = \Gamma_{ij}^{-1}$, which takes the form [108,109]

$$\Gamma_{ij} \equiv \left\langle \frac{\partial h}{\partial \vartheta^i} \middle| \frac{\partial h}{\partial \vartheta^j} \right\rangle \Bigg|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}_{\rm bf}}. \tag{10}$$

The marginalized one-dimensional errors are then given by the diagonal elements, $\Delta\vartheta^i = \sqrt{C_{ii}}$. [145] The approximation holds for large SNR.

We use the publicly available package GWBENCH [146] to calculate the measurement errors. GWBENCH is an easy-to-use FIM analysis tool for ground-based detectors that implements finite-difference derivatives to estimate the approximate measurement errors. Other recent FIM analysis codes for CBCs are GWFast [32] and GWFish [147]. The waveform models are internally referenced from the LALSuite [148] libraries. While the IMRPhenomXPHM model is directly present in LALSuite, the SEOBNRv5PHM model is interfaced through the pySEOBNR package [149] within LALSuite. For IMRPhenomXPHM, the DEFAULT model in LALSuite implements a multibanding approach [137] for faster waveform computation. However, we turn this off in our FIM analysis because we find that the output of the last frequency bin has some randomness associated to it. This approach is harmless in a Monte Carlo sampling of the likelihood since the amplitude in that frequency bin is subdominant and does not contribute to the integral of Eq. (7). However, a FIM analysis involves taking waveform derivatives with respect to binary parameters, and the

randomness manifests as a delta function that dominates the integral in Eq. (10).

2. Systematic biases

A further assumption in the FIM formalism is that there are no mismodeling errors,; that is, the signal is accurately represented by the model waveform, and errors are only due to a measurement process using detectors with finite sensitivity. In reality, we do not know the true GW waveform, and we use various approximate models to faithfully represent the true signal. As such, there is a source of error arising from a difference between the signal and the waveform model used to represent the signal (template). As a result, the parameters that maximize the likelihood are biased from the true parameters of the GW signal by $\delta\boldsymbol{\vartheta}^i$. This mismodeling error, henceforth called bias $\delta\boldsymbol{\vartheta}^i$, is given by [104,109,150],

$$\delta\boldsymbol{\vartheta}^i = C^{ij}\langle\partial_j h|\delta h\rangle|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}_{\rm bf}}, \tag{11}$$

at the leading order, where $\delta h = h_s - h$, with h_s being the true signal. In practice, the true signal is not known, so this formula can only be used if the true signal is replaced by some fiducial reference model, here taken to be SEOBNRv5PHM.

3. Waveform alignment

We now discuss a few subtleties in the use and applicability of Eq. (11) for the estimation of biases. Note that the bias is directly proportional to the waveform difference, δh. In part due to different conventions for some extrinsic parameters, δh can be artificially large when evaluated at the same value of all parameters, but it can be significantly reduced by changing the values of certain extrinsic parameters, such as the global phase and time shift, while keeping the intrinsic parameters fixed. In Appendix A, we describe a toy model that illustrates how a simple time shift can cause biases in physical parameters to become large. Since Eq. (11) is derived under the LSA, large waveform differences stretch the formula beyond its domain of validity, resulting in unreliable estimates. However, large uncertainties in the extrinsic parameters are typically not problematic for scientific applications of GW observations; thus, if by changing only a subset of the extrinsic parameters we can bring the waveform difference back into the range of validity of the LSA, we should do so in order to improve the accuracy of the inferred results.

Waveform-accuracy studies in the literature that use δh as a metric to quantify waveform differences, and estimate expected biases, have typically followed this approach and minimized δh over the extrinsic parameters [82,151] (alignment). However, to the best of our knowledge, many studies employing Eq. (11) to estimate the bias either neglect this aspect and naively use the difference between waveform models to estimate the systematic bias, or they do not discuss it. The incorrect use leads to unreasonably large estimated biases, particularly for the luminosity distance. Therefore, we describe here how we implement the alignment in the bias formula.

Using Eq. (8), we define the unfaithfulness or mismatch between two waveforms h_1 and h_2 as

$$\mathcal{M} = \min_{\boldsymbol{\lambda}}\left\{1 - \frac{\langle h_1|h_2\rangle}{\sqrt{\langle h_1|h_1\rangle\langle h_2|h_2\rangle}}\right\}, \tag{12}$$

where the minimization is performed on a subset of the binary's parameters, which we denote $\boldsymbol{\lambda}$. For nonprecessing waveform models employing only the dominant quadrupolar mode, $\boldsymbol{\lambda} = \{\psi, t_c\}$. In this case, ψ is degenerate with $\phi_{\rm ref}$, so we need to consider only one of them. On the other hand, since we are considering spin-precessing waveform models, $\boldsymbol{\lambda} = \{\psi, t_c, \phi_{\rm ref}, \phi_{\rm JL}\}$, where $\phi_{\rm JL}$ is a rotation of the in-plane spin angles. For spin-precessing waveform models, some studies have chosen to minimize the mismatch over the reference frequency instead of in-plane spin rotations [152,153]. However, in this study, we choose to optimize the mismatch by rotating the in-plane spin components [51,65], thus keeping the reference frequency fixed at $f_{\rm ref} = 20$ Hz.

Starting with the set of parameters $\boldsymbol{\vartheta}$, we find the parameters $\bar{\boldsymbol{\lambda}}$ that minimize $\mathcal{M}$ in Eq. (12) for the detector network being considered. The minimization over the polarization angle ψ is performed analytically while the coalescence time t_c is optimized by convolving the two waveforms utilizing the convolution theorem [116,154,155]. The reference phase $\phi_{\rm ref}$ and in-plane spin rotations $\phi_{\rm JL}$ are optimized numerically by using standard optimization algorithms. Having found the parameters that minimize Eq. (12), $\bar{\boldsymbol{\lambda}}$, we have a new set of parameters $\bar{\boldsymbol{\vartheta}}_{\rm bf}$, where the parameters $\boldsymbol{\lambda} = \{\psi, t_c, \phi_{\rm ref}, \phi_{\rm JL}\}$ have been replaced by the values obtained through Eq. (12). We use this set of parameters to compute the FIM, as well as the δh in Eq. (11). Therefore, the alignment procedure modifies Eq. (11) to

$$\delta\boldsymbol{\vartheta}^i = C^{ij}(\bar{\boldsymbol{\vartheta}}_{\rm bf})\langle\partial_j h(\bar{\boldsymbol{\vartheta}}_{\rm bf})|h_s(\boldsymbol{\vartheta}) - h(\bar{\boldsymbol{\vartheta}}_{\rm bf})\rangle. \tag{13}$$

If the parameters $\boldsymbol{\lambda}$ are uncorrelated with the other binary parameters, the bias formula Eq. (11) should give $\delta\boldsymbol{\lambda} = 0$. However, in general, that is not the case and, therefore, $\delta\boldsymbol{\lambda} \neq 0$. Thus, the total bias for $\boldsymbol{\lambda}$ is $\Delta\boldsymbol{\lambda} = \delta\boldsymbol{\lambda} + (\boldsymbol{\lambda}_s - \bar{\boldsymbol{\lambda}})$, where $\boldsymbol{\lambda}_s - \bar{\boldsymbol{\lambda}}$ is the difference between the parameters $\boldsymbol{\lambda}$ of the fiducial signal and those obtained after the optimization procedure.

Note that we chose to modify the template in Eq. (11) following the optimization procedure, Eq. (12). Under the LSA, we are free to modify the signal evaluating the template at the fiducial parameters. However, we notice a slightly better agreement of the bias with full Bayesian

results when modifying the template because the Bayesian analyses are performed using the fiducial parameters as the values of the synthetic-injected signal; thus, we find the systematic bias to be more sensitive to small changes in the injected values compared to the measurement errors.

Lastly, we note that $\mathcal{M}$ could already be close to the minimum for certain pairs of waveform models at a given set of parameters out of the box. In such cases, the optimization procedure will have a minimal effect on the total bias, and one could simply use the bias formula as it is. However, the total bias $\Delta\boldsymbol{\lambda}$ would be the same regardless of whether one chooses to perform the initial optimization or not, even though the output of the bias formula will not be the same. Correspondingly, the net bias does not depend sensitively on the precision of the optimization routine as the bias formula compensates for it. Therefore, it is more prudent to compare the net biases rather than the optimized values following the initial minimization. We verify our minimization procedure using a brute-force 4D minimization algorithm and find that, while $\bar{\boldsymbol{\lambda}}$ is slightly different between the two minimization routines, $\Delta\boldsymbol{\lambda}$ remains the same.

In the following, we calculate the systematic biases with and without the optimization procedure outlined above. We compare these estimates with the full Bayesian-analysis results for a subset of the parameters and find agreement with the Bayesian analysis when using the optimization procedure in Eq. (12).

C. Comparing Bayesian and linear-signal approximation analyses

As a representative case to compare results between the Bayesian method and the LSA with and without the optimization procedure of Eq. (12), we consider a BBH with parameters given in Table I (see Sec. VI B), denoted as Binary 1 in the *O5* detector network. In Fig. 3, we show the posterior distributions for selected parameters, namely, $\mathcal{M}_c$, ν, D_L, and χ_1, using the Bayesian analysis and the LSA estimate for the errors and biases computed from Eqs. (10) and (11), respectively. The full Bayesian posterior estimates are shown in green. The LSA posterior distributions are multidimensional normal distributions centered at the biased value with the covariance matrix given by Eq. (10). The curves in orange show the distributions when the biases are estimated by minimizing the mismatch between the waveforms [see Eq. (13)], while the ones in brown are the estimates when the minimization is not performed [as it is typically done in the literature; Eq. (11)]. Since the covariance matrix is approximately the same in a neighborhood, the posterior widths are similar. However, the predicted bias differs substantially between the two procedures. The effect on the estimation for the distance bias is especially noticeable, with the traditional method predicting an approximately 50% bias when it is unbiased in actuality. This effect would be of particular importance

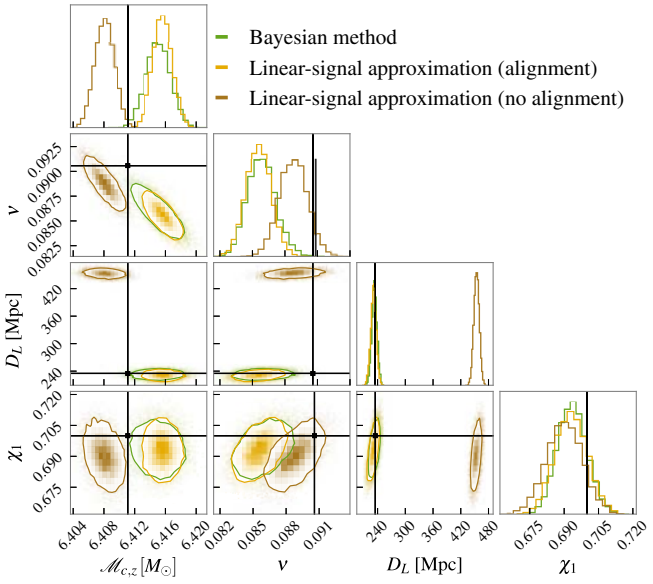

FIG. 3. Comparison of the posterior distributions for the chirp mass, symmetric mass ratio, luminosity distance, and primary spin magnitude for Binary 1 with parameters given in Table I and in the *O5* detector network. The distributions obtained from a Bayesian parameter estimation using Bilby are shown in green. The estimates from the LSA with and without the minimization procedure [Eq. (12)] (alignment) are shown in orange and brown, respectively. The black cross-hairs show the true injected value. The parameter estimation is performed by injecting a SEOBNRv5PHM signal and recovering it with the IMRPhenomXPHM waveforms. The Bayesian posteriors are accurately represented by the LSA when the alignment is enforced.

for cosmology studies where the traditional estimates found in the literature would be overly pessimistic. The contours show the 90% credible intervals of the parameters.

We now briefly discuss the validity of the LSA. Even though we find excellent agreement between the LSA and the Bayesian analysis for this fiducial case, it is important to keep in mind that the estimates are approximate. Particularly, both the FIM and the bias formula [Eq. (13)] are derived under the assumption that a waveform model can be expanded linearly in its parameters. While the FIM approximation improves with increasing SNR, with higher-order contributions scaling as $\mathcal{O}(1/\mathrm{SNR})$, the bias is independent of the SNR, both in the linear approximation and the full likelihood. For the LSA, this case can be easily gauged from the bias equation, Eq. (13), which is independent of the distance and/or simple scaling of the PSD. For the Bayesian analysis, one can conclude from Eq. (7) that a simple scaling of the PSD will not affect the stationary points. Therefore, the point in the parameter space where the likelihood peaks remains constant, which means that the error in the bias computation is also constant. In addition, note that we are interested in the bias in units of the statistical errors. Hence, while the measurement becomes better with improved sensitivity

(or larger SNR), the error in the systematic bias estimated using LSA becomes more important. *A priori* it is difficult to know the range of the optimal place where both approximations hold. However, the event shown in Fig. 3 has an SNR of around 75, and we also observe similar agreement in the *A*# network where the event has an SNR of around 220 (see Table I), prompting us to make the reasonable assertion that the LSA is most trustworthy for such ranges of the SNR. We were not able to directly compare the Bayesian results in the XG network with the LSA estimates because, as we explain in Sec. III A, the former assumed an L-shaped interferometer for ET while the latter was performed using a triangular ET configuration. We would also like to stress that one would expect the LSA to hold when the mismatch between two waveform models is not too large. For the case illustrated above, we find the mismatch $\mathcal{M} \sim 3\%$. However, the binaries that are considered in Secs. IV and V can have much larger mismatches, and a more detailed analysis is required to quantify the validity of the LSA as a function of the mismatch, which is beyond the scope of this study.

D. Hierarchical Bayesian analysis

We now discuss the method that we employ in Sec. IV to understand the impact of the biases on the inference of the properties of the BBH population. Given a set of $N_{\rm obs}$ observed data $\{d_i\}$, we can estimate the underlying distribution of parameters that generated it through a hierarchical Bayesian analysis. We denote by ϑ ($\subset \boldsymbol{\vartheta}$) the set of parameters, whose distribution we wish to infer. Assuming a form for the number density of observed events, $[(\mathrm{d}N)/(\mathrm{d}\vartheta)](\Lambda)$, that depends on hyperparameters Λ, the posterior on the latter is given by [156,157]

$$p(\Lambda|\{d_i\}) \propto \pi(\Lambda) e^{-N(\Lambda)} \prod_{i=1}^{N_{\rm obs}} \int \frac{\mathrm{d}N}{\mathrm{d}\vartheta}(\Lambda) \frac{p(\vartheta|d_i)}{\pi_{\rm PE}(\vartheta)} \mathrm{d}\vartheta, \quad (14)$$

where $p(\vartheta|d_i)$ is the single-event posterior, $\pi_{\rm PE}(\vartheta)$ is the prior used for parameter estimation, $\pi(\Lambda)$ is the prior on the hyperparameters, and $N(\Lambda)$ is the total number of events, defined as

$$N(\Lambda) = \int \frac{\mathrm{d}N}{\mathrm{d}\vartheta}(\Lambda) \mathrm{d}\vartheta. \quad (15)$$

In the analysis of real data, the above equation must be modified to include selection effects. We interpret $\mathrm{d}N/\mathrm{d}\vartheta$ as the rate density of the full population and modify the argument of the exponential to $p_{\rm det}(\Lambda)N(\Lambda)$, where $p_{\rm det}(\Lambda)$ is the probability of detection of a source, averaged over the population model. In this work, we do not seek to perform a full astrophysical inference. Thus, we treat the distribution of events that pass the cutoff in SNR as the population of interest and investigate the impact of systematic effects on the shape of this distribution. This method is not strictly equivalent to performing the inference on the observed population, as this would require us to "renormalize" the likelihood to the observable portion of the parameter space. Here, instead, we neglect the selection process, reconstruct $\mathrm{d}N/\mathrm{d}\vartheta$ as the distribution of events that pass the cutoff, and compare how this changes under the effect of systematic biases. Let us mention that systematic errors might also bias our estimation of the selection function. The latter is usually evaluated by performing injections with GW templates, and if the true signal that generates the observed data differs from our templates, we might estimate the probability of detecting an event wrong. In the analysis performed here, we instead approximate the selection as a hard cut on the intrinsic SNR of the source. In this model, selection is now defined on the source parameters, not the data, and the above equation can be used directly; however, $\mathrm{d}N/\mathrm{d}\vartheta$ must now be interpreted as the rate density in this observed portion of the population. This approach, which is common in the literature, ignores the fuzziness at the detection horizon that arises from instrumental noise, but it will give quantitatively reliable and unbiased results, provided the data are simulated from the same model. In Eq. (14), we use the proportionality symbol instead of the equality one because we have omitted numerical factors that depend on the observed data $\{d_i\}$ but not on Λ, i.e., the individual event evidence and the overall model evidence. These factors are required to perform model selection but are unimportant when the goal is to obtain the posterior distribution on Λ.

We perform a hierarchical Bayesian analysis for each source parameter separately—i.e., χ_1, q, $\cos\theta_1$, and $\mathcal{M}_c$—so that $\mathrm{d}N/\mathrm{d}\vartheta$ is a one-dimensional function. This approach yields optimistic measurements for the number densities as compared to the full inference but allows us to have a quick assessment of the impact of systematic biases on population inference. Adopting the approach of Toubiana *et al.* [158], we describe the number density of observed events, $\mathrm{d}N/\mathrm{d}\vartheta(\Lambda)$, as a piecewise linear function. The extremities of the ϑ range over which we perform the inference are fixed and determined by the minimum and maximum samples present in the data. Thus, our hyperparameters are as follows: the values of the number densities at the extremities, the number of knots, their positions, and the value of the number density at the knots. The number density at any point is then obtained by linear interpolation. We stress that the number of knots is a free parameter of the model, and it is inferred by using a reversible-jump Markov chain Monte Carlo algorithm [159]. In this way, the complexity of the model is determined by the data themselves.

For a given detector network, we perform population inference on a mock catalog with systematic biases and on one without, generated as follows.

(1) We draw the parameters ϑ_0 from the population model described in Sec. IV and select those with SNR above a given threshold.
(2) We compute the measurement error and the systematic bias for all observable events using the LSA, as described in Sec. III B 2.
(3) For the catalog with systematic biases, we shift the true parameters by $\delta\vartheta_{\rm bf}$ to obtain the biased parameters, $\vartheta_{\rm bf}$.
(4) For each event ϑ_i, we attribute a measurement error σ_i drawn randomly among the set of computed measurement errors, allowing for replacement.
(5) We draw a noisy measurement $\vartheta_{n,i}$ of each event from a Gaussian centered at ϑ_i ($\vartheta_{{\rm bf},i}$ for the biased catalog), with the standard deviation given by the error drawn in step 4.

Under the LSA, the posterior distribution on ϑ is a truncated Gaussian:

$$p(\vartheta|\vartheta_{n,i}) = \frac{2\exp\left[-\frac{1}{2}\frac{(\vartheta-\vartheta_{n,i})^2}{\sigma_i^2}\right]}{\sqrt{2\pi}\sigma_i\left[\mathrm{erf}\left(\frac{\vartheta_{\rm max}-\vartheta_{n,i}}{2\sigma_i}\right)+\mathrm{erf}\left(\frac{\vartheta_{n,i}-\vartheta_{\rm min}}{2\sigma_i}\right)\right]}, \quad (16)$$

where $\vartheta_{\rm min}$ and $\vartheta_{\rm max}$ are the boundaries of the prior domain on ϑ. The purpose of the randomization of the errors (step 4) is to remove the dependency on ϑ from the standard deviation entering the posterior distribution. If we were to use the corresponding value predicted by the FIM for each event, we would have to account for the complicated dependency of σ on ϑ, and the posterior on ϑ would no longer be a Gaussian, requiring us to go beyond quadratic order in the LSA. Moreover, σ would also depend on the remaining parameters in $\boldsymbol{\vartheta}$, and, by performing the inference on a single parameter, we would not be accounting for this dependence correctly, making our analysis not self-consistent. However, we observe that, for $\vartheta = q$ or χ_1 or $\cos\theta_1$, the amount by which the estimated uncertainty in the parameter varies over the range of our priors is small, so we expect our procedure to yield realistic results for those parameters. [160] Step 5 is crucial to make sure our mock catalog is self-consistent from the statistical point of view. Working in the so-called zero-noise approximation is valid for performing parameter estimation on single mock events because it is a fair realization of the noise in the detector. On the other hand, having zero noise for all events is no longer a fair realization, and it would be valid only if all the events were perfectly measured. Note that, in steps 3 and 5, we allow the biased parameters and the noisy ones to be outside of the prior range. The rationale is that those steps are meant to mimic the behavior of the likelihood function in the presence of systematic biases and noise, which, as a function, does not contain information on the physically allowed range of a given parameter. The posterior, in turn, is truncated to the prior range, as given explicitly in Eq. (16).

In the hierarchical Bayesian analysis, we take the parameter estimation prior $\pi_{\rm PE}(\vartheta)$ to be flat in ϑ. Thus, each of the integrands in Eq. (14) is the product of a Gaussian with a piecewise linear function, and we can perform the integration analytically. This approach allows us to evade problems related to having an insufficient number of samples when performing Monte Carlo integration [161], and it speeds up the analysis.

IV. SYSTEMATIC BIASES IN THE BBH POPULATION

In this section, we study the effect of systematic biases on a BBH population. We use the GWTC-3 results [5,14] only for the distribution of masses. We explore the impact of systematic biases on this LVK-like population, considering the three detector networks introduced in Sec. II C. We also perform a hierarchical Bayesian inference of the underlying population where we reweight our population distribution to the current LVK distribution of astrophysical BBHs.

A. LVK-like population

We simulate 10^5 binaries in each of the detector networks described in Sec. II C up to a redshift $z = 3$ using the SEOBNRv5PHM waveform model. Following Borhanian and Sathyaprakash [27], this result is around the expected number of BBH mergers per year. We choose a network SNR threshold of 12 for detection and use it to identify the subset of simulated binaries that are in the population observed by each network.

The redshift distribution for the population is drawn from a probability distribution given by

$$p(z) \propto \frac{dV_c}{dz}\frac{1}{1+z}\psi(z), \quad (17)$$

where dV_c/dz is a comoving volume element per unit redshift and $\psi(z)$ is the SFR, which is taken to be [113]

$$\psi(z) = 0.015\frac{(1+z)^{2.7}}{1+[(1+z)/2.9]^{5.6}}M_\odot\ \mathrm{yr}^{-1}\,\mathrm{Mpc}^{-3}. \quad (18)$$

The distribution of the primary source-frame mass is assumed to follow the Power Law + Peak model of Abbott *et al.* [14], with the parameters fixed to their maximum likelihood values. For completeness and ease of reference, we elucidate the model here. The Power Law + Peak model is given by

$$\begin{aligned} p(m_1|\lambda_{\rm peak},\alpha_m,m_{\rm min},m_{\rm max},\mu_m,\sigma_m,\delta_m) \\ \propto [(1-\lambda_{\rm peak})\mathrm{PL}(m_1|\alpha_m,m_{\rm max}) + \lambda_{\rm peak}\mathcal{N}(m_1|\mu_m,\sigma_m)] \\ \times S(m_1|m_{\rm min},\delta_m), \end{aligned} \quad (19)$$

where $\lambda_{\rm peak}$ gives the weight of the peak component, $\mathrm{PL}(m_1|\alpha_m, m_{\rm min}, m_{\rm max})$ is a normalized power-law distribution with spectral index α_m truncated to the range $[m_{\rm min}, m_{\rm max}]$, $\mathcal{N}(m_1|\mu_m, \sigma_m)$ is a normalized Gaussian distribution with mean μ_m and width σ_m, and finally, $S(m_1|m_{\rm min}, \delta_m)$ is a smoothing function defined by

$$S(m|m_{\rm min}, \delta_m) = \begin{cases} 0; & \text{if } m < m_{\rm min}, \\ [f(m - m_{\rm min}, \delta_m) + 1]^{-1}; & \text{if } m_{\rm min} \leq m \leq m_{\rm min} + \delta_m, \\ 1; & \text{if } m > m_{\rm min} + \delta_m, \end{cases} \tag{20}$$

with

$$f(m, \delta_m) = \exp\left(\frac{\delta_m}{m} + \frac{\delta_m}{m - \delta_m}\right), \tag{21}$$

where δ_m regulates the sharpness of the smoothing function. The maximum likelihood values for the fit to GWTC-3 [14] are $\lambda_{\rm peak} = 0.02$, $\alpha_m = -3.5$, $m_{\rm min} = 4.8 M_\odot$, $m_{\rm max} = 83 M_\odot$, $\mu_m = 34 M_\odot$, $\sigma = 1.9 M_\odot$, and $\delta_m = 5.4 M_\odot$. The mass-ratio distribution is modeled using a power law with a smoothing function, and it takes the form

$$p(q|\beta_q, m_1, m_{\rm min}, \delta_m) \propto q^{\beta_q} S(q m_1|m_{\rm min}, \delta_m), \tag{22}$$

where the maximum likelihood value for the spectral index is $\beta_q = 0.76$. We show the component mass distributions in Fig. 4.

The analyses performed on GWTC-3 [14] suggest a broad distribution for the spin magnitude, peaking around 0.2 and falling off to 0 for large spins. However, we are particularly interested in estimating systematic biases for large-spin systems, so we draw the magnitude uniformly between 0 and 1. The analyses on systematic effects are performed on this population; in particular, we compute the measurement uncertainties and systematic biases within the LSA for these parameters, but when performing the hierarchical Bayesian analysis, we generate the mock catalog by performing importance sampling on this flat

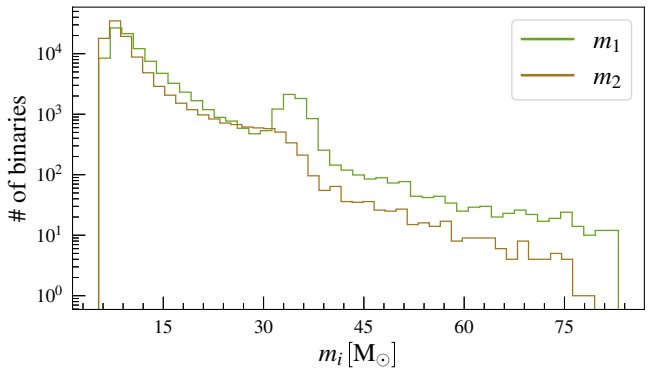

FIG. 4. Distribution of component masses for a population of 10^5 BBHs following the astrophysical distribution as determined by the LVK Collaboration.

population to obtain a spin distribution in agreement with the results of the DEFAULT model of Abbott *et al.* [14]. Finally, we assume the spins' orientation to be distributed isotropically, which is in qualitative agreement with the results on GWTC-3.

The location and orientation of the binary in the plane of the sky is assumed to be randomly distributed. Therefore, the declination angle δ, right ascension α, and inclination angle ι follow the distributions $\cos\delta \in \mathcal{U}[-1, 1]$, $\alpha \in \mathcal{U}[0, 2\pi]$, and $\cos\theta_{\rm JN} \in \mathcal{U}[-1, 1]$, respectively. The polarization angle ψ and the coalescence phase ϕ_c are also drawn from a uniform distribution, $\psi, \phi_c \in \mathcal{U}[0, 2\pi]$.

We report the SNR distribution of the 10^5 binaries simulated in the three detector networks in Fig. 5. We shade the region below the SNR threshold of 12 in gray. The tails of the three distributions exhibit the $\propto \rho^{-4}$ dependence of the rate of mergers per unit redshift in accordance with the uniform in comoving-volume distribution of sources in the nearby universe. On the other hand, the peak of the distribution correlates with the peak of the SFR, while the initial slope depicts the first generation of stars following the dark ages. While we are limited by the sensitivity of the current detector networks and their upgrades in our ability to observe GWs from the mergers

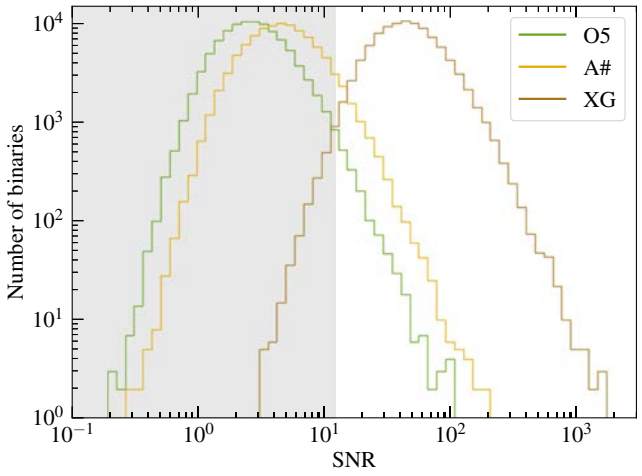

FIG. 5. Distribution of the SNRs of the 10^5 BBHs simulated in the three detector networks computed using the IMRPhenomXPHM model. The SNR distribution is similar when using the SEOBNRv5PHM model. In shaded gray, we indicate the region with network-SNR threshold below 12.

of the first stellar-origin BBHs, the *XG* network will enable us to study BBH mergers immediately following reionization.

B. Systematic bias

In the following, we first discuss the systematic biases for the individual events in the LVK-like population. Then, we carry out a hierarchical Bayesian inference of the population distributions by reweighting the parameter distributions to the LVK distribution. Finally, we determine the type of binaries more likely to be biased, which motivates our analysis in Sec. V.

1. Variation with mismatch

Armed with the biases in the GW parameters and their measurement errors, in Fig. 6, we report their ratio $|\delta\vartheta/\Delta\vartheta|$ as a function of the mismatch $\mathcal{M}$ between SEOBNRv5PHM and IMRPhenomXPHM for the chirp mass, symmetric mass ratio, primary spin magnitude, and luminosity distance. We recall that the biases are computed using Eq. (13). The SNR for the binaries of the population are portrayed using a color scale with lighter colors representing smaller SNR and vice versa. A value of $|\delta\vartheta/\Delta\vartheta| > 1$ indicates that systematic biases are larger than the typical size of statistical errors. A common feature for all the parameters is a direct correlation between $|\delta\vartheta/\Delta\vartheta|$ and the mismatch. This feature is intuitive because a larger mismatch implies a greater difference between the two waveform models and, therefore, larger biases, assuming the measurement errors do not vary significantly with the mismatch, which we find to be broadly true for the population. On the other hand, the color scale shows that the loudness of a signal is not a guarantee for a dominant systematic effect with quieter signals exhibiting significant systematic biases particularly when the mismatch is greater, which is especially true for the *O5* and *A#* networks. We provide a complementary plot of $|\delta\vartheta/\Delta\vartheta|$ as a function of the SNR in Fig. 24 in Appendix D for the interested reader.

In this section and in other places where measurements of the spin using LSA are discussed, we examine the primary spin magnitude instead of the popular variables $\chi_{\rm eff}$ and χ_p. This choice is made for the practical reason that the Fisher matrix is evaluated in χ_1. It is not trivial to transform the likelihood to other variables since it can extend beyond the prior bounds for a parameter. In fact, imposing priors, particularly in the presence of large biases, is also nontrivial since the true posterior distribution need not simply be a truncated likelihood distribution (in the case of a uniform prior with boundaries) peaking outside the prior boundary but rather peaking around the subdominant maximum of the likelihood that is within the bounds. However, we do not have any information about such features since we are working under the LSA. Hence, in order to avoid such difficulties and keep the analysis simple, we ignore the effects of priors for individual binaries. Furthermore, since χ_1 is better measured than χ_2 in a majority of the cases, it is reasonable to consider that its distribution is closer to a

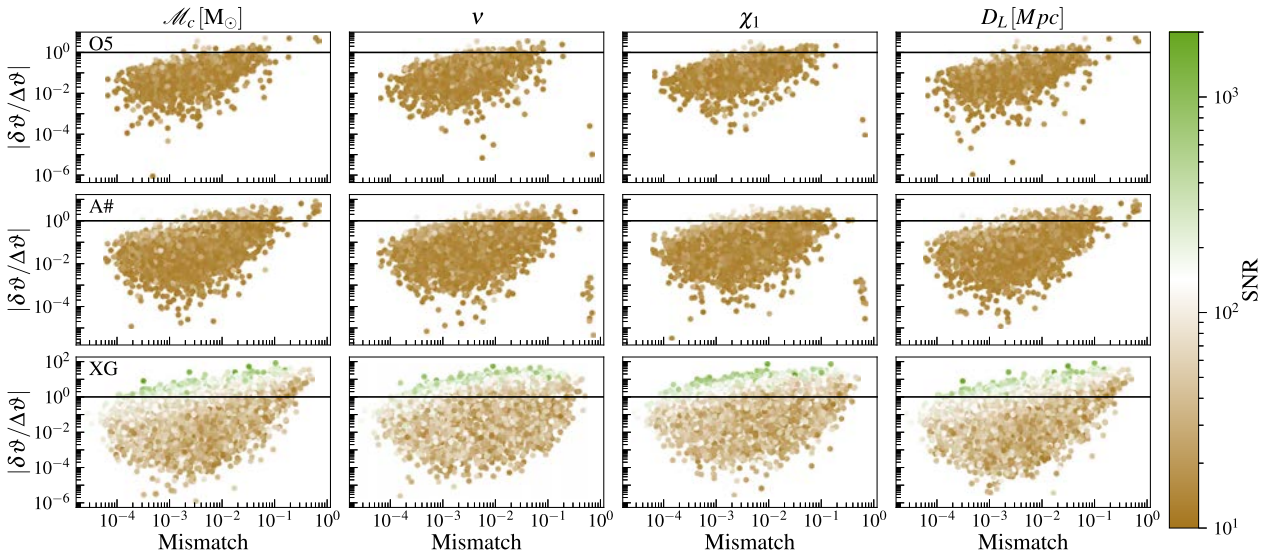

FIG. 6. Ratio between the systematic bias and statistical errors, $\delta\vartheta/\Delta\vartheta$, for source-frame chirp mass, symmetric mass ratio, primary spin magnitude, and luminosity distances as a function of the mismatch between SEOBNRv5PHM and IMRPhenomXPHM waveform models for a population of BBH mergers as observed by the LVK. A network SNR threshold of 12 was imposed on the 10^5 binaries in the population, resulting in detected events of around 1800, 8100, and 99,000 in the *O5* (top), *A#* (middle), and *XG* (bottom) networks, respectively. The SNRs of the binaries are depicted using the different color scales. The outlier events in the top and middle panels having the largest mismatches are the heaviest and most distant events with redshifted total mass greater than $400M_\odot$ and redshift greater than 2.

normal distribution and, hence, better described under the LSA.

A few events appear as outliers in the figure with large mismatches but extremely small $|\delta\vartheta/\Delta\vartheta|$ for the *O5* and *A#* networks. These events are the heaviest and most distant, with redshifted total masses greater than $400M_\odot$ and redshifts greater than 2. As such, these signals are extremely short, consisting of only the merger ringdown. The FIM in these cases is close to singular because the signal does not contain much information, resulting in extremely large errors. This result suppresses the ratio for every parameter except the luminosity distance and the chirp mass. For the former, this is because the ratio is directly proportional to the mismatch and, hence, large; for the latter, the effect of the luminosity distance dominates in the conversion from detector-frame chirp mass to source-frame chirp mass. As a result, even though the events appear as outliers for the detector-frame chirp mass, they follow the behavior of the luminosity distance for the source-frame parameter. The LSA approximation is not reliable for such cases, and we should resort to full Bayesian analyses. Nevertheless, we include these binaries in the figure to show their existence in the population observable by these two networks.

Figure 6 also shows that the number of binaries with biased parameters as a fraction of the detected population increases with improved detector sensitivity. This finding can be simply understood as due to the independence of the overall scale of the PSD in estimating the parameter bias [see Eq. (11)] while the covariance is inversely proportional to it. While an improving detector sensitivity leads to an increase in the total number of binaries that are significantly biased, the increase in the biased fraction has to do with the finite number of stellar origin BBH mergers in the universe. Note that because of different PSD shapes, interferometer designs, and minimum frequencies, the three rows of Fig. 6 are not simply shifted versions of one another. Nevertheless, even for the *XG* network, only a minority of events are biased, with the biased fraction ranging from 10%–25% depending on the parameter. For detectors of the current generation, the biased fraction is even smaller, with only about 2% and 2.5% of binaries significantly biased for the *O5* and *A#* networks, respectively. This finding suggests that biases in the parameter estimation will only be of importance for extraordinary individual events rather than for inferring general characteristics of the population. We check this case more carefully in the next section.

It is also important to realize that a larger value of $|\delta\vartheta/\Delta\vartheta|$ does not necessarily mean a larger value of the absolute bias. For instance, as we will see in Sec. VI, the *XG* network quite often has smaller absolute biases due to improved low-frequency sensitivity, where waveform models agree to a greater extent. However, the improved sensitivity reduces the measurement error more than the decrease in the systematic bias, resulting in a larger value of $|\delta\vartheta/\Delta\vartheta|$. We will see the effect of this result in the next section.

2. Inferred distributions

Figure 7 shows the inferred number density of events for χ_1, $\cos\theta_1$, and q for the three detector networks. Solid lines indicate the mean of the number density, and colored bands are the 90% confidence intervals. The results shown in black correspond to an unphysical scenario where all observed events are perfectly measured. In this case, the only source of uncertainty is the Poisson error due to the finite number of events. The black curves serve as guidance to indicate the underlying distribution on which we perform the inference. In brown, we show the case with measurement error and no bias, following the procedure outlined in Sec. III D. The brown and black shaded areas overlap, with the red encompassing the blue most of the time due to the inclusion of measurement errors, indicating that our procedure yields unbiased results. Thus, differences between the cases without (in brown) and with (in green) biases are due to waveform systematics.

We recall that, when performing the hierarchical Bayesian analysis, we resample the flat χ_1 distribution into the distribution inferred by the LVK Collaboration [14] by means of importance sampling. For O5, the biased and nonbiased distributions are mostly compatible. However, for *A#* and *XG*, we observe that, when including bias, the inferred χ_1 distribution is broadened, with the peak being shifted to lower values. The broadening is a consequence of the occurrence of large biases, while the shift happens because the systematic bias typically increases with χ_1 (see Fig. 8 and the associated discussion). Events with large spin are more shifted, with a small preference for shifts towards lower χ_1, than events with small spins. From the astrophysical point of view, the shift of the peak is rather negligible, but the tail of high spin events would lead to an overestimation of the number of BHs with high spins, by up to 2 orders of magnitude, potentially challenging formation scenarios.

For the tilt angle, we observe an excess at the ends due to events with more precession ($\cos\theta_1 \sim 0$) being more biased than those with aligned spins. We note that the *XG* population is less biased than the *A#* one. This feature is a consequence of the improvement at detectors at low frequencies, which increases the proportion of inspiral signal that is observed, where waveform models agree best, and yielding a less-biased estimate of precession effects. We observe a similar behavior for the q distribution. Asymmetric events are more biased than nearly equal-mass ones, shifting the overall distribution to $q \sim 1$. As for $\cos\theta_1$, the *XG* population is less biased than the *A#* one. At first glance, this finding might seem in contradiction with Fig. 6, which shows that the ratio between the systematic bias and statistical error on ν tends to be larger in the *XG* case. However, as further discussed in Sec. VI,

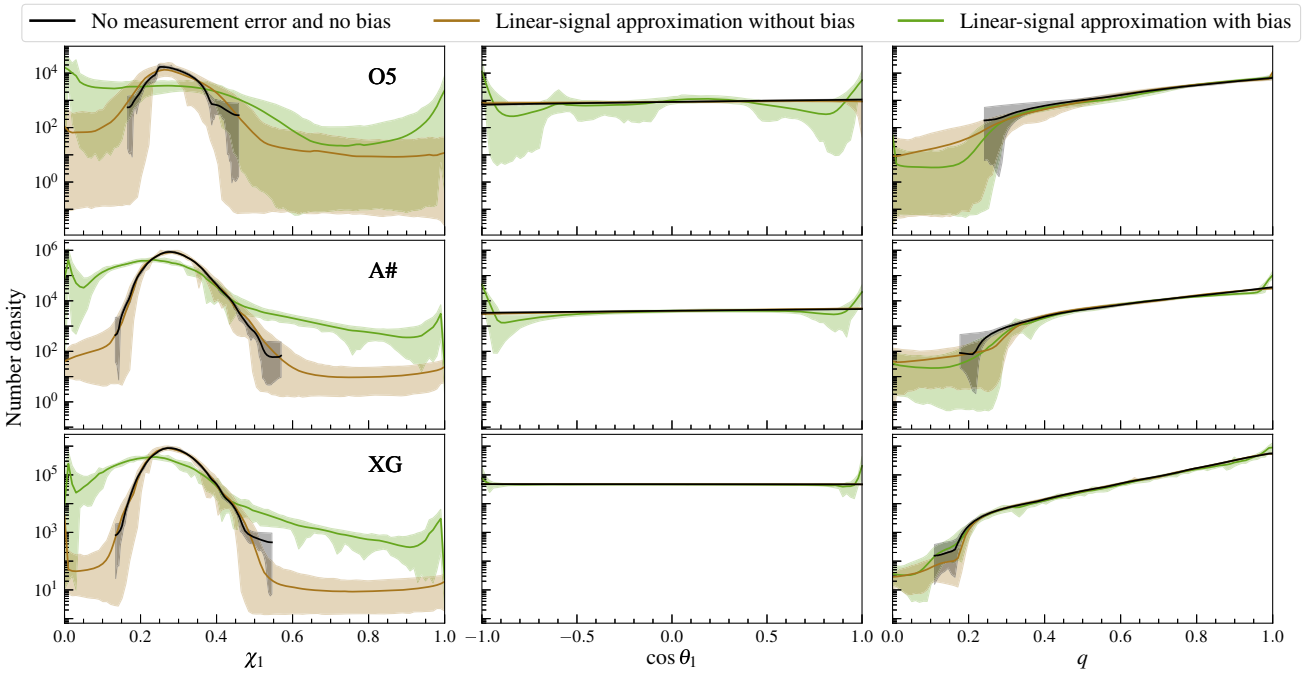

FIG. 7. Inferred distribution of χ_1, $\cos\theta_1$, and q for different detector networks, in the case where the parameters are measured perfectly (black), where there is measurement error but no bias (brown), and where there is measurement error and bias (green). The solid lines show the mean value in a given bin, and the shaded areas show the 90% confidence intervals. The "no measurement error and no bias" case does not correspond to any realistic scenario and is there to show the underlying distribution of observed events, accounting for Poisson errors. For χ_1, when including biases, we observe a shift towards smaller values, together with a broadening of the distribution, particularly for *A#* and *XG*. For $\cos\theta_1$, we observe an excess of events at both ends and for q at around 1. The distributions in the *XG* case for those two parameters are less biased because the bias for individual events is smaller in *XG*.

when comparing the full posteriors, in many cases the result in the *XG* case is closer to the true value than in the *A#* case. The ratio between the systematic bias and statistical error is larger for *XG* because the measurement error decreases more than the bias (in relative terms); however, both errors decrease, and the fact that the absolute bias is smaller ends up reducing the bias at the level of the population inference. We also perform hierarchical Bayesian analysis on $\mathcal{M}_c$ and find no bias at the level of the population, as expected given that this parameter is typically less biased.

The narrowing of the error band for next-generation detectors is more important for $\cos\theta_1$ and q than for χ_1 because, in the astrophysical model we use, the spin magnitude distribution is more concentrated in a narrow region than the distribution of the inclination angle and mass distribution. As a rule of thumb, the error in a bin goes as $\sqrt{N_b}$, where N_b is the number of events in the bin. The expectation value of N_b can be related to the total number of events N as $\langle N_b \rangle = N p_b$, with p_b being the probability of the bin. Therefore, in the bins with low probability, the number of events needed for the error in the bin to go below a threshold is larger, explaining why the error on the spin magnitude distribution remains large in the tails of the distribution. Moreover, for the inclination angle distribution, the improvement is also driven by the reduction in parameter estimation errors.

Finally, let us stress again that those results were obtained using the LSA for the measurement error and the systematic bias. As explained in Sec. III D, we allow the biased estimate of the parameters to be outside of the physical range, with the idea that this would mimic the likelihood behavior: It is reasonable that the likelihood of an event $\chi_1 \sim 0$ seems to peak at $\chi_1 = -0.1$, and when performing parameter estimation, we would observe a truncated distribution due to the physical prior. However, in some cases, our formula predicts biases that are orders of magnitude outside of the physical range (e.g., $\chi_1 \sim -10$), most likely indicating that the LSA should not be trusted. Indeed, the LSA relies on the quadratic approximation to the likelihood, which should hold only in a region around the peak of the likelihood, with a better agreement at high SNRs. Thus, the reason for our estimates of the bias with *O5* (*A#*) being so much larger than with *XG* might also be due to the invalidity of the LSA in some cases. However, observing more of the inspiral certainly contributes, as discussed in more detail in Sec. VI. Overall, we expect the results shown here for *XG* to be the most reliable.

3. Which binaries are likely biased?

Having explored the effect of waveform systematics on the full detected LVK-like population and studied the

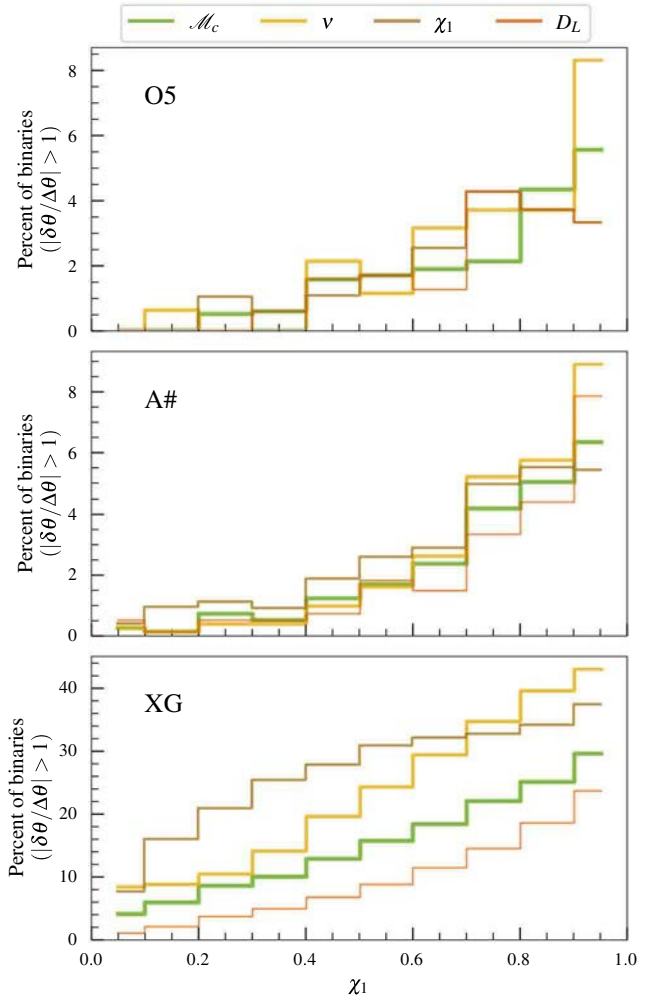

FIG. 8. Percentage of binaries in each χ_1 bin with $|\delta\vartheta/\Delta\vartheta| > 1$ for different parameters. The top, middle, and bottom panels show the percentages for the *O5*, *A#*, and *XG* networks, respectively.

inferred population properties via hierarchical Bayesian inference, we turn our attention to the subset of binaries with significant parameter biases. In the following, we identify the properties of binaries that have a greater susceptibility to systematic biases. To accomplish this goal, we explore the dependence of the systematic bias as a direct function of the binary parameters. In Fig. 8, we show the percentage of binaries in each χ_1 bin with $|\delta\vartheta/\Delta\vartheta| > 1$ for various parameters, such as the chirp mass, symmetric mass ratio, primary spin magnitude, and luminosity distance. It is immediately clear that the number of binaries with biased parameters increases with increasing χ_1. Notice that for current detectors and their upgrades, only a tiny fraction of binaries ($\lesssim$1%) have biased parameters when $\chi_1 < 0.4$. Even for highly spinning binaries, the biased percent is less than 10%. In contrast, we observe that a relatively large fraction (10%–25%) of the binaries have biased parameters in the *XG* network even when $\chi < 0.4$.

Before ending this section, we remark that the results for the systematic biases of the LVK-like population have been obtained by comparing two state-of-the-art quasicircular, spin-precessing, multipolar waveform models. However, when assessing the accuracy of the waveform models, more robust and definitive results can be achieved when comparing models to NR waveforms. We plan to carry out such a study in the near future, although it will be limited by the number of NR waveforms and their length.

V. SYSTEMATIC BIASES ACROSS BINARY PARAMETER SPACE

In Sec. IV, the general properties of systematic bias across an LVK-like BBH population were explored. Here, we consider an agnostic BBH population in order to explore a wider region of the binary parameter space and identify the regions with greater susceptibility to systematic biases.

We sample uniformly in the total redshifted mass, $M^z \in \mathcal{U}(10, 200)[M_\odot]$, and inverse mass ratio, $1/q \in \mathcal{U}(1, 30)$. However, we impose a constraint on the mass of the lighter object, $m_2 \geq 5M_\odot$, and only select those binaries that satisfy this constraint. This approach results in a nonuniform distribution in the two masses. Regardless, in this section, our interest is not in any particular distribution of parameters but rather in the coverage of the parameter space. Exploring the region of large inverse mass ratio is interesting since the waveform models considered here are expected to differ more in this part of the parameter space due to differences in the models' calibration. For SEOBNRv5PHM, SXS NR simulations for aligned-spin systems at $1/q \geq 15$ and a nonspinning simulation at $1/q \geq 30$ were included in the calibration of the model, and second-order gravitational self-force information was incorporated, improving the reliability of the model at large $1/q$ [52]. An example of the different behavior of the models for large $1/q$ is shown in Fig. 21 of Pompili *et al.* [50], where the differences in parameter recovery for a large inverse mass-ratio NR simulation is shown between the aligned-spin versions of SEOBNRv5PHM and IMRPhenomXPHM, with SEOBNRv5PHM being more reliable in the recovery of the parameters.

Because of the lack of calibration to precessing NR waveforms in both waveform models, we expect waveform models to have greater differences for large χ_p values as evident from the mismatch plot of Fig. 10 in Ramos-Buades *et al.* [51]. Therefore, it is of interest to understand the behavior of systematic biases in these parts of parameter space. Thus, we create a sample of binaries having a uniform distribution in χ_p (the LVK-like population, in turn, disfavors large values of χ_p). For this purpose, we generate a large set of samples for the spin magnitudes and tilt angles from the precessing prior and retain a subset of these samples such that the resulting distribution in χ_p is

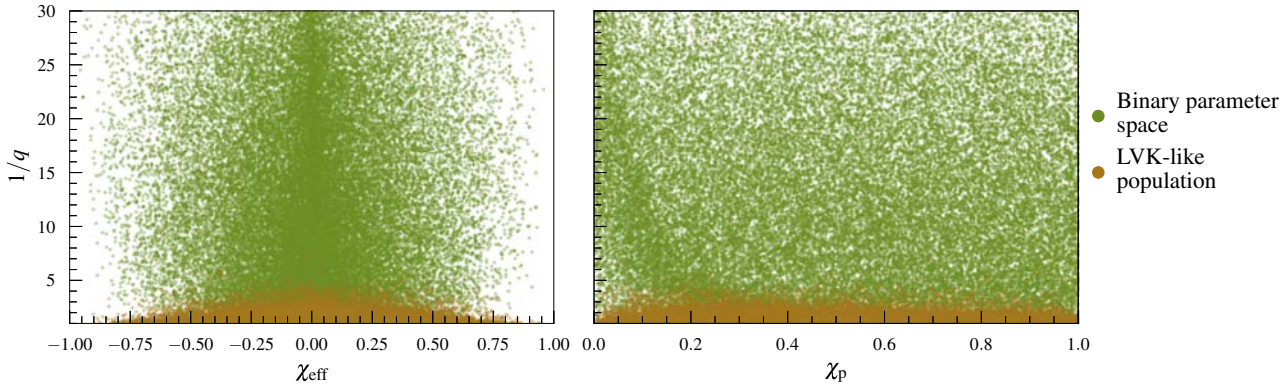

FIG. 9. Comparison of the LVK-like population and distribution of exploratory binaries.

uniform. This selection procedure has a negligible effect on the distribution of spin magnitude and tilt of the secondary companion while giving greater weight to large and in-plane spin for the primary companion. We draw 50,000 binaries using this procedure to cover the binary parameter space. A comparison in the $1/q - \chi_{\rm eff}$ and $1/q - \chi_p$ planes between the LVK-like and the agnostic population is shown in Fig. 9.

The distributions of all other parameters are the same as for the LVK-like population, except for the distance, which is kept fixed at $D_L = 235$ Mpc while computing the measurement errors and biases. However, errors have a simple scaling with the distance for given redshifted masses while the biases remain unaffected, and we use this scaling to obtain results at other distances.

A. Bias horizon

We compute the biases and the measurement errors on the parameter set $\boldsymbol{\vartheta}$ for the 50,000 binaries considered in this section under the LSA, as was done in Sec. IV. The ratio $\delta\vartheta/\Delta\vartheta$ of systematic errors to statistical errors is a function of the distance to a binary through the dependence of the statistical error $\Delta\vartheta \sim D_L$. We exploit this dependence to calculate the distance at which the ratio $\delta\vartheta/\Delta\vartheta = 1$ for any given parameter. Since the measurement errors for a binary with given redshifted parameters increase with its distance, systematic biases will become less important the farther away the binary is located. Hence, the distance at which $\delta\vartheta/\Delta\vartheta = 1$ is a measure of the bias horizon, i.e., the maximum distance up to which systematic biases dominate statistical errors. Given that the biases and errors for each binary parameter are different, the bias horizons for different GW parameters are also different.

We show the bias horizon for the D_L parameter in Fig. 10 for the *O5* (left) and *A#* (right) networks, while that for the *XG* network is shown in Fig. 11. The distribution of the 50,000 binaries in the 4D space given by $\{M^z, 1/q, \chi_{\rm eff}, \chi_p\}$ is illustrated by projecting them into 2D subspaces. The bias horizon for each binary is shown by the color bar. The bias horizon increases with increasing detector sensitivity, implying that systematic biases will be prevalent up to greater distances. In particular, the majority of the binaries in the *XG* network have D_L bias horizons exceeding 25 Gpc ($z \approx 3$), the distance around which the first stars formed. This finding is qualitatively different from the conclusions of Sec. IV B, in particular, Fig. 6, where all the binaries have $z \leq 3$ but around 75% of them are not systematic-error dominated. This result is due to the different distributions of the parameters in this section compared to the LVK-like population. Particularly, the majority of the binaries considered in this section are highly asymmetric with large redshifted masses. Moreover, a uniform distribution in χ_p leads to a large fraction of highly spinning binaries. From Fig. 8, we gather that systematic biases are more prevalent for such binaries. Note that the boundary in the $\chi_{\rm eff} - \chi_p$ space is physical, and it is a result of the maximum value of the spin magnitude being 1. When comparing the biases for a given binary for different detector networks, it is important to keep in mind that a larger value of the ratio $\delta\vartheta/\Delta\vartheta$ for a better detector network need not necessarily be due to a bigger $\delta\vartheta$ but rather a much more precise measurement (see Sec. IV B for further discussion).

We also observe that the importance of systematic biases depends on the parameter space inhabited by the binaries. For instance, from the $\chi_{\rm eff} - \chi_{\rm p}$ space, it is clear that binaries with small spins have smaller D_L bias horizon compared to binaries with large $\chi_{\rm eff}$ and/or $\chi_{\rm p}$, with the binaries lying on the parameter space boundary having the largest D_L bias horizon. Similarly, one can also observe that, for positive $\chi_{\rm eff}$, the bias horizon at large $\chi_{\rm p}$ is higher than for $\chi_{\rm eff} < 0$. This result can be intuited from Fig. 13 of Pompili *et al.* [50], which shows that the mismatch is larger for positive $\chi_{\rm eff}$ compared to negative $\chi_{\rm eff}$. Highly precessing binaries with aligned spins have greater D_L bias horizon compared to similarly highly precessing binaries but with antialigned spins. Along the same lines, we also observe that binaries with large total masses and inverse mass ratios have larger D_L bias horizons. Note that we intentionally do not discuss properties relating to the

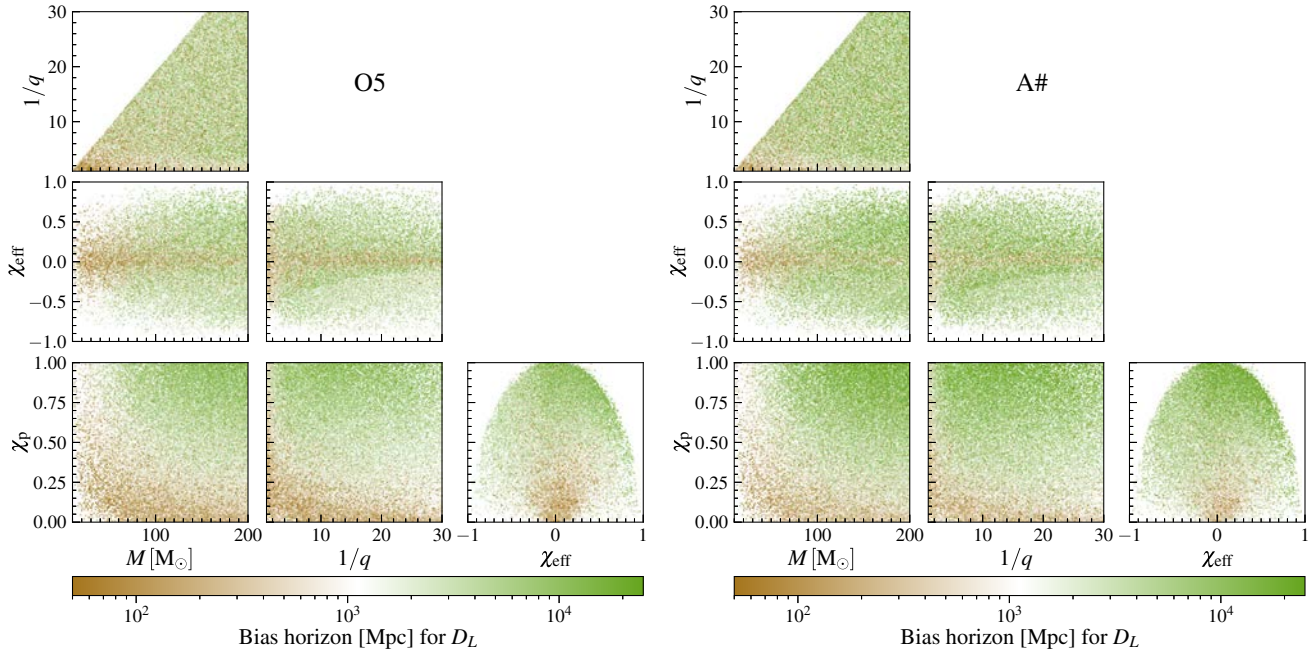

FIG. 10. Distribution of the 50,000 binaries in the parameter space represented in Fig. 9, with the color scale showing the distance to which the D_L parameter is biased ($\delta D_L/\Delta D_L \geq 1$) for the *O5* (left) and *A#* (right) networks. Systematic biases become less important if a binary is at a larger distance since measurement precision decreases with distance. Therefore, a large bias horizon signifies that a given parameter (D_L in this case) is measured well enough for systematic biases to be important even at such large distances. Notice that the binaries are biased up to a greater distance in the *A#* network compared to the *O5* network due to its greater sensitivity and the resultant improvement in measurement precision.

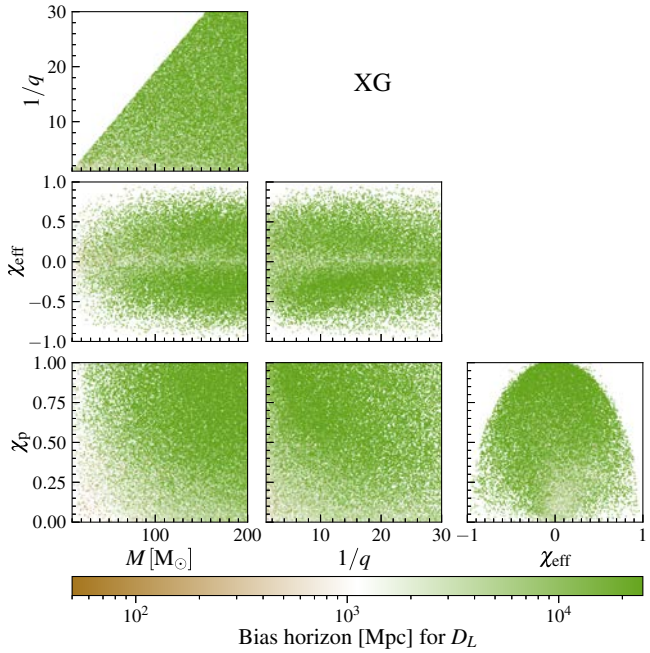

FIG. 11. Same as Fig. 10 for the *XG* network. BBHs observed with the *XG* network are biased up to a greater distance than those observed with either the *O5* or *A#* network, due to its greater sensitivity and the resultant improvement in measurement precision, with a majority of the binaries having a bias horizon greater than or equal to 25 Gpc ($z \approx 3$) beyond which stellar-origin BBHs are not expected to exist.

distribution of binaries in the parameter space since they do not follow any physically motivated parameter distributions. We report the bias horizon for χ_1 in Figs. 26 and 27 of Appendix F. They broadly show the same dependence across the parameter space, although the quantitative values of the bias horizon are different for each binary parameter.

VI. IMPACT OF SYSTEMATICS ON THE SCIENCE OF INDIVIDUAL EVENTS

Until now, we have discussed the effect of systematic biases for the LVK-like population in Sec. IV and explored systematic biases across parameter space within the LSA in Sec. V. In recent years, a number of studies have emphasized the science objectives that can be accomplished using GWs in the near future [29,30,162]. In this section, we consider a few of those science applications. We then handpick three binaries (see Table I) with very relevant science potential and discuss the effects of systematic biases on various science objectives. The majority of the results in this section are obtained using a full Bayesian analysis except where indicated otherwise.

A. Science objectives

In the following, we introduce the science applications that will be considered in this section.

(1) *Cosmology*: There exists a tension, at the level of 4.4σ, between the value of the Hubble-Lemaître parameter H_0 measured at high redshift from the cosmic microwave background [117] and measured using the local distance ladder comprising Cepheid variables and type-Ia supernova [163]. The emergence of GW astronomy provides an avenue for an independent measurement of the Hubble-Lemaître parameter, which could importantly contribute to resolving this tension. Indeed, GWs have already provided multiple independent measurements, albeit not yet at an accuracy to resolve the tension [19,20,164–166].

A measurement of H_0 requires both the luminosity distance and redshift of a source to be estimated. GW observations provide the former, but additional data or assumptions are required to provide the latter. For GW observations accompanied with an EM counterpart, primarily binaries containing NSs, the redshift information is provided by spectroscopic and/or photometric observations of the host galaxy [167,168]. In BNS or neutron-star–black-hole (NSBH) observations, NS tides can be used to provide an independent redshift measurement [169–171] and, thereby, infer H_0 [172–175]. Alternatively, features in the mass distribution of compact binaries can be exploited [176–179]. Finally, a statistical measurement of the redshift using galaxy catalogs [36,180–184] or galaxy cross-correlation techniques [185–187] can be employed.

Here, the focus is on the last technique, which can be applied to BBH systems but relies on an accurate measurement of the distance and sky position. An indirect effect on cosmological inference due to inaccurate determination of the mass distribution will be discussed in the following subsections.

Events that have the smallest volumetric uncertainties are the most informative systems for statistical measurements using galaxy catalogs [20,182]. This feature can be intuitively understood as follows: If there is a single galaxy in the localization volume of a GW event and one assumes that the event originated in a galaxy, then there is unit probability that the identified galaxy hosted the event and the redshift is known as well as the galaxy redshift. On the other hand, if the volumetric localization of the GW event is poor and there are a large number of galaxies that are potential hosts, the redshift distribution would essentially be uniform, obtaining contributions from each possible host. All three events studied here are prime candidates for such a method due to their asymmetrical component masses, which make distance estimates more precise than analogous comparable mass mergers. This feature results from a greater contribution from subdominant harmonics that break the distance-inclination degeneracy. Spin precession also helps in breaking this degeneracy since it mixes different modes in the inertial frame.

(2) *Lower mass gap*: The nature of compact objects with masses between $2M_\odot$ and $3M_\odot$ can have wide-ranging consequences in fundamental physics—from the physics of nuclear matter to primordial BH formation mechanisms—and astrophysics—from hierarchical formation probabilities to the proportion of rapidly spinning NSs. The LVK Collaboration has observed two events where one of the components of the binary unambiguously lies in this mass range—GW190814 [188] and GW200210_092254 [5]. Tidal effects on GW waveforms in highly asymmetric mergers hosting candidate NSs, such as the Binary 1 system, are minimal. Therefore, indirect constraints on the nature of the secondary and the binary's formation history are derived indirectly from their mass and spin measurements.

The existence of ultraheavy NSs has consequences for the nuclear equation of state at a few times the nuclear saturation density [189], with the possibility of nontrivial structures in the speed-of-sound relation and related phase transition phenomena [190,191] or rapidly rotating NSs stabilized against collapse by its rotation [192–195]. On the other hand, BHs in this mass range will inform the primordial BH formation scenarios [196–198] and hierarchical mergers in dense environments [199–203]. It is also proposed that the secondary gains mass due to accretion, either prior to a supernova explosion [204] or following it [205]. In all of these scenarios, accurate measurements of the mass and spin are essential to further the discussion.

(3) *PISN mass gap*: In the mass range of about 50–$120M_\odot$, there is expected to be a dearth of stellar origin BHs because main sequence stars with masses heavier than around $120M_\odot$ have core temperatures that facilitate electron-positron pair production, leading to a decrease in radiation pressure in the core of the star, causing explosive oxygen burning, and a resultant disruption of the entire star. This process, known as a PISN process, does not leave behind a remnant, thereby producing a dearth of BHs above around $60M_\odot$. However, if the mass of the main sequence star is greater than around $250M_\odot$, all the heavy elements undergo photodisintegration, first to alpha particles and then further. This process reduces the radiation pressure, causing

the star to implode and form a BH with mass greater than around $120M_\odot$.

The determination of the boundaries of this mass gap can inform the physics of PISN, such as the $^{12}C(\alpha,\gamma)^{16}O$ reaction rate [35,206], and the role of stellar rotation [207]. Other astrophysical processes—such as mass reversal, mass growth due to accretion, and hierarchical mergers—can result in the formation of BHs that populate this mass gap.

(4) *Spin morphology*: The spin distribution of BHs in binaries provides crucial information on their formation channels. Upcoming observing runs of LVK detectors and XG observatories will measure the spins of compact binaries to ever greater precision, which will help to constrain the spin distribution of astrophysical BBH populations and their formation channels. For instance, binaries formed via isolated evolution tend to have their spins aligned with the orbital angular momentum, while those formed dynamically are likely to have an isotropic spin distribution. Similarly, hierarchical formation is expected to produce larger spins compared to stellar collapse. However, given that the spin measurements are expected to be precise, it is crucial for them to be accurate as well to allow unbiased inference of the source properties of the underlying population. According to our LVK-like population, only 50% of the events detected in *O5* will have $\Delta\chi_1 < 0.8$ ($\Delta\theta_1 < 85°$), while it is $\Delta\chi_1 < 0.1$ ($\Delta\theta_1 < 10°$) in *XG*. We discuss the indirect effects of the spin measurement on the inference of the nature of the secondary component of the Binary 1 system in Sec. VI C 2. In the following, we discuss the systematic biases on the spin for another binary system.

The origin of massive BBHs, particularly those filling the upper mass gap, can be traced using their effective spin $\chi_{\rm eff}$ and spin-precession χ_p parameters. A hierarchical formation mechanism leads to large component spins since the remnant of the previous merger is expected to be spinning.

While the $\chi_{\rm eff}$ and χ_p parameters can broadly inform and differentiate between an isolated and dynamical formation channel, an accurate measurement of the tilts of the two spin vectors with respect to the orbital angular momentum, θ_1 and θ_2, and their relative orientations in the orbital plane, ϕ_{12}, provide detailed knowledge of the formation mechanisms and spin distributions of the BBH merger population. Precessing binaries exist in different spin morphologies [208–211] due to spin-orbit resonances [212]. Therefore, precessing binaries can exist in subpopulations characterized by their spin morphology depending on their tilt angles at formation [213–218]. Recent efforts have been made to better understand the spin-precession dynamics [219] and the ability of current detectors to measure spin-precession effects on the waveform [220]. There have also been efforts to probe whether the BBHs detected by the LVK Collaboration are associated with a particular spin morphology [221], as well as studies on the capability of current detectors at improved sensitivities to distinguish different spin morphologies [222,223].

(5) *Remnant quantities*: The properties of the remnant BH following a BBH merger can be determined from its binary parameters. However, the nonlinear merger makes a fully analytical calculation intractable. As such, estimates of the final mass (M_f) and spin (χ_f) of the remnant include information from NR. Several studies have proposed fits for the remnant quantities for nonprecessing binaries [224,225]. In the precessing case, the final mass is found to agree very well using a nonprecessing formula, but the same does not hold true for the final spin where the in-plane spin components are important [226,227]. Some simple arguments to include in-plane contributions have also been proposed in the literature [228–230]. Such arguments have been used to augment the nonprecessing final spin estimates [224,225]. A surrogate model for the final spin using NR has also been proposed for moderate mass ratios [227]. In this paper, the reported final mass estimates are the average of the nonprecessing fits in Refs. [224,225]. The final spin is computed by averaging the estimates of Refs. [224,225,230] with in-plane spin augmentations for the nonprecessing fits of Refs. [224,225].

Such estimates of the remnant quantities are important not only for modeling the ringdown in inspiral-merger-ringdown waveforms but also for probing the dynamics of the merger and the nature of the remnant. The ringdown of a Kerr BH is described by quasinormal modes whose complex frequencies are determined from the properties of the final stationary BH. Therefore, the ringdown signal can be used to estimate the mass and spin of the final remnant independently. The remnant quantities can be affected by deviations from GR that can modify the radiated energy and angular momentum, as well as an exotic remnant, which will have a modified spectrum, resulting in inconsistent estimates from the ringdown signal and NR-inspired fits. Such consistency tests have been performed for GW150914 [21,231] and GW190521 [232,233]. With improving detector sensitivities, such tests form an important part of null hypothesis tests of GR.

(6) *Maximal BH spins*: Obtaining accurate spin measurements from astrophysical BHs is a nontrivial

endeavor for which different electromagnetic approaches exist. In the context of stellar-mass BHs, it is possible to apply the continuum fitting method or reflection spectroscopy to x-ray observations [234,235]. However, modeling the astrophysical environment is nontrivial and can introduce systematic effects. GW measurements provide a novel way to measure BH spins and, because of their vacuum environment, might be easier and more robust to model. While the spin distribution of BBH systems can provide valuable information about their formation channels, an individual event with very high spins, especially if close to extremal Kerr, would be of great scientific interest. For instance, because of the cosmic censorship hypothesis, no naked singularities should exist, which implies that BHs should not spin above $a > 1$. Moreover, quasinormal modes of a rapidly rotating BH become long-lived, which could lead to turbulence phenomena [236].

B. Handpicked binary black holes

We pick three asymmetric and precessing systems, out of which one has low total mass and two have large total masses. These systems have some precedence in the current LVK GWTC, though they are not the most common events. We pick them due to their science potential and to explore the parts of the parameter space that are expected to be considerably affected by systematic biases. The parameters of these systems, their SNRs in the different detector networks, and their science objectives are listed in Table I. We now discuss the properties of these systems and their analogs in the LVK GWTC.

(i) Binary 1: highly asymmetric, spin-precessing, low total-mass binary. This system is modeled after GW190814 and has masses, distance, and inclination compatible with the measured values [188]. While GW190814 has no measurable spin precession, the system we consider is highly precessing. It is a reasonable choice because a dynamical capture or hierarchical merger is widely accepted as a possible formation channel for GW190814, and this channel can produce highly precessing binaries. The merger rate of such systems is estimated to be 7^{+16}_{-6} Gpc^{-3} yr^{-1}. Hence, even a pessimistic merger rate of 1 Gpc^{-3} yr^{-1} equates to more than 400 such mergers within a redshift of 3 each year. With planned and future observatories, a handful of these mergers will be observed with large SNRs.

(ii) Binary 2: rather asymmetric, nonprecessing, high total-mass binary. Among the LVK observations, the candidate events GW190403_051519 [6] and GW200208_222617 [5] have properties similar to this system. Both the candidates are asymmetric binaries with large effective spins, indicating that the spins are aligned to the orbital angular momentum. The median of the posterior distribution of the primary spin magnitude for GW190403_051519 is $\chi_1 = 0.89$ while the probability that the primary spin

TABLE I. Properties of the three systems that we study using Bayesian analysis in Sec. VI. The top row lists the intrinsic parameters of the systems, the middle row enumerates the SNR of the systems in the three detector networks used, and the last row illustrates the science applications associated with each system.

		Binary 1: highly asymmetric, spin-precessing, low total mass	*Binary 2:* rather asymmetric, nonprecessing, high total mass	*Binary 3:* rather asymmetric, spin-precessing, high total mass
Parameters	$m_1[M_\odot]$	23.2	61.8	61.8
	$m_2[M_\odot]$	2.6	9.5	9.5
	χ_1	0.7	0.9	0.9
	χ_2	0.4	0.8	0.3
	$\theta_1[^\circ]$	40	0	140
	$\theta_2[^\circ]$	40	0	120
	$\chi_{\rm eff}$	0.51	0.89	−0.43
	$\chi_{\rm p}$	0.45	0	0.77
Network SNR	*O5*	75.3	222	119
	A#	137	405	219
	XG	1040	3150	2490
Science cases	Cosmology	✓	✓	✓
	Lower mass gap	✓	✗	✗
	PISN mass gap	✗	✓	✓
	Spin morphology	✓	✗	✓
	Remnant quantities	✓	✓	✓
	Maximally spinning secondary	✗	✓	✓

of GW200208_222617 is $\chi_1 > 0.8$ is 51%. The primary components for both candidates also have a considerable probability of being in the upper mass gap. While these events were marginal detections with moderate astrophysical significance, their science potential is immense, and confident detections of similar systems in the future will be invaluable. From a modeling perspective, it is also an interesting region of the parameter space to probe because different state-of-the-art waveform models have significant mismatches for binaries with large aligned spins. This feature is only partly due to the relatively sparse NR coverage in this region of the parameter space. Binaries with spins aligned with the orbital angular momentum merge at higher frequencies [237], where the accuracy of the PN approximation—used as a baseline in all models—degrades. As a result, this limits the accuracy that models can achieve even after calibrating to NR, assuming no fitting errors.

(iii) Binary 3: rather asymmetric, spin-precessing, high total-mass binary. This system has the same parameters as Binary 2 except for the spins, which are now misaligned with respect to the orbital angular momentum. The origin and associated formation channels for such systems through stellar evolution are highly uncertain [238–242]. Nonetheless, the current observations provide sufficient evidence for the existence of such systems; therefore, it is essential to understand whether current waveform models have the necessary accuracy to study such mergers.

For all three events, we perform a full Bayesian analysis, placing them at a luminosity distance of 235 Mpc. While BBH mergers at this distance are certainly possible and have been observed, they are few in number, and the majority of events will originate from larger distances. On the other hand, the SNR for events that are sufficiently far away will be small, and the systematic biases will be inconsequential compared to the measurement errors. It is therefore desirable to know the maximum distance up to which a given science objective will be affected due to biases in parameter estimation. However, it is infeasible to repeat the full parameter estimation calculation for multiple distances. For that reason, we model the posterior distribution using its median and covariance, which is equivalent to the LSA approximation in the large SNR limit. In this limit, the bias is independent of the strength of the signal while the covariance increases with the square of the distance. We compute the bias horizon for the three sources, and they are shown in Fig. 12. A cursory look at the SNRs required for a parameter to be biased reveals that biases could be present for much smaller SNRs than what is

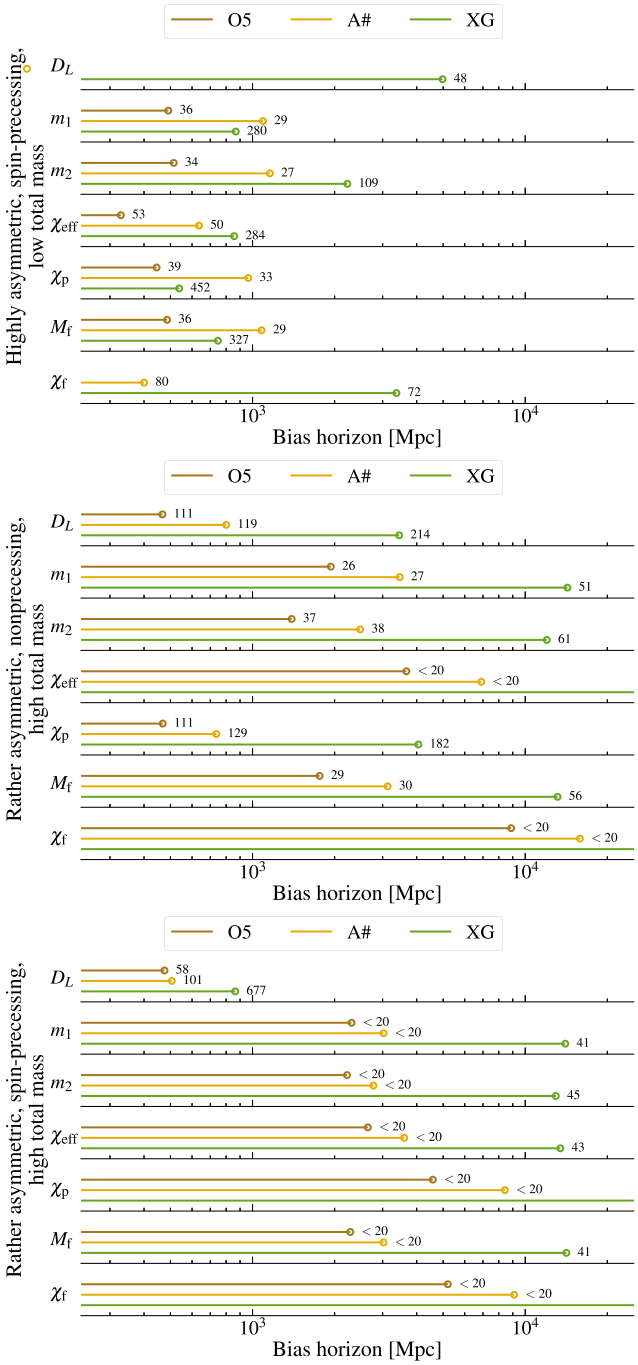

FIG. 12. Bias horizon for the most relevant parameters of the three systems (from top to bottom panels) for the three different networks. The colored lines indicate up to which distance the systematic error of a parameter is outside the 90% credible interval of the posterior and thus describes biased parameters. The circles at the end of each line indicate the bias horizon, which is the distance where it is at 90%. The number at the end of each line reports the SNR at that distance. Note that there is a cutoff at 25 Gpc (lines without circle) and that we only explicitly show SNRs down to 20.

simulated here. Below, we will discuss the implications of these results in the context of various science applications. The posteriors for events with low SNR are not well approximated by a normal distribution. The simple scaling argument employed here fails for such cases. Therefore, for parameters for which the SNRs at the bias horizon are < 20<20, we do not quote the exact value but rather denote it with an inequality sign. Similarly, we do not show the exact projected distances if they are larger than 25 Gpc ($z \sim 3$). Note that there is an implicit assumption in this scaling argument that the detector-frame masses are constant. For large distances, the source-frame masses would be materially smaller. Therefore, the binaries considered may no longer be appropriate for science objectives related to the source-frame mass. If, in turn, these binaries are placed at a larger distance while keeping the source-frame quantities constant, they will have a larger detector-frame total mass and, consequently, fewer GW cycles in the detectable band. Since the differences between various waveform models are greater closer to the merger, one expects that the bias will be larger for such systems compared to the scaled binaries. For instance, it can be seen from Figs. 10 and 11 that the bias horizon is generally greater for larger total detector-frame masses. In this sense, the reported scaled numbers can be considered conservative estimates.

C. Impact on science objectives

In the rest of this section, we investigate the effect of systematic bias on various science objectives. Even though we isolate different science objectives to discuss them individually, this is an artificial separation since the science is interconnected and so are the biases. Therefore, some of the same results may be discussed in different sections in slightly different ways. For instance, a biased measurement of the maximum NS mass not only has consequences for nuclear physics but also for cosmology. Similarly, different representations of the same parameter space may be helpful in highlighting different aspects of the science. For instance, we refer to the magnitude and tilt of the spin vector when discussing lower mass-gap events, effective spin and spin-precession parameters when discussing astrophysical formation channels of heavy BBHs, and the relative orientations of the spin vectors when talking about spin morphologies. They are all different slices of the same spin space. However, the different parametrizations are useful for discussing and highlighting different aspects of the spin space.

1. Cosmology

The luminosity distance and sky localization posteriors for the three simulated systems in the different detector networks are shown in Fig. 13. An illustration of an

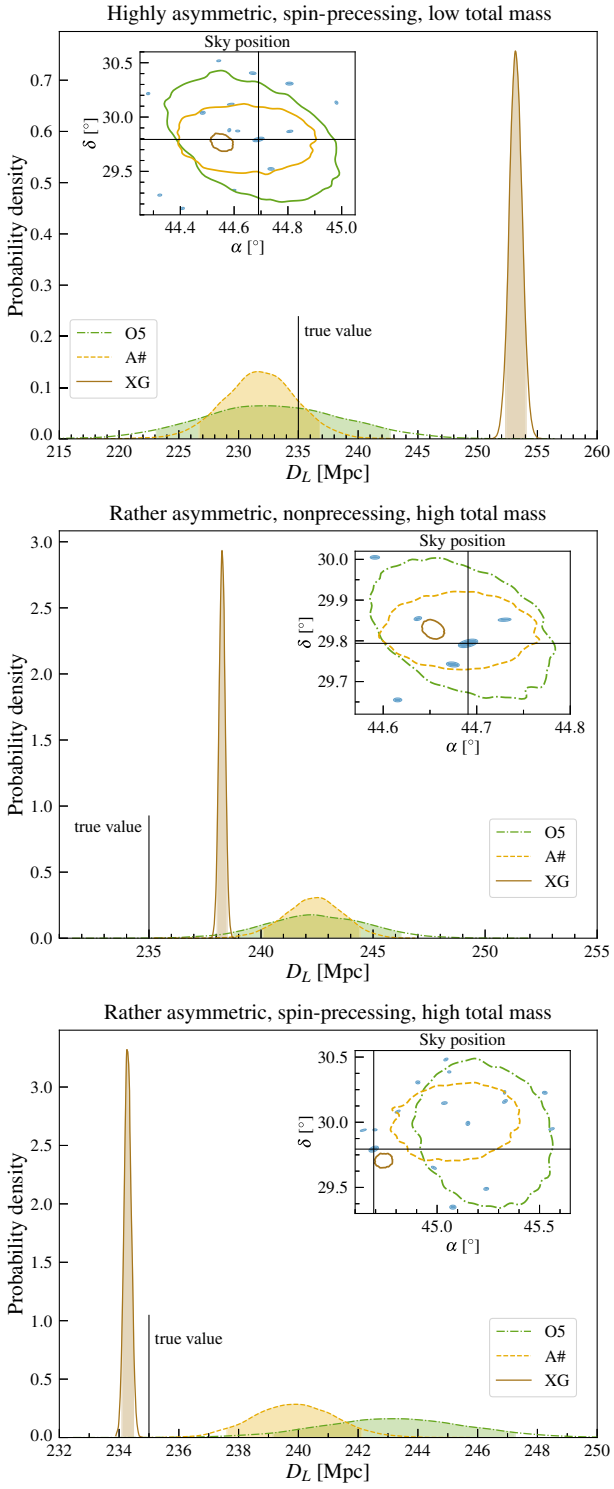

FIG. 13. Distance posteriors for the Binary 1, Binary 2, and Binary 3 systems (from top to bottom) in the three detector networks considered in this study, The shaded region shows the 90% credible interval. The contours in the inset show the 90% credible region for the corresponding sky position measurement. The true values are denoted by black lines.

expected distribution of galaxies in a volume uncertainty region is depicted as blue ellipses. The volume uncertainty includes galaxies with redshift between $z_{\rm min} = H_{0,\rm min}/D_{L,\rm max}$ and $z_{\rm max} = H_{0,\rm max}/D_{L,\rm min}$, where $H_{0,\rm min} = 35\ {\rm km\,s^{-1}\,Mpc^{-1}}$ and $H_{0,\rm max} = 140\ {\rm km\,s^{-1}\,Mpc^{-1}}$, and $D_{L,\rm min}$ and $D_{L,\rm max}$ are the edges of the 90% credible interval of the D_L posterior. The sky patch is the size of the inset panel that contains the sky location posterior. The areal density of luminous L* galaxies in the local Universe is taken to be 0.07 $\rm deg^{-2}$ at a distance of 100 Mpc [143].

While the localization volume of the Binary 1 system (upper panel) is accurately measured for current detectors at design sensitivity and their next upgrade, both the distance and the sky position are biased for the *XG* network. With a sky position measurement precision of less than 0.1 $\rm deg^{-2}$ at 90% credibility, at most a single galaxy is expected to be in a volumetric cone up to the distance to the event [36]. An inaccurate GW measurement will completely miss the host galaxy in this specific example or result in the identification of the wrong host in general. From Fig. 12, it is clear that such a system will give a biased distance estimate for sources at distances up to around 35 times farther, having an SNR greater than or around 30 in the XG network.

For the Binary 2 system (middle panel), while *O5* and *A#* networks accurately recover the sky position again, the distance is biased for all three networks. A bias in the distance measurement towards larger values will cause the H_0 measurement to be systematically biased towards smaller values. As in the case of the previous system, the precision in the sky position measurement for the *XG* network implies the existence of a single galaxy in the volume uncertainty region on average. Figure 12 tells us that the distance bias is important in *O5* and *A#* networks for SNRs around 70, which implies distances of around 800 Mpc and 1300 Mpc, respectively, while the distance at which $|\delta\vartheta/\Delta\vartheta| = 1$ for the *XG* network is around 5700 Mpc. Even at a distance that is around 7 (around 4) times the maximum distances in *O5* (*A#*), the event in *XG* will have an SNR of 130.

The measurement of both the distance and sky position is inaccurate for the Binary 3 system (lower panel) for all three detector networks under consideration. However, the distance bias, particularly for the *XG* network, is smaller than the previous systems. On the other hand, the sky position is biased even for *O5* and *A#* networks. Because of the smaller distance biases, we see from Fig. 12 that the bias horizon for the *XG* network is also smaller, resulting in a very loud SNR around 400. Meanwhile, since the SNRs in the three detector networks for this system are smaller than the Binary 2 system, an SNR of around 35 in the *O5* network is sufficient for a material distance bias.

Given that a 2% measurement of the Hubble constant will resolve the Hubble tension, all three events in the *XG* network can single-handedly resolve the tension, while the Binary 2 and the Binary 3 events can do so even in an *A#* network, at a simulated distance of $D_L = 235$ Mpc.

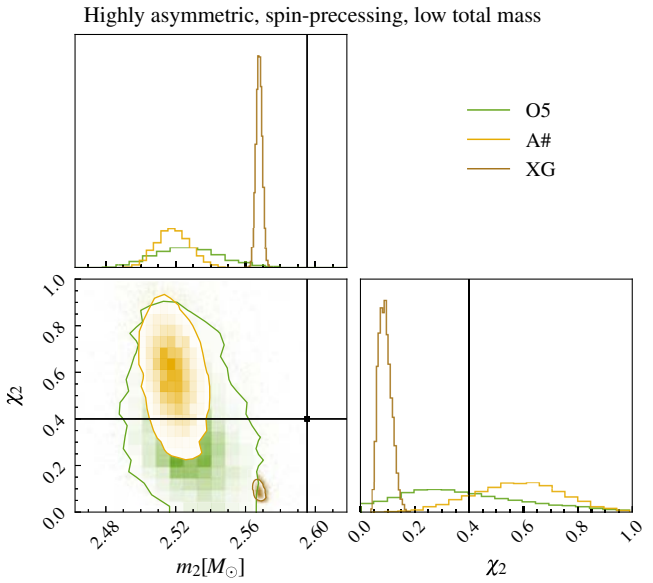

FIG. 14. Corner plot of the posterior distributions for the mass and dimensionless spin magnitude of the secondary companion of the Binary 1 system in the three detector networks in which the system is simulated. The shaded region in the 1D posteriors and the contours in the 2D space denote the 90% credible region. The black lines show the true injected value. A smaller minimum frequency for the *XG* network ($f_{\rm low}$= 5 Hz) compared to the *O5* and *A#* networks ($f_{\rm low}$= 10 Hz) results in a smaller bias for the mass.

2. Lower mass gap

In Fig. 14, the recovered distributions for the mass and dimensionless spin magnitude of the secondary companion of the Binary 1 system in the three detector networks are shown. Since this is a highly asymmetric merger, the spin of the secondary is poorly measured, as is clear from the χ_2 posteriors in the *O5* and *A#* networks. Nevertheless, the measurement precision improves significantly in the XG network, both due to its broad sensitivity improvement and a smaller minimum frequency. However, the measured value is materially inaccurate, predicting a much lower value than the injection. This result would affect inference on hierarchical formation channels and accretion-induced mass growth of a NS since both predict a large component spin. Instead, primordial BH formation scenarios that predict small spins would be favored [198]. The measurement bias in the component mass would also have impacts on constraints on the speed-of-sound relation in NSs [190,191] and inference on the effect of rotation on NS radii [195].

More directly, the upper edge in the mass distribution of astrophysical NSs can be used to determine the redshift and, thereby, estimate H_0 [176–179]. Hence, an inaccurate

determination of the edge of the mass distribution could also bias cosmological parameter estimation. This effect can be quantified using some simple calculations. Assume that the NS mass distribution is uniform and that this system lies at the edge of that distribution. It can be shown that the error in the determination of the upper edge of the mass distribution is given by $\Delta m_{\rm max} = \max\left[\sqrt{\sigma_m R_o/(N-1)}, R_o/(N-1)\right]$, where $R_o = (m_{{\rm max},o} - m_{{\rm min},o})$, σ_m is the typical mass measurement uncertainty for systems near the upper edge, N is the number of observations, and $m_{{\rm max},o}$ and $m_{{\rm min},o}$ are the maximum and minimum observed NS masses, respectively [243,244]. The posterior peak of the secondary companion in *A#* network is around $0.08M_\odot$ away from the true value. It is easy to calculate from the above equation that with $N \sim 20$ observations and $m_{\rm min} = 1M_\odot$, the measurement error in determining the upper edge of the mass distribution is $\Delta m_{\rm max} \sim 0.08M_\odot$.

The merger rate of a Binary 1 system within a distance of 235 Mpc is 1 every 3 years. Several studies have estimated that upcoming GW observing runs are expected to detect tens of BNS mergers per year [27,32]. Therefore, a systematic bias in the mass and spin measurement can bias the inference of NS properties and cosmological parameters in the near future.

3. PISN mass gap

We report the measurements of the primary mass for the Binary 2 and Binary 3 systems in Fig. 15 in the three detector networks. While the biases in the aligned spin binary are less than in the precessing case, the measured values for both are significantly different from the injected value.

Furthermore, knowledge of the BH mass spectrum, particularly the edge of the mass distribution, can be used for cosmological inference as with the NS mass distribution. Following similar calculations and considering the simplified case where the BH mass distribution follows a power law, $p(m) \propto m^{\alpha_m}$, with power law index $\alpha_m = -3.5$ and sharp cutoffs at $m_{\rm min} = 10M_\odot$ and $m_{\rm max} = 65M_\odot$, it can be shown that the error in the determination of the upper cutoff is given by

$$\Delta m_{\rm max} = \frac{\left(m_{{\rm max},o}^{\alpha_m+1} N - m_{{\rm min},o}^{\alpha_m+1}\right)^{\frac{1}{\alpha_m+1}}}{(N-1)^{\frac{1}{\alpha_m+1}}} - m_{{\rm max},o}, \quad (23)$$

until $\Delta m_{\rm max} \sim \sigma_m$, the individual event mass-measurement uncertainty. This uncertainty for $N = 10^3$ events is then $\Delta m_{\rm max} \approx 3M_\odot$. From Fig. 15, it can be observed that the systematic error becomes dominant, even for the *O5* network, which is expected to observe $\mathcal{O}(10^3)$ events every year.

4. Spin morphology

In Fig. 16, the posterior distributions of $\chi_{\rm eff}$ and $\chi_{\rm p}$ are depicted for the Binary 3 system in the three detector networks. It is immediately clear that the posterior distributions for both spin parameters and in all three detector networks are far from the true value. However, it is also noticeable that the absolute value of the bias is larger for the current networks and their upgrades compared to the future *XG* network. Spin precession measurements are enabled by low-frequency sensitivity since the modulations in the GW amplitude due to spin precession occur on these timescales.

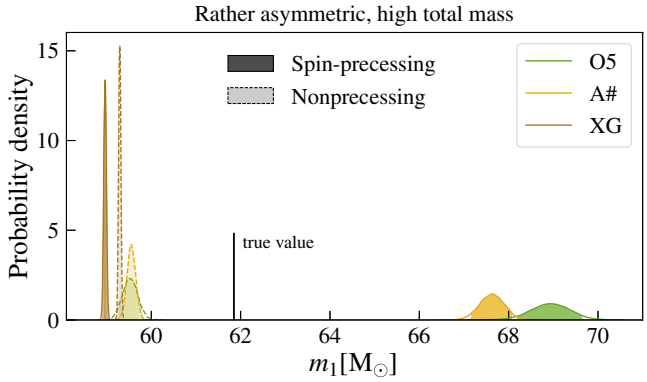

FIG. 15. Posterior distributions of the primary mass for the Binary 2 (solid lines) and Binary 3 (dashed lines) systems in the three detector networks. The true injected value is shown by the vertical black line. The filled regions depict the 90% credible interval.

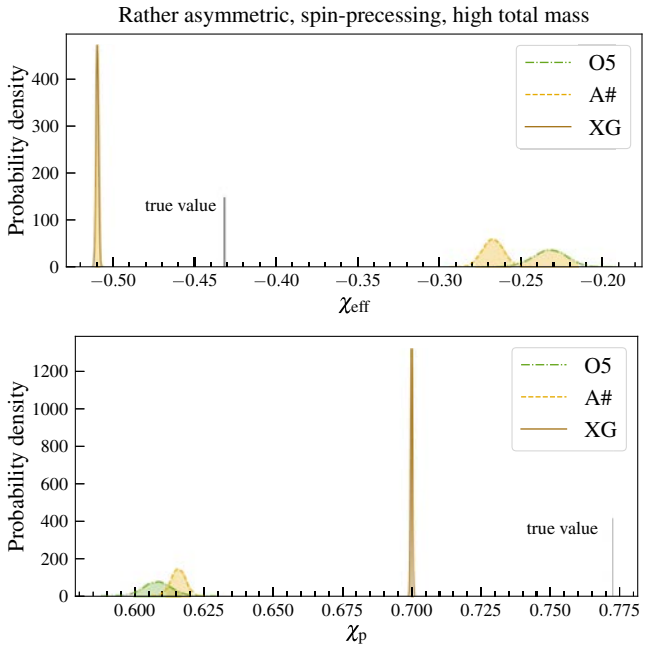

FIG. 16. Posterior probability distribution of the $\chi_{\rm eff}$ (left) and χ_p (right) parameters for the Binary 3 system in the three detector networks. The true value of the injection is shown by the black vertical line. The filled regions show the 90% credible interval.

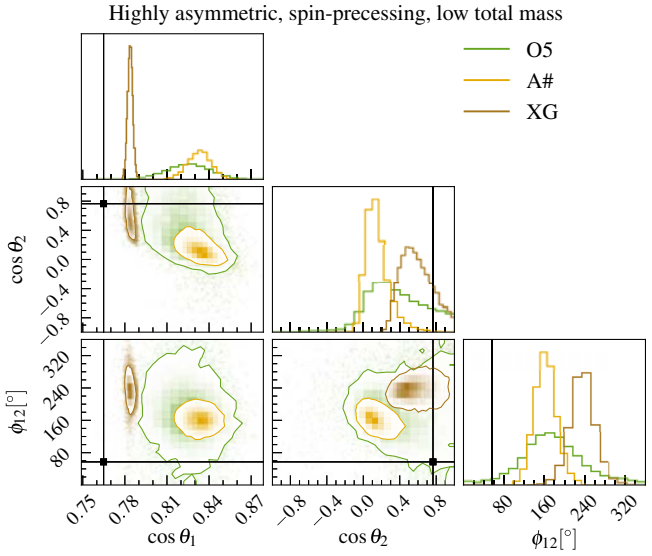

FIG. 17. Posterior distributions of the cosine of the tilt angles $\cos\theta_1$ and $\cos\theta_2$ and the relative in-plane spin angle ϕ_{12} for the Binary 1 system. The black cross-hairs show the true injected value.

Moreover, different waveform models agree to a greater degree at lower frequencies because this regime is closely informed by PN calculations in the various models. Therefore, a smaller minimum frequency and a better low-frequency sensitivity enable a more precise and accurate measurement of the $\chi_{\rm eff}$ and $\chi_{\rm p}$ parameters in the *XG* network. Even so, the parameters are significantly biased.

The posterior distributions of the parameters characterizing the spin morphology for the Binary 1 and Binary 3 systems are reported in Figs. 17 and 18. Let us analyze the Binary 1 system first. The tilt angles determine which morphology the binary falls into. We observe that both the tilt angles are recovered inaccurately. While the median bias for θ_2 is greater than that for θ_1, with the median value of θ_2 being 1.5–2 times the injected value, the poor measurement accuracy due to the highly asymmetric and low total mass nature of the binary results in the median value of θ_1 being farther away from the true value, when expressed as a multiple of *sigma* (the measurement error). The parameter ϕ_{12}, which characterizes the spin morphology, is also heavily biased, with the best-fit median values 2–3 times the injected value depending on the detector network.

Now, considering the Binary 3 system, it is observed that the measurement accuracy increases due to the greater total mass and a resultant higher SNR. Similar to the Binary 1 system, the systematic bias in the θ_1 is smaller in absolute terms compared to θ_2, whose median inferred value is 0.4–0.6 the injected value. Interestingly, ϕ_{12} is less biased for the *O5* and *A#* networks. The estimates in the *XG* network are not only more biased but also completely different from the *O5* and *A#* networks, revealing the sensitivity of the measurement to the minimum frequency. Specifically, the measured value of ϕ_{12} in the *XG* network is consistent with 0° while the injected value is around 200°.

In this section, we do not explore the Binary 2 system in detail because it is a nonprecessing binary. Nevertheless, we report that the spin-precession parameter $\chi_{\rm p}$ is accurately recovered to be 0, and systematic biases do not lead to an inaccurate inference of an aligned-spin system as a spin-precessing system.

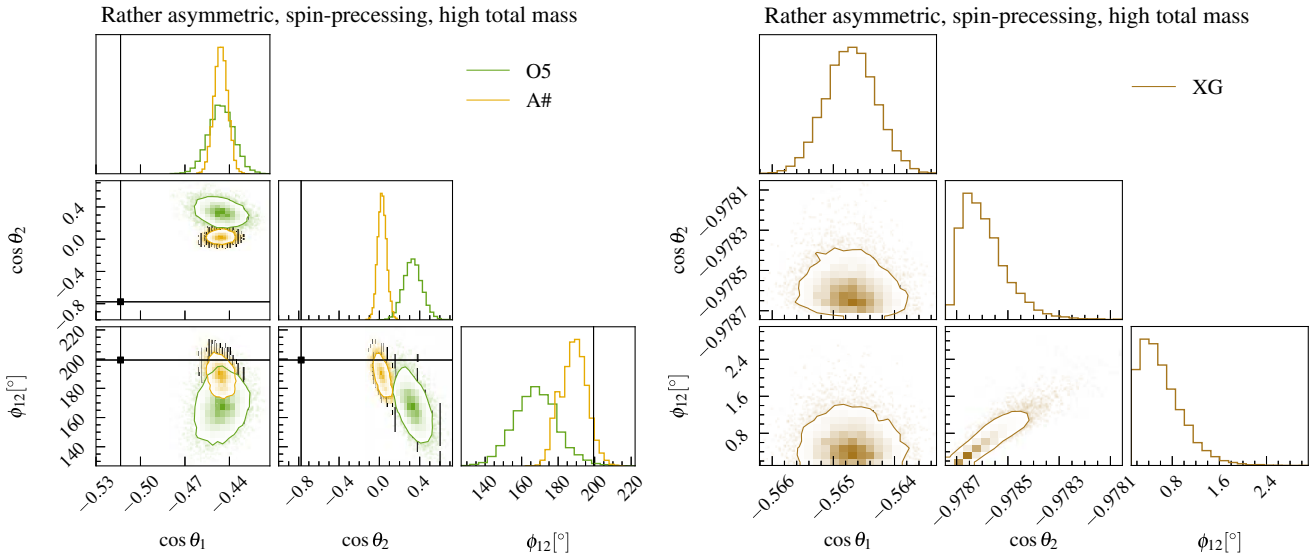

FIG. 18. Posterior distributions of the cosine of the tilt angles $\cos\theta_1$ and $\cos\theta_2$ and the relative in-plane spin angle ϕ_{12} for the Binary 3 system in the *O5* and *A#* networks (left) and the *XG* network (right). The black cross-hairs show the true injected value. The *XG* network is shown separately because it is significantly different from the other networks, and incorporating them in the same figure distorts its appearance.

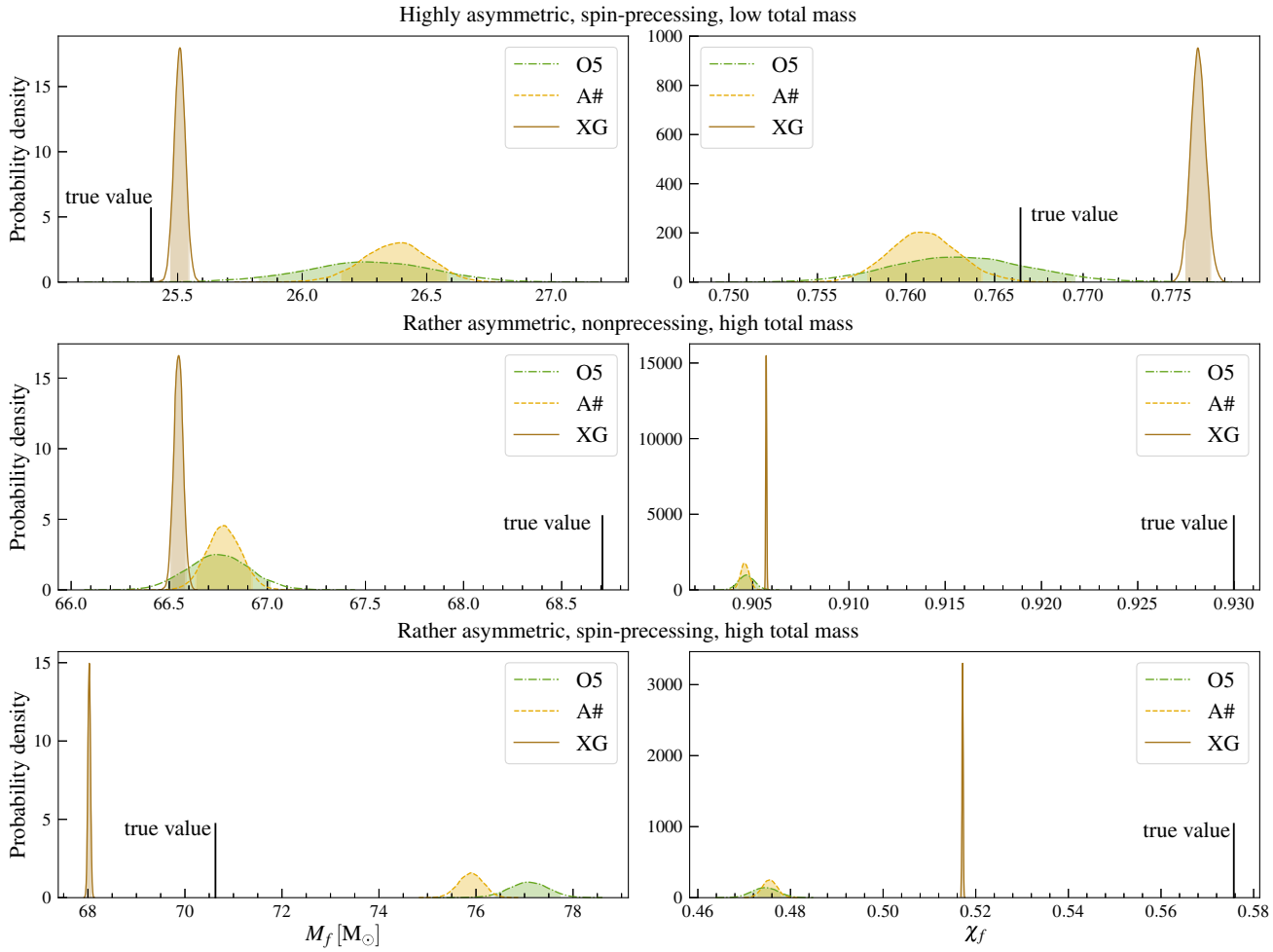

FIG. 19. Final mass (left) and final spin (right) posteriors for the Binary 1, Binary 2, and Binary 3 systems (from top to bottom) in the three detector networks considered in this study. The true values are denoted by black lines.

5. Remnant quantities

We report the inferred remnant quantities for the Binary 1, Binary 2, and Binary 3 systems (top to bottom) in the three detector networks in Fig. 19. Unsurprisingly, there is significant bias in most of the cases. Since we reported the biases in various mass and spin quantities in the preceding sections, the biases in the remnant quantities are expected. It can also be observed from the figure that the remnant properties of high-mass systems are better measured compared to the low-mass system. Consequently, the biases for Binary 2 and Binary 3 as a multiple of the measurement error are larger.

We illustrate the impact of these biases on inspiral-merger-ringdown consistency tests using GW150914 as a gauge. The 90% confidence interval ranges on the oscillation frequency and damping time of the fundamental quasinormal mode of GW150914, with a ringdown SNR of around 8.5, are about 29 Hz and 5.6 ms, respectively [21]. In comparison, the biases in the fundamental quasinormal mode calculated from the remnant quantities range between around 0.2 Hz and 30 Hz for the oscillation frequency and around 0.02 ms and 0.6 ms for the damping time across the different systems and networks considered here. Therefore, the consistency tests would fail for most of the cases considered here.

6. Maximal BH spin

In Fig. 20, the posterior distributions on the secondary dimensionless spin magnitude χ_2 are reported for the Binary 2 (left panel) and Binary 3 (right panel) systems. We observe that χ_2 is biased for both systems. While the effect of χ_2 on the GW waveform is subdominant due to the high mass asymmetry and smaller magnitude compared to χ_1, which results in a poorer measurement of this parameter compared to χ_1, the significantly biased recovery hints at systematic differences in the modeling of the effects of the parameter on the GW waveform. Perhaps, more interestingly, the secondary is measured to be maximally spinning, which is of great importance because such a measurement in a real event would be revolutionary. We do not show the

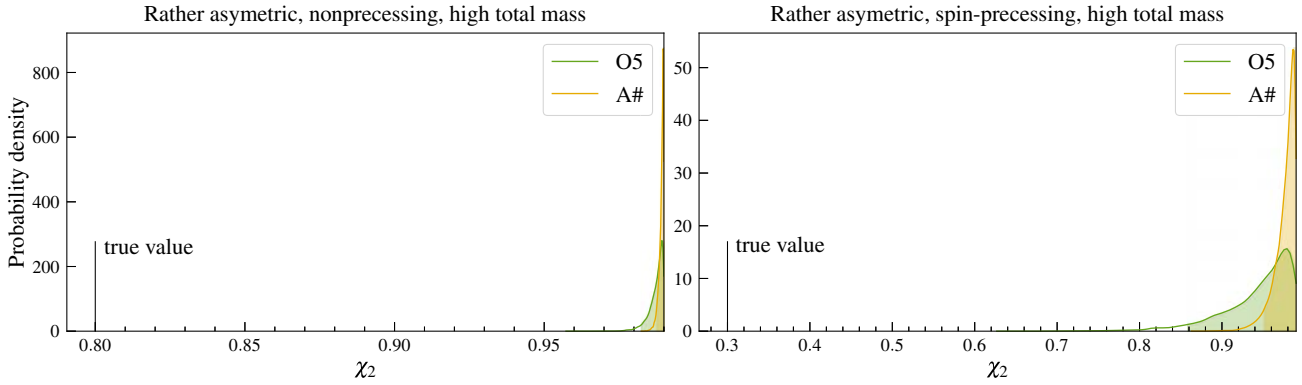

FIG. 20. Probability distributions of the secondary dimensionless spin magnitude χ_2 in the Binary 2 (left) and Binary 3 (right) systems in the *O5* and *A#* networks. The injected signal is generated using SEOBNRv5PHM while the template model is IMRPhenomXPHM. The estimates are biased for both the systems and in both the detector networks, but the more interesting feature is that the recovered χ_2 distributions imply a maximally spinning secondary. We do not show the posteriors in the *XG* network because they do not rail against the boundary, albeit still biased.

results for the *XG* detectors here because χ_2 is not inferred to be maximally spinning.

As shown in Fig. 10, the measurement of χ_2 will be biased even for low SNR, albeit the prior will start becoming important at such SNRs.

VII. DISCUSSION AND CONCLUSION

In this work, we have studied systematic biases arising in state-of-the-art BBH waveform models used by the LVK Collaboration, in particular, the quasicircular, spin-precessing, multipolar IMRPhenomXPHM and SEOBNRv5PHM models. With increasing detector sensitivities of the current LVK network, its future upgrades, and XG detectors ahead, quantifying waveform systematics becomes central in achieving the promising science goals of GW astronomy. Unbiased parameter estimation is crucial for a range of applications, from individual events to the entire BBH population. Moreover, it is also vital for precision tests of GR as the prevailing theory of gravity.

Throughout this work, we have assumed that the true signal can be represented by SEOBNRv5PHM, and we modeled it with IMRPhenomXPHM. Although neither model represents exact solutions, using them in this way for injection-recovery studies allows one to explore a wide range of the BBH parameter space, for which accurate NR simulations are not yet available.

To quantify the bias in parameter estimation, we utilized statistical tools from LSA and full Bayesian analysis (see Sec. III). While LSA is approximate and relies on large-SNR events, it can forecast results for a large number of events, which is what we are interested in. Here, the two main methods are the FIM to approximate measurement errors and the bias formula to predict the systematic errors. The full Bayesian analysis is computationally expensive but more reliable, which makes it suitable for understanding selected events in detail. With these methods, we studied biases and performed hierarchical inference on the BBH population in Sec. IV. We explored vast parts of the BBH parameter space in Sec. V and investigated in detail the impact of systematics on the science cases of individual events in Sec. VI. In the following, we summarize our main results.

Our first result is mainly on the "use and abuse" of the widely used bias formula based on the LSA. By comparing with full Bayesian results, we explicitly demonstrated that the direct application of the bias formula without a "waveform alignment" procedure, as described in Eq. (12), can yield unreliable results. To our knowledge, this case has not been discussed in the literature yet, and was not explained in the main references for the bias formula [104,109]. We find good agreement with full Bayesian results only after performing alignment and proper bookkeeping of the adjusted parameters. When not correcting for it, one would predict biases in the BBH population that could overestimate the actual bias by 2 orders of magnitude, which is particularly important when comparing waveforms of different families. Even after such caretaking, it is important to keep in mind that the LSA is an approximate estimate of the biases and statistical errors. We include discussions on the validity of the estimates and possible shortcomings of the method.

Regarding the LVK-like BBH population studied in Sec. IV, we find that the vast majority of events measured with the *O5* and *A#* networks are not biased, around 2.5% of the events, whereas, for *XG* detectors, up to 25% of the events are expected to show biases. This finding is most clearly reported in Fig. 6, which relates the mismatch of each event with the ratio of systematic-to-statistical error. Although the bias relative to the statistical error is larger for *XG* detectors, the absolute bias *per se* is smaller thanks to the detectors' improvement at low frequencies, which allows us to better observe the signal in the inspiral regime, where waveform models agree best. Thus, when combining

all events through a hierarchical Bayesian analysis on the observed population, we find that the impact of systematic biases is more pronounced for *A#* than for *XG* detectors. As illustrated in Fig. 7, the most biased parameter is the magnitude of the spin of the primary, with the recovered distribution peaking at smaller values while exhibiting a tail at large spin magnitudes that is absent in the injected population. The latter feature, in particular, could erroneously guide astrophysical formation scenarios into explaining the existence of a sizable number of BHs with large spins. We stress that the bias estimates obtained with the LSA might be overly pessimistic since it relies on the quadratic approximation to the likelihood, which holds only in the high SNR regime. Thus, the impact on the population might be exaggerated for *O5* and *A#* networks, but we expect our results for *XG* to be more accurate.

When spanning the possible BBH parameter space in Sec. V, we sampled events with uniform distribution $\chi_{\rm p} \in [0,1]$, uniform in total detector-frame mass $\in [10,200] M_\odot$, and uniform in inverse mass ratio $1/q \in [1,30]$. Although this sample does not represent the LVK BBH population, exploring the challenging parts of the parameter space is important since GW events in these regions may be discovered with more sensitive detectors in the future. Using the LSA, for each parameter, we computed the bias horizon, which describes the maximum distance up to which an event is systematically biased, after which the SNR is low enough that the statistical error is larger than the systematic one. Since most of these binaries are difficult to model, analytically and numerically, not surprisingly, the biases are more important than for the LVK-like population of Sec. IV.

Furthermore, our detailed analysis of selected events summarized in Table I in Sec. VI has several important findings depending on the science cases.

Focusing on cosmological implications in Sec. VI C 1, we found that distances and sky localization can be significantly biased. Here, one would be unable to infer the correct value of the Hubble-Lemaître parameter and thus would not resolve the Hubble-Lemaître tension. Biases in the sky position can be sufficiently large to prevent the correct identification of the host galaxy, which would be drastic since single *XG* events have the potential to determine H_0 to a few percent. Furthermore, those biases may also affect the determination of H_0 from stacking GW events, requiring a dedicated future study.

When studying the lower mass gap in Sec. VI C 2, we reported that the estimate of the secondary mass for the highly asymmetric, spin-precessing, low total-mass system would be underestimated in all networks. The spin of the secondary would be significantly underestimated in the *XG* network. This result could lead to wrong estimates of the upper edge of the NS mass distribution, which would inflict further biases for studies of the equation of state and, again, H_0. The PISN mass gap was investigated in Sec. VI C 3. Here, we showed that the estimate of the upper mass gap through the primary mass by the high total-mass binaries is strongly biased even for *O5*. Although precession changes the primary mass posteriors of both binaries significantly, the injected value for m_1 is not recovered within the 90% credible interval for any network.

In Sec. VI C 4, we looked into the spin morphology and found that *XG* detectors are not always more prone to biases than *O5* and *A#*. By extending the detectors' bandwidth to lower frequencies where the waveform models are more similar, the recovery of spin parameters can become closer to the injected values, at least for the low total-mass system. The remnant quantities were studied in Sec. VI C 5, where we reported the final mass and final spin of the remnant for all three "golden" binaries. Both upper mass-gap events have significant measurement bias, excluding the injected values far outside the 90% credible interval. For the highly asymmetric and precessing binary with low total mass, the overall biases are not as strong, besides the final spin in *XG*, which is also outside the 90% credible interval. By performing an independent ringdown analysis, at least for *XG*, one may conclude violations from the Kerr hypothesis and thus violations from GR (not performed in this work).

In Sec. VI C 6, we observed that the two high total-mass events predict posteriors for χ_2 that rail against the maximum spin of a BH for *O5* and *A#*. If not identified as systematic bias, this prediction would certainly have important consequences for astrophysical formation channels in explaining such high spins in BBH systems, as well as theoretical interest in extremal Kerr BHs and eventually exotic compact objects.

In summary, depending on the binary's parameters, biases can be present for the upcoming LVK O5 run and can affect crucial science. As expected, systematics become even more relevant with increasing detector sensitivity; thus, they are important for future *XG* detectors. The fact that many exciting science cases can be jeopardized by biases underlines the importance of improving existing waveform models. It also motivates the need to include modeling error estimates when performing parameter estimation, even if it will inevitably broaden our posteriors. Lastly, much more work would be needed to more robustly quantify the waveform systematics—for example, by employing as signals NR and NRSur waveforms, where available, and extending the current study to binaries on generic orbits, notably, BBHs on eccentric orbits.

The main limitation of our work is the use of the LSA for the population analysis. While it is the most readily available and feasible way to conduct a study like ours, one can be critical about its validity across the parameter space and view the ensuing conclusions with a grain of salt. In the future, we intend to carry out population-scale studies using modern data analysis tools, such as DINGO [245], that allow for rapid evaluations of the posteriors.

Note added. Recently, we became aware of a complementary study, Ref. [89], that focuses on assessing waveform systematics for XG detectors using two quasicircular aligned-spin models.

ACKNOWLEDGMENTS

We wish to thank Antoni Ramos Buades, Nihar Gupte, Serguei Ossokine, and Michael Pürrer for collaboration during the early part of this project. We also thank Veome Kapil, Luca Reali, and Emanuele Berti for useful discussions. We thank Ish Mohan Gupta for his comments during the LIGO review. S. H. V. acknowledges funding from the Deutsche Forschungsgemeinschaft (DFG), Project No. 386119226. The authors are grateful for computational resources provided by the Hypatia computer cluster at the Max Planck Institute for Gravitational Physics in Potsdam, the Gwave computing cluster at Penn State University, and the LIGO Laboratory, supported by National Science Foundation Grants No. PHY-0757058 and No. PHY-0823459. Some of the results in this paper have been obtained using the PESUMMARY package [246].

APPENDIX A: TOY MODEL

We consider a simple toy model to illustrate the effect of nonuniform parameter definitions across waveform models on the systematic bias calculated under LSA using Eq. (11). For this example, we take the IMRPhenomXPHM model as both the signal and the template. However, we shift the signal by τ, evaluating it at $t_c = \tau$, while the template is evaluated at $t_c = 0$. Note that for a pair of arbitrary waveform models, we do not know of the shift τ *a priori* and, hence, compute the bias at $t_c = 0$. In a Bayesian analysis, this would simply mean that the likelihood distribution for t_c peaks at $t_c = \tau$, for a noiseless injection, without impacting any physical parameter.

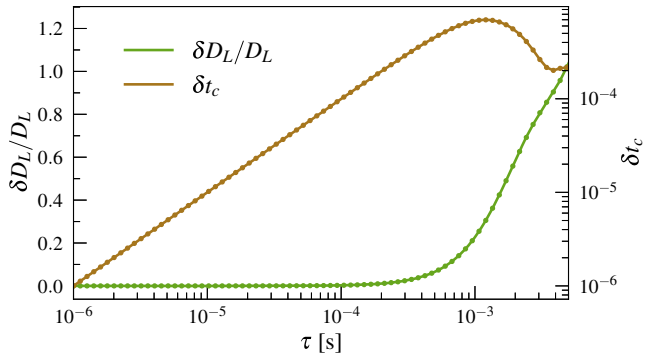

FIG. 21. Bias in the luminosity distance and time of coalescence parameters as a function of a time shift in the signal for the simple toy model considered in Appendix A. For very small values of the time shift parameter τ, the estimates of the bias formula are reliable, but they start deviating for τ values that are still small.

Let us now examine the predictions of the bias formula Eq. (11). We consider a reduced two-dimensional parameter space, $\boldsymbol{\vartheta} = \{D_L, t_c\}$. Here, the bias for the two parameters can be calculated analytically. The distance bias is given by

$$\begin{aligned}\delta D_L &= \frac{D_L^2}{\langle h|h\rangle}\langle -h/D_L|h(t_c=\tau)-h(t_c=0)\rangle \\ &= \frac{D_L^2}{\langle h|h\rangle}\langle -h/D_L|h(t_c=0)e^{-2\pi i f\tau}-h(t_c=0)\rangle \\ &= 0+\mathcal{O}(\tau^2),\end{aligned} \tag{A1}$$

while the t_c bias reduces to

$$\begin{aligned}\delta t_c &= \frac{1}{\langle 2\pi i f h|2\pi i f h\rangle}\langle -2\pi i f h|h(t_c=\tau)-h(t_c=0)\rangle \\ &= \frac{-2\pi}{\langle 2\pi i f h|2\pi i f h\rangle}\langle i f h|h(t_c=0)e^{-2\pi i f\tau}-h(t_c=0)\rangle \\ &= \tau+\mathcal{O}(\tau^2).\end{aligned} \tag{A2}$$

In Fig. 21, we show the bias in the luminosity distance and time of coalescence calculated using Eq. (11). We see that the estimates receive contributions from the quadratic and higher-order terms of Eqs. (A1) and (A2) for small values of the time shift τ. However, these corrections are not physical since D_L should not be biased for simple time shifts of the signal and the bias in t_c should simply correspond to the value of the time shift. While the incorrect estimates for the t_c bias are not of physical consequence in most cases, it is readily seen from the figure that the D_L bias can be incorrectly estimated to very large values, impacting the outlook on science applications like cosmology where the D_L parameter is crucial.

APPENDIX B: EFFECT OF $f_{\rm low}$ FOR HIGHER MODES IN THE SEOBNRv5PHM SIGNAL

When generating an SEOBNRv5PHM waveform, the minimum frequency refers to the frequency of the $(l, m) = (2, 2)$ harmonic. Higher harmonics in the given time segment occur at a higher frequency, which is a feature of all time-domain waveforms. Since phenomenological waveforms are constructed in the frequency domain, all the harmonics are present at any given frequency. Thus, in analyzing a GW signal generated using SEOBNRv5PHM with IMRPhenomXPHM (with the same minimum frequency), the template contains the higher harmonics at frequencies lower than where the same is present in the signal. For instance, if $f_{\rm low} = 10$ Hz for both SEOBNRv5PHM and IMRPhenomXPHM, the $(l, m) = (3, 3), (4, 4)$ harmonics for the SEOBNRv5PHM signal start at 15 Hz and 20 Hz, respectively.

All analyses in the paper are performed by taking the same $f_{\rm low}$ for both SEOBNRv5PHM and IMRPhenomXPHM

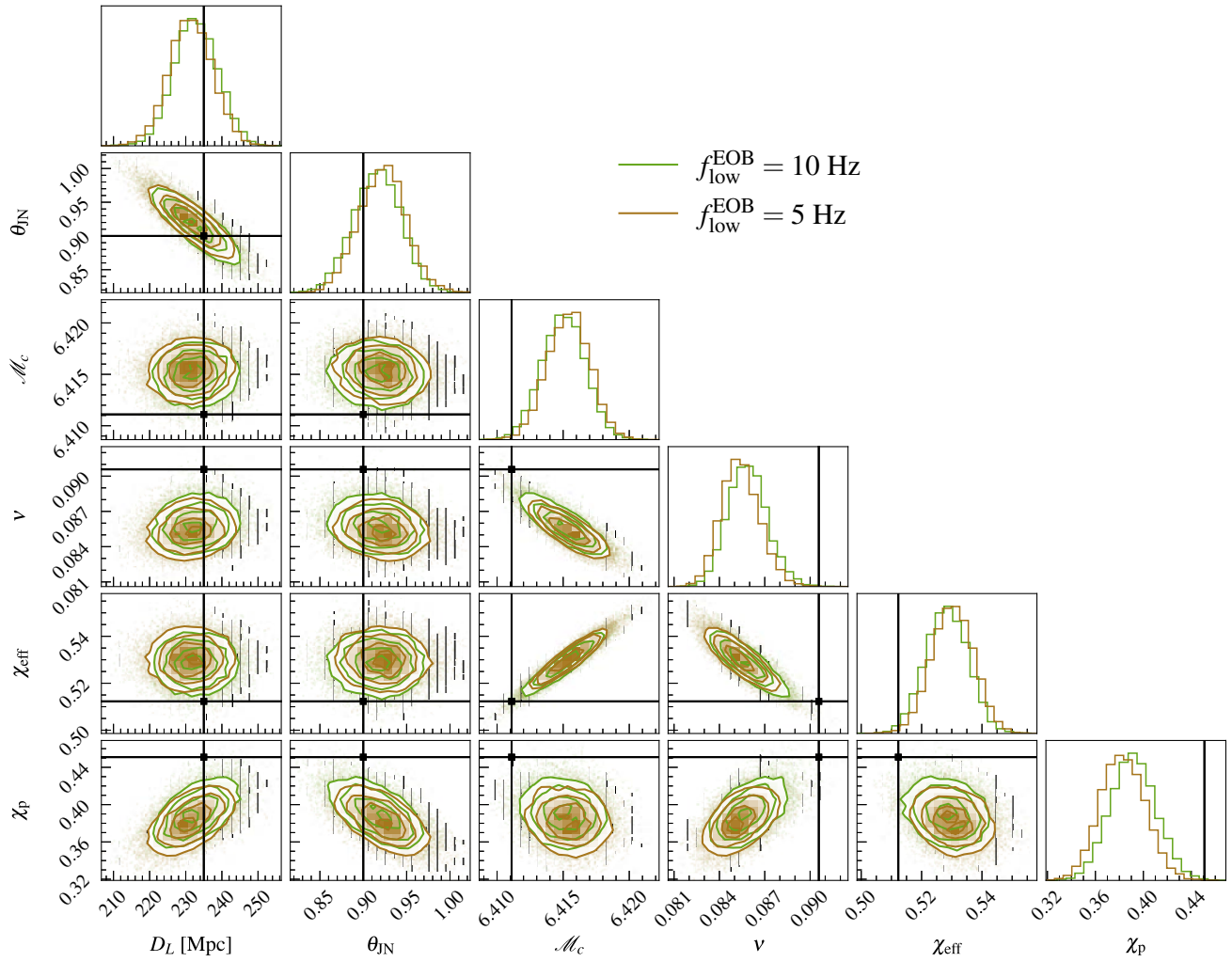

FIG. 22. Posterior distributions of select parameters for the Binary 1 system in the *O5* network with different minimum frequencies at which the SEOBNRv5PHM signal is generated. The starting frequency for the analysis is 10 Hz for both cases. The indistinguishability of the distributions in the two cases shows that the absence of certain subdominant harmonics in the signal at low frequencies for the case depicted in green is inconsequential to the analysis.

waveforms. We verify in Fig. 22 that this approach does not affect any of the results of the paper. The figure shows the posterior distributions of select parameters of the Binary 1 system for two cases simulated in the *O5* network. For the distributions plotted in orange, the SEOBNRv5PHM waveform is generated starting from $f_{\rm low} = 5$ Hz, while the analysis is completed using IMRPhenomXPHM setting $f_{\rm low} = 10$ Hz. On the other hand, the SEOBNRv5PHM waveform is generated with $f_{\rm low} = 10$ Hz for the case in green and an identical analysis setting as the previous case. We remark that the posterior distributions in the two cases are indistinguishable.

APPENDIX C: EFFECT OF $f_{\rm low}$ ON PARAMETER ESTIMATION AND BIAS

We briefly discuss the impact of the minimum frequency $f_{\rm low}$ used in analyzing a GW signal. This case is of particular importance given the excess low-frequency noise in the Advanced LIGO and Advanced Virgo detectors limiting the low-frequency cutoff to 20 Hz instead of the predicted 10 Hz.

We generate a signal corresponding to the Binary 1 system in the *O5* network at two starting frequencies, $f_{\rm low} = 10$ and 20 Hz, with the SEOBNRv5PHM model. This signal is then analyzed with the respective starting frequencies used in its generation, with the IMRPhenomXPHM model as the template. The posterior distributions in the $\mathcal{M}_c - \chi_{\rm eff}$ parameter space are reported in Fig. 23. The distributions on the other nonderivative parameters are not reported since they are similar in both cases. It is observed that a lower minimum frequency leads to a smaller measurement error and bias for the chirp mass. Since $\chi_{\rm eff}$ has a large (positive) correlation with $\mathcal{M}_c$, as seen from the figure, it is also affected in a similar manner.

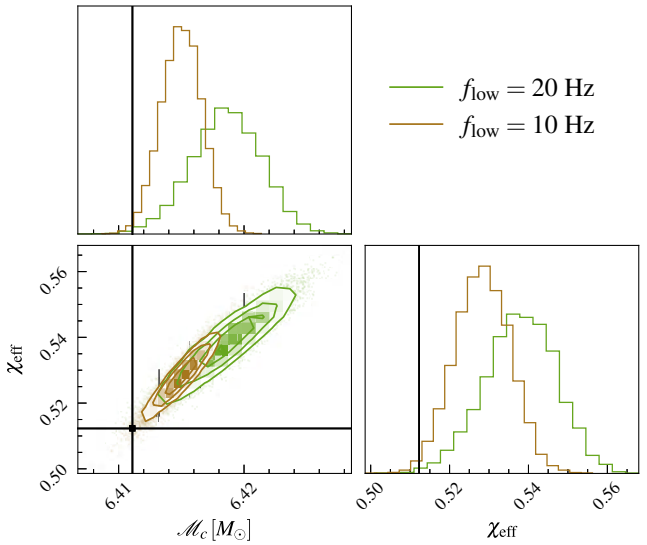

FIG. 23. Posterior distributions in the $\mathcal{M}_c - \chi_{\rm eff}$ parameter space for the Binary 1 system in the *O5* network with different minimum frequencies, $f_{\rm low}$. Note that a smaller minimum frequency results in a greater measurement precision and a smaller bias.

The number of cycles in a GW waveform is inversely related to the minimum frequency, and $\mathcal{M}_c$ is the leading-order contributor to this relation. Hence, the large number of GW cycles between 10 Hz and 20 Hz leads to a better measurement precision for $\mathcal{M}_c$. It is also the part of the waveform where different models are in better agreement since all models have to reproduce the PN limit leading to a more accurate measurement. Thus, the absolute magnitude of the $\mathcal{M}_c$ bias in the *XG* network, which has $f_{\rm low} = 5$ Hz, is smaller than the *O5* and *A#* networks.

APPENDIX D: DEPENDENCE OF $|\delta\vartheta/\Delta\vartheta|$ ON THE SNR

In the following, we report complementary results for the same data as presented in Fig. 6. Figure 24 shows how the ratio $|\delta\vartheta/\Delta\vartheta|$ depends on the SNR (instead of the mismatch), and the color bar now indicates the mismatch. The panels are structured similarly to those of Fig. 6. Overall, the cumulation of biased events (ratio larger than 1) depends strongly on the mismatch, although outliers exist for all networks. In all cases, the number of events with a ratio much smaller than 1 decreases as a function of the SNR, although strongly biased events exist even for small SNRs. These results underline the importance of improving waveform modeling to reduce mismatches for all detector networks, not only for *XG*.

APPENDIX E: DISTRIBUTION OF POPULATION PARAMETERS

In the panels of Fig. 25, we show the distribution of parameters of detectable events from the LVK-like BBH population for the different networks. The detectability criterion is SNR > 12, which acts as a strong filter for *O5* and *A#* but only very mildly impacts *XG*. The parameters $a_1, a_2, \cos\theta_1$, and $\cos\theta_2$ are distributed uniformly for all networks, which agrees well with the population. However, the shape of the distributions describing the parameters M_c, ν, D_L, and $\cos\theta_{\rm JN}$ is more complicated and changes throughout the networks. This result implies a significant

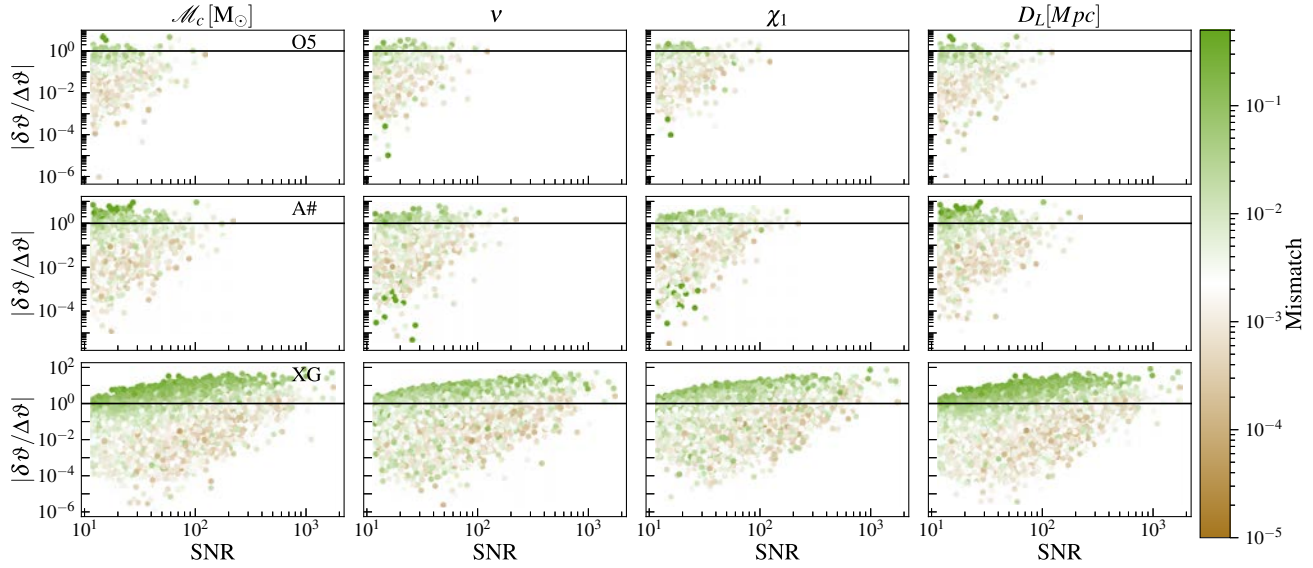

FIG. 24. Ratio between the systematic bias and statistical errors for select parameters as a function of the SNR for a population of BBH mergers as observed by the LVK. A network SNR threshold of 12 was imposed on the 10^5 binaries in the population, resulting in around 1800 and 8100 in the *O5* (top) and *A#* (bottom) networks, respectively. The color bar depicts the mismatch between SEOBNRv5PHM and IMRPhenomXPHM waveform models.

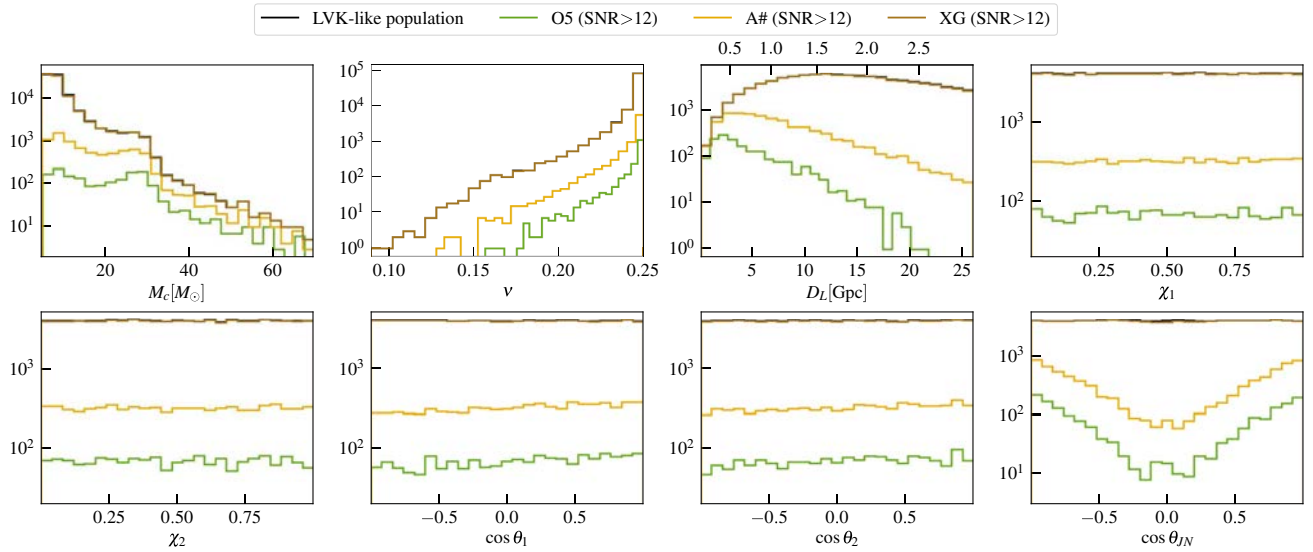

FIG. 25. Distribution of select parameters of the LVK-like BBH population and of the detectable population in the three detector networks. The black line overlaps almost entirely with the detectable population in the *XG* network.

selection bias unless one uses *XG*. Note that $\cos\theta_{\rm JN}$ is sampled from a uniform distribution but shows strong selection bias for *O5* and *A#*.

APPENDIX F: BIAS IN χ_1 WHEN SCANNING THE PARAMETER SPACE

In Figs. 26 and 27, we show the bias horizon for the parameter χ_1. The binaries are distributed as described in Sec. V.

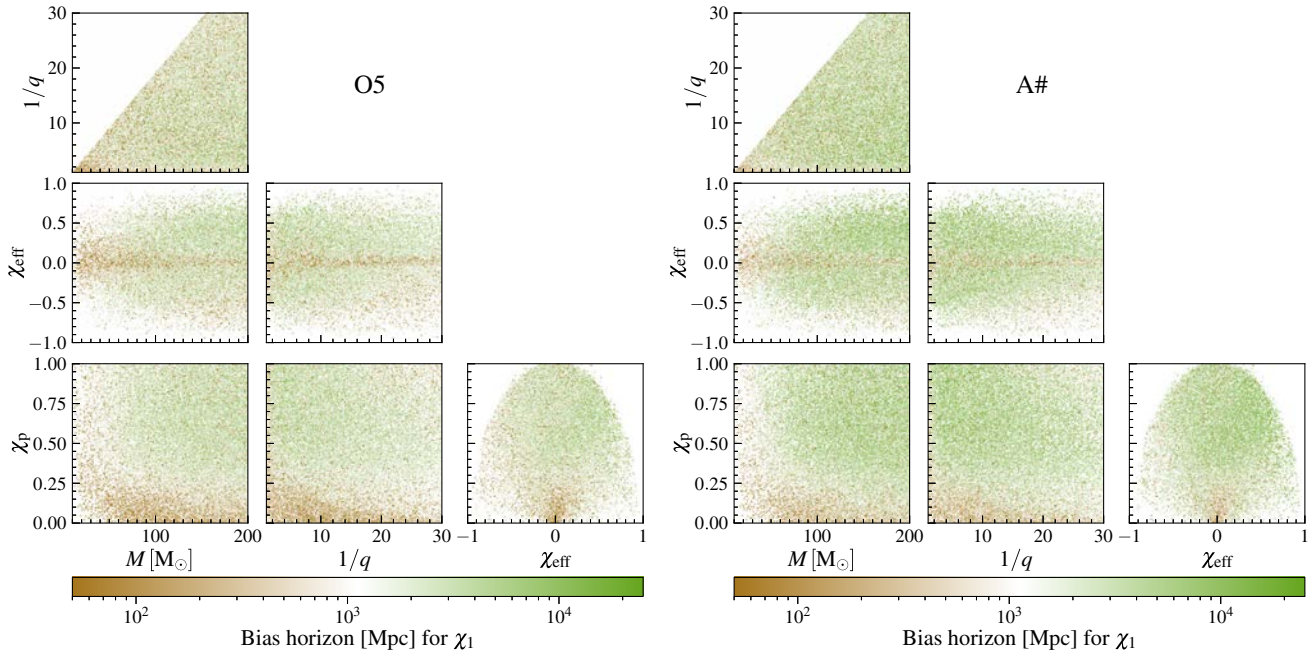

FIG. 26. Distribution of the 50,000 binaries in the parameter space represented in Fig. 9, with the color scale showing the distance to which the χ_1 parameter is biased ($\delta\chi_1/\Delta\chi_1 \geq 1$) for the *O5* (left) and *A#* (right) networks. Systematic biases become less important if a binary is at a larger distance since measurement precision decreases with distance. Therefore, a large bias horizon signifies that a given parameter (χ_1 in this case) is measured well enough for systematic biases to be important even at such large distances. Notice that the binaries are biased up to a greater distance in the *A#* network compared to the *O5* network due to its greater sensitivity and the resultant improvement in measurement precision.

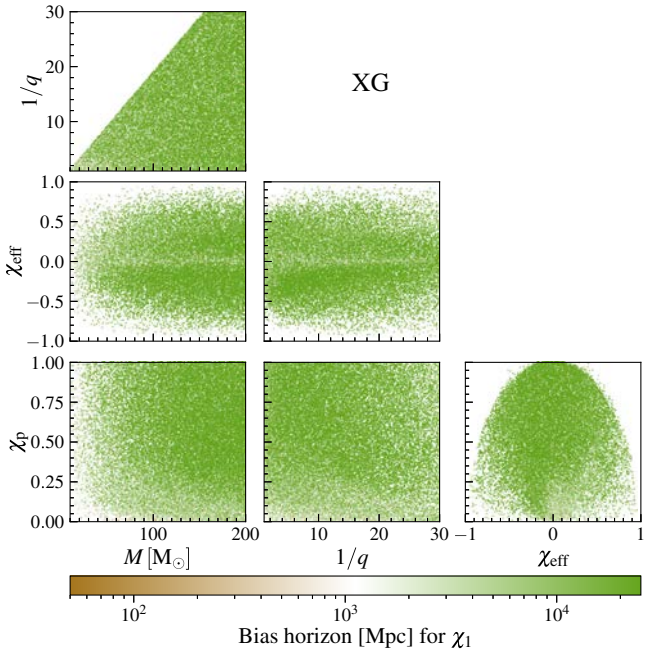

FIG. 27. Same as Fig. 26 for the *XG* network. BBHs observed with the *XG* network are biased up to a greater distance than those observed with either the *O5* or *A#* network due to its greater sensitivity and the resultant improvement in measurement precision, with a majority of the binaries having a bias horizon greater than or equal to 25 Gpc ($z \approx 3$), beyond which stellar-origin BBHs are not expected to exist.

[1] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), *Observation of gravitational waves from a binary black hole merger*, Phys. Rev. Lett. **116**, 061102 (2016).

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

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

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

[5] R. Abbott *et al.* (KAGRA, Virgo, and LIGO Scientific 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).

[6] 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).

[7] A. G. Abac *et al.* (LIGO Scientific, Virgo, KAGRA, and Virgo Collaborations), *Observation of gravitational waves from the coalescence of a* 2.5–4.5$M_\odot$ *compact object and a neutron star*, Astrophys. J. Lett. **970**, L34 (2024).

[8] T. Venumadhav, B. Zackay, J. Roulet, L. Dai, and M. Zaldarriaga, *New search pipeline for compact binary mergers: Results for binary black holes in the first observing run of Advanced LIGO*, Phys. Rev. D **100**, 023011 (2019).

[9] T. Venumadhav, B. Zackay, J. Roulet, L. Dai, and M. Zaldarriaga, *New binary black hole mergers in the second observing run of Advanced LIGO and Advanced Virgo*, Phys. Rev. D **101**, 083030 (2020).

[10] B. Zackay, L. Dai, T. Venumadhav, J. Roulet, and M. Zaldarriaga, *Detecting gravitational waves with disparate detector responses: Two new binary black hole mergers*, Phys. Rev. D **104**, 063030 (2021).

[11] S. Olsen, T. Venumadhav, J. Mushkin, J. Roulet, B. Zackay, and M. Zaldarriaga, *New binary black hole mergers in the LIGO-Virgo O3a data*, Phys. Rev. D **106**, 043009 (2022).

[12] A. K. Mehta, S. Olsen, D. Wadekar, J. Roulet, T. Venumadhav, J. Mushkin, B. Zackay, and M. Zaldarriaga, *New binary black hole mergers in the LIGO-Virgo O3b data*, Phys. Rev. D **111**, 024049 (2025).

[13] D. Wadekar, J. Roulet, T. Venumadhav, A. K. Mehta, B. Zackay, J. Mushkin, S. Olsen, and M. Zaldarriaga, *New black hole mergers in the LIGO-Virgo O3 data from a gravitational wave search including higher-order harmonics*, arXiv:2312.06631.

[14] 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).

[15] M. Fishbach, D. E. Holz, and W. M. Farr, *Does the black hole merger rate evolve with redshift?*, Astrophys. J. Lett. **863**, L41 (2018).

[16] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), *GW170817: Measurements of neutron star radii and equation of state*, Phys. Rev. Lett. **121**, 161101 (2018).

[17] D. Radice, A. Perego, F. Zappa, and S. Bernuzzi, *GW170817: Joint constraint on the neutron star equation of state from multimessenger observations*, Astrophys. J. Lett. **852**, L29 (2018).

[18] S. De, D. Finstad, J. M. Lattimer, D. A. Brown, E. Berger, and C. M. Biwer, *Tidal deformabilities and radii of neutron stars from the observation of GW170817*, Phys. Rev. Lett. **121**, 091102 (2018); **121**, 259902(E) (2018).

[19] B. P. Abbott *et al.* (LIGO Scientific, Virgo, 1M2H, Dark Energy Camera GW-E, DES, DLT40, Las Cumbres Observatory, VINROUGE, MASTER Collaborations), *A gravitational-wave standard siren measurement of the Hubble constant*, Nature (London) **551**, 85 (2017).

[20] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), *Constraints on the cosmic expansion history from GWTC–3*, Astrophys. J. **949**, 76 (2023).

[21] 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).

[22] 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).

[23] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), *Tests of general relativity with GWTC-3*, arXiv:2112.06861.

[24] M. Punturo *et al.*, *The Einstein Telescope: A third-generation gravitational wave observatory*, Classical Quantum Gravity **27**, 194002 (2010).
[25] M. Maggiore *et al.*, *Science case for the Einstein telescope*, J. Cosmol. Astropart. Phys. 03 (2020) 050.
[26] D. Reitze *et al.*, *Cosmic explorer: The U.S. Contribution to gravitational-wave astronomy beyond LIGO*, Bull. Am. Astron. Soc. **51**, 035 (2019), https://baas.aas.org/pub/2020n7i035/release/1.
[27] S. Borhanian and B. S. Sathyaprakash, *Listening to the Universe with next generation ground-based gravitational-wave detectors*, Phys. Rev. D **110**, 083040 (2024).
[28] M. Evans *et al.*, *A horizon study for cosmic explorer: Science, observatories, and community*, arXiv:2109.09882.
[29] I. Gupta *et al.*, *Characterizing gravitational wave detector networks: From* $a^\sharp$ *to cosmic explorer*, Classical Quantum Gravity **41**, 245001 (2024).
[30] M. Branchesi *et al.*, *Science with the Einstein telescope: A comparison of different designs*, J. Cosmol. Astropart. Phys. 07 (2023) 068.
[31] S. Bogdanov *et al.*, *Snowmass 2021 Cosmic Frontier White Paper: The dense matter equation of state and QCD phase transitions*, in Snowmass 2021 (2022), arXiv:2209.07412.
[32] F. Iacovelli, M. Mancarella, S. Foffa, and M. Maggiore, *Forecasting the detection capabilities of third-generation gravitational-wave detectors using GWFast*, Astrophys. J. **941**, 208 (2022).
[33] S. E. Woosley and A. Heger, *The pair-instability mass gap for black holes*, Astrophys. J. Lett. **912**, L31 (2021).
[34] K. Belczynski *et al.*, *The effect of pair-instability mass loss on black hole mergers*, Astron. Astrophys. **594**, A97 (2016).
[35] R. Farmer, M. Renzo, S. de Mink, M. Fishbach, and S. Justham, *Constraints from gravitational wave detections of binary black hole mergers on the* $^{12}C(\alpha,\gamma)^{16}O$ *rate*, Astrophys. J. Lett. **902**, L36 (2020).
[36] S. Borhanian, A. Dhani, A. Gupta, K. G. Arun, and B. S. Sathyaprakash, *Dark sirens to resolve the Hubble–Lemaître tension*, Astrophys. J. Lett. **905**, L28 (2020).
[37] P. T. H. Pang, J. C. Bustillo, Y. Wang, and T. G. F. Li, *Potential observations of false deviations from general relativity in gravitational wave signals from binary black holes*, Phys. Rev. D **98**, 024019 (2018).
[38] E. Maggio, H. O. Silva, A. Buonanno, and A. Ghosh, *Tests of general relativity in the nonlinear regime: A parametrized plunge-merger-ringdown gravitational waveform model*, Phys. Rev. D **108**, 024043 (2023).
[39] Q. Hu and J. Veitch, *Accumulating errors in tests of general relativity with gravitational waves: Overlapping signals and inaccurate waveforms*, Astrophys. J. **945**, 103 (2023).
[40] A. Toubiana, L. Pompili, A. Buonanno, J. R. Gair, and M. L. Katz, *Measuring source properties and quasinormal mode frequencies of heavy massive black-hole binaries with LISA*, Phys. Rev. D **109**, 104019 (2024).
[41] S. A. Bhat, P. Saini, M. Favata, and K. G. Arun, *Systematic bias on the inspiral-merger-ringdown consistency test due to neglect of orbital eccentricity*, Phys. Rev. D **107**, 024009 (2023).
[42] P. Saini, S. A. Bhat, M. Favata, and K. G. Arun, *Eccentricity-induced systematic error on parametrized tests of general relativity: Hierarchical Bayesian inference applied to a binary black hole population*, Phys. Rev. D **109**, 084056 (2024).
[43] P. Narayan, N. K. Johnson-McDaniel, and A. Gupta, *Effect of ignoring eccentricity in testing general relativity with gravitational waves*, Phys. Rev. D **108**, 064003 (2023).
[44] A. Buonanno and T. Damour, *Effective one-body approach to general relativistic two-body dynamics*, Phys. Rev. D **59**, 084006 (1999).
[45] A. Buonanno and T. Damour, *Transition from inspiral to plunge in binary black hole coalescences*, Phys. Rev. D **62**, 064015 (2000).
[46] T. Damour, P. Jaranowski, and G. Schafer, *On the determination of the last stable orbit for circular general relativistic binaries at the third post-Newtonian approximation*, Phys. Rev. D **62**, 084011 (2000).
[47] T. Damour, *Coalescence of two spinning black holes: An effective one-body approach*, Phys. Rev. D **64**, 124013 (2001).
[48] A. Buonanno, Y. Chen, and T. Damour, *Transition from inspiral to plunge in precessing binaries of spinning black holes*, Phys. Rev. D **74**, 104005 (2006).
[49] A. Ramos-Buades, A. Buonanno, M. Khalil, and S. Ossokine, *Effective-one-body multipolar waveforms for eccentric binary black holes with nonprecessing spins*, Phys. Rev. D **105**, 044035 (2022).
[50] L. Pompili *et al.*, *Laying the foundation of the effective-one-body waveform models SEOBNRv5: Improved accuracy and efficiency for spinning nonprecessing binary black holes*, Phys. Rev. D **108**, 124035 (2023).
[51] A. Ramos-Buades, A. Buonanno, H. Estellés, M. Khalil, D. P. Mihaylov, S. Ossokine, L. Pompili, and M. Shiferaw, *Next generation of accurate and efficient multipolar precessing-spin effective-one-body waveforms for binary black holes*, Phys. Rev. D **108**, 124037 (2023).
[52] M. van de Meent, A. Buonanno, D. P. Mihaylov, S. Ossokine, L. Pompili, N. Warburton, A. Pound, B. Wardell, L. Durkan, and J. Miller, *Enhancing the SEOBNRv5 effective-one-body waveform model with second-order gravitational self-force fluxes*, Phys. Rev. D **108**, 124038 (2023).
[53] M. Khalil, A. Buonanno, H. Estelles, D. P. Mihaylov, S. Ossokine, L. Pompili, and A. Ramos-Buades, *Theoretical groundwork supporting the precessing-spin two-body dynamics of the effective-one-body waveform models SEOBNRv5*, Phys. Rev. D **108**, 124036 (2023).
[54] A. Nagar *et al.*, *Time-domain effective-one-body gravitational waveforms for coalescing compact binaries with nonprecessing spins, tides and self-spin effects*, Phys. Rev. D **98**, 104052 (2018).
[55] A. Nagar, G. Pratten, G. Riemenschneider, and R. Gamba, *Multipolar effective one body model for nonspinning black hole binaries*, Phys. Rev. D **101**, 024041 (2020).
[56] A. Nagar, G. Riemenschneider, G. Pratten, P. Rettegno, and F. Messina, *Multipolar effective one body waveform model for spin-aligned black hole binaries*, Phys. Rev. D **102**, 024077 (2020).

[57] R. Gamba, S. Akçay, S. Bernuzzi, and J. Williams, *Effective-one-body waveforms for precessing coalescing compact binaries with post-Newtonian twist*, Phys. Rev. D **106**, 024020 (2022).

[58] A. Nagar, P. Rettegno, R. Gamba, S. Albanesi, A. Albertini, and S. Bernuzzi, *Analytic systematics in next generation of effective-one-body gravitational waveform models for future observations*, Phys. Rev. D **108**, 124018 (2023).

[59] F. Pretorius, *Evolution of binary black hole spacetimes*, Phys. Rev. Lett. **95**, 121101 (2005).

[60] M. Campanelli, C. O. Lousto, P. Marronetti, and Y. Zlochower, *Accurate evolutions of orbiting black-hole binaries without excision*, Phys. Rev. Lett. **96**, 111101 (2006).

[61] J. G. Baker, J. Centrella, D.-I. Choi, M. Koppitz, and J. van Meter, *Gravitational wave extraction from an inspiraling configuration of merging black holes*, Phys. Rev. Lett. **96**, 111102 (2006).

[62] Y. Pan, A. Buonanno, J. G. Baker, J. Centrella, B. J. Kelly, S. T. McWilliams, F. Pretorius, and J. R. van Meter, *A data-analysis driven comparison of analytic and numerical coalescing binary waveforms: Nonspinning case*, Phys. Rev. D **77**, 024014 (2008).

[63] P. Ajith *et al.*, *Phenomenological template family for black-hole coalescence waveforms*, Classical Quantum Gravity **24**, S689 (2007).

[64] 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).

[65] 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).

[66] H. Estellés, M. Colleoni, C. García-Quirós, S. Husa, D. Keitel, M. Mateu-Lucena, M. d. L. Planas, and A. Ramos-Buades, *New twists in compact binary waveform modeling: A fast time-domain model for precession*, Phys. Rev. D **105**, 084040 (2022).

[67] J. E. Thompson, E. Hamilton, L. London, S. Ghosh, P. Kolitsidou, C. Hoy, and M. Hannam, *PhenomXO4a: A phenomenological gravitational-wave model for precessing black-hole binaries with higher multipoles and asymmetries*, Phys. Rev. D **109**, 063012 (2024).

[68] M. Hannam *et al.*, *The samurai project: Verifying the consistency of black-hole-binary waveforms for gravitational-wave detection*, Phys. Rev. D **79**, 084025 (2009).

[69] I. Hinder *et al.*, *Error-analysis and comparison to analytical models of numerical waveforms produced by the NRAR Collaboration*, Classical Quantum Gravity **31**, 025012 (2014).

[70] M. Boyle *et al.*, *The SXS Collaboration catalog of binary black hole simulations*, Classical Quantum Gravity **36**, 195006 (2019).

[71] O. Rinne, L. T. Buchman, M. A. Scheel, and H. P. Pfeiffer, *Implementation of higher-order absorbing boundary conditions for the Einstein equations*, Classical Quantum Gravity **26**, 075009 (2009).

[72] L. T. Buchman, H. P. Pfeiffer, M. A. Scheel, and B. Szilagyi, *Simulations of non-equal mass black hole binaries with spectral methods*, Phys. Rev. D **86**, 084033 (2012).

[73] L. T. Buchman, M. D. Duez, M. Morales, M. A. Scheel, T. M. Kostersitz, A. M. Evans, and K. Mitman, *Numerical relativity multimodal waveforms using absorbing boundary conditions*, Classical Quantum Gravity **41**, 175011 (2024).

[74] T. Chu, H. Fong, P. Kumar, H. P. Pfeiffer, M. Boyle, D. A. Hemberger, L. E. Kidder, M. A. Scheel, and B. Szilagyi, *On the accuracy and precision of numerical waveforms: Effect of waveform extraction methodology*, Classical Quantum Gravity **33**, 165001 (2016).

[75] K. Mitman *et al.*, *Adding gravitational memory to waveform catalogs using BMS balance laws*, Phys. Rev. D **103**, 024031 (2021).

[76] J. Blackman, S. E. Field, C. R. Galley, B. Szilágyi, M. A. Scheel, M. Tiglio, and D. A. Hemberger, *Fast and accurate prediction of numerical relativity waveforms from binary black hole coalescences using surrogate models*, Phys. Rev. Lett. **115**, 121102 (2015).

[77] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, L. E. Kidder, and H. P. Pfeiffer, *Surrogate model of hybridized numerical relativity binary black hole waveforms*, Phys. Rev. D **99**, 064045 (2019).

[78] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, D. Gerosa, L. C. Stein, L. E. Kidder, and H. P. Pfeiffer, *Surrogate models for precessing binary black hole simulations with unequal masses*, Phys. Rev. Res. **1**, 033015 (2019).

[79] J. Yoo *et al.*, *Numerical relativity surrogate model with memory effects and post-Newtonian hybridization*, Phys. Rev. D **108**, 064027 (2023).

[80] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), *Effects of waveform model systematics on the interpretation of GW150914*, Classical Quantum Gravity **34**, 104002 (2017).

[81] M. Pürrer and C.-J. Haster, *Gravitational waveform accuracy requirements for future ground-based detectors*, Phys. Rev. Res. **2**, 023151 (2020).

[82] Q. Hu and J. Veitch, *Assessing the model waveform accuracy of gravitational waves*, Phys. Rev. D **106**, 044042 (2022).

[83] T. Islam, A. Vajpeyi, F. H. Shaik, C.-J. Haster, V. Varma, S. E. Field, J. Lange, R. O'Shaughnessy, and R. Smith, *Analysis of GWTC-3 with fully precessing numerical relativity surrogate models*, arXiv:2309.14473.

[84] Note that some differences are likely to be attributed to sampler issues rather than waveform systematics.

[85] A. Puecher, A. Samajdar, G. Ashton, C. Van Den Broeck, and T. Dietrich, *Comparing gravitational waveform models for binary black hole mergers through a hypermodels approach*, Phys. Rev. D **109**, 023019 (2024).

[86] A. Z. Jan, A. B. Yelikar, J. Lange, and R. O'Shaughnessy, *Assessing and marginalizing over compact binary coalescence waveform systematics with RIFT*, Phys. Rev. D **102**, 124069 (2020).

[87] J. S. Read, *Waveform uncertainty quantification and interpretation for gravitational-wave astronomy*, Classical Quantum Gravity **40**, 135002 (2023).

[88] C. B. Owen, C.-J. Haster, S. Perkins, N. J. Cornish, and N. Yunes, *Waveform accuracy and systematic uncertainties in current gravitational wave observations*, Phys. Rev. D **108**, 044018 (2023).
[89] V. Kapil, L. Reali, R. Cotesta, and E. Berti, *Systematic bias from waveform modeling for binary black hole populations in next-generation gravitational wave detectors*, Phys. Rev. D **109**, 104043 (2024).
[90] V. Varma, P. Ajith, S. Husa, J. C. Bustillo, M. Hannam, and M. Pürrer, *Gravitational-wave observations of binary black holes: Effect of nonquadrupole modes*, Phys. Rev. D **90**, 124004 (2014).
[91] V. Varma and P. Ajith, *Effects of nonquadrupole modes in the detection and parameter estimation of black hole binaries with nonprecessing spins*, Phys. Rev. D **96**, 124024 (2017).
[92] M. K. Singh, S. J. Kapadia, A. Vijaykumar, and P. Ajith, *Impact of higher harmonics of gravitational radiation on the population inference of binary black holes*, Astrophys. J. **971**, 23 (2024).
[93] M. Favata, C. Kim, K. G. Arun, J. C. Kim, and H. W. Lee, *Constraining the orbital eccentricity of inspiralling compact binary systems with Advanced LIGO*, Phys. Rev. D **105**, 023003 (2022).
[94] H.-S. Cho, *Systematic bias due to eccentricity in parameter estimation for merging binary neutron stars*, Phys. Rev. D **105**, 124022 (2022).
[95] A. Samajdar and T. Dietrich, *Waveform systematics for binary neutron star gravitational wave signals: Effects of the point-particle baseline and tidal descriptions*, Phys. Rev. D **98**, 124030 (2018).
[96] A. Samajdar and T. Dietrich, *Waveform systematics for binary neutron star gravitational wave signals: Effects of spin, precession, and the observation of electromagnetic counterparts*, Phys. Rev. D **100**, 024046 (2019).
[97] R. Gamba, M. Breschi, S. Bernuzzi, M. Agathos, and A. Nagar, *Waveform systematics in the gravitational-wave inference of tidal parameters and equation of state from binary neutron star signals*, Phys. Rev. D **103**, 124015 (2021).
[98] G. Pratten, P. Schmidt, and N. Williams, *Impact of dynamical tides on the reconstruction of the neutron star equation of state*, Phys. Rev. Lett. **129**, 081102 (2022).
[99] P. Kolitsidou, J. E. Thompson, and M. Hannam, *Impact of antisymmetric contributions to signal multipoles in the measurement of black-hole spins*, Phys. Rev. D **111**, 024050 (2025).
[100] D. Ferguson, K. Jani, P. Laguna, and D. Shoemaker, *Assessing the readiness of numerical relativity for LISA and 3G detectors*, Phys. Rev. D **104**, 044037 (2021).
[101] A. Jan, D. Ferguson, J. Lange, D. Shoemaker, and A. Zimmerman, *Accuracy limitations of existing numerical relativity waveforms on the data analysis of current and future ground-based detectors*, Phys. Rev. D **110**, 024023 (2024).
[102] 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).
[103] 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).
[104] E. E. Flanagan and S. A. Hughes, *Measuring gravitational waves from binary black hole coalescences: 2. The waves' information and its extraction, with and without templates*, Phys. Rev. D **57**, 4566 (1998).
[105] K. Chatziioannou, A. Klein, N. Yunes, and N. Cornish, *Constructing gravitational waves from generic spin-precessing compact binary inspirals*, Phys. Rev. D **95**, 104004 (2017).
[106] A. Toubiana and J. R. Gair, *Indistinguishability criterion and estimating the presence of biases*, arXiv:2401.06845.
[107] We refer the reader to Secs. II A, II B, and III B 3 for discussions on GW parameters, waveform models, and maximization of overlaps between waveforms.
[108] L. S. Finn and D. F. Chernoff, *Observing binary inspiral in gravitational radiation: One interferometer*, Phys. Rev. D **47**, 2198 (1993).
[109] C. Cutler and M. Vallisneri, *LISA detections of massive black hole inspirals: Parameter extraction errors due to inaccurate template waveforms*, Phys. Rev. D **76**, 104018 (2007).
[110] M. Vallisneri, *Use and abuse of the Fisher information matrix in the assessment of gravitational-wave parameter-estimation prospects*, Phys. Rev. D **77**, 042001 (2008).
[111] H.-S. Cho, E. Ochsner, R. O'Shaughnessy, C. Kim, and C.-H. Lee, *Gravitational waves from black hole-neutron star binaries: Effective Fisher matrices and parameter estimation using higher harmonics*, Phys. Rev. D **87**, 024004 (2013).
[112] I. Harry and A. Lundgren, *Failure of the Fisher matrix when including tidal terms: Considering construction of template banks of tidally deformed binary neutron stars*, Phys. Rev. D **104**, 043008 (2021).
[113] P. Madau and M. Dickinson, *Cosmic star formation history*, Annu. Rev. Astron. Astrophys. **52**, 415 (2014).
[114] We adopt the convention $m_1 \geq m_2$.
[115] P. Schmidt, I. W. Harry, and H. P. Pfeiffer, *Numerical relativity injection infrastructure*, arXiv:1703.01076.
[116] B. S. Sathyaprakash and S. V. Dhurandhar, *Choice of filters for the detection of gravitational waves from coalescing binaries*, Phys. Rev. D **44**, 3819 (1991).
[117] N. Aghanim *et al.* (Planck Collaboration), *Planck 2018 results. VI. Cosmological parameters*, Astron. Astrophys. **641**, A6 (2020); **652**, C4(E) (2021).
[118] P. Ajith *et al.*, *Inspiral-merger-ringdown waveforms for black-hole binaries with non-precessing spins*, Phys. Rev. Lett. **106**, 241101 (2011).
[119] L. Santamaria *et al.*, *Matching post-Newtonian and numerical relativity waveforms: Systematic errors and a new phenomenological model for non-precessing black hole binaries*, Phys. Rev. D **82**, 064016 (2010).
[120] P. Schmidt, F. Ohme, and M. Hannam, *Towards models of gravitational waveforms from generic binaries II:*

Modelling precession effects with a single effective precession parameter, Phys. Rev. D **91**, 024043 (2015).
[121] L. M. Thomas, P. Schmidt, and G. Pratten, *New effective precession spin for modeling multimodal gravitational waveforms in the strong-field regime*, Phys. Rev. D **103**, 083022 (2021).
[122] D. Gerosa, M. Mould, D. Gangardt, P. Schmidt, G. Pratten, and L. M. Thomas, *A generalized precession parameter χ_p to interpret gravitational-wave data*, Phys. Rev. D **103**, 064067 (2021).
[123] G. Ashton *et al.*, *Bilby: A user-friendly Bayesian inference library for gravitational-wave astronomy*, Astrophys. J. Suppl. Ser. **241**, 27 (2019).
[124] I. M. Romero-Shaw *et al.*, *Bayesian inference for compact binary coalescences with Bilby: Validation and application to the first LIGO–Virgo gravitational-wave transient catalogue*, Mon. Not. R. Astron. Soc. **499**, 3295 (2020).
[125] D. A. Brown, P. Kumar, and A. H. Nitz, *Template banks to search for low-mass binary black holes in advanced gravitational-wave detectors*, Phys. Rev. D **87**, 082004 (2013).
[126] C. Capano, Y. Pan, and A. Buonanno, *Impact of higher harmonics in searching for gravitational waves from nonspinning binary black holes*, Phys. Rev. D **89**, 102003 (2014).
[127] I. Harry, J. C. Bustillo, and A. Nitz, *Searching for the full symphony of black hole binary mergers*, Phys. Rev. D **97**, 023004 (2018).
[128] R. Cotesta, A. Buonanno, A. Bohé, A. Taracchini, I. Hinder, and S. Ossokine, *Enriching the symphony of gravitational waves from binary black holes by tuning higher harmonics*, Phys. Rev. D **98**, 084028 (2018).
[129] C. Kalaghatgi, M. Hannam, and V. Raymond, *Parameter estimation with a spinning multimode waveform model*, Phys. Rev. D **101**, 103004 (2020).
[130] We note that, since beginning our work, there have been a few important updates on phenomenological models [67,131], which included NR calibration to the precessing sector, a more faithful ringdown model, and improvements to the spin-precessing equations. However, we do not expect that our results would change substantially if we used those new waveform models.
[131] M. Colleoni, F. A. R. Vidal, N. K. Johnson-McDaniel, T. Dietrich, M. Haney, and G. Pratten, *IMRPhenomXP_NRTidalv2: An improved frequency-domain precessing binary neutron star waveform model*, Phys. Rev. D **111**, 064025 (2025).
[132] A. Buonanno, Y. Chen, and M. Vallisneri, *Detecting gravitational waves from precessing binaries of spinning compact objects: Adiabatic limit*, Phys. Rev. D **67**, 104025 (2003); **74**, 029904(E) (2006).
[133] 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).
[134] M. Boyle, R. Owen, and H. P. Pfeiffer, *A geometric approach to the precession of compact binaries*, Phys. Rev. D **84**, 124011 (2011).
[135] R. O'Shaughnessy, B. Vaishnav, J. Healy, Z. Meeks, and D. Shoemaker, *Efficient asymptotic frame selection for binary black hole spacetimes using asymptotic radiation*, Phys. Rev. D **84**, 124002 (2011).
[136] 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).
[137] C. García-Quirós, S. Husa, M. Mateu-Lucena, and A. Borchers, *Accelerating the evaluation of inspiral–merger–ringdown waveforms with adapted grids*, Classical Quantum Gravity **38**, 015006 (2021).
[138] B. P. Abbott *et al.* (KAGRA, LIGO Scientific, and Virgo Collaborations), *Prospects for observing and localizing gravitational-wave transients with Advanced LIGO, Advanced Virgo and KAGRA*, Living Rev. Relativity **23**, 3 (2018).
[139] V. Srivastava, D. Davis, K. Kuns, P. Landry, S. Ballmer, M. Evans, E. D. Hall, J. Read, and B. S. Sathyaprakash, *Science-driven tunable design of cosmic explorer detectors*, Astrophys. J. **931**, 22 (2022).
[140] The A+, V+, ET, and CE sensitivity curves in this work are those used in Ref. [27] while the A# sensitivity curve is taken from https://dcc.ligo.org/LIGO-T2300041/public.
[141] J. S. Speagle, *DYNESTY: A dynamic nested sampling package for estimating Bayesian posteriors and evidences*, Mon. Not. R. Astron. Soc. **493**, 3132 (2020).
[142] L. P. Singer and L. R. Price, *Rapid Bayesian position reconstruction for gravitational-wave transients*, Phys. Rev. D **93**, 024013 (2016).
[143] L. P. Singer *et al.*, *Going the distance: Mapping host galaxies of LIGO and Virgo sources in three dimensions using local cosmography and targeted follow-up*, Astrophys. J. Lett. **829**, L15 (2016).
[144] 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).
[145] C_{ii} is the ith element of C_{ij}.
[146] S. Borhanian, *GWBENCH: A novel Fisher information package for gravitational-wave benchmarking*, Classical Quantum Gravity **38**, 175014 (2021).
[147] U. Dupletsa, J. Harms, B. Banerjee, M. Branchesi, B. Goncharov, A. Maselli, A. C. S. Oliveira, S. Ronchini, and J. Tissino, *GWFish: A simulation software to evaluate parameter-estimation capabilities of gravitational-wave detector networks*, Astron. Comput. **42**, 100671 (2022).
[148] LIGO Scientific, Virgo, and KAGRA Collaborations, *LVK Algorithm Library—LALSuite, Free software (GPL)* (2018), 10.7935/GT1W-FZ16.
[149] D. P. Mihaylov, S. Ossokine, A. Buonanno, H. Estelles, L. Pompili, M. Pürrer, and A. Ramos-Buades, *pySEOBNR: A software package for the next generation of effective-one-body multipolar waveform models*, arXiv:2303.18203.
[150] This expression first appeared in Ref. [104], but it is often referred to as the Cutler-Vallisneri formula after a later paper [109], which was the first to explore its implications.
[151] T. Damour, A. Nagar, and M. Trias, *Accuracy and effectualness of closed-form, frequency-domain waveforms for non-spinning black hole binaries*, Phys. Rev. D **83**, 024006 (2011).

[152] 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).
[153] S. Ossokine *et al.*, *Multipolar effective-one-body waveforms for precessing binary black holes: Construction and validation*, Phys. Rev. D **102**, 044055 (2020).
[154] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, *FINDCHIRP: An algorithm for detection of gravitational waves from inspiraling compact binaries*, Phys. Rev. D **85**, 122006 (2012).
[155] A. Buonanno, B. R. Iyer, E. Ochsner, Y. Pan, and B. S. Sathyaprakash, *Comparison of post-Newtonian templates for compact binary inspiral signals in gravitational-wave detectors*, Phys. Rev. D **80**, 084043 (2009).
[156] I. Mandel, W. M. Farr, and J. R. Gair, *Extracting distribution parameters from multiple uncertain observations with selection biases*, Mon. Not. R. Astron. Soc. **486**, 1086 (2019).
[157] S. Vitale, D. Gerosa, W. M. Farr, and S. R. Taylor, *Inferring the properties of a population of compact binaries in presence of selection effects*, in *Handbook of Gravitational Wave Astronomy*, edited by C. Bambi, S. Katsanevas, and K. D. Kokkotas (Springer, Singapore, 2020), pp. 1–60.
[158] A. Toubiana, M. L. Katz, and J. R. Gair, *Is there an excess of black holes around $20M_\odot$? Optimizing the complexity of population models with the use of reversible jump MCMC.*, Mon. Not. R. Astron. Soc. **524**, 5844 (2023).
[159] N. Karnesis, M. L. Katz, N. Korsakova, J. R. Gair, and N. Stergioulas, *Eryn: A multi-purpose sampler for Bayesian inference*, Mon. Not. R. Astron. Soc. **526**, 4814 (2023).
[160] In this way, we are not respecting the properties of the noise since we associate a random error to a given event. Crucially, this is performed for one-dimensional distributions, so, in practice, we project all the errors onto a one-dimensional space, inducing a large scatter in the value of measurement errors for a given value of ϑ. Thus, for parameters that do not exhibit a strong trend between measurement errors and the true parameter—as we observed to be the case for the mass ratio and spin angles and magnitude—our procedure should be close enough to what we would obtain in the full case, and it allows us to also remove the dependency of the error on the other parameters.
[161] C. Talbot and J. Golomb, *Growing pains: Understanding the impact of likelihood uncertainty on hierarchical Bayesian inference for gravitational-wave astronomy*, Mon. Not. R. Astron. Soc. **526**, 3495 (2023).
[162] M. Evans *et al.*, *Cosmic Explorer: A submission to the NSF MPSAC ngGW subcommittee*, arXiv:2306.13745.
[163] A. G. Riess *et al.*, *A comprehensive measurement of the local value of the Hubble constant with 1 km* $1\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$ *uncertainty from the Hubble Space Telescope and the SH0ES Team*, Astrophys. J. Lett. **934**, L7 (2022).
[164] M. Fishbach *et al.* (LIGO Scientific and Virgo Collaborations), *A standard siren measurement of the Hubble constant from GW170817 without the electromagnetic counterpart*, Astrophys. J. Lett. **871**, L13 (2019).
[165] M. Soares-Santos *et al.* (DES, LIGO Scientific, and Virgo Collaborations), *First measurement of the Hubble constant from a dark standard siren using the dark energy survey galaxies and the LIGO/Virgo binary–black-hole merger GW170814*, Astrophys. J. Lett. **876**, L7 (2019).
[166] A. Palmese *et al.* (DES Collaboration), *A statistical standard siren measurement of the Hubble constant from the LIGO/Virgo gravitational wave compact object merger GW190814 and dark energy survey galaxies*, Astrophys. J. Lett. **900**, L33 (2020).
[167] H.-Y. Chen, M. Fishbach, and D. E. Holz, *A two per cent Hubble constant measurement from standard sirens within five years*, Nature (London) **562**, 545 (2018).
[168] I. Gupta, *Using grey sirens to resolve the Hubble–Lemaître tension*, Mon. Not. R. Astron. Soc. **524**, 3537 (2023).
[169] C. Messenger and J. Read, *Measuring a cosmological distance-redshift relationship using only gravitational wave observations of binary neutron star coalescences*, Phys. Rev. Lett. **108**, 091101 (2012).
[170] C. Messenger, K. Takami, S. Gossan, L. Rezzolla, and B. S. Sathyaprakash, *Source redshifts from gravitational-wave observations of binary neutron star mergers*, Phys. Rev. X **4**, 041004 (2014).
[171] W. Del Pozzo, T. G. F. Li, and C. Messenger, *Cosmological inference using only gravitational wave observations of binary neutron stars*, Phys. Rev. D **95**, 043502 (2017).
[172] D. Chatterjee, A. Hegade K. R., G. Holder, D. E. Holz, S. Perkins, K. Yagi, and N. Yunes, *Cosmology with love: Measuring the Hubble constant using neutron star universal relations*, Phys. Rev. D **104**, 083528 (2021).
[173] T. Ghosh, B. Biswas, and S. Bose, *Simultaneous inference of neutron star equation of state and the Hubble constant with a population of merging neutron stars*, Phys. Rev. D **106**, 123529 (2022).
[174] A. Dhani, S. Borhanian, A. Gupta, and B. Sathyaprakash, *Cosmography with bright and Love sirens*, arXiv:2212.13183.
[175] B. Shiralilou, G. Raaijmakers, B. Duboeuf, S. Nissanke, F. Foucart, T. Hinderer, and A. R. Williamson, *Measuring the Hubble constant with dark neutron star–black hole mergers*, Astrophys. J. **955**, 149 (2023).
[176] D. F. Chernoff and L. S. Finn, *Gravitational radiation, inspiraling binaries, and cosmology*, Astrophys. J. Lett. **411**, L5 (1993).
[177] S. R. Taylor and J. R. Gair, *Cosmology with the lights off: Standard sirens in the Einstein telescope era*, Phys. Rev. D **86**, 023502 (2012).
[178] W. M. Farr, M. Fishbach, J. Ye, and D. Holz, *A future percent-Level measurement of the Hubble expansion at redshift 0.8 with Advanced LIGO*, Astrophys. J. Lett. **883**, L42 (2019).
[179] J. M. Ezquiaga and D. E. Holz, *Spectral sirens: Cosmology from the full mass distribution of compact binaries*, Phys. Rev. Lett. **129**, 061102 (2022).
[180] B. F. Schutz, *Determining the Hubble constant from gravitational wave observations*, Nature **323**, 310 (1986).
[181] W. Del Pozzo, *Inference of the cosmological parameters from gravitational waves: Application to second generation interferometers*, Phys. Rev. D **86**, 043011 (2012).

[182] N. Muttoni, D. Laghi, N. Tamanini, S. Marsat, and D. Izquierdo-Villalba, *Dark siren cosmology with binary black holes in the era of third-generation gravitational wave detectors*, Phys. Rev. D **108**, 043543 (2023).

[183] R. Gray *et al.*, *Cosmological inference using gravitational wave standard sirens: A Mock data analysis*, Phys. Rev. D **101**, 122001 (2020).

[184] R. Gray *et al.*, *Joint cosmological and gravitational-wave population inference using dark sirens and galaxy catalogues*, J. Cosmol. Astropart. Phys. 12 (2023) 023.

[185] M. Oguri, *Measuring the distance-redshift relation with the cross-correlation of gravitational wave standard sirens and galaxies*, Phys. Rev. D **93**, 083511 (2016).

[186] S. Mukherjee, B. D. Wandelt, S. M. Nissanke, and A. Silvestri, *Accurate precision cosmology with redshift unknown gravitational wave sources*, Phys. Rev. D **103**, 043520 (2021).

[187] T. Ghosh, S. More, S. Bera, and S. Bose, *Bayesian framework to infer the Hubble constant from cross-correlation of individual gravitational wave events with galaxies*, arXiv:2312.16305.

[188] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), *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).

[189] I. Tews, P. T. H. Pang, T. Dietrich, M. W. Coughlin, S. Antier, M. Bulla, J. Heinzel, and L. Issa, *On the nature of GW190814 and its impact on the understanding of supranuclear matter*, Astrophys. J. Lett. **908**, L1 (2021).

[190] H. Tan, J. Noronha-Hostler, and N. Yunes, *Neutron star equation of state in light of GW190814*, Phys. Rev. Lett. **125**, 261104 (2020).

[191] H. Tan, T. Dore, V. Dexheimer, J. Noronha-Hostler, and N. Yunes, *Extreme matter meets extreme gravity: Ultraheavy neutron stars with phase transitions*, Phys. Rev. D **105**, 023018 (2022).

[192] E. R. Most, L. J. Papenfort, L. R. Weih, and L. Rezzolla, *A lower bound on the maximum mass if the secondary in GW190814 was once a rapidly spinning neutron star*, Mon. Not. R. Astron. Soc. **499**, L82 (2020).

[193] N.-B. Zhang and B.-A. Li, *GW190814's secondary component with mass* 2.50–2.67$M_\odot$ *as a superfast pulsar*, Astrophys. J. **902**, 38 (2020).

[194] V. Dexheimer, R. O. Gomes, T. Klähn, S. Han, and M. Salinas, *GW190814 as a massive rapidly rotating neutron star with exotic degrees of freedom*, Phys. Rev. C **103**, 025808 (2021).

[195] C. J. Krüger and S. H. Völkel, *Rapidly rotating neutron stars: Universal relations and EOS inference*, Phys. Rev. D **108**, 124056 (2023).

[196] K. Vattis, I. S. Goldstein, and S. M. Koushiappas, *Could the* 2.6$M_\odot$ *object in GW190814 be a primordial black hole?*, Phys. Rev. D **102**, 061301(R) (2020).

[197] S. Clesse and J. Garcia-Bellido, *GW190425, GW190521 and GW190814: Three candidate mergers of primordial black holes from the QCD epoch*, Phys. Dark Universe **38**, 101111 (2022).

[198] E. Bianchi, A. Gupta, H. M. Haggard, and B. S. Sathyaprakash, *Small spins of primordial black holes from random geometries: Bekenstein-Hawking entropy and gravitational wave observations*, arXiv:1812.05127.

[199] D. Gerosa and E. Berti, *Are merging black holes born from stellar collapse or previous mergers?*, Phys. Rev. D **95**, 124046 (2017).

[200] A. Gupta, D. Gerosa, K. G. Arun, E. Berti, W. M. Farr, and B. S. Sathyaprakash, *Black holes in the low mass gap: Implications for gravitational wave observations*, Phys. Rev. D **101**, 103036 (2020).

[201] M. Safarzadeh, A. S. Hamers, A. Loeb, and E. Berger, *Formation and merging of mass gap black holes in gravitational wave merger events from wide hierarchical quadruple systems*, Astrophys. J. Lett. **888**, L3 (2020).

[202] D. Gerosa and M. Fishbach, *Hierarchical mergers of stellar-mass black holes and their gravitational-wave signatures*, Nat. Astron. **5**, 749 (2021).

[203] Z. Doctor, D. Wysocki, R. O'Shaughnessy, D. E. Holz, and B. Farr, *Black hole coagulation: Modeling hierarchical mergers in black hole populations*, Astrophys. J. **893**, 35 (2019).

[204] M. Zevin, M. Spera, C. P. L. Berry, and V. Kalogera, *Exploring the lower mass gap and unequal mass regime in compact binary evolution*, Astrophys. J. Lett. **899**, L1 (2020).

[205] M. Safarzadeh and A. Loeb, *Formation of mass gap objects in highly asymmetric mergers*, Astrophys. J. Lett. **899**, L15 (2020).

[206] 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).

[207] M. Mapelli, M. Spera, E. Montanari, M. Limongi, A. Chieffi, N. Giacobbo, A. Bressan, and Y. Bouffanais, *Impact of the rotation and compactness of progenitors on the mass of black holes*, Astrophys. J. **888**, 76 (2020).

[208] M. Kesden, D. Gerosa, R. O'Shaughnessy, E. Berti, and U. Sperhake, *Effective potentials and morphological transitions for binary black-hole spin precession*, Phys. Rev. Lett. **114**, 081103 (2015).

[209] D. Gerosa, M. Kesden, U. Sperhake, E. Berti, and R. O'Shaughnessy, *Multi-timescale analysis of phase transitions in precessing black-hole binaries*, Phys. Rev. D **92**, 064016 (2015).

[210] D. Gerosa, U. Sperhake, and J. Vošmera, *On the equal-mass limit of precessing black-hole binaries*, Classical Quantum Gravity **34**, 064004 (2017).

[211] K. S. Phukon, A. Gupta, S. Bose, and P. Jain, *Effect of orbital eccentricity on the dynamics of precessing compact binaries*, Phys. Rev. D **100**, 124008 (2019).

[212] J. D. Schnittman, *Spin-orbit resonance and the evolution of compact binary systems*, Phys. Rev. D **70**, 124020 (2004).

[213] M. Kesden, U. Sperhake, and E. Berti, *Final spins from the merger of precessing binary black holes*, Phys. Rev. D **81**, 084054 (2010).

[214] M. Kesden, U. Sperhake, and E. Berti, *Relativistic suppression of black hole recoils*, Astrophys. J. **715**, 1006 (2010).

[215] E. Berti, M. Kesden, and U. Sperhake, *Effects of post-Newtonian spin alignment on the distribution of black-hole recoils*, Phys. Rev. D **85**, 124049 (2012).

[216] D. Gerosa, M. Kesden, E. Berti, R. O'Shaughnessy, and U. Sperhake, *Resonant-plane locking and spin alignment in stellar-mass black-hole binaries: A diagnostic of compact-binary formation*, Phys. Rev. D **87**, 104028 (2013).

[217] D. Gerosa, E. Berti, R. O'Shaughnessy, K. Belczynski, M. Kesden, D. Wysocki, and W. Gladysz, *Spin orientations of merging black holes formed from the evolution of stellar binaries*, Phys. Rev. D **98**, 084036 (2018).

[218] N. Steinle and M. Kesden, *Signatures of spin precession and nutation in isolated black-hole binaries*, Phys. Rev. D **106**, 063028 (2022).

[219] D. Gerosa, G. Fumagalli, M. Mould, G. Cavallotto, D. P. Monroy, D. Gangardt, and V. De Renzis, *Efficient multi-timescale dynamics of precessing black-hole binaries*, Phys. Rev. D **108**, 024042 (2023).

[220] G. Pratten, P. Schmidt, R. Buscicchio, and L. M. Thomas, *Measuring precession in asymmetric compact binaries*, Phys. Rev. Res. **2**, 043096 (2020).

[221] V. Varma, S. Biscoveanu, M. Isi, W. M. Farr, and S. Vitale, *Hints of spin-orbit resonances in the binary black hole population*, Phys. Rev. Lett. **128**, 031101 (2022).

[222] S. Kulkarni, N. K. Johnson-McDaniel, K. S. Phukon, N. V. Krishnendu, and A. Gupta, *Inferring spin tilts of binary black holes at formation with plus-era gravitational wave detectors*, Phys. Rev. D **109**, 043002 (2024).

[223] N. K. Johnson-McDaniel, K. S. Phukon, N. V. Krishnendu, and A. Gupta, *Distinguishing binary black hole precessional morphologies with gravitational wave observations*, Phys. Rev. D **108**, 103003 (2023).

[224] J. Healy and C. O. Lousto, *Remnant of binary black-hole mergers: New simulations and peak luminosity studies*, Phys. Rev. D **95**, 024037 (2017).

[225] D. Keitel *et al.*, *The most powerful astrophysical events: Gravitational-wave peak luminosity of binary black holes as predicted by numerical relativity*, Phys. Rev. D **96**, 024006 (2017).

[226] N. K. Johnson-McDaniel, A. Gupta, P. Ajith, K. Keitel, O. Birnholtz, F. Ohme, and S. Husa, *Determining the final spin of a binary black hole system including in-plane spins: Method and checks of accuracy*, Technical Report No. T1600168, LIGO, 2016.

[227] V. Varma, D. Gerosa, L. C. Stein, F. Hébert, and H. Zhang, *High-accuracy mass, spin, and recoil predictions of generic black-hole merger remnants*, Phys. Rev. Lett. **122**, 011101 (2019).

[228] L. Rezzolla, E. Barausse, E. Nils Dorband, D. Pollney, C. Reisswig, J. Seiler, and S. Husa, *On the final spin from the coalescence of two black holes*, Phys. Rev. D **78**, 044002 (2008).

[229] E. Barausse and L. Rezzolla, *Predicting the direction of the final spin from the coalescence of two black holes*, Astrophys. J. Lett. **704**, L40 (2009).

[230] F. Hofmann, E. Barausse, and L. Rezzolla, *The final spin from binary black holes in quasi-circular orbits*, Astrophys. J. Lett. **825**, L19 (2016).

[231] 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).

[232] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), *Properties and astrophysical implications of the* $150m_{\odot}$ *binary black hole merger GW190521*, Astrophys. J. Lett. **900**, L13 (2020).

[233] H. Siegel, M. Isi, and W. M. Farr, *Ringdown of GW190521: Hints of multiple quasinormal modes with a precessional interpretation*, Phys. Rev. D **108**, 064008 (2023).

[234] S. N. Zhang, W. Cui, and W. Chen, *Black hole spin in x-ray binaries: Observational consequences*, Astrophys. J. Lett. **482**, L155 (1997).

[235] L. W. Brenneman and C. S. Reynolds, *Constraining black hole spin via x-ray spectroscopy*, Astrophys. J. **652**, 1028 (2006).

[236] H. Yang, A. Zimmerman, and L. Lehner, *Turbulent black holes*, Phys. Rev. Lett. **114**, 081101 (2015).

[237] M. Campanelli, C. O. Lousto, and Y. Zlochower, *Spinning-black-hole binaries: The orbital hang up*, Phys. Rev. D **74**, 041501(R) (2006).

[238] D. Gerosa, M. Kesden, R. O'Shaughnessy, A. Klein, E. Berti, U. Sperhake, and D. Trifirò, *Precessional instability in binary black holes with aligned spins*, Phys. Rev. Lett. **115**, 141102 (2015).

[239] K. Belczynski, R. E. Taam, E. Rantsiou, and M. van der Sluys, *Black hole spin evolution: Implications for short-hard gamma ray bursts and gravitational wave detection*, Astrophys. J. **682**, 474 (2008).

[240] M. Zaldarriaga, D. Kushnir, and J. A. Kollmeier, *The expected spins of gravitational wave sources with isolated field binary progenitors*, Mon. Not. R. Astron. Soc. **473**, 4174 (2018).

[241] K. Belczynski *et al.*, *Evolutionary roads leading to low effective spins, high black hole masses, and O1/O2 rates for LIGO/Virgo binary black holes*, Astron. Astrophys. **636**, A104 (2020).

[242] L. A. C. van Son, S. E. de Mink, F. S. Broekgaarden, M. Renzo, S. Justham, E. Laplace, J. Moran-Fraile, D. D. Hendriks, and R. Farmer, *Polluting the pair-instability mass gap for binary black holes through super-Eddington accretion in isolated binaries*, Astrophys. J. **897**, 100 (2020).

[243] R. C. H. Cheng and N. A. K. Amin, *Estimating parameters in continuous univariate distributions with a shifted origin*, J. R. Stat. Soc. Ser. B **45**, 394 (1983).

[244] J. M. Ezquiaga and D. E. Holz, *Jumping the gap: Searching for LIGO's biggest black holes*, Astrophys. J. Lett. **909**, L23 (2021).

[245] M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno, and B. Schölkopf, *Real-Time gravitational wave science with neural posterior estimation*, Phys. Rev. Lett. **127**, 241103 (2021).

[246] C. Hoy and V. Raymond, *PESUMMARY: The code agnostic parameter estimation summary page builder*, SoftwareX **15**, 100765 (2021).