High-mass-ratio challenge in gravitational waveform modeling

Parthapratim Mahapatra,[1,*] Jonathan E. Thompson,[2] Edward Fauchon-Jones,[1] and Mark Hannam[1]
[1]*School of Physics and Astronomy, Cardiff University, Cardiff CF24 3AA, United Kingdom*
[2]*Mathematical Sciences and STAG Research Centre, University of Southampton, Southampton SO17 1BJ, United Kingdom*

(Received 30 March 2026; accepted 12 August 2026; published 15 September 2026)

Binary black hole (BBH) mergers detected via gravitational waves are addressing key open questions in astrophysics, cosmology, and fundamental physics. Our scientific conclusions rely on extracting accurate source parameters, for which we require accurate signal modeling. It is well known that current BBH waveform models need to be improved for high-mass-ratio, strongly precessing systems, and in this paper we provide a concrete illustration of this issue, showing that the degradation in model performance is substantially more severe than might have been anticipated. We present numerical relativity (NR) simulations of precessing BBH systems with a mass ratio of 18 and a dimensionless spin of 0.8 on the larger black hole (with the smaller black hole nonspinning), covering five values of spin misalignment. We assess the accuracy of state-of-the-art waveform models in this region of parameter space by computing the standard mismatch between the models and the NR waveforms. We find that all current waveform models often exhibit significant mismatches ($\gtrsim 0.1$), indicating poor performance in this regime. We also perform limited parameter estimation using a subset of state-of-the-art waveform models, injecting these NR simulations as signals into the three-detector LIGO-Virgo network. In some cases we find errors in mass measurements of over 100%, dramatically illustrating that substantial improvements are required in existing waveform models. The numerical simulations presented here will be valuable for calibrating future BBH waveform models in this region of parameter space.

DOI: 10.1103/r917-75sf

I. INTRODUCTION

It has now been a decade since the first landmark detection of gravitational waves (GWs) in 2015, an event that inaugurated a new era of observational astronomy [1]. Over this period, advances in detector sensitivity and network coordination have dramatically expanded the reach of GW science [2–8]. The 2015-era detector network could observe only a few stellar-mass black hole (BH) mergers per year, whereas today's detectors routinely observe hundreds, transforming what was once extraordinary into routine [8–11]. These observations have already provided key insights into the strong-field nature of gravity [12–14], the local expansion rate of the Universe [15,16], the physics of ultradense matter [17], and the astrophysical population of compact objects [18]. Looking ahead, next-generation ground-based observatories such as the Einstein Telescope [19,20] and Cosmic Explorer [21] are projected to detect thousands of mergers annually, effectively surveying nearly all stellar-origin black-holes in the observable Universe [22]. Complementing these efforts, the space-based Laser Interferometer Space Antenna (LISA) will open a new frequency window, granting access to massive black-hole mergers and expanding the multiband potential of GW astronomy [23,24]. Together, these advances are transforming GW astronomy into an increasingly important component of fundamental physics, astrophysics, and cosmology.

The scientific potential of these observations relies critically on accurate theoretical waveform models of the GW signals as predicted by general relativity (GR) [25–27]. Source properties are inferred by convolving the observed GW data with these model waveforms, within a Bayesian inference framework that systematically identifies the set of model parameters that best reproduce the observed signal [28,29]. In this approach, the likelihood of the observed data given a particular waveform model is combined with our prior expectations of the source population to compute a posterior probability distribution, providing both the most likely source parameter values and their associated uncertainties [29–31]. Accurate waveform models are essential because they encode the detailed multipolar structure of the signals, ideally including higher-order modes, precession effects, and eccentricity, which are critical to precisely

*Contact author: MahapatraP@cardiff.ac.uk

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

extracting the masses, spins, distances, and orbital properties of the binary [25–27]. Inaccuracies in these models can lead to biased parameter estimates and limit our ability to achieve the science goals of current and future GW detectors, making continued improvements in waveform modeling a key priority for the future of GW astronomy.

The gravitational dynamics of a compact binary can be divided into three stages: inspiral, merger, and ringdown. Due to the complexity of Einstein's equations, there are no exact analytic solutions for the full two-body problem of coalescing binary systems of compact objects. During the early inspiral, Einstein's equations can be solved perturbatively using the post-Newtonian approximation [32] and resummation techniques within the effective-one-body approach [33], while key features of the ringdown are described by black hole perturbation theory [34]. The late inspiral and merger phases, lying between these two regimes, require full solutions of the Einstein equations. After decades of effort to solve the binary black hole (BBH) problem in the merger phase, the first complete numerical relativity (NR) simulations spanning a single plunge-orbit, merger, and ringdown were achieved in 2005 [35–37]. Since then, numerous independent NR codes [38–45] have been developed, enabling simulations over multiple orbits and accommodating more challenging configurations, including high mass ratios, rapidly spinning black holes, and highly eccentric binaries [46–51].

A BBH system in a quasicircular orbit is described by eight intrinsic parameters: the redshifted (detector-frame) component masses of the black holes, m_1 and m_2, and their respective spin vectors, $\vec{S}_1$ and $\vec{S}_2$.[1] In the absence of an intrinsic mass or length scale in vacuum GR, the binary dynamics can be written in terms of the dimensionless combination fM, where f is the GW frequency and $M = m_1 + m_2$ is the redshifted total mass of the system (with $G = c = 1$ throughout this paper). Since the total mass sets the overall frequency scale, it can be conveniently factored out of NR waveforms. It is further convenient to introduce the dimensionless spin vector of each black hole as $\vec{\chi}_i = \vec{S}_i/m_i^2$ $(i = 1, 2)$. We also define the mass ratio as $q = m_1/m_2 \geq 1$ by convention. In addition to these intrinsic parameters, the full gravitational waveform observed at the detector depends on seven extrinsic parameters: the inclination angle $\theta_{\rm JN}$ (angle between the line of sight and the total angular momentum); the luminosity distance $D_{\rm L}$; sky location (right ascension α and declination δ); polarization angle ψ; coalescence time, $t_{\rm c}$; and coalescence phase, $\phi_{\rm c}$. For precessing BBHs, the angular parameters are always specified at some reference frequency $f_{\rm ref}$.

Unfortunately, in most cases NR simulations are too computationally expensive to be used directly in data-analysis applications such as parameter estimation (PE) of BBHs from GW observations. Consequently, several fast-to-evaluate approximate models have been developed that generate complete gravitational waveforms that capture all three stages of binary evolution. For instance, the recent GWTC-4.0 catalog of compact binary mergers reported by the LIGO-Virgo-KAGRA (LVK) Collaboration [8], covering the first part of the fourth observing run, relied on waveform models from three main families: `Phenom` [52,53], `SEOBNR` [54], and `NRSurrogate` [55,56]. The `Phenom` and `SEOBNR` families combine analytic and numerical input to construct complete inspiral-merger-ringdown models, calibrating a theoretically motivated ansatz for the merger-ringdown regime against NR data. In contrast, the `NRSurrogate` models are constructed entirely from NR simulations, interpolating across the available NR data. The state-of-the-art waveform models for quasi-circular precessing BBH signals from these families include: `IMRPhenomXPNR` (`XPNR`) [57], `IMRPhenomTPHM` (`TPHM`) [58], `SEOBNRv5PHM` (`v5PHM`) [59], and `NRSur7dq4` [60]. The `XPNR` model computes the GW strain in the frequency domain, whereas the `TPHM`, `v5PHM`, and `NRSur` models operate directly in the time domain.

In the `XPNR` model, inspiral double-spin precession is described by evolving the SpinTaylor equations [61], and phenomenological fits calibrated to single-spin precessing NR simulations (covering mass ratios up to 8 and dimensionless spins on the larger BH up to 0.8) are applied for the merger and ringdown [62], while the model also incorporates the mode asymmetry of the dominant multipole [63]. In contrast, `TPHM` and `v5PHM` are calibrated to NR simulations only for aligned-spin binaries [64,65] and model precession by extending post-Newtonian or effective-one-body results [66,67].[2] All three models include the $(\ell, |m|) = (2, 2), (2, 1), (3, 3), (4, 4)$ multipolar harmonics, with `XPNR` additionally including the (3, 2) mode, `TPHM` additionally including the (5, 5) mode, and `v5PHM` additionally including the (3, 2), (4, 3), and (5, 5) modes. The `NRSur` model is fully calibrated to NR waveforms across the BBH parameter space up to spin magnitudes $\chi_{1,2} = 0.8$ and mass ratios of 4, with extrapolation extending to $\chi_{1,2} = 1$ and mass ratios of 6. By design, it includes all multipoles up to $\ell = 4$ as well as precessional features such as mode asymmetry. In this work, we consider a subset of state-of-the-art waveform models. `XPNR`, `v5PHM`, and `NRSur7dq4` are currently used in LVK analyses, but only `XPNR` and `v5PHM` can be evaluated at high mass ratios. We also include `TPHM`, which is being used in LISA studies [69].

There are multiple extensive catalogs of BBH NR simulations, covering different regions of the parameter

[1]Throughout, we denote the more massive BH in a binary as the primary and labeled with a "1."

[2]We note that a recent update of the `v5PHM` model incorporates multipole mode asymmetries (`SEOBNRv5PHM_w/asym`) [68]. In this work, we use the earlier version, and this choice is expected to have a negligible impact on our conclusions, as explicitly demonstrated in the Appendix.

space [49–51,70–75]. We now have somewhat systematic coverage of quasicircular precessing simulations, extending to mass ratios up to 8 and dimensionless spins up to $|\vec{\chi}| \leq 0.8$ [49,51]. NR simulations with mass ratios beyond 8 that are sufficiently long and accurate for gravitational waveform modeling remain sparse. As future GW detectors become more sensitive, they are expected to observe a wider variety of sources, including high mass-ratio binaries and strongly precessing systems. This highlights the need for more systematic NR simulations covering high-mass-ratio binaries with diverse spin orientations.

Here we report NR simulations of highly asymmetric, precessing BBH systems with a mass ratio of 18 and a dimensionless spin magnitude of 0.8 on the larger BH, while the smaller BH is nonspinning. The simulations span five spin misalignment angles, θ_{LS_1}, defined as the angle between the spin vector of the larger BH and the Newtonian orbital angular momentum: 30°, 60°, 90°, 120°, and 150°. These simulations are primarily intended to facilitate the development of precessing gravitational waveform models calibrated against NR waveforms in the high mass-ratio and strongly precessing regimes.

These new simulations also allow us to demonstrate the critical need for NR calibration of waveform models in the high-q, high-precession regime by comparing current models to our new simulations in two ways. The first is to quantify the accuracy of current models by calculating the standard mismatch between the models and NR waveforms. We find that all existing models generally exhibit substantial mismatches of $\gtrsim 0.1$ (i.e., matches $\lesssim 0.9$), suggesting that they are of poor accuracy in this regime.

Although these mismatches are at least an order of magnitude higher than those typical in the models' calibration region, poor mismatches do not necessarily translate into parameter biases, as discussed in Ref. [76], and demonstrated in practice for GW observations in Ref. [7]. To test the models' performance for parameter recovery [25,77], we additionally perform targeted PE using a subset of state-of-the-art waveform models, injecting these NR simulations as signals into the three-detector LIGO-Virgo network. The results confirm that significant improvements in waveform modeling are required to achieve reliable parameter inference for highly asymmetric, strongly precessing binaries.

The outline of the rest of the paper is as follows. In Sec. II, we briefly outline the details of the methods used to assess the accuracy of the waveform models. In Sec. III, we briefly describe the BAM code [38,39] used to generate the NR simulations of BBH systems presented here, outline the initial data and properties of these simulations, and perform a waveform accuracy analysis to validate them. In Secs. IV and V, we assess the accuracy of current BBH waveform models by comparing them with these NR simulations and highlight the need for improved modeling of highly asymmetric, strongly precessing BBH systems. We present our conclusions and thoughts in Sec. VI.

II. METHODS FOR EVALUATING WAVEFORM ACCURACY

We begin by outlining two complementary methods to quantify waveform accuracy in the context of gravitational wave astronomy: (1) the mismatch, which quantifies the disagreement between waveforms with respect to a detector's noise curve, and (2) parameter estimation, which, given an NR waveform that we assume to be a proxy for the true signal, allows us to quantify how well a set of waveform models will recover individual binary source parameters.

Neither measure is ideal. The mismatch has the advantage of providing a single measure of "distance" between two waveforms (see, e.g., Refs. [76,78]), but, when evaluated between a fiducial "true" waveform and a model, does not tell us how that difference will manifest as loss of signal power, or as parameter biases, or which parameters will be biased and by how much. We can learn these by performing parameter estimation on the true waveform using the model, but parameter estimation is typically far more computationally expensive than computing mismatches, and we can only assess the model's performance for discrete instances of the binary's inclination, polarization, sky position, etc.

In this work we use the mismatch to quantify the overall accuracy of our NR simulations, and both the mismatch and parameter estimation to assess the accuracy of current waveform models in the high-mass-ratio, high-precession regime.

A. Mismatches

To quantify the agreement or disagreement between two waveforms—for example, a model template h_t and an NR signal h_s—we define the standard noise-weighted inner product as [78],

$$\langle h_t | h_s \rangle = 4\mathrm{Re} \int_{f_{\min}}^{f_{\max}} \frac{\tilde{h}_t(f)\tilde{h}_s^*(f)}{\tilde{S}_n(f)} \mathrm{d}f, \tag{1}$$

where the tilde denotes a Fourier transform, the waveforms are expressed as functions of frequency f, the detector is sensitive in the frequency range $f \in [f_{\min}, f_{\max}]$, and $\tilde{S}_n(f)$ is the one-sided power spectral density (PSD) of the detector [79,80]. The optimal signal-to-noise ratio (SNR) of a GW signal h is given by $\rho = |h| = \sqrt{\langle h|h \rangle}$. A commonly used measure to evaluate the accuracy of the waveform model is the *match*, defined as the noise-weighted inner product between two normalized waveforms, maximized over a subset of template parameters θ_{opt} described below,

$$\mathcal{M}(h_t, h_s) = \max_{\theta_{\mathrm{opt}}} \frac{\langle h_t | h_s \rangle}{|h_t||h_s|}. \tag{2}$$

The match is equal to unity when the two waveforms are identical, aside from an overall amplitude scaling. To quantify

the difference between two waveforms, it is convenient to define the *mismatch*,

$$\mathfrak{M}(h_t, h_s) = 1 - \mathcal{M}(h_t, h_s). \tag{3}$$

A lower mismatch indicates better agreement between the two waveforms. In GW data analysis, assuming that the statistical likelihood is approximately Gaussian (which is valid in the high-SNR limit), waveform modeling systematics will not bias the PE if the model's mismatch with the true signal (typically represented by NR simulations that serve as proxies for the true signals) is smaller than $\chi_k^2(1-p)/2\rho^2$, where $\chi_k^2(1-p)$ is the chi-square value at probability p for k degrees of freedom [81–83]. For single-parameter measurements (i.e., $k = 1$), this criterion for the 90% credible interval reduces to $\mathfrak{M} \leq 1.35/\rho^2$. As an example, for a signal with SNR $\rho = 30$, the mismatch between the waveform model and the NR simulation must be $\mathfrak{M} \lesssim 1.5 \times 10^{-3}$ to prevent the true parameter from lying outside the 90% credible interval of the marginalized 1D posterior distribution, in the absence of noise. However, this is only a conservative criterion: a model will not report biased parameters if it meets this criterion, but failing the criterion does not necessarily imply a bias. For a detailed discussion of the proper use and interpretation of this mismatch criterion, we refer interested readers to Ref. [76].

We now explain the set of parameters $\theta_{\rm opt}$ that are optimized to obtain the maximized match. Shifting $t_{\rm c}$ simply moves the binary's coalescence time, while adjusting $\phi_{\rm c}$ changes only the binary's initial orbital orientation. For nonprecessing waveforms that include only the quadrupolar modes, ψ is degenerate with $\phi_{\rm c}$. In addition, the inclination, luminosity distance, and the sky-position angles are mutually degenerate, since variations in these parameters merely rescale the overall amplitude of the waveform. Thus, the waveform depends on these parameters only through an overall phase and an overall amplitude.

In the precessing case, the source's orientation relative to the detector evolves as the orbit precesses. Consequently, the symmetries present in the nonprecessing case no longer apply; both the inclination and the polarization angle must be explicitly considered. We follow the procedure for computing the optimized match for precessing BBH systems outlined in Refs. [84,85] and implemented in Ref. [62] (see Sec. XI A therein for further details). For signals from nonprecessing BBH systems that consist only of the dominant $(\ell, |m|) = (2, 2)$ modes, an analytical expression can be derived for the match maximized over $D_{\rm L}$, $\phi_{\rm c}$, ψ, α, and δ of the model template h_t. Reference [85] provides an analytical expression for the SNR maximized over these parameters [see Eq. (27) and Secs. III and IV for a detailed derivation of Eq. (27)]. Dividing the maximized SNR by the signal norm, i.e., $\langle h_s | h_s \rangle$, yields the match maximized over $D_{\rm L}$, $\phi_{\rm c}$, ψ, α, and δ of the model template h_t. Since the coalescence time $t_{\rm c}$ parametrizes time translations of h_t, the match can be further maximized over $t_{\rm c}$ in a computationally efficient manner using an inverse fast Fourier transform routine [86,87].[3] It should be noted that all parameters are defined at the reference frequency $f_{\rm ref}$.

This procedure can be extended to templates that include higher-order modes and precession, allowing estimation of the match maximized over $D_{\rm L}$, ψ, α, δ, and $t_{\rm c}$ of the model template h_t. For templates with higher-order modes, each (ℓ, m) mode carries an $m\phi_{\rm c}$ dependency; thus, we further optimize the match numerically over $\phi_{\rm c}$. We also numerically optimize the match over rotations of the in-plane spin components of the template at the reference frequency, following Refs. [62,89]. In order to perform the numerical optimizations, we use the Nelder-Mead algorithm as implemented in the SciPy Python package [90]. While computing the optimized match, we fix the masses and other spin parameters (leaving only the first Euler angle, or precession angle, free), as well as the inclination angle $\theta_{\rm LN}$ (angle between the line of sight and the orbital angular momentum at the reference frequency $f_{\rm ref}$) to be the same for both the template h_t and the signal h_s.

Here, we adopt the `aLIGOZeroDetHighPower` PSD of advanced LIGO [91] with $f_{\rm min} = \max\{20\ {\rm Hz}, 1.35 \times f_{\rm min}^{\rm NR}\}$ and $f_{\rm max} = 2048$ Hz, where $f_{\rm min}^{\rm NR}$[4] is the frequency of the (2, 2)-mode in Hz at the beginning of the NR waveform, measured in a frame whose z-axis is instantaneously aligned with the orbital angular momentum. We compute the optimized match over $D_{\rm L}$, $\phi_{\rm c}$, ψ, α, and δ of the model template for optimally oriented signals (i.e., $\alpha = \delta = 0$ for h_s) in a single detector. We consider representative values of the redshifted total mass M. For each mass, we adopt 13 equally spaced inclination angles from $\theta_{\rm LN} = 0$ to π. For each combination of $(M, \theta_{\rm LN})$, we evaluate the optimized match over a grid of signal phase and polarization values, $\phi_{\rm c,s}$ and $\psi_{\rm s}$: eight equally spaced $\phi_{\rm c,s}$ values from 0 to 2π and 13 equally spaced $\psi_{\rm s}$ values from 0 to π. For each combination, we record the maximum and minimum optimized match values.

The number of GW cycles in the frequency range of 20–2048 Hz can vary significantly between different binary configurations, even for a fixed total mass M. The mismatch depends strongly on the total cycles, since in general it is easier to optimize Eq. (2) for a less complex signal; this point is illustrated in Ref. [92]. This should be borne in mind when attempting to compare mismatches between our five configurations, where, for a binary total mass of

[3]We note that, when using the inverse fast Fourier transform routine for timeshift optimization, the resolution of the discrete timestep is increased by padding the frequency-domain data. This is particularly important for signals that differ only slightly in the linear-in-frequency contributions to their phases [87,88].

[4]$f_{\rm min}^{\rm NR}$ is conventionally the same as the reference GW frequency $f_{\rm ref}$, at which the binary parameters are defined.

$150M_\odot$, the number of GW cylces in band varies between $\sim$30 ($\theta_{\mathrm{LS}_1} = 30°$) and $\sim$5 ($\theta_{\mathrm{LS}_1} = 150°$).

Since lower matches are typically associated with edge-on systems that have reduced SNR and consequently a lower likelihood of source detection, we then follow Refs. [62,84,89,93] and also evaluate the SNR-weighted match in this work. The SNR-weighted match is obtained by averaging the optimized match over different signal phases and polarizations, with each contribution weighted by the corresponding SNR, thereby accounting for the detectable volume,

$$\mathcal{M}_{\mathrm{w}} = \left(\frac{\sum_{\psi_{\mathrm{s}},\phi_{\mathrm{c,s}}} \mathcal{M}^3 \langle h_{\mathrm{s}} | h_{\mathrm{s}} \rangle^{3/2}}{\sum_{\psi_{\mathrm{s}},\phi_{\mathrm{c,s}}} \langle h_{\mathrm{s}} | h_{\mathrm{s}} \rangle^{3/2}} \right)^{1/3}, \tag{4}$$

where the summation runs over different values of the coalescence phase and polarization of the source, i.e., in our case, of the NR waveform. The SNR-weighted mismatch is defined as $\mathfrak{M}_{\mathrm{w}} = 1 - \mathcal{M}_{\mathrm{w}}$. We note that an identical mismatch calculation was also employed in the waveform systematics study of the heaviest BBH merger event, GW231123, reported to date by the LVK Collaboration [7].

B. Parameter estimation

We further assess model accuracy by investigating the parameter biases incurred during Bayesian parameter estimation. This offers a direct view of how waveform inaccuracies translate into biases in the inferred source properties. Bayesian inference is the process of estimating the parameters $\vec{\theta}$ that characterize a signal, given a hypothesis $\mathcal{H}$ specifying the signal model h (here, a waveform model for the GW signal in GR) and observed data d (here, GW data) corresponding to that signal. Consider the time series d measured by the ground-based GW interferometers. Under the hypothesis that a GW signal exists in the data, the data can be modeled as $d = s + n$, with s being the true signal and n the detector noise. The *posterior probability distribution* on model parameters $\vec{\theta}$, can be obtained via Bayes's theorem

$$p(\vec{\theta}|d, \mathcal{H}) = \frac{\mathcal{L}(d|\vec{\theta}, \mathcal{H}) \pi(\vec{\theta}|\mathcal{H})}{\mathcal{Z}(d|\mathcal{H})}, \tag{5}$$

where, $\pi(\vec{\theta}|\mathcal{H})$ is the *prior probability* distribution of $\vec{\theta}$, $\mathcal{L}(d|\vec{\theta}, \mathcal{H})$ is the *likelihood function*, and $\mathcal{Z}(d|\mathcal{H}) \equiv \int \mathcal{L}(d|\vec{\theta}, \mathcal{H}) \pi(\vec{\theta}|\mathcal{H}) d\vec{\theta}$ is the *Bayesian evidence* in favor of the hypothesis $\mathcal{H}$ (i.e., the signal model h). Under the assumption that the detector noise n is wide-sense stationary and Gaussian, the likelihood function is proportional to [79,80]

$$\mathcal{L}(d|\vec{\theta}, \mathcal{H}) \propto \exp\left\{ -\frac{1}{2} \langle d - h(\vec{\theta}) | d - h(\vec{\theta}) \rangle \right\}. \tag{6}$$

For quasicircular, precessing BBHs, evaluating the posterior distribution in GW astronomy involves a 15-dimensional integral, making direct calculation challenging. A common approach to reduce this complexity is to marginalize analytically over selected parameters [94–97]; in this work, we adopt this approach for the luminosity distance [96,98]. Due to the high dimensionality of the evidence integral, stochastic sampling methods are commonly used to generate samples from the posterior distribution. A variety of software packages exist for performing Bayesian inference in GW astronomy, many of which implement the nested sampling algorithm. Here, we use the Bayesian parameter inference package Bilby [99,100] with the Dynesty [101] sampler, interfaced via the Bilby-PIPE wrapper [100], to generate discrete samples for the probability distribution function $p(\vec{\theta}|d, h)$.

Our sampler setup closely follows that used by the LVK Collaboration in their production analyses [8,102], employing the `rwalk` algorithm implemented in Bilby. All analyses use 1000 live points and an average of 60 accepted steps per Markov chain Monte Carlo iteration. We also adopt the prior probability distribution $\pi(\vec{\theta}|h)$ following the standard setup used in the publicly released LVK analysis [8,102]. Uniform priors are assigned for the redshifted component masses and spin magnitudes of the individual BHs, while the spin orientations and other angular parameters are sampled isotropically. The luminosity distance D_{L} prior corresponds to sources uniformly distributed in comoving volume and source-frame time. These choices provide broad, uninformative priors that encompass the relevant parameter space while preserving computational efficiency. Overall, these chosen priors are broad and uninformative, ensuring that the posterior support lies well within the prior bounds while maintaining computational efficiency.

We first generate a synthetic GW-signal injection from the reported NR simulations following the LVK NR injection infrastructure [103], using a fiducial redshifted total mass and the extrinsic parameters specified below. All injections assume the "zero-noise" approximation in a three-detector LIGO-Virgo network [104–106], employing the predicted noise power spectral densities at design sensitivity: `aLIGOZeroDetHighPower` for the LIGO detectors [91], and `AdvVirgo` for the Virgo detector [107]. In this setup, d corresponds to the time-series strain data generated from the NR simulations. The zero-noise approximation enables us to isolate potential systematic biases and avoid statistical fluctuations due to specific noise realizations. For all NR injection analyses reported here, we adopt the following parameter values for the source's sky location and polarization: $\alpha = 240°$, $\delta = 45°$, and $\psi = 60°$.

We also choose an optimal network SNR of ∼50 for the injected NR signals, adjusting the luminosity distance for each injection accordingly.

We then perform Bayesian parameter estimation using the three state-of-the-art waveform models `XPNR`, `TPHM`, and `v5PHM`. A lower cutoff frequency of 20 Hz is adopted when evaluating the likelihood integral in Eq. (6), while the upper cutoff frequency varies between systems, primarily depending on the redshifted total mass. These frequency limits are selected to retain the signal content while avoiding unnecessary computational overhead from excessively high frequency cutoffs. We set both the injected NR signal and the recovery template waveform to start at $f_{\rm min}^{\rm NR}$, to avoid missing frequency content from higher-order multipoles [108]. This is particularly important for Bayesian PE analyses that use frequency-domain waveform approximants. The value of $f_{\rm min}^{\rm NR}$ varies across NR simulations and is determined by the characteristics of the NR waveform and the redshifted total mass. Other analysis settings—such as the signal duration and sampling frequency—are consistent with standard practices used in LVK studies [8,102], and are adjusted for each system based on the physical characteristics of the corresponding injection. The sampler configuration has been described previously and remains the same across all analyses. Lastly, we note that we usually generate synthetic GW signals from NR simulations using all the available modes (i.e., $\ell \leq 4$ modes) and perform the PE recovery using all the available modes in the waveform approximant, unless otherwise specified.

We compare the true injected parameters $\vec{\theta}_{\rm Inj}$ of the NR signal with the posterior distribution inferred from the PE analysis to assess how well the model captures the underlying signal. In the zero-noise approximation, the inferred posterior distribution should peak at the true parameters $\vec{\theta}_{\rm Inj}$ when the model perfectly describes the NR signal i.e., $s(\vec{\theta}_{\rm Inj}) = h(\vec{\theta}_{\rm Inj})$.[5] If the model does not faithfully represent the underlying NR signal, the resulting posterior distribution can exhibit systematic biases, i.e., the posterior may peak at values different from $\vec{\theta}_{\rm Inj}$. In GW astronomy, it is common practice to regard a posterior distribution as biased when the marginalized one-dimensional distribution excludes the true parameter value from its 90% credible interval.

III. NR SIMULATIONS

We present a set of NR simulations of high-mass-ratio, high-spin, highly precessing BBHs, motivated by the lack of simulations in this region of parameter space. An earlier set of simulations of non-precessing binaries at mass-ratio 18 was used a decade ago to calibrate the aligned-spin BBH waveform model `IMRPhenomD` [109,110], as well as the aligned-spin sector of successor `Phenom` models [111]. The most recent SXS catalog contains precessing binaries up to mass-ratio 15, but there is only one configuration with a primary spin of 0.8, all others are below 0.5 [51]. Simulations at larger mass ratios are also available (for example, in the RIT catalog [112]) but are restricted to aligned-spin configurations. A set of highly precessing, high-mass-ratio simulations enables an initial assessment of how well current BBH waveform models extrapolate into this region of parameter space and provides guidance for first attempts to extend modelling efforts for precessing binaries to higher mass ratios.

Our ability to cover the high-q region in NR simulations is limited by the high computational cost of simulations in this part of parameter space. The finest length scale that must be resolved in our simulations, which also determines the timestep, is dictated by the smallest BH, while the overall time scale of the simulation is determined by the total mass of the binary; this introduces an approximate factor of q in the scaling of the computational cost. Since the time to merger from a given start frequency also scales as q, the computational cost is expected to scale at least as q^2; see, for example, Ref. [113]. In addition, the code we use, BAM [38,39], is based on Cartesian boxes, which implies a prohibitive computational cost in maintaining sufficient angular resolution for wave extraction as we extract further from the source. At the extraction radii that are feasible in our simulations $\sim 100M$, we risk significant errors in both the amplitude and phasing of subdominant modes. In an earlier set of precessing-binary simulations, which extended up to mass-ratio 8, we found that the wave extraction error was the dominant error source [49]. For both of these reasons, we consider only a limited set of simulations and prioritize the most strongly precessing configurations.

We simulate five configurations at mass-ratio 18, with a spin of 0.8 on the larger BH (the smaller BH is non-spinning), and, as in Ref. [49], a set of five spin misalignment angles. The simulations are $\sim 3000M$ in length; these are roughly 50% longer than the earlier BAM simulations used for waveform modeling [49,110], in order to capture as large a frequency range as feasible with the computational resources available to us. From a given starting frequency, the total inspiral time scales as approximately $1/\eta$ [113,114], where $\eta = \frac{q}{(1+q)^2}$ is the symmetric mass-ratio. For a binary with $q = 18$, $\eta \approx 0.05$, so the total inspiral time will be roughly five times that of a comparable equal-mass system, with $\eta = 0.25$ if starting at the same orbital frequency. For this reason, although our simulations are longer than in the earlier BAM catalog, the starting frequencies are higher.

We construct our initial data, find low-eccentricity parameters, and set up the computational grids following the same

[5]While sampling error and the finite number of posterior samples typically introduce only negligible shifts, parameter degeneracies can in principle lead to much larger offsets, though they are not expected to significantly affect the results reported here.

TABLE I. Initial-data parameters, binary properties, and remnant properties of each NR simulation. $M\omega_{\rm orb}$ is the binary's initial orbital frequency, $t_{\rm M}$ is the time to merger, and $N_{\rm orb}$ is the number of orbits to merger. See text for more details.

Name	q	χ_1	$\theta_{\rm LS_1}$ (°)	$\chi_{\rm eff}$	$\chi_{\rm p}$	D/M	e ($\times 10^{-3}$)	$M\omega_{\rm orb}$ ($\times 10^{-2}$)	$t_{\rm M}/M$	$N_{\rm orb}$	M_f	χ_f	$V_{\rm kick}$ (km s^{-1})
CF_81	18	0.80	30	0.66	0.40	8.33	1.2	3.4	3074	24.6	0.990	0.814	53
CF_82	18	0.80	60	0.38	0.69	8.93	1.5	3.2	3227	23.4	0.992	0.787	148
CF_83	18	0.80	90	0.00	0.80	9.70	2.0	2.9	3265	20.4	0.994	0.729	171
CF_84	18	0.80	120	−0.38	0.69	10.47	1.3	2.6	3150	17.3	0.995	0.645	117
CF_85	18	0.80	150	−0.66	0.40	11.05	2.4	2.5	3014	15.2	0.996	0.561	34

procedures and criteria as in Ref. [49]. Specifically, we generate initial data by using the Bowen-York analytic solution to the momentum constraints for two boosted, spinning black holes [115] and solve the Hamiltonian constraint in the puncture construction [116] with a pseudospectral solver [117]. We evolve the data using the standard moving-puncture treatment [36,37] of the Baumgarte-Shapiro-Shibata-Nakamura decomposition [118,119] of the $3+1$ Einstein equations [120]. We use sixth-order-accurate spatial finite differencing in the bulk [39], and fourth-order Runge-Kutta time integration with a Courant factor of 0.25. To adequately resolve each BH, the smallest box following each BH should have a width between 1.2 and 1.5 times the maximum coordinate diameter of the apparent horizon. We also require sufficient resolution to resolve the (4, 4) multipole in the ringdown, with a grid spacing of $\leq 1/(20 f_{\rm RD})$, where $f_{\rm RD}$ is the ringdown frequency of the dominant (2,2) mode. Within these specifications, the basic indicator of numerical resolution is the number of points in each direction on the finest refinement level. For the simulations reported in Ref. [49], the standard box size on the finest level was 96^3 points, while here we have used 120^3 points in an attempt to achieve comparable phase accuracy over a longer simulation. We discuss accuracy further in Sec. III A; we find that the computational cost to achieve accuracy comparable to earlier simulations is prohibitive, although our new simulations are accurate enough for initial modeling purposes and for a preliminary assessment of the accuracy of current waveform models.

Selected properties of the five NR simulations are listed in Table I. All configurations have mass-ratio $q = 18$, primary spin magnitude $\chi_1 = 0.8$ and no spin on the secondary. The primary's spin is tilted with respect to the Newtonian orbital angular momentum (i.e., the normal to the orbital plane) by $\theta_{\rm LS_1}$. We report the standard effective spin parameters, the effective inspiral spin $\chi_{\rm eff} \equiv (\chi_1 \cos\theta_{\rm LS_1} + q\chi_2 \cos\theta_{\rm LS_2})/(1+q)$ [53] and $\chi_p \equiv \max(\chi_1 \sin\theta_{\rm LS_1}, q\frac{3+4q}{4+3q}\chi_2 \sin\theta_{\rm LS_2})$ [121], as calculated from the initial parameters of the simulation. The separation at the beginning of each simulation is D/M. The eccentricity e is estimated over the interval $[200, 1000]M$ of the simulation using the method described in [122]. The orbital frequency $M\omega_{\rm orb}$ is derived from the puncture dynamics at the relaxed time $t = 50M$. The time at merger, t_M, is defined as the duration from the relaxed time until the separation between the two punctures falls to $0.5M$. The number of orbits, $N_{\rm orb}$, is calculated from the relaxed time, where $M\omega_{\rm orb}$ is reported up to t_M. The mass and spin of the remnant BH, M_f and χ_f, are calculated from the apparent horizon of the final black hole. The radiated linear momentum is calculated as described in Ref. [38], and and the recoil (or kick) velocity $V_{\rm kick}$ of the remnant BH is obtained by integrating the radiated linear momentum from the relaxed time until the end of the simulation.

The GW content of the binary system is extracted at finite radius from the source using the Newman-Penrose scalar ψ_4 [123], following the procedures outlined in Ref. [124]. The observable GW strain h is obtained by performing a double time integration of ψ_4.[6] Here, we have extracted up to $\ell = 4$ multipole moments of the gravitational-wave strain; the memory modes (i.e., multipole moments with $m = 0$) are not included. Figure 1 shows plus polarization of the gravitational-wave strain in the time domain from the five reported NR simulations for an illustrative binary with redshifted total mass $150M_\odot$, $\theta_{\rm LN} = 30°$, and $D_L = 300$ Mpc (polarization angle set to zero). The reported waveforms reveal strikingly rich modulations, arising from the combined effects of spin precession and higher-harmonic contributions, highlighting the intricate structure of high mass-ratio precessing BBH signals.

As noted earlier, the total number of inspiral cycles from a given starting frequency scales approximately as $\sim 1/\eta$ [126,127]. For a binary with mass ratio $q = 18$, this results in roughly five times more cycles than a nearly equal-mass system from the same starting frequency. For example, GW190521, which was a nearly equal-mass binary with a total redshifted mass of $150M_\odot$ and some evidence of spin-induced orbital precession, had only a few cycles in the LIGO-Virgo band [4]. In contrast, at $150M_\odot$ our CF_83 simulation includes almost ten GW cycles from 20 Hz to merger; see the third panel of Fig. 1. In this configuration the spin lies entirely in the orbital plane, so there is no

[6]In practice, when required for applications such as waveform modeling, we compute it in the frequency domain by dividing the Fourier transform of ψ_4 by ω^2, where ω is the angular frequency [125].

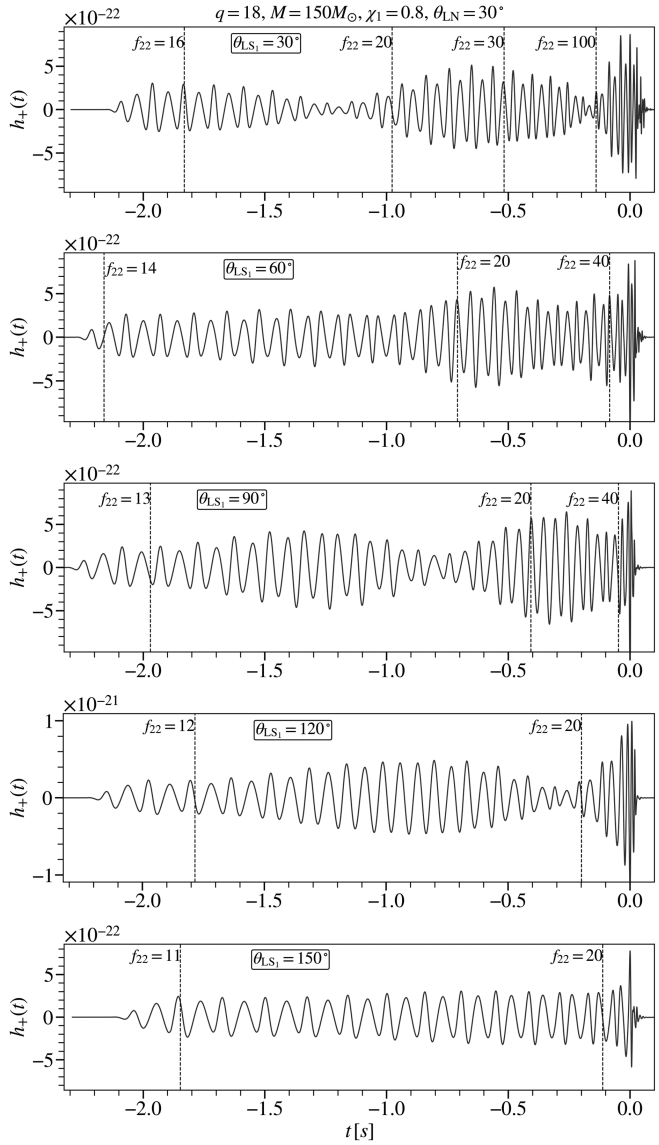

FIG. 1. Plus GW polarization of each of our NR simulations, with redshifted total mass $150M_\odot$, $\theta_{LN} = 30°$, and $D_L = 300$ Mpc. The five panels show each choice of spin misalignment: $\theta_{LS_1} = 30°$, $60°$, $90°$, $120°$, $150°$. The waveforms have been tapered to remove junk radiation.

aligned-spin component and we expect the inspiral rate to be similar to a nonspinning binary with the same masses.

These configurations also clearly illustrate the "orbital hang-up" effect [127,128]: when BH spins are coaligned with the orbital angular momentum, the binary inspirals for longer before merger, producing more GW cycles. Antialigned spins, in contrast, lead to faster mergers with fewer cycles. This effect is evident in our simulations. Considering the instant of binary separation when $f_{22} = 20$ Hz in all five cases shown in Fig. 1, we find that the number of cycles to merger decreases significantly as the spin misalignment angle with the orbital angular momentum increases (top to bottom panels in Fig. 1), with the inspiral shortening due to the orbital hang-up effect.

A. Accuracy of NR simulations

To evaluate the accuracy of the NR simulation data at such a large mass ratio, we study the configuration `CF_81`, with parameters $\{q, \chi_1, \theta_{LS_1}\} = \{18, 0.8, 30°\}$. The two primary sources of error in our NR waveforms are the finite numerical resolution of the simulation and the finite radius at which the waveform data are extracted. To evaluate the impact of finite resolution, we performed three simulations of this configuration at low, medium, and high resolutions. These resolutions correspond to $N = \{120, 144, 172\}$ grid points within the refinement boxes that surround the punctures. The width of the smallest refinement box around each BH is typically of the order of $\sim 2m$, where m is the mass of the BH. The waveform data are extracted at a distance of $90M$ from the source for all three resolutions. To quantify errors in the NR waveforms arising from the finite resolution of the simulation, we estimate the mismatch between the highest-resolution ($N = 172$) waveform and those at medium ($N = 144$) and lower ($N = 120$) resolutions.

In the mismatch calculation, we adopt a redshifted total mass of $M = 150M_\odot$ and estimate the optimized mismatch as outlined in Sec. II A. Note that here we cannot optimize the mismatch over rotations of the in-plane spin components of the template. Therefore, the match is optimized over all extrinsic parameters, excluding the binary inclination angle θ_{LN}, while the intrinsic parameters (masses and spin parameters) are kept fixed according to the NR simulation configuration. We present the mismatches as a function of the inclination angle θ_{LN} in Fig. 2. Solid markers indicate the SNR-weighted mismatch, while vertical lines denote the range of mismatches across the sampled polarizations and phases. For example, for $\theta_{LN} = \pi/3$, the smallest mismatch between the $N = 120$ and $N = 172$ NR waveforms is 0.016, while the smallest mismatch between the $N = 144$ and $N = 172$ waveforms is 0.001. The average mismatch is 0.03 between the $N = 120$ and $N = 172$ simulations, and 0.002 between the $N = 144$ and $N = 172$ simulations. We note that these are much larger than the estimate of the mismatch error for the simulations in Ref. [49], which was $\sim 10^{-4}$ (see Fig. 8 of Ref. [49]) for a binary with total mass $100M_\odot$.

Adopting the Richardson extrapolation scheme, Ref. [49] derived a relation [their Eq. (20)] describing how the mismatch between two NR waveforms is expected to scale with an expansion parameter (e.g., numerical resolution or waveform extraction radius) characterizing the NR waveforms, assuming a given convergence order (see also Ref. [129]). If we perform simulations with two different numerical resolutions characterized by expansion parameters $\Delta_1 = 1/N_1$ and $\Delta_2 = 1/N_2$ (where the moving boxes around the punctures have N^3 points) and assume a convergence order n, then the mismatch between the waveforms h_1 and h_2, computed from the two simulations, is given by

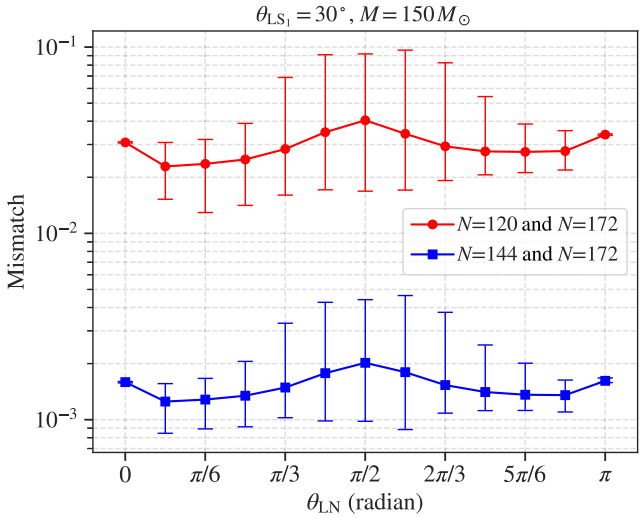

FIG. 2. Mismatch between NR waveforms for CF_81 configuration ($\{q, \chi_1, \theta_{LS_1}\} = \{18, 0.8, 30°\}$) at different numerical resolutions as a function of the binary inclination angle θ_{LN} (angle between the line of sight and the initial orbital angular momentum) at redshifted total mass $150M_\odot$. The mismatch is optimized over luminosity distance, coalescence time and phase, sky location, and polarization. Each solid marker denotes the SNR-weighted mismatch, while the vertical lines show the range between the minimum and maximum mismatch obtained across the sampled signal polarizations and phases (see text for details).

$$\mathfrak{M}(h_1, h_2) = \kappa(\Delta_1^n - \Delta_2^n)^2, \tag{7}$$

where κ is a constant. From the two average mismatch values quoted above from Fig. 2, we find a consistent value of κ for $n \approx 6$, which implies sixth-order convergence for our simulations. The code uses sixth-order accurate spatial finite differencing, but fourth-order accurate Runge-Kutta timestepping. At sufficiently high resolution, we expect the fourth-order timestepping error to dominate. Observing apparent sixth-order convergence may indicate that the timestepping error does not yet dominate, or—more likely given the simulation length—that the system is not yet in the fully convergent regime.

Assuming sixth-order convergence, the mismatch between the waveform from the $N = 172$ simulation and the true waveform at infinite resolution ($N \to \infty$) is estimated to be $\sim 6 \times 10^{-4}$, while for the waveform from the $N = 120$ simulation it is ~ 0.04. This suggests that the $N = 120$ simulation is not suitable for high-precision studies. Nevertheless, we will see later that, despite its lower resolution, the $N = 120$ simulation captures the broad physical features of the waveform reasonably well. The $N = 172$ simulation, by contrast, is likely of comparable accuracy to simulations in the previous BAM catalog [49].

To further evaluate the impact of waveform extraction at finite radius, we consider the $N = 120$ simulation and extract the waveform at three different radii from the source: $R_{ext} = 70M$, $80M$, and $90M$. To quantify errors in the NR waveforms arising from finite-radius extraction, we perform a mismatch study similar to the previous analysis of finite-resolution errors, but now comparing the waveform extracted at $R_{ext} = 90M$ with those extracted at $R_{ext} = 70M$ and $80M$. The results are shown in Fig. 3.

As anticipated, the mismatches between the waveforms extracted at $R_{ext} = 80M$ and $R_{ext} = 90M$ are smaller than those between $R_{ext} = 70M$ and $R_{ext} = 90M$, typically by a factor of $\sim 2 - 5$. The average mismatch is approximately 2×10^{-4} for the $70M$ vs $90M$ mismatches, and 7×10^{-5} for the $80M$ vs $90M$ mismatches. For comparison, these values are again notably larger than the estimated mismatch error reported in Ref. [49], which was $\sim 10^{-5}$ (see Fig. 9 of Ref. [49]) for a binary with total mass $100M_\odot$.

From this mismatch study, we infer first-order convergence for the finite-radius extraction error, i.e., the error due to finite extraction radius falls of as $\sim 1/R_{ext}$. Assuming first-order convergence, we estimate the mismatch between the waveform extracted at $R_{ext} = 90M$ and the true waveform at $R_{ext} \to \infty$ to be $\sim 4 \times 10^{-3}$. Using Eq. (23) of Ref. [49], we estimate the mismatch between our NR waveforms—computed at finite resolution ($N = 120$) and extraction radius ($R_{ext} = 90M$)—and the true waveform at infinite resolution and extraction radius to be $(\sqrt{\mathfrak{M}_{\text{resolution}}} + \sqrt{\mathfrak{M}_{\text{extraction}}})^2 \approx (\sqrt{0.04} + \sqrt{0.004})^2 \approx 0.07$. This corresponds to an upper bound of $\sim 7\%$ on the mismatch error of the reported NR waveforms. The overall mismatch error for the higher-resolution simulations is 0.016 for $N = 144$ and 0.007 for $N = 172$. (For the NR simulations in Ref. [49], the overall mismatch

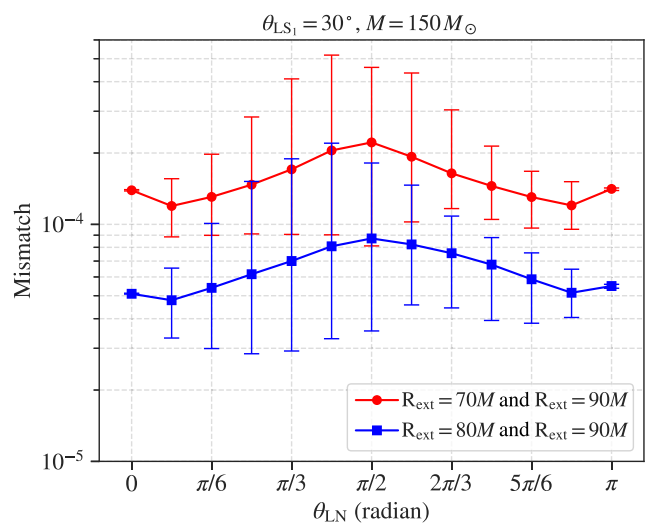

FIG. 3. Mismatch between NR waveforms for CF_81 configuration in Table I at different extraction radii (as indicated in the legend) as a function of the binary inclination angle θ_{LN} at redshifted total masses of $150M_\odot$. The parameters over which the mismatch is optimized, as well as the plotting style and interpretation, are the same as in Fig. 2.

error was estimated at $\sim 4 \times 10^{-3}$ for a binary with total mass $100 M_\odot$.)

Nominally, these mismatch errors are large. The conservative indistinguishability criterion ($\mathfrak{M} \leq 1.35/\rho^2$) [83] suggests that the $N = 120$ waveform could be distinguished from a true waveform at any SNR above 5, i.e., any detectable signal. (For the $N = 172$ waveform, it is any SNR above 14.) However, as illustrated in Ref. [76], this criterion is generally too conservative, and the actual indistinguishability SNR typically occurs at much higher values. We will see in Sec. V that this is also the case for these waveforms; in fact, the two highest resolutions are almost completely indistinguishable at an SNR of 50, and would likely not be distinguishable even at an SNR of 100.

To achieve comparable overall mismatch error to the $q \leq 8$ precessing-binary simulations in the earlier BAM catalog [49] (0.004), we would require each of our simulations to be performed with higher resolution than our $N = 172$ simulation (0.007). With our current methods the computational cost would be prohibitive. The $N = 120$ simulation required 1.6M core hours, while the $N = 172$ simulation, by far the most expensive we have performed to date, required almost an order of magnitude more computational resources, 13.2M CPU hours. We will discuss the implications of this further in Sec. VI.

IV. MISMATCHES AGAINST CURRENT MODELS

We now quantify the accuracy of current waveform models with respect to our new NR simulations. Specifically, we assess the accuracy of three contemporary waveform models (`XPNR`, `TPHM`, and `v5PHM`) by computing the mismatches following the procedure outlined in Sec. II A. We again consider the redshifted total mass of $M = 150 M_\odot$ for the mismatch calculation. We choose this mass to ensure that the detector's sensitivity band contains a full signal for all $\ell \leq 4$ modes.

In Fig. 4, we show the plus polarizations for the NR simulations `CF_81` and `CF_85` (which are used in the parameter estimation studies in Sec. V), choosing a redshifted total mass of $150 M_\odot$ and $\theta_{\rm LN} = 30°$, with the polarization angle set to zero and the luminosity distance fixed at 300 Mpc. We superimpose the NR waveform with predictions from the three waveform models. We optimize over the reference time, reference phase, and in-plane spin components to identify the best-matching template. All models look significantly different compared to the NR waveform.

We then quantify these differences using the mismatch metric described in Sec. II A, providing a more systematic measure of the waveform models' accuracies across the parameter space covered by our NR simulations. The mismatches are optimized over luminosity distance, coalescence time, and phase, sky location, polarization, and precession angle. The optimized mismatch is computed across a grid of signal inclination, polarization, and

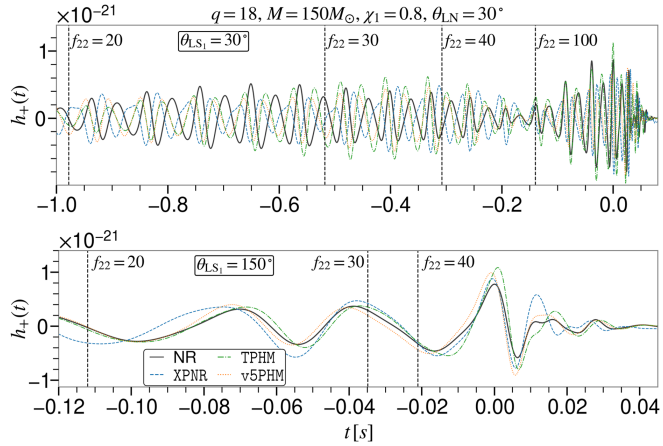

FIG. 4. Plus polarization of NR simulations and three contemporary waveform models, `XPNR`, `TPHM`, and `v5PHM`. Redshifted total mass is $150 M_\odot$, $\theta_{\rm LN} = 30°$, and $D_L = 300$ Mpc (with the polarization angle chosen to be zero). Top: NR simulation `CF_81` ($\theta_{\rm LS_1} = 30°$). Bottom: `CF_85` ($\theta_{\rm LS_1} = 150°$). The `XPNR` (blue dashed), `TPHM` (green dash-dotted), and `v5PHM` (orange dotted) models show significant disagreement with the underlying NR data (solid black).

reference phase values, with each parameter varied over its full physical range as described in Sec. II A. For each NR-model comparison, this procedure yields a collection of optimized mismatch values spanning all sampled parameter combinations. In our results we show the full range of these values.

The mismatches are shown in Fig. 5. In the top panel we restrict to the $\ell = 2$ modes in both the waveform models and the NR simulations. In the bottom panel we include all available modes in the waveform models and modes up to $\ell = 4$ in the NR simulations. We have not considered the memory modes (modes with $m = 0$) from the NR simulations; the available modes in each model are listed in Sec. I.

We find that the mismatches are poor for all configurations. For all models there are configurations where the mismatch is as high as 0.4, i.e., a match of 0.6; these are extremely poor matches that, in the comparable-mass regime, would disqualify a model from use in parameter estimation. High mismatches are found even when restricting to only the $\ell = 2$ modes. Overall we see some improvement for larger values of $\theta_{\rm LS_1}$, but recall that these configurations have large antialigned spin components, leading to merger at lower frequencies (see Fig. 1), and therefore far fewer GW cycles from the mismatch starting frequency $f = 20$ Hz. Note also that these mismatches are typically much larger than the mismatch errors of the NR waveforms. For the $\theta_{\rm LS_1} = 30°$ configuration, the one configuration where we have full NR error estimates (Sec. III A), the mismatch error is 0.007 for the highest-resolution simulation, well below the smallest mismatch (0.02) against any model. We have also found that for this

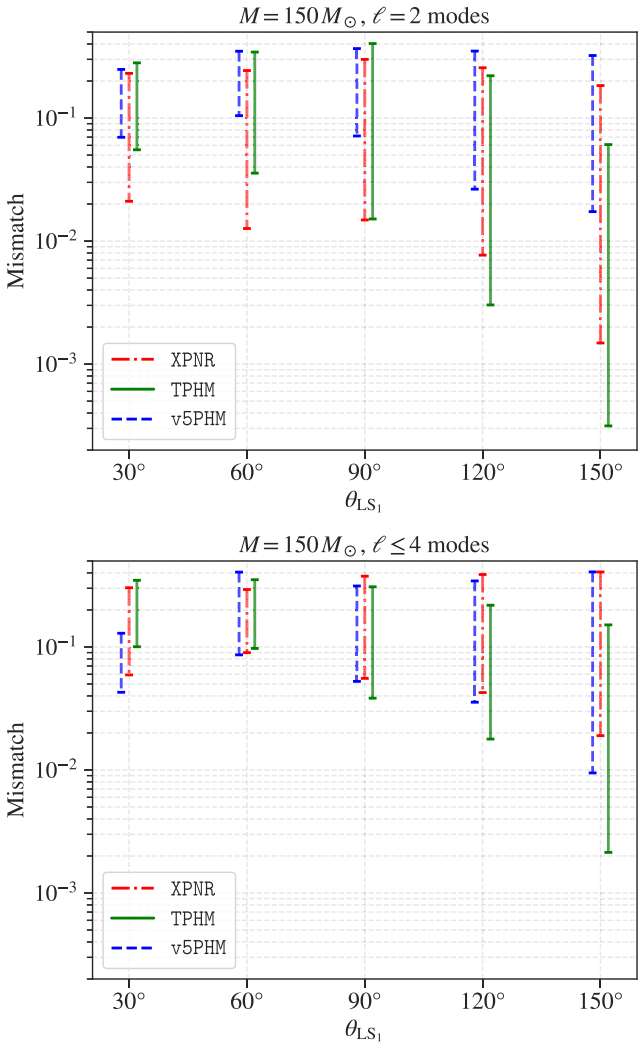

FIG. 5. Mismatches against three current models. The spin misalignment angle θ_{LS_1} of each simulation is indicated on the x-axis. The range of values corresponds to variations in inclination, polarization and reference phase. Top: mismatches restricted to $\ell = 2$ modes of both the NR waveforms and the waveform models. Bottom: all available modes in both NR and model.

configuration the range of mismatches against the models do not vary significantly with NR resolution.

We also show the optimized mismatch as a function of the binary orbital inclination angle θ_{LN}, for two redshifted total masses in the Appendix.

The mismatches suggest that none of the current models is able to accurately capture the full complexity of the GW signals in our NR simulations, even when considering the dominant quadrupolar modes. The discrepancies become more pronounced when higher-order modes are included, suggesting that the treatment of these modes in the waveform models remains inadequate, although it is by no means the only source of error. (Note that for aligned-spin systems, the worst mismatches against NR waveforms in Fig. 16 of Ref. [130], which includes configurations at $q = 18$, is ~0.01.) We also see that the NR-model mismatches are typically much larger than the mismatch uncertainties in the NR waveforms, indicating that the large mismatches cannot be ascribed to NR errors. We draw the clear conclusion that significant improvements are required in the modeling of high-mass-ratio precessing binaries.

V. PARAMETER ESTIMATION WITH CURRENT MODELS

Mismatches are a standard metric to assess the accuracy of waveform models, but they do not tell us which parameters (if any) will be biased when the model is used to measure the source properties following an observation. Therefore, to assess the performance of current models in practice, we have also carried out Bayesian PE on synthetic BBH signals [29] generated from our NR simulations, to assess how well the models recover the source properties. The details of the PE analysis and settings are described in Sec. II B.

We first consider the NR simulation `CF_81` listed in Table I, with $\theta_{LS_1} = 30°$, and generate two synthetic GW signals with a redshifted total mass of $150 M_\odot$, and an SNR of 50. For the first signal, we set the injected value $\theta_{JN} = 0°$, by adjusting θ_{LN} and the reference phase accordingly. For the second signal, we set $\theta_{JN} = 48°$ and choose the reference phase such that $\theta_{LN} = 60°$ at the reference frequency. Although both configurations have the same intrinsic amount of precession, we expect the former to exhibit a much weaker observable precession imprint on the signal because its line of sight is aligned with the total angular momentum. We perform PE on both signals using the `XPNR`, `TPHM`, and `v5PHM` waveform models. The posterior distributions of the component masses (m_1 and m_2) and the primary spin components, both aligned with the orbital angular momentum ($\chi_1^{\parallel} \equiv \chi_1 \cos\theta_{LS_1}$) and in the orbital plane ($\chi_1^{\perp} \equiv \chi_1 \sin\theta_{LS_1}$) at the reference frequency, are shown in Fig. 6. The left column displays the results for the $\theta_{JN} = 0°$ configuration, while the right column displays those for $\theta_{JN} = 48°$. The top panels show the posteriors on the masses, while the bottom panels show the distributions of the primary spin parameters. We find that the mass and spin parameters are tightly constrained in terms of their statistical uncertainties (posterior widths), owing to the presence of precession-induced modulations and the substantial higher-harmonic content,[7] both of which help break several parameter degeneracies.

From the results in Fig. 6 we see that no current model reliably infers the correct parameters for the $\theta_{LS_1} = 30°$ system. For the $\theta_{JN} = 0°$ orientation `XPNR` correctly recovers the masses, but the primary spin components are biassed. The true masses are outside the 90% credible

[7]For the second signal, the optimal injected SNRs in the $\ell = 2$, 3, and 4 multipoles are ~45, 19, and 8, respectively.

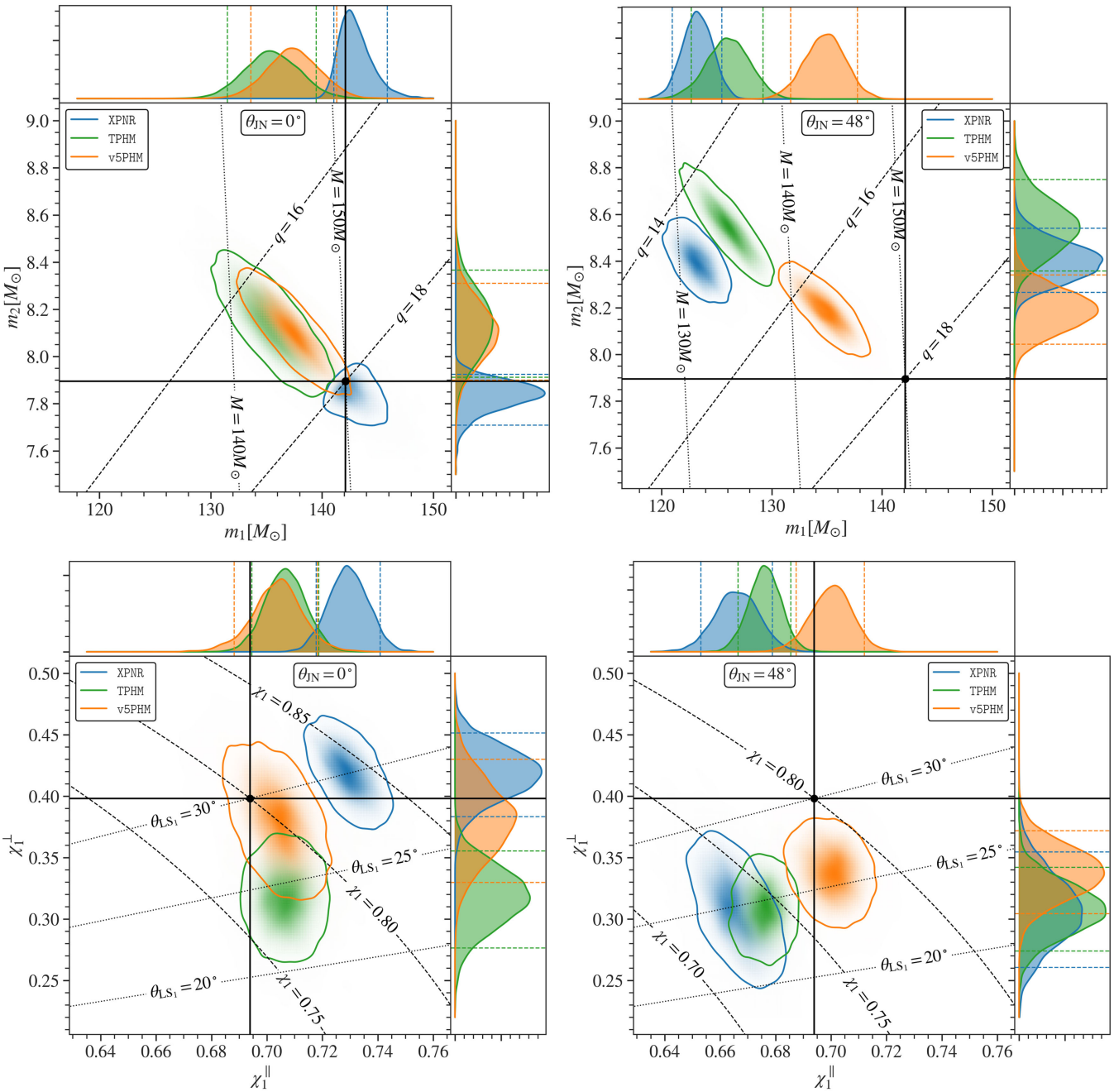

FIG. 6. Mass and primary spin measurements of two signals generated from the $\theta_{LS_1} = 30°$ simulation (CF_81 in Table I) with a redshifted total mass of $150M_\odot$ and SNR 50, using the XPNR, TPHM, and v5PHM waveform models. Left: $\theta_{JN} = 0°$. Right: $\theta_{JN} = 48°$. Top panels: component masses (m_1 and m_2). Bottom panels: primary spin components along the orbital angular momentum ($\chi_1^\parallel \equiv \chi_1 \cos\theta_{LS_1}$) and in the orbital plane ($\chi_1^\perp \equiv \chi_1 \sin\theta_{LS_1}$), at the reference frequency. The black dot and lines indicate the true (injected) parameters. The contours and dashed lines represent the 90% credible regions. The black dashed and dotted curves denote contours of constant mass ratio and redshifted total mass (top), and of primary spin magnitude and the angle between the primary spin vector and the Newtonian orbital angular momentum, θ_{LS_1} (bottom panel).

interval for both other models, although v5PHM correctly infers the spin components. The biases at orientation $\theta_{JN} = 48°$ are more severe due to the relatively stronger precessional modulations in the injected signal: no model correctly infers either the masses or the primary spin components. In this case the bias is not small: the errors in the individual-mass recovery are over 10%, and the recovered mass ratio can be as low as $q = 14$.

From the bottom panel of Fig. 5, we see that the minimum possible mismatch between the XPNR model and the NR waveform for the $\theta_{LS_1} = 30°$ case is 0.06, which corresponds to a 1D bias SNR of about 5. However,

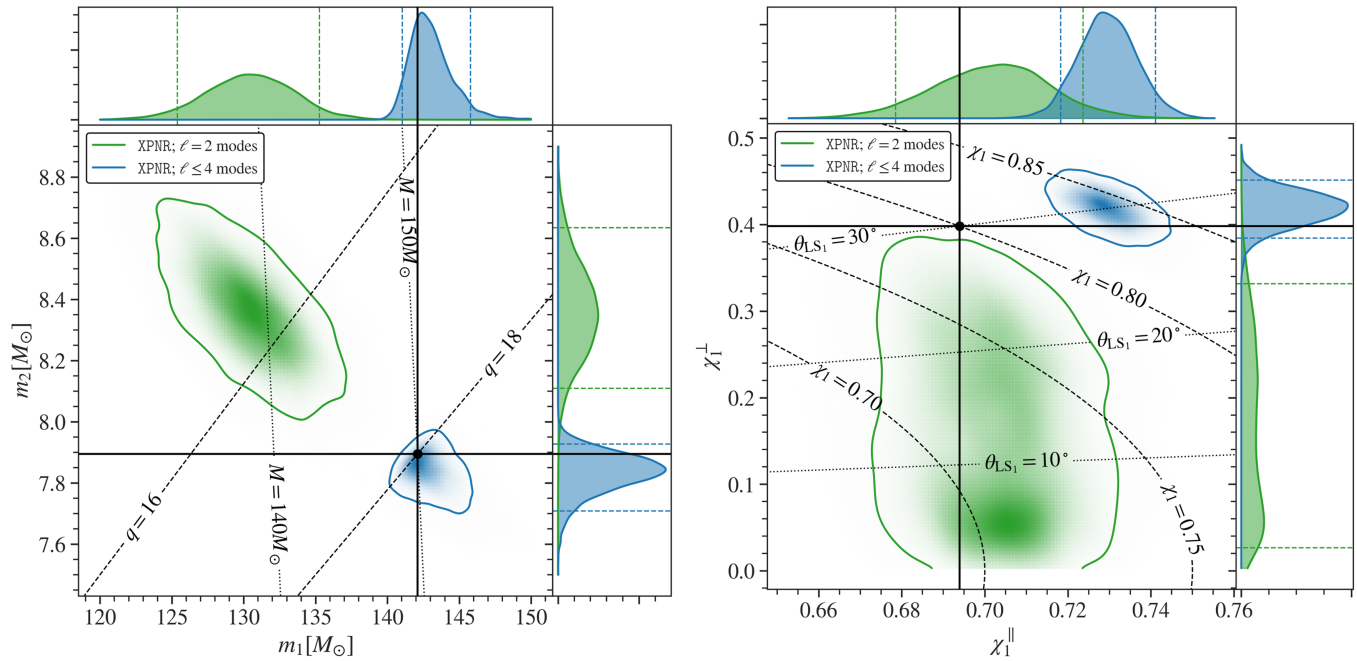

FIG. 7. Mass and primary spin measurements for the same $\theta_{LS_1} = 30°$, $\theta_{JN} = 0°$ signal as in Fig. 6. In the green results, only the quadrupolar modes (i.e., $\ell = 2$ modes) are used for both the NR injection and the parameter estimation with the XPNR model. In the blue results, the full NR signal is again recovered with the full XPNR model.

for the $\theta_{JN} = 0°$ case, we find that even at an SNR of 50, the marginalized 1D posteriors of m_1, m_2, and $\chi_1^\perp$ still enclose the true injected values. This was also noted in our earlier discussion: the indistinguishability-based SNR estimate is overly conservative, as suggested in Ref. [76]. We further find that the bias SNR varies across different parameters. Similar trends persist across the other PE results presented here as well as in the subsequent discussion.

In Fig. 7 we examine whether the biases can be ascribed to poor modelling of the higher modes. Here we restrict ourselves to the XPNR model (the only model calibrated to precessing NR simulations among the three models considered) and the $\theta_{JN} = 0°$ signal, and recover a signal that contains only the $\ell = 2$ modes with a corresponding $\ell = 2$ model, and compare the results with the recovery of the full signal with the full model. We find that restricting to $\ell = 2$ modes broadens the posteriors, as we might expect with a signal lacking rich higher-mode structure, but we also see that the bias is now *worse*. This illustrates that the models are of poor accuracy even at the level of only the $\ell = 2$ modes.

In Fig. 8 we consider the $\theta_{LS_1} = 150°$ system (NR simulation CF_85 of Table I), with orientations $\theta_{JN} = 0°$ and $\theta_{JN} = 80°$. The $\theta_{JN} = 0°$ case is representative of relatively weak precessional modulation in the signal seen by the detector, whereas the $\theta_{JN} = 80°$ case is representative of relatively strong precessional modulation. This configuration, CF_85, exhibits (compared to CF_81) far stronger precession, but, due to the large antialigned spin components, merges at a low frequency and includes only a small number of GW cycles in the frequency range of our analysis (see Figs. 1 and 4). Despite the small number of measurable GW cycles and the relatively good agreement by eye in Fig. 4, we still see large biases with all models for at least one orientation. Now the error in the recovery of the secondary mass can be as large as 100%, and the mass ratio can be recovered below $q = 10$. We note that, although in all cases the mismatches were not a useful indicator of the level of bias, the mismatches do allow us to correctly predict that the most accurate model for this configuration is TPHM.

We note that the posteriors for other parameters are also biased, although the magnitude of the bias varies across parameters. The mass and spin parameters shown here are neither the only nor necessarily the most or least affected parameters; they are presented simply as representative examples of the observed parameter biases.

Finally, we aim to understand how NR waveform errors map onto the waveform model manifold, i.e., how sensitive waveform models are to the numerical errors of NR simulations in the parameter estimation process. For this purpose, we generate three synthetic GW signals from the NR waveforms of configuration CF_81 in Table I, corresponding to varying numerical resolutions $N = 120$, 144, 172, at a redshifted total mass of $150M_\odot$ and $\theta_{JN} = 48°$, choosing the reference phase so that $\theta_{LN} = 60°$. We then perform parameter estimation on these signals using the XPNR waveform model. The posterior distributions of the component masses and the primary spin components $\vec{\chi}_1^\parallel$ and $\vec{\chi}_1^\perp$ at the reference frequency are shown in Fig. 9.

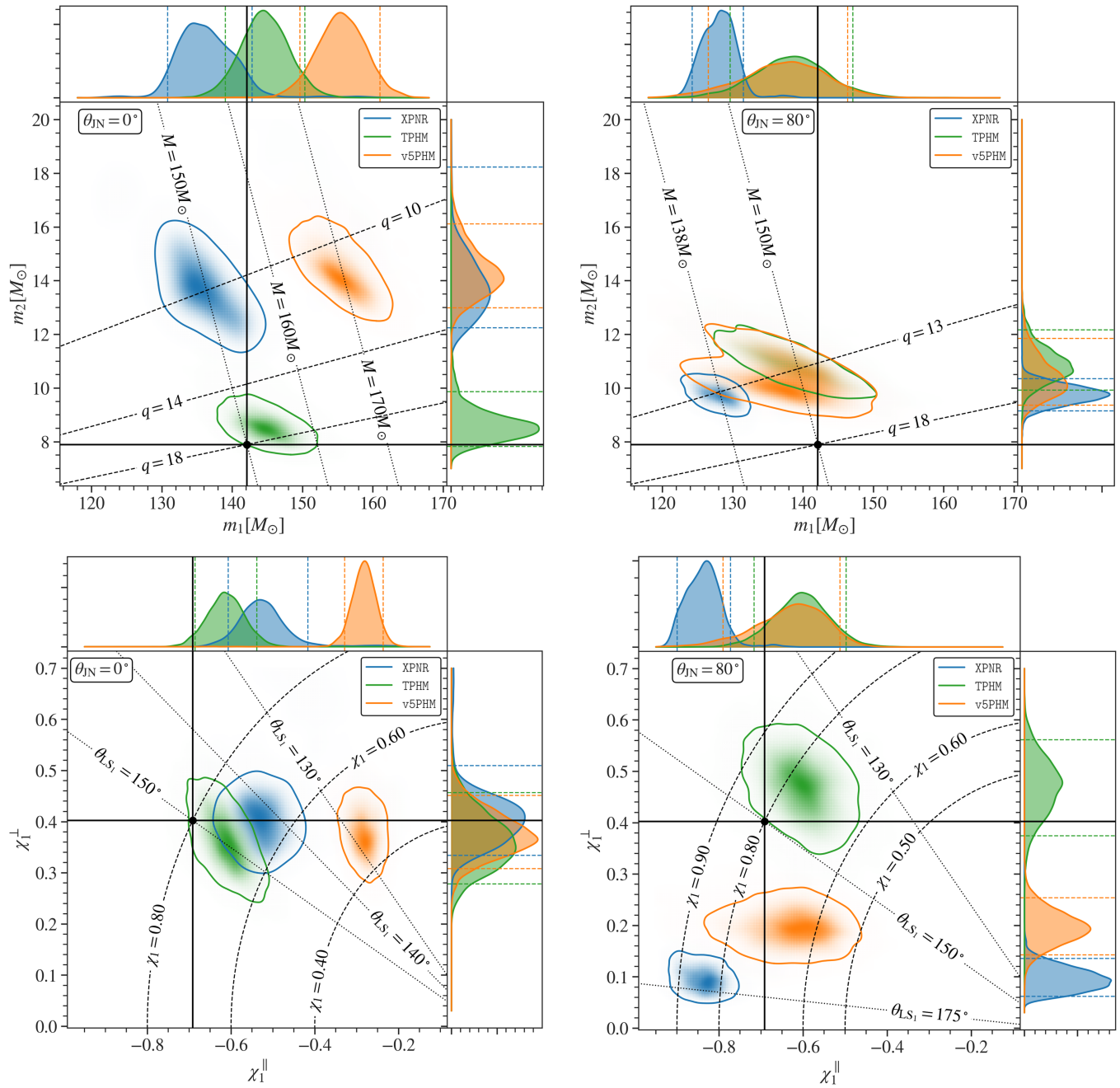

FIG. 8. Mass and primary spin measurements for the CF_85 ($\theta_{LS_1} = 150°$) configuration with redshifted total mass $150M_\odot$ and SNR 50, for orientations $\theta_{JN} = 0°$ (left) and $\theta_{JN} = 80°$ (right).

We see that the posteriors are only slightly shifted for the signals corresponding to the three NR resolutions, and the $N = 144$ and $N = 172$ results are almost identical. This is in stark contrast to what we might expect from the mismatches between the NR waveforms, which suggest that the $N = 120$ configuration would be distinguishable from the $N = 172$ configuration at an SNR of only ~9. These results suggest that the error in the NR waveforms are predominantly orthogonal to the physical signal manifold, and do not incur parameter biases. This may have implications for the NR accuracy necessary for model construction, which we will return to in the discussion.

We note that we did not consider all current models in this study, for example the TEOBResumS models [131–133]. However, although their construction differs from the XPNR, TPHM and v5PHM models, they do not include any additional information: they use the same inspiral ingredients from post-Newtonian theory or the effective-one-body framework, employ similar techniques to treat the merger and ringdown, and also lack calibration to precessing

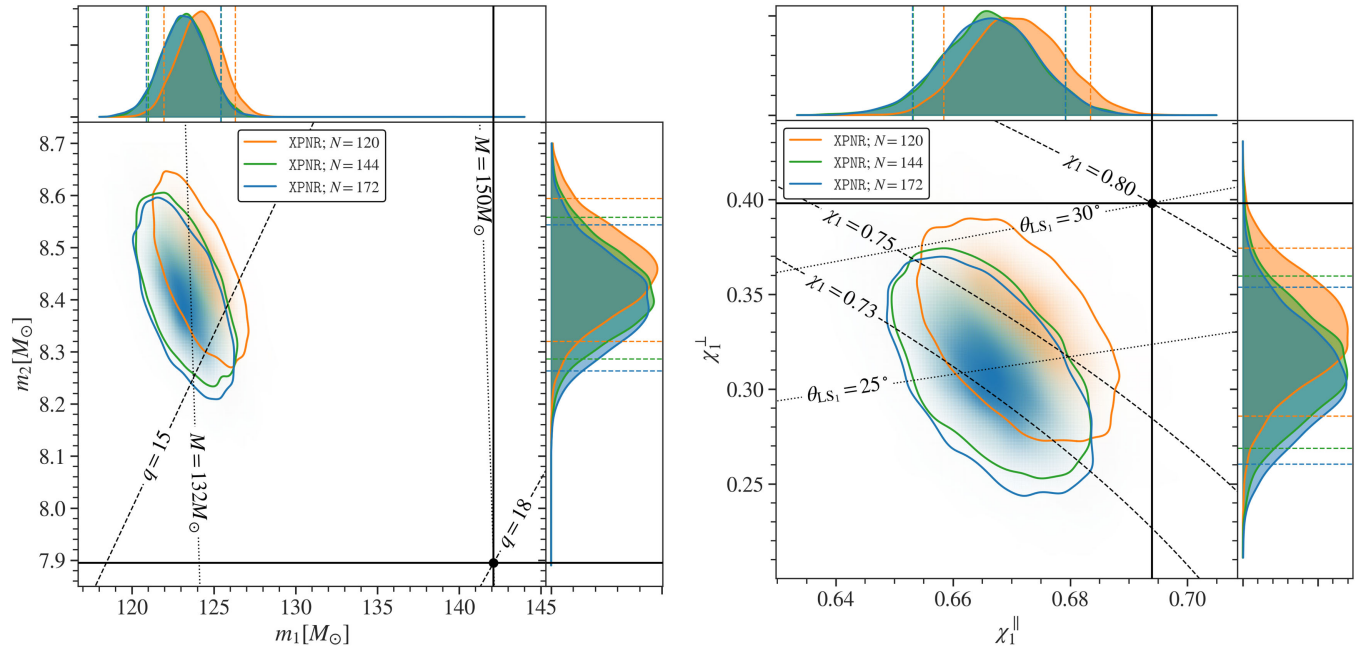

FIG. 9. Mass and primary spin measurements for the $\theta_{\mathrm{LS}_1} = 30°$ system with redshifted total mass $150M_\odot$ and orientation $\theta_{\mathrm{JN}} = 48°$, from the NR simulations at three resolutions, $N = 120$, $N = 144$, and $N = 172$, recovered with the XPNR model.

NR waveforms. We therefore expect that our conclusions apply to all current models: they are not suitable for GW astronomy involving high-mass-ratio, high-spin binaries with significant precession. This limitation may be less pressing for current detectors, where binaries with large mass ratios and high spins are expected to be rare, although waveform systematics can already be relevant for some current observations [7,134]. We expect better-tuned models in the near future, assisted by these NR waveforms. Nevertheless, these results starkly illustrate that models cannot be trusted far outside their region of calibration and validation.

VI. DISCUSSION

We have presented a series of five NR simulations of black-hole binaries with a high mass-ratio ($q = 18$) and significant precession. These extend the catalog presented in Ref. [49], and are a first step in filling a significant gap in simulation parameter-space coverage. In each simulation the larger black hole has a dimensionless spin of 0.8, and the simulations span five spin misalignment angles. The simulations cover $\sim 3000M$ before merger (between 30 and 49 GW cycles), which make them ideal for extending current precessing models into the high-mass-ratio regime, and for assessing the accuracy of current models.

Estimates of the accuracy of the NR waveforms, and comparisons (both through mismatches and parameter-estimation examples) allow us to draw a number of conclusions, and raise several important points.

We preface our conclusions with the caveat that our statements are restricted to high-mass binaries, where the late inspiral and merger-ringdown constitute significant power in the signal. The length of our NR simulations preclude statements about binaries with lower masses. However, we note that if we assume a minimum black-hole mass of $5M_\odot$, then the lowest total mass for a $q = 18$ binary will be $95M_\odot$, and masses much lower than we have considered are unlikely.

Our main conclusion is that current state-of-the-art waveform models for generic-spin binary cannot be trusted in measurements of high-mass-ratio high-precession binaries. Across a set of four signals that would be observed with SNR $\rho = 50$ in the Advanced LIGO detector at design sensitivity, we typically find that even the most basic property, the masses of the binaries, are recovered with errors up to $\sim 100\%$; a mass-ratio 18 system can be measured to have a mass ratio of less than 10. (See Fig. 8 for the most striking example.) No model is consistently more or less accurate than any other. The primary spin magnitude shows biases where the true value is up to 20% away from the 90% credible region of the measurement, and the lower bound on the measured spin magnitude is half of the true value. The spin orientation can be biased by up to 20°, although this is sufficient to determine if the spin is predominantly aligned or antialigned with the binary's orbital angular momentum.

Although these results are striking, they should perhaps not be surprising. Of the three models we considered, precession dynamics are tuned to NR simulations only in XPNR, and even in that case only up to $q = 8$, and only for the dominant multipole in the coprecessing frame. We have also seen in earlier work that current models show significant disagreement at high mass ratios, even in the nonprecessing sector [64]. As a precessing-binary example,

Ref. [135] presents detailed results for a highly precessing binary with mass-ratio $q = 10$, with mass-measurement errors of up to 10%. In that example the binary had a total mass of $\sim 25 M_\odot$, so the signal was dominated by the inspiral, where we might expect current models (all based on PN and EOB inspiral results) to be in better agreement. Based on those results, we should not be surprised that systems with higher mass (i.e., more signal power in the poorly modeled merger regime), and higher mass ratio would incur even larger biases. However, the study in Ref. [135] did not involve NR simulations, and could therefore only consider disagreements between models, and could not make absolute statements about the biases that would be observed from a true BBH signal. In this work we *have* used fully general-relativistic NR results, and have been able to conclusively show that current models are almost entirely inadequate for measurements of high-mass-ratio highly precessing binaries.

Of the ~200 binaries detected to date in LVK observations, the largest well-measured mass ratio is $q = 10$ [5], so we can expect larger mass ratios to be extremely rare with current detectors, and high-mass-ratio binaries with high spins are expected to be even rarer. However, intermediate-mass-ratio binaries are an important source for next-generation ground-based detectors and the space-based detector LISA, and for those observations we need far more accurate models. And to produce more accurate models, we need high-mass-ratio NR simulations.

This raises the question of NR accuracy requirements and computational cost. There is still no robust method to determine minimal NR or model accuracy requirements for GW astronomy. Our new NR waveforms provide a striking illustration of this issue. The mismatch errors for our base-resolution simulation are very high (~0.04), suggesting that the lowest and highest resolution waveforms would be distinguishable in an observation at SNRs above ~5. In PE examples, however, we find the three NR resolutions are *not* clearly distinguishable, even at an SNR of 50. (See Fig. 9.) This illustrates the degree to which NR errors can be orthogonal to waveform variations with respect to physical parameters.

Let us consider one way in which this could happen, to illustrate how NR errors may be orthogonal to measurement biases. The phase error in an NR simulation increases more rapidly as the binary approaches merger; as the black holes orbit more rapidly the numerical error per timestep increases. A slower or faster inspiral will be partially degenerate with a change in total mass (which is an overall scaling in the time), and with mass ratio and aligned spins, where the changes in inspiral rate are frequency dependent [83]. However, after merger the signal has a fixed ringdown frequency, and the even low-resolution simulations tend to capture this correctly, so that the phase error plateaus after merger. This change is *not* degenerate with a change in total mass or mass ratio, since these will change the final mass and spin, and therefore the ringdown frequency. In a signal with comparable power in both the inspiral and merger-ringdown, we may find that, although the NR waveform has the wrong phasing, the phase errors are not clearly degenerate with a biased mass and spins; in a search or parameter-estimation exercise such a signal may be recovered with a lower SNR, but not necessarily with a significant bias in the masses and spins. This illustrates the difficulty in specifying the NR accuracy requirements for future detectors. Our results suggests that, while ideally more accurate NR waveforms than those presented here will ultimately be necessary to calibrate waveform models, these first simulations may be sufficient to produce models that will be reliable for signals of moderate SNRs (<100). This requires more study (we may find, for example, that with accurate models these NR waveforms *are* easily distinguishable), and the full determination of NR and waveform accuracy requirements remains an open question.

We now turn to computational cost. The most accurate simulation we produced as part of our convergence series, the $N = 172$ simulation of the $\theta_{\rm LS} = 30°$ configuration (with a mismatch error of 0.007), is still less accurate than typical simulations in the earlier BAM catalog [49] (0.004), yet the computational cost is far higher. The $N = 172$ simulation alone cost 13M core hours, while the *entire catalog* of 80 simulations in Ref. [49] cost 25M core hours. This illustrates the well-known problem that, even though high-mass-ratio simulations do not present any known technical challenge (e.g., proof-of-principle moving-puncture simulations have been performed at mass ratios up to ~1000 [136]), binary simulations that include a sufficient number of orbits before merger are prohibitively expensive with current codes and methods. If the NR community is to produce accurate NR simulations in the intermediate-mass-ratio regime ($10 < q < 1000$) then improved codes and methods are crucial. In addition it is also crucial that we reach a better understanding of the relationship between NR accuracy and GW measurement accuracy.

ACKNOWLEDGMENTS

The authors thank Jannik Mielke for useful comments as the LIGO-Virgo-KAGRA internal reviewer, and Lorenzo Pompili and Carlos Lousto for their comments. P. M. and M. H. acknowledge the Science and Technology Facilities Council (STFC) for support through Grants No. ST/V005618/1 and No. UKRI2489. J. T. acknowledges support from the NASA LISA Preparatory Science Grant No. 20-LPS20-0005. This research used the supercomputing facilities at Cardiff University operated by Advanced Research Computing at Cardiff (ARCCA) on behalf of the Cardiff Supercomputing Facility and the HPC Wales and Supercomputing Wales (SCW) projects. We acknowledge the support of the latter, which is part funded by the European Regional Development Fund (ERDF) via the Welsh Government. The authors are also grateful for

computational resources provided by the Cardiff University and support by STFC Grants No. ST/I006285/1, No. ST/V005618/1 and No. UKRI2489. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants No. PHY-0757058 and No. PHY-0823459. This manuscript has the LIGO preprint No. P2600112. Various plots and analyses in this paper were made using Python software packages LALSuite [137], PyCBC [138], PESUMMARY [139], MATPLOTLIB [140], NumPy [141], and SciPy [90].

DATA AVAILABILITY

The data that support the findings of this article are not publicly available upon publication because it is not technically feasible and/or the cost of preparing, depositing, and hosting the data would be prohibitive within the terms of this research project. The data are available from the authors upon reasonable request.

APPENDIX

Sec. IV and Fig. 5 of the main text present only the minimum and maximum values, and the overall range, of optimized mismatches between the waveform models and the NR waveforms. This Appendix provides additional mismatch results that complement the main text. In particular, we present the optimized mismatch as a function of the binary orbital inclination angle $\theta_{\rm LN}$, for two redshifted total masses, $M = 150M_\odot$ and $250M_\odot$.

We first focus on the dominant quadrupolar $\ell = 2$ modes in both the waveform models and the NR simulations, and compute the corresponding optimized mismatches as described in the main text (see Secs. II A and IV for more details). Figure 10 presents the mismatches as a function of $\theta_{\rm LN}$, for the two redshifted total masses mentioned above. It employs the five precessing NR simulations introduced in the main text, each representing a different spin misalignment configuration. Solid markers indicate the SNR-weighted mismatch, while horizontal lines denote the range of mismatches across the sampled polarizations and phases.

We find that, for the first three considered values of $\theta_{\rm LS_1}$, the SNR-weighted mismatches are typically greater than 10^{-2}, corresponding to a bias SNR of ≥ 12 for single-parameter measurements, according to the conservative mismatch criterion ($\mathfrak{M} \leq 1.35/\rho^2$) mentioned above. The lowest mismatch is found to be 3×10^{-3}, corresponding to a bias SNR of 21. For the latter two values of $\theta_{\rm LS_1}$, the SNR-weighted mismatches are generally above 10^{-3}, corresponding to a bias SNR of ≥ 37. The mismatches are typically smaller for $\theta_{\rm LS_1} = 150°$ compared to $\theta_{\rm LS_1} = 30°$, and for $\theta_{\rm LS_1} = 120°$ compared to $\theta_{\rm LS_1} = 60°$. This occurs because the number of cycles within a given frequency band decreases significantly as the spin-orbit misalignment angle increases due to the hang-up effect. We also find that $M = 150M_\odot$ systems generally yield worse mismatch compared to $250M_\odot$ cases, because lower-mass

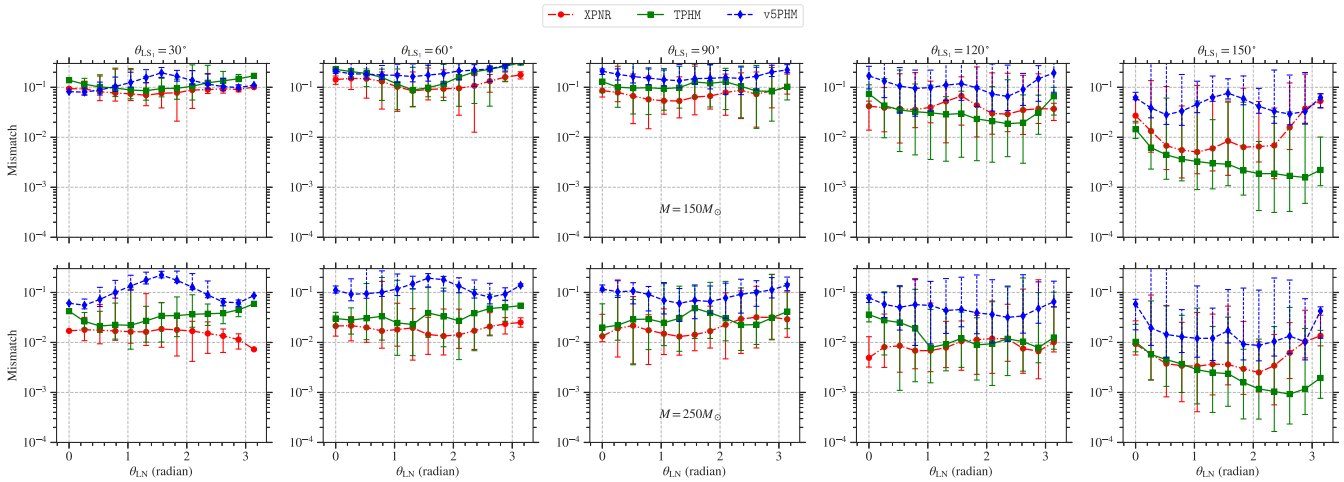

FIG. 10. Mismatch between the dominant quadrupolar $(\ell, |m|) = (2, 2)$ modes of five reported NR simulations and the corresponding modes of three state-of-the-art waveform models (shown in the legend) as a function of the binary inclination angle—quantified by the angle between the line of sight and the Newtonian orbital angular momentum—at redshifted total masses of $150M_\odot$ (top row) and $250M_\odot$ (bottom row). The mismatch shown here is optimized over luminosity distance, coalescence time and phase, sky location, polarization, and precession angle. The reported NR simulations represent high mass-ratio, precessing BBH systems with a mass ratio of 18 and a dimensionless spin magnitude of 0.8 on the larger BH (the smaller BH being nonspinning), and the five simulations correspond to spin misalignment angles of 30°, 60°, 90°, 120°, and 150°, with each column in the figure representing one of these angles. Each solid marker shows the SNR-weighted mismatch, and the horizontal lines indicate the range from minimum to maximum mismatch across the considered signal polarizations and phases.

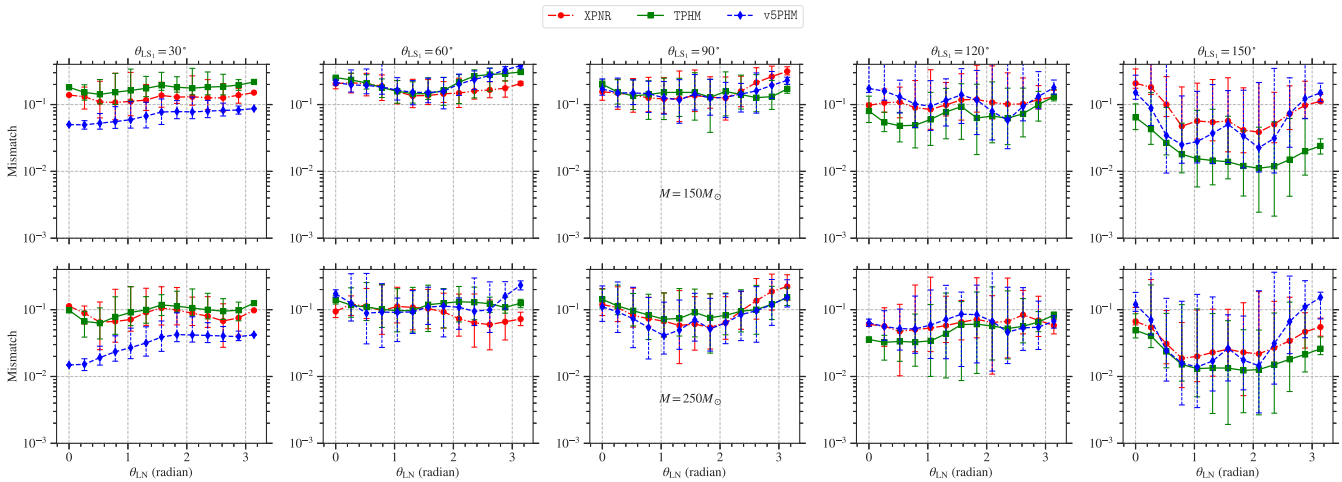

FIG. 11. Mismatch between NR simulations and waveform models including all available modes, analogous to Fig. 10 but extending beyond the dominant quadrupolar modes.

systems will have more GW cycles in the LIGO frequency band.

Next, we focus on all the available modes in the waveform models and on modes up to $\ell = 4$ in the NR simulations, and compute the corresponding mismatches. We have not considered the memory modes (modes with $m = 0$) from the NR simulations (the available modes in the three waveform models are listed in Sec. I of the main text). We present the results in Fig. 11. By comparing the results shown in Fig. 10 and Fig. 11, we find that the mismatches between the waveform models and the NR simulations become larger when higher-order modes are included in addition to the dominant quadrupolar modes. Here, we find that SNR–weighted mismatches are always greater than 10^{-2}, corresponding to a bias SNR of ≥ 12, while the lowest mismatch is 2×10^{-3}, corresponding to a bias SNR of 26 for single-parameter measurements, according to the conservative mismatch criterion. All the adopted waveform models perform roughly equally poorly in terms of mismatch, and none can be considered superior to the others. All other trends remain largely similar to those shown in Fig. 10. Current models require significant improvement in the region of parameter space covered by the reported NR simulations.

In addition to the mismatch studies presented above, we also assess the impact of the recent update to the `v5PHM` model that incorporates multipole-mode asymmetries (`SEOBNRv5PHM_w/asym`) [68] in this Appendix. The results presented in the main text were obtained using the earlier version of the model, `v5PHM`, which does not include these effects. Here, we investigate whether the inclusion of multipole-mode asymmetries affects the parameter estimation results or the conclusions of this work. To this end, we perform additional parameter estimation analyses using the updated `SEOBNRv5PHM_w/asym` waveform (`v5PHM_asym`) and compare the results with those obtained using the `v5PHM` model. For this comparison, we focus on a subset of the configurations shown in Fig. 8. The new results obtained with `v5PHM_asym` are presented in Fig. 12, where we explicitly show both waveform versions, labeled as `v5PHM` and `v5PHM_asym`. We find that the inclusion of mode asymmetries leads to only minimal changes in the recovered source parameters, with no significant impact on the overall results. Consequently, our main conclusion—that current models are not suitable for GW astronomy involving high-mass-ratio, high-spin binaries with significant precession—remains unchanged.

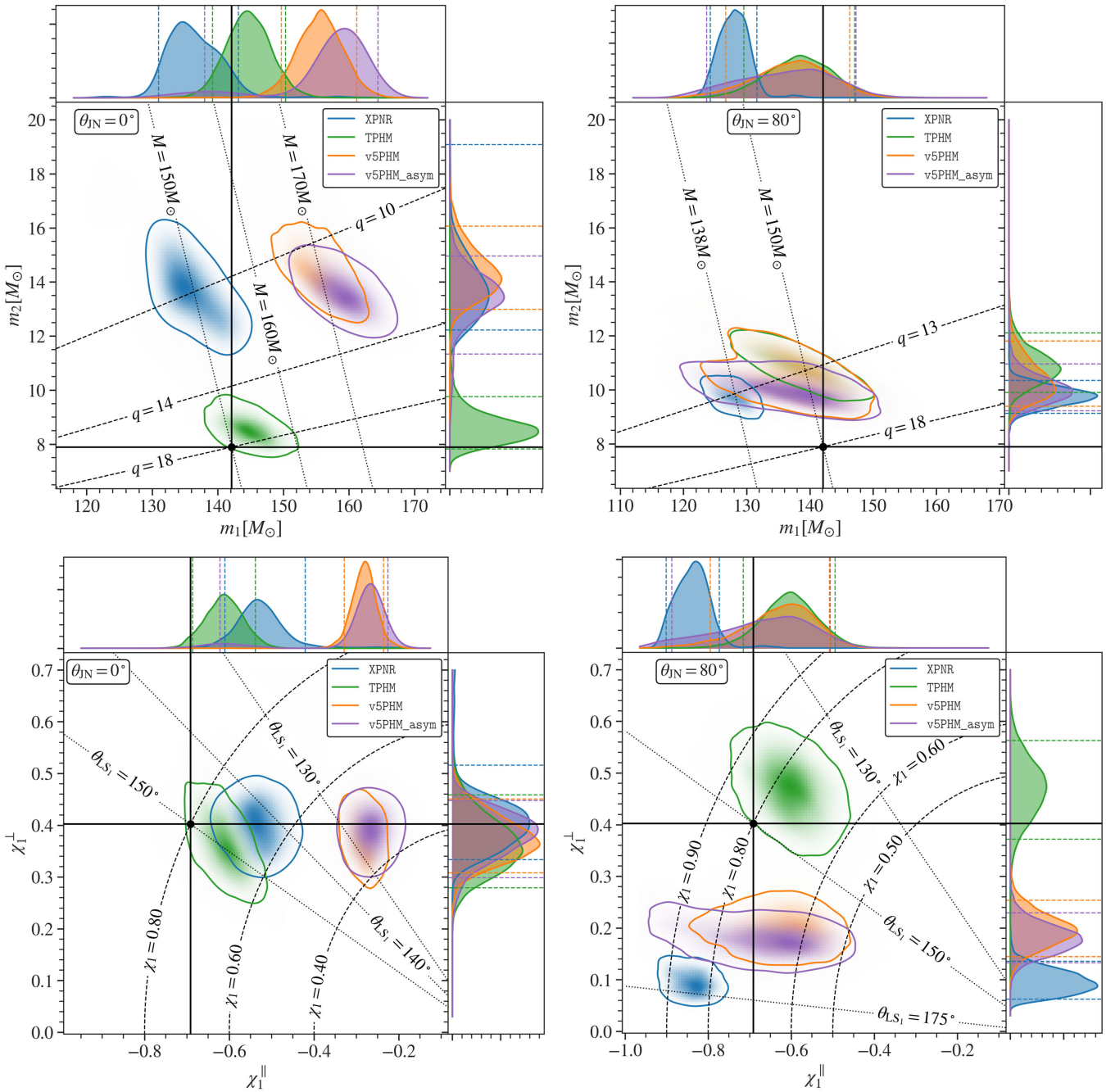

FIG. 12. Mass and primary spin measurements for the `CF_85` ($\theta_{LS_1} = 150°$) configuration with redshifted total mass $150M_\odot$ and SNR 50, for orientations $\theta_{JN} = 0°$ (left) and $\theta_{JN} = 80°$ (right). Similar to Fig. 8 of the main text, results are additionally shown for the `v5PHM_asym` model that includes multipole-mode asymmetries. We find only minimal differences between `v5PHM` and `v5PHM_asym`, with no significant impact on the recovered parameters and the conclusions of this work.

[1] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. Lett. **116**, 061102 (2016).

[2] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. Lett. **119**, 161101 (2017).

[3] B. P. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration, Fermi-GBM Collaboration, and INTEGRAL Collaboration), Astrophys. J. Lett. **848**, L13 (2017).

[4] R. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. Lett. **125**, 101102 (2020).
[5] R. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Astrophys. J. Lett. **896**, L44 (2020).
[6] A. G. Abac *et al.* (LIGO Scientific Collaboration, Virgo Collaboration, KAGRA Collaboration, and VIRGO Collaboration), Astrophys. J. Lett. **970**, L34 (2024).
[7] A. G. Abac *et al.* (LIGO Scientific Collaboration, VIRGO Collaboration, and KAGRA Collaboration), Astrophys. J. Lett. **993**, L25 (2025).
[8] A. G. Abac *et al.* (LIGO Scientific Collaboration, VIRGO Collaboration, and KAGRA Collaboration), Astrophys. J. Lett. **1004**, L22 (2026).
[9] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. X **9**, 031040 (2019).
[10] R. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. X **11**, 021053 (2021).
[11] R. Abbott *et al.* (KAGRA Collaboration, VIRGO Collaboration, and LIGO Scientific Collaboration), Phys. Rev. X **13**, 041039 (2023).
[12] B. P. Abbott *et al.* (LIGO Scientific Collaboration and VIRGO Collaboration), Phys. Rev. Lett. **116**, 221101 (2016); **121**, 129902(E) (2018).
[13] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. Lett. **123**, 011102 (2019).
[14] R. Abbott *et al.* (LIGO Scientific Collaboration, VIRGO Collaboration, and KAGRA Collaboration), Phys. Rev. D **112**, 084080 (2025).
[15] B. P. Abbott *et al.* (MASTER Collaborations), Nature (London) **551**, 85 (2017).
[16] A. G. Abac *et al.* (LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration), Astrophys. J. Lett. **1007**, L17 (2026).
[17] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. Lett. **121**, 161101 (2018).
[18] A. G. Abac *et al.* (LIGO Scientific Collaboration, VIRGO Collaboration, and KAGRA Collaboration), Astrophys. J. Lett. **1005**, L51 (2026).
[19] B. Sathyaprakash, M. Abernathy, A. Acernese, F., P.-S. N., K. Arun *et al.*, Classical Quantum Gravity **29**, 124013 (2012).
[20] M. Punturo *et al.*, Classical Quantum Gravity **27**, 194002 (2010).
[21] D. Reitze *et al.*, Bull. Am. Astron. Soc. **51**, 035 (2019).
[22] I. Gupta *et al.*, Classical Quantum Gravity **41**, 245001 (2024).
[23] P. Amaro-Seoane *et al.*, arXiv:1702.00786.
[24] M. Colpi *et al.* (LISA Collaboration), arXiv:2402.07571.
[25] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Classical Quantum Gravity **34**, 104002 (2017).
[26] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. Lett. **116**, 241102 (2016).
[27] T. D. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. X **6**, 041014 (2016).
[28] T. Bayes, Phil. Trans. **53**, 370 (1763).
[29] J. Veitch *et al.*, Phys. Rev. D **91**, 042003 (2015).
[30] C. Rover, R. Meyer, and N. Christensen, Classical Quantum Gravity **23**, 4895 (2006).
[31] M. van der Sluys, V. Raymond, I. Mandel, C. Rover, N. Christensen, V. Kalogera, R. Meyer, and A. Vecchio, Classical Quantum Gravity **25**, 184011 (2008).
[32] L. Blanchet, Living Rev. Relativity **17**, 2 (2014).
[33] T. Damour and A. Nagar, Phys. Rev. D **79**, 081503 (2009).
[34] M. Sasaki and H. Tagoshi, Living Rev. Relativity **6**, 6 (2003).
[35] F. Pretorius, Phys. Rev. Lett. **95**, 121101 (2005).
[36] M. Campanelli, C. O. Lousto, P. Marronetti, and Y. Zlochower, Phys. Rev. Lett. **96**, 111101 (2006).
[37] J. G. Baker, J. Centrella, D.-I. Choi, M. Koppitz, and J. van Meter, Phys. Rev. Lett. **96**, 111102 (2006).
[38] B. Bruegmann, J. A. Gonzalez, M. Hannam, S. Husa, U. Sperhake, and W. Tichy, Phys. Rev. D **77**, 024027 (2008).
[39] S. Husa, J. A. Gonzalez, M. Hannam, B. Bruegmann, and U. Sperhake, Classical Quantum Gravity **25**, 105006 (2008).
[40] M. A. Scheel, H. P. Pfeiffer, L. Lindblom, L. E. Kidder, O. Rinne, and S. A. Teukolsky, Phys. Rev. D **74**, 104006 (2006).
[41] D. A. Hemberger, M. A. Scheel, L. E. Kidder, B. Szilágyi, G. Lovelace, N. W. Taylor, and S. A. Teukolsky, Classical Quantum Gravity **30**, 115001 (2013).
[42] F. Herrmann, I. Hinder, D. Shoemaker, and P. Laguna, in New Frontiers in Numerical Relativity (NFNR 2006) (2006).
[43] Y. Zlochower, J. G. Baker, M. Campanelli, and C. O. Lousto, Phys. Rev. D **72**, 024021 (2005).
[44] U. Sperhake, Phys. Rev. D **76**, 104015 (2007).
[45] F. Loffler *et al.*, Classical Quantum Gravity **29**, 115001 (2012).
[46] C. O. Lousto and Y. Zlochower, Phys. Rev. Lett. **106**, 041101 (2011).
[47] U. Sperhake, V. Cardoso, C. D. Ott, E. Schnetter, and H. Witek, Phys. Rev. D **84**, 084038 (2011).
[48] M. A. Scheel, M. Giesler, D. A. Hemberger, G. Lovelace, K. Kuper, M. Boyle, B. Szilágyi, and L. E. Kidder, Classical Quantum Gravity **32**, 105009 (2015).
[49] E. Hamilton *et al.*, Phys. Rev. D **109**, 044032 (2024).
[50] D. Ferguson *et al.*, Phys. Rev. D **112**, 044043 (2025).
[51] M. A. Scheel *et al.*, Classical Quantum Gravity **42**, 195017 (2025).
[52] P. Ajith *et al.*, Phys. Rev. D **77**, 104017 (2008); **79**, 129901 (E) (2009).
[53] P. Ajith *et al.*, Phys. Rev. Lett. **106**, 241101 (2011).
[54] A. Buonanno, Y. Pan, J. G. Baker, J. Centrella, B. J. Kelly, S. T. McWilliams, and J. R. van Meter, Phys. Rev. D **76**, 104049 (2007).
[55] S. E. Field, C. R. Galley, J. S. Hesthaven, J. Kaye, and M. Tiglio, Phys. Rev. X **4**, 031006 (2014).
[56] J. Blackman, S. E. Field, C. R. Galley, B. Szilágyi, M. A. Scheel, M. Tiglio, and D. A. Hemberger, Phys. Rev. Lett. **115**, 121102 (2015).
[57] E. Hamilton *et al.*, Phys. Rev. D **113**, 084055 (2026).
[58] H. Estellés, M. Colleoni, C. García-Quirós, S. Husa, D. Keitel, M. Mateu-Lucena, M. d. L. Planas, and A. Ramos-Buades, Phys. Rev. D **105**, 084040 (2022).
[59] A. Ramos-Buades, A. Buonanno, H. Estellés, M. Khalil, D. P. Mihaylov, S. Ossokine, L. Pompili, and M. Shiferaw, Phys. Rev. D **108**, 124037 (2023).

[60] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, D. Gerosa, L. C. Stein, L. E. Kidder, and H. P. Pfeiffer, Phys. Rev. Res. **1**, 033015 (2019).
[61] M. Colleoni, F. A. R. Vidal, C. García-Quirós, S. Akçay, and S. Bera, Phys. Rev. D **111**, 104019 (2025).
[62] E. Hamilton, L. London, J. E. Thompson, E. Fauchon-Jones, M. Hannam, C. Kalaghatgi, S. Khan, F. Pannarale, and A. Vano-Vinuales, Phys. Rev. D **104**, 124027 (2021).
[63] S. Ghosh, P. Kolitsidou, and M. Hannam, Phys. Rev. D **109**, 024061 (2024).
[64] L. Pompili *et al.*, Phys. Rev. D **108**, 124035 (2023).
[65] H. Estellés, S. Husa, M. Colleoni, D. Keitel, M. Mateu-Lucena, C. García-Quirós, A. Ramos-Buades, and A. Borchers, Phys. Rev. D **105**, 084039 (2022).
[66] M. Khalil, A. Buonanno, H. Estelles, D. P. Mihaylov, S. Ossokine, L. Pompili, and A. Ramos-Buades, Phys. Rev. D **108**, 124036 (2023).
[67] H. Estellés, A. Ramos-Buades, S. Husa, C. García-Quirós, M. Colleoni, L. Haegel, and R. Jaume, Phys. Rev. D **103**, 124060 (2021).
[68] H. Estellés, A. Buonanno, R. Enficiaud, C. Foo, and L. Pompili, Phys. Rev. D **113**, 044049 (2026).
[69] C. García-Quirós, S. Tiwari, and S. Babak, Phys. Rev. D **112**, 064017 (2025).
[70] A. H. Mroue *et al.*, Phys. Rev. Lett. **111**, 241104 (2013).
[71] M. Boyle *et al.*, Classical Quantum Gravity **36**, 195006 (2019).
[72] J. Healy, C. O. Lousto, Y. Zlochower, and M. Campanelli, Classical Quantum Gravity **34**, 224001 (2017).
[73] J. Healy, C. O. Lousto, J. Lange, R. O'Shaughnessy, Y. Zlochower, and M. Campanelli, Phys. Rev. D **100**, 024021 (2019).
[74] J. Healy and C. O. Lousto, Phys. Rev. D **102**, 104018 (2020).
[75] J. Healy and C. O. Lousto, Phys. Rev. D **105**, 124010 (2022).
[76] J. E. Thompson, C. Hoy, E. Fauchon-Jones, and M. Hannam, Phys. Rev. D **112**, 064011 (2025).
[77] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. D **100**, 064064 (2019).
[78] B. J. Owen, Phys. Rev. D **53**, 6749 (1996).
[79] L. S. Finn, Phys. Rev. D **46**, 5236 (1992).
[80] C. Cutler and E. E. Flanagan, Phys. Rev. D **49**, 2658 (1994).
[81] L. Lindblom, B. J. Owen, and D. A. Brown, Phys. Rev. D **78**, 124020 (2008).
[82] S. T. McWilliams, B. J. Kelly, and J. G. Baker, Phys. Rev. D **82**, 024014 (2010).
[83] E. Baird, S. Fairhurst, M. Hannam, and P. Murphy, Phys. Rev. D **87**, 024035 (2013).
[84] P. Schmidt, F. Ohme, and M. Hannam, Phys. Rev. D **91**, 024043 (2015).
[85] I. Harry, S. Privitera, A. Bohé, and A. Buonanno, Phys. Rev. D **94**, 024012 (2016).
[86] R. Balasubramanian, B. S. Sathyaprakash, and S. V. Dhurandhar, Phys. Rev. D **53**, 3033 (1996); **54**, 1860(E) (1996).
[87] F. Ohme, Bridging the Gap between Post-Newtonian Theory and Numerical Relativity in Gravitational-Wave Data Analysis, Ph.D. thesis, Potsdam University, 2012.
[88] P. Ajith *et al.*, Classical Quantum Gravity **29**, 124001 (2012); **30**, 199401(E) (2013).
[89] G. Pratten *et al.*, Phys. Rev. D **103**, 104056 (2021).
[90] P. Virtanen *et al.*, Nat. Methods **17**, 261 (2020).
[91] LIGO-Virgo-KAGRA Collaboration, https://dcc.ligo.org/LIGO-T2000012/public (2022).
[92] K. Mitman, L. C. Stein, M. Boyle, N. Deppe, L. E. Kidder, H. P. Pfeiffer, and M. A. Scheel, Classical Quantum Gravity **42**, 117001 (2025).
[93] S. Khan, K. Chatziioannou, M. Hannam, and F. Ohme, Phys. Rev. D **100**, 024059 (2019).
[94] J. Veitch and W. D. Pozzo, Analytic Marginalisation of Phase Parameter, Technical Report Tech. Rep. T1300326 (2013), https://dcc.ligo.org/LIGO-T1300326.
[95] W. M. Farr, Marginalisation of the time and phase parameters in cbc parameter estimation, 2014.
[96] L. P. Singer and L. R. Price, Phys. Rev. D **93**, 024013 (2016).
[97] E. Thrane and C. Talbot, Pub. Astron. Soc. Aust. **36**, e010 (2019).
[98] L. P. Singer *et al.*, Astrophys. J. Lett. **829**, L15 (2016).
[99] G. Ashton *et al.*, Astrophys. J. Suppl. Ser. **241**, 27 (2019).
[100] I. M. Romero-Shaw *et al.*, Mon. Not. R. Astron. Soc. **499**, 3295 (2020).
[101] J. S. Speagle, Mon. Not. R. Astron. Soc. **493**, 3132 (2020).
[102] A. G. Abac *et al.* (LIGO Scientific Collaboration, VIRGO Collaboration, and KAGRA Collaboration), Astrophys. J. Lett. **1004**, L21 (2026).
[103] P. Schmidt, I. W. Harry, and H. P. Pfeiffer, arXiv:1703.01076.
[104] F. Acernese, M. Agathos *et al.*, Classical Quantum Gravity **32**, 024001 (2014).
[105] J. Aasi *et al.*, Classical Quantum Gravity **32**, 074001 (2015).
[106] B. P. Abbott *et al.* (KAGRA Collaboration, LIGO Scientific Collaboration, and Virgo Collaboration), Living Rev. Relativity **19**, 1 (2016).
[107] LIGO-Virgo-KAGRA Collaboration, https://dcc.ligo.org/LIGO-P1200087-v42/public (2022).
[108] R. Ursell, C. Hoy, I. Harry, and L. K. Nuttall, Mon. Not. R. Astron. Soc. **547**, stag398 (2026).
[109] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez Forteza, and A. Bohé, Phys. Rev. D **93**, 044007 (2016).
[110] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. Jiménez Forteza, and A. Bohé, Phys. Rev. D **93**, 044006 (2016).
[111] L. London, S. Khan, E. Fauchon-Jones, C. García, M. Hannam, S. Husa, X. Jiménez-Forteza, C. Kalaghatgi, F. Ohme, and F. Pannarale, Phys. Rev. Lett. **120**, 161102 (2018).
[112] C. O. Lousto and J. Healy, Phys. Rev. Lett. **125**, 191102 (2020).
[113] N. Afshordi *et al.* (LISA Consortium Waveform Working Group), Living Rev. Relativity **28**, 9 (2025).
[114] L. Blanchet, Living Rev. Relativity **27**, 4 (2024).
[115] J. M. Bowen and J. W. York, Jr., Phys. Rev. D **21**, 2047 (1980).
[116] S. Brandt and B. Bruegmann, Phys. Rev. Lett. **78**, 3606 (1997).

[117] M. Ansorg, B. Bruegmann, and W. Tichy, Phys. Rev. D **70**, 064011 (2004).
[118] M. Shibata and T. Nakamura, Phys. Rev. D **52**, 5428 (1995).
[119] T. W. Baumgarte and S. L. Shapiro, Phys. Rev. D **59**, 024007 (1998).
[120] J. W. York, Jr., in *Workshop on Sources of Gravitational Radiation* (1978), pp. 83–126.
[121] P. Schmidt, F. Ohme, and M. Hannam, Phys. Rev. D **91**, 024043 (2015).
[122] S. Husa, M. Hannam, J. A. Gonzalez, U. Sperhake, and B. Bruegmann, Phys. Rev. D **77**, 044037 (2008).
[123] E. Newman and R. Penrose, J. Math. Phys. (N.Y.) **3**, 566 (1962).
[124] J. G. Baker, M. Campanelli, and C. O. Lousto, Phys. Rev. D **65**, 044001 (2002).
[125] C. Reisswig and D. Pollney, Classical Quantum Gravity **28**, 195015 (2011).
[126] K. G. Arun, B. R. Iyer, B. S. Sathyaprakash, and P. A. Sundararajan, Phys. Rev. D **71**, 084008 (2005); **72**, 069903(E) (2005).
[127] L. E. Kidder, Phys. Rev. D **52**, 821 (1995).
[128] L. E. Kidder, C. M. Will, and A. G. Wiseman, Phys. Rev. D **47**, R4183 (1993).
[129] D. Ferguson, K. Jani, P. Laguna, and D. Shoemaker, Phys. Rev. D **104**, 044037 (2021).
[130] C. García-Quirós, M. Colleoni, S. Husa, H. Estellés, G. Pratten, A. Ramos-Buades, M. Mateu-Lucena, and R. Jaume, Phys. Rev. D **102**, 064002 (2020).
[131] A. Nagar *et al.*, Phys. Rev. D **98**, 104052 (2018).
[132] S. Akcay, R. Gamba, and S. Bernuzzi, Phys. Rev. D **103**, 024014 (2021).
[133] R. Gamba, S. Akçay, S. Bernuzzi, and J. Williams, Phys. Rev. D **106**, 024020 (2022).
[134] A. G. Abac *et al.* (LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration), Astrophys. J. Lett. **993**, L21 (2025).
[135] A. Dhani, S. H. Völkel, A. Buonanno, H. Estelles, J. Gair, H. P. Pfeiffer, L. Pompili, and A. Toubiana, Phys. Rev. X **15**, 031036 (2025).
[136] C. O. Lousto and J. Healy, Classical Quantum Gravity **40**, 09LT01 (2023).
[137] LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration, LVK Algorithm Library—LALSuite, Free software (GPL), 2018.
[138] A. Nitz *et al.*, gwastro/pycbc: v2.3.3 release of pycbc, 2024.
[139] C. Hoy and V. Raymond, SoftwareX **15**, 100765 (2021).
[140] J. D. Hunter, Comput. Sci. Eng. **9**, 90 (2007).
[141] C. R. Harris *et al.*, Nature (London) **585**, 357 (2020).