Assessing the model waveform accuracy of gravitational waves

Qian Hu[1,*] and John Veitch[1,†]

[1] *Institute for Gravitational Research, School of Physics and Astronomy, University of Glasgow, Glasgow, G12 8QQ, United Kingdom*

(Dated: August 17, 2022)

With the improvement in sensitivity of gravitational wave (GW) detectors and the increasing diversity of GW sources, there is a strong need for accurate GW waveform models for data analysis. While the current model accuracy assessments require waveforms generated by numerical relativity (NR) simulations as the "true waveforms", in this paper we propose an assessment approach that does not require NR simulations, which enables us to assess model accuracy everywhere in the parameter space. By measuring the difference between two waveform models, we derive a necessary condition for a pair of waveform models to both be accurate, for a particular set of parameters. We then apply this method to the parameter estimation samples of the Gravitational-Wave Transient Catalogs GWTC-3 and GWTC-2.1, and find that the waveform accuracy for high signal-to-noise ratio events in some cases fails our assessment criterion. Based on analysis of real events' posterior samples, we discuss the correlation between our quantified accuracy assessments and systematic errors in parameter estimation. We find waveform models that perform worse in our assessment are more likely to give inconsistent estimations. We also investigate waveform accuracy in different parameter regions, and find the accuracy degrades as the spin effects go up, the mass ratio deviates from one, or the orbital plane is near-aligned to the line of sight. Furthermore, we make predictions of waveform accuracy requirements for future detectors and find the accuracy of current waveform models should be improved by at least 3 orders of magnitude, which is consistent with previous works.

I. INTRODUCTION

Over 90 gravitational wave (GW) events have been detected since 2015 [1–12] by Advanced LIGO [13] and Advanced Virgo [14], all of them are from compact binary coalescences (CBCs), where GW waveforms can be modeled by various methods. The data analysis for CBCs such as signal searching [15] and parameter estimation [16–19], are based on the these waveform models of CBCs. Therefore, inaccurate waveforms may cause systematic errors in the scientific interpretation of GW data [20].

For binary black holes, the evolution can be divided into 3 stages: inspiral, merger, and ringdown, while binary neutron stars and neutron star black hole binaries may exhibit tidal disruption prior to the formation of a final black hole or hypermassive neutron star. The post-Newtonian (PN) expansion gives a good approximation of the inspiral stage [21], and black hole ringdown can be described by quasi-normal modes [22]. Other than theoretical approximations that only give the waveform of one of the stages, the most accurate GW waveforms of the whole process of CBCs are generated by numerical relativity (NR) [23–25], where the Einstein Field Equations are solved numerically. However, NR waveforms are so expensive to compute that the latest SXS NR waveform catalog contains less than 2000 waveforms [25]. Besides, NR waveforms are generally short: the inspiral stage is usually calculated for only the last few cycles of the binary (there are exceptions, e.g., Ref. [26]), plus their sparsity in the parameter space, it is impractical to use them directly in data analysis. Coverage of the parameter space is also uneven, as NR waveforms with unequal masses and high spins are more difficult to compute.

These NR waveforms are therefore used to tune more tractable approaches to waveform modeling, with the aim of minimising the difference between the full NR model and the approximate model. Several methods exist to compute GW waveforms rapidly, for example, the `IMRPhenom` [27–31] family, the `SEOBNR` [32–37] family, `TEOBResumS` family [38–40], and surrogate models [41–45] like `NRSur` family. These waveform approximants originate from different ideas of approximation or interpolation, and are calibrated with NR waveforms or hybridized waveforms of NR simulation and PN approximation. They are widely used in GW data analysis. The state-of-art waveform models from the two families mentioned above, `IMRPhenomXPHM` [31] and `SEOBNRv4PHM` [37], are employed in the latest third Gravitational-Wave Transient Catalog (GWTC-3) to extract source properties [3]. `NRSur` and `TEOBResumS` waveforms are also used in several analyses of LIGO-Virgo data release [1, 2, 46, 47].

While no approximate waveform will be perfect, we are interested in the question of whether these approximate waveforms are accurate *enough* for the analysis of data from current and future gravitational wave detectors. Ref. [20] gives an accuracy standard of a waveform model used in data analysis under a particular detector noise curve.. It calculates the difference between a model and the "true waveform", which in practice is often represented by NR simulations due to their high accuracy, and states the waveform difference should lie

* q.hu.2@research.gla.ac.uk
† John.Veitch@glasgow.ac.uk

within the unit ball centered on the true waveform. Here, the waveform difference is regarded as a vector, and its length can be calculated by the noise-weighted inner product with itself. If the length is less than the unit radius, the detector could not distinguish the model and the true waveform, thus the waveform model is accurate enough for data analysis. The assessment against NR simulations is widely used in the waveform community [31, 37, 41, 43, 48, 49].

However, as mentioned before, the number of NR waveforms is limited, and the assessment against NR is only available on the parameter grids where NR simulations are available. With the improvement of detector sensitivity and accumulating observation time, the diversity of GW sources will increase, and they may be located in parts of the parameter space where waveform approximants have poor or unknown performance. In fact, several intriguing GW events like this have been revealed in GWTC-3. GW191219_163120 has mass ratio estimated outside of where the waveform models have been calibrated, which results in the uncertainties in its $p_{\rm astro}$ [3, 50]. Parameter estimation of GW200129_065458 shows notable inconsistencies between the results from two different waveform models `IMRPhenomXPHM` and `SEOBNRv4PHM` [3, 50], which is a source of systematic uncertainty on the presence of orbital precession in this system [51]. Assessing and mitigating waveform systematics for current and future detectors has received considerable attention in recent works [48, 52–56].

To assess the waveform accuracy in the regions where NR waveforms are not available, we need an alternative approach. In this work, we address this problem by extending the method of [20] using the triangle inequality in the noise-weighted inner product space. Instead of calculating the difference between one waveform model and NR simulations, we calculate the difference between two waveform models. We will derive a necessary condition of a pair of the waveform models are both accurate enough, a violation of which means *at least* one of the waveform models is not accurate. Although we can not tell whether one model is inaccurate or both are, the violation of this condition still gives information of waveform model validity to a certain degree, especially when the violation is strong. The model-pair assessment does not require NR simulations, and can be performed anywhere in the parameter space as long as the models are able to generate GW waveforms.

We will discuss three types of GW waveforms: binary black hole (BBH) waveform, neutron star-black hole (NSBH) waveform, and binary neutron star (BNS) waveform, for compact binaries are the main sources of current GW detection. For BBH waveform, we focus on `IMRPhenomXPHM` and `SEOBNRv4PHM` which are used in GWTC-3 and GWTC-2.1 data analysis. We assess their accuracy on the parameter estimation samples of GWTC-3 and GWTC-2.1. We find only part of the samples can pass our assessment, and the overall accuracy performance is on the edge of our criterion. Further analysis and simulations shows the inaccurate samples are basically located in the low mass ratio region (we define mass ratio $q < 1$), the high spin region and the edge-on region ($\theta_{\rm JN} \sim \pi/2$). Based on this, we conclude that waveform accuracy should be improved by at least 3-4 orders of magnitude for the 3rd generation GW detectors, which is consistent with previous works [48]. Besides, thanks to the sufficient amount of BBH events, we are able to perform a population-level analysis on the relation between the difference of the waveform models and the posterior sample inconsistency the different waveforms lead to. We find events with less than 40% posterior samples that can meet our accuracy standard tend to have inconsistent results from `IMRPhenomXPHM` and `SEOBNRv4PHM`. For NSBH and BNS waveform models, we perform similar but simpler analysis, as most of them do not include higher modes and precession effects, which may constrain their validness in data analysis.

This paper is organized as follows. In Sec. II, we introduce our accuracy assessment method, including assessment for detector response in Sec. II A and normalization of the waveform difference and its relation to overlap (or mismatch) in Sec. II B. In Sec. III, we apply our method on the 3 types of waveforms mentioned above. Results of BBH waveforms are showed in Sec. III A; NSBH and BNS waveform are showed in Sec. III B. In Sec. IV we summarize our methods and conclusions.

II. ASSESSING WAVEFORM ACCURACY

In this section we will introduce the waveform accuracy standard proposed in Ref. [20], then extend it to model-pair case. We will discuss different standards for the detector response and for waveforms at a fixed signal-to-noise ratio (SNR), which reflects the intrinsic accuracy in the parameter space.

A. Assessment of the detector response

We will use frequency domain waveforms $h_i(f)$, where f means frequency, $i = 0$ denotes the true waveform, and $i = 1, 2$ denotes the 1st and 2nd waveform models. We define inner product between two frequency series as follows:

$$(a \mid b) = 4 \int_0^{+\infty} \frac{a^*(f) b(f)}{S_n(f)} df, \tag{1}$$

where *star* means complex conjugate, and $S_n(f)$ is the power spectral density (PSD) of the detector which is defined as

$$< n^*(f) n(f') >= \frac{1}{2} S_n(f) \delta(f - f'). \tag{2}$$

Here $< \cdots >$ denotes ensemble average and n is the detector noise. Note that Eq. 1 defines an inner product

space, in which frequency series can be treated as vectors. We can define the length (or norm) of a vector:

$$\|a\| = \sqrt{(a|a)}. \tag{3}$$

Some literature defines the inner product as the real part of Eq. 1, but there is no difference between two definitions when it comes to the length. As other inner product spaces, the Cauchy-Schwarz inequality and the triangle inequality hold:

$$\|a\|^2\|b\|^2 \geq |(a|b)|^2 \tag{4}$$

$$\|a\| + \|b\| \geq \|a \pm b\| \geq |\, \|a\| - \|b\| \,| \,. \tag{5}$$

A model waveform can be thought as "accurate enough" when the detector can not distinguish it from the real one. Ref. [20] constructs a waveform family H to quantify the detector's ability to measure the difference between the model and the real waveform, which will be used and extended in this section. Let h_0 be the the true waveform, h_1 be the waveform given by the first model, and $\delta h_1 = h_1 - h_0$ represents their difference. We construct the following waveform family of the first model

$$H_1(\lambda) = (1-\lambda)h_0 + \lambda h_1 = h_0 + \lambda \delta h_1, \quad 0 < \lambda < 1, \tag{6}$$

where λ is a parameter which interpolates between the two models. If the measurement error on λ is greater than the length of its domain of definition (i.e. the parametric distance between real and model waveforms), we can claim the detector is not able to distinguish the waveforms, thus the model is accurate enough. The error σ_λ is given by [57, 58]

$$\sigma_\lambda^{-2} = \left(\frac{\partial H_1}{\partial \lambda} \middle| \frac{\partial H_1}{\partial \lambda}\right) = (\delta h_1 \mid \delta h_1). \tag{7}$$

Therefore, the accuracy standard for a waveform model is

$$\|\delta h_1\|^2 = (\delta h_1 \mid \delta h_1) < 1. \tag{8}$$

Eq. 8 implies the waveform difference should lie within a unit ball in the inner product space, any violation of which means the model is not accurate enough. Since, $< n|n >= 1$ [59], another way to understand Eq. 8 is that if the distance to the real waveform is longer than the length of detector noise, the detector will be able to tell the error of the model. From this angle, the waveform we are considering here should be the detector response, i.e.,

$$\begin{aligned} h_0 &= F_+ h_0^+ + F_\times h_0^\times \\ h_1 &= F_+ h_1^+ + F_\times h_1^\times, \end{aligned} \tag{9}$$

where $F_+, F_\times$ are antenna response functions determined by source sky direction and detector orientation [60]. $+, \times$ denote plus and cross polarizations of GWs. We only consider polarizations under Einstein's general relativity in this work.

We note that some works [20, 59, 61] propose a less stringent criterion than Eq. 8 by changing the upper limit to $2\epsilon\rho^2$, where ρ is SNR and ϵ is the maximum tolerated fractional loss in SNR which needs to be appropriately chosen for detection. In this work, we focus more on the waveform systematics in measurement rather than in detection, so we keep using Eq. 8, i.e., the strict distinguishability criterion.

However, to compute $\|\delta h_1\|$, the true waveform h_0 is needed, which is usually replaced by the computationally expensive NR simulations that can not span all over the parameter space. As a result, the uncertainties are unknown for the waveforms out of the model's calibration range and this may cause some unknown systematic errors in data analysis. An example is GW191219_163120 [3], of which mass ratio is estimated to be out of the waveform calibration region (≤ 0.041) so that there are potential uncertainties in its $p_{\rm astro}$.

To avoid being limited by the true waveform h_0, we introduce another waveform model h_2 to be paired with h_1. Although δh_1 and δh_2 are unknown, their difference $\delta h_1 - \delta h_2$ can be easily calculated:

$$\begin{aligned} \Delta &= \delta h_1 - \delta h_2 \\ &= (h_1 - h_0) - (h_2 - h_0) \\ &= h_1 - h_2. \end{aligned} \tag{10}$$

Assuming both of two waveforms are accurate, i.e., they both satisfy Eq. 8, we can obtain an upper limit of $\|\Delta\|$ using the triangle inequality:

$$\|\Delta\| \leq \|\delta h_1\| + \|\delta h_2\| < 2. \tag{11}$$

Eq. 11 is a necessary condition if h_1 and h_2 are both accurate. That is to say, if Eq. 11 is violated, at least one of the waveform models does not satisfy Eq. 8.

We illustrate possible cases for the $\|\Delta\|$ in Eq. 11 in the vector plots Fig. 1, in which waveforms are treated as vectors in the noise-weighted inner product space. The black circle denotes the sphere of radius 2. Vectors δh_1 and δh_2 denote the difference between the real waveform h_0 and the models h_1, h_2, respectively, and different line styles denote different possibilities. δh_i lies in the circle means the i-th model is accurate and satisfies Eq. 8. If the length of Δ is greater than the upper limit 2 (the diameter of the black circle), as shown in case I, *at least* one of the waveform model errors can not be put inside the circle, i.e., it does not meet the accuracy standard. However, $\|\Delta\| < 2$ does not mean both of the waveforms are accurate, as shown in Case II. Small $\|\Delta\|$ only implies the two models give similar predictions of the waveform but can not guarantee their accuracy. The key idea of this method is: if two waveforms have significant difference, they can not both be correct.

Eq. 11 is not a strong criterion; it can not tell which waveform causes the violation (case I), and may miss

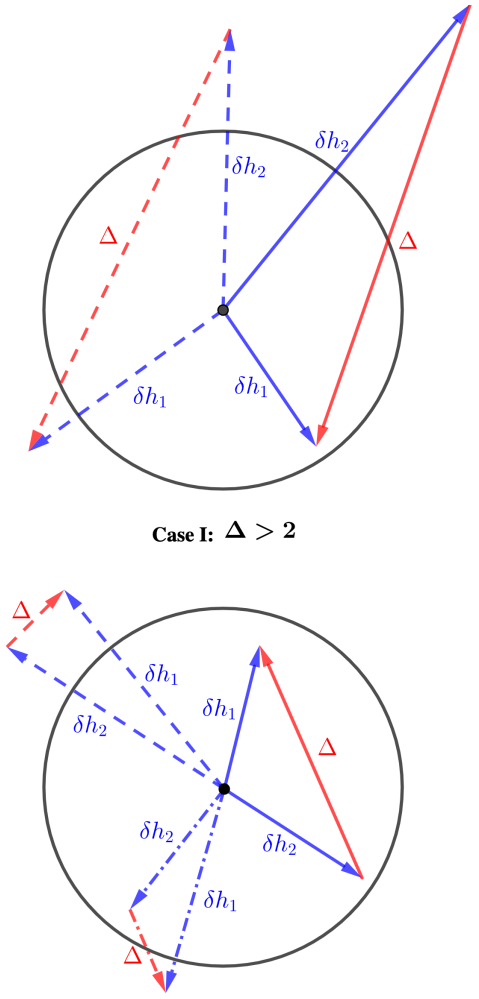

FIG. 1. Vector plots to illustrate all cases of Δ. Blue vectors are the difference between waveform models and the real waveform, and black circles represent the sphere of radius 2, the upper limit of length of δh_i if h_i is accurate ($i = 1, 2$). Red vectors are Δ, the difference between two waveform models (defined in Eq. 10). Different line styles denote different possibilities. In Case I, $\|\Delta\|$ exceeds the upper limit given by Eq. 11, so at least one in h_1, h_2 is not accurate enough. In Case II, $\|\Delta\|$ satisfies Eq. 11, there may be 0, 1, 2 inaccurate waveforms, corresponding to solid line, dotted-dashed line and dashed line, respectively. We can not determine the accuracy of a waveform pair in Case II.

some waveform errors (case II). Despite this, we suppose it still gives certain information about the correctness of waveform modeling. If $\|\Delta\| > 2$, the waveform pair should become less reliable; if $\|\Delta\| \gg 2$, the systematic errors in waveform models should not be neglected as it is highly possible that either of the waveforms is accurate, or one of them has seriously deviated. If $\|\Delta\| < 2$, no evidence of waveform inaccuracy is found by this approach, although we could not exclude the possibility that two waveforms have large but similar errors. The advantage of this method is that it can be performed everywhere in the parameter space, as long as waveform models work in that region.

Eq.11 can be extended to a detector network by defining inner product between matrices (whose elements are frequency series):

$$\mathbf{C} = (\mathbf{D}|\mathbf{B}) \Rightarrow C_{jk} = \sum_{p=1}^{n} (D_{jp} \mid B_{pk}), \tag{12}$$

where $\mathbf{D}$ is an $m \times n$ matrix, $\mathbf{B}$ is an $n \times l$ matrix and the result $\mathbf{C}$ is an $m \times l$ matrix. The signal of the network can be denoted as a column vector $\mathbf{h} = (h^{(1)}, h^{(2)}, \ldots, h^{(\mathrm{N_d})})^{\mathrm{T}}$, where superscript (k) denotes the k-th detector and $\mathrm{N_d}$ is the number of detectors in the network. We can also subtract two waveform models, and define $\mathbf{h_1} - \mathbf{h_2} = \Delta_{\mathrm{net}}$. The norm of Δ_{net} can be calculated

$$\begin{aligned} \|\Delta_{\mathrm{net}}\|^2 &= (\delta\mathbf{h}^{\mathbf{T}}|\delta\mathbf{h}) = \sum_k (\delta h^{(k)}|\delta h^{(k)}) \\ &= \sum_k \left(\Delta^{(k)}\right)^2 < 4\mathrm{N_d}, \end{aligned} \tag{13}$$

where $F_+^{(k)}, F_\times^{(k)}$ are the antenna response functions of the k-th detector. In practice, we can weight the Δ by the number of detectors:

$$\Delta'_{\mathrm{net}} = \frac{\Delta_{\mathrm{net}}}{\sqrt{\mathrm{N_d}}}, \tag{14}$$

so that the Δ'_{net} will have an upper limit of 2 if the waveforms are both accurate enough.

B. Assessment at fixed SNR

The two accuracy standards we proposed, Eq. 11 and Eq. 14, are related to the SNR, as the length of Δ is proportional to the amplitude of GWs. It is reasonable that the higher the SNR is, the easier it is for detectors to distinguish different waveforms, and the more important systematic errors will be in data analysis. However, SNR depends on not only intrinsic parameters, but also extrinsic parameters that trivially modulate the amplitude. It is the phase evolution that is critical to reveal physical properties of the source, and is the intrinsic characteristic of a GW waveform [62] We therefore normalize the Δ with SNR to eliminate the impacts from amplitudes. The optimal SNR is defined as $\rho = \sqrt{(h|h)}$ [57], which is also proportional to the amplitude of GWs like Δ. Thus we have $\Delta \propto \rho$. In fact, we have two waveforms to calculate Δ. The normalization factor is chosen as the geometric mean of SNRs from two waveforms, i.e., $\rho_0 = \sqrt{\rho_1 \rho_2}$. Take Eq. 11 as an example, the normalized Δ is

$$\|\Delta_{\rm SNR=1}\|^2 = \frac{(\delta h_1 - \delta h_2|\delta h_1 - \delta h_2)}{\sqrt{(h_1|h_1)(h_2|h_2)}} = \frac{(h_1 - h_2|h_1 - h_2)}{\sqrt{(h_1|h_1)(h_2|h_2)}}, \quad (15)$$

and we simply have

$$\|\Delta_{\rm SNR=\rho_0}\| = \rho_0 \|\Delta_{\rm SNR=1}\|. \quad (16)$$

Eq. 16 can be used to evaluate waveform accuracy at a fixed SNR. Note the threshold of $\|\Delta_{\rm SNR=\rho_0}\|$ is always 2.

The normalized $\|\Delta\|$ can be related to the current waveform evaluation variable, overlap $\mathcal{O}$, which is defined as

$$\mathcal{O}(h_1, h_2) = \Re \frac{(h_1|h_2)}{\sqrt{(h_1|h_1)(h_2|h_2)}}, \quad (17)$$

where $\Re$ means the real part. $\mathcal{O}$ is between 0 and 1, the higher value represents higher similarities between waveforms h_1 and h_2. One can define mismatch $\mathcal{M} = 1 - \mathcal{O}$. Overlap (or the equivalent mismatch) is widely used to assess the correctness of GW waveforms. The state-of-art models of `IMRPhenom` and `SEOBNR` families can achieve mismatches between 10^{-5} and 10^{-1} compared with NR simulations [31, 37], with precession effects and higher modes being taken into consideration. Overlap between two waveform models $\mathcal{O}(h_1, h_2)$ and the length of normalized waveform difference $\|\Delta\|$ have the following relation:

$$\|\Delta_{\rm SNR=1}\|^2 = \frac{\rho_1}{\rho_2} + \frac{\rho_2}{\rho_1} - 2\mathcal{O}(h_1, h_2) \approx 2(1 - \mathcal{O}), \quad (18)$$

where $\rho_i = \sqrt{(h_i|h_i)}$, $i = 1, 2$. $\|\Delta_{\rm SNR=1}\|$ will decrease with the increase of overlap, and a pair of identical waveforms give $\mathcal{O} = 1$ and $\|\Delta\| = 0$.

We should mention that the inner product in the calculation of waveform difference Δ (as well as overlap [49]), should be minimized (or, for overlap, maximized) over an arbitrary phase ϕ_0 and time shift t_0, in order to eliminate the kinematical difference between models [59]. Considering the sensitive frequency band of current GW detectors, the inner product is integrated from 20 Hz to 2048 Hz throughout this paper.

III. APPLICATIONS

In this section we will apply the accuracy standard Eq. 14 and Eq. 16 to GW waveforms from 3 types of compact binary coalescence: BBH, NSBH, and BNS. We employ the assessment on the GWTC parameter estimation samples and parameter grids we generate; the former aims to investigate whether faulty waveforms were used in GW data analysis and possible systematic error caused by waveform errors, while the latter explores waveforms' performances in different regions of the parameter space. Throughout this paper, we ignore the calibration error, which can cause our waveforms to be slightly different from those used in GWTC-3 and GWTC-2.1 parameter estimation. The calibration error is typically < 4 degrees in phase and $< 7\%$ in amplitude [63], and it acts on both waveform models, so ignoring it will not have large impacts on our results.

A. BBH waveforms

BBH mergers are the most frequent GW events at this stage: Among all 91 GW candidates (36 in GWTC-3 [3], 44 in GWTC-2.1 [4] and 11 in GWTC-1 [1]), over 80 of them are confirmed to be BBH events. In the latest data release from LIGO-Virgo-KAGRA (LVK) collaboration, waveform models `IMRPhenomXPHM` and `SEOBNRv4PHM` are used for analysis of all the BBH events, including re-analysis of GWTC-1 events published in GWTC-2.1. Due to the low SNR of current NSBH events, the resolution of tidal deformability is poor and no strong sign of matter effects is revealed in data analysis. Besides, higher modes and spin precession effects are more important than matter effects for waveform modeling of NSBHs [64], so `IMRPhenomXPHM` and `SEOBNRv4PHM` are also employed on NSBH events to extract physical information.

For all the 89 BBH and NSBH events, we use the cosmologically reweighted parameter estimation posterior samples from GWTC-3 and GWTC-2.1 data release and calculate $\|\Delta'_{\rm net}\|$ (Eq. 14) of the waveform models mentioned above. We use the mixture of `IMRPhenomXPHM` and `SEOBNRv4PHM` samples in most events, but in some events `SEOBNRv4PHM` samples are not provided [4], so we use `IMRPhenomXPHM` samples to calculate $\|\Delta'_{\rm net}\|$ between `IMRPhenomXPHM` and `SEOBNRv4PHM`. Samples we use are the same as GWTC-3 [3] and GWTC-2.1 [4]. For each sample, we generate the waveform (including the detector response) for both models, then apply a time and phase shift on one of them to minimize Eq. 14. The minimized $\|\Delta'_{\rm net}\|$ is the waveform difference we refer to in the following discussion.

When $\|\Delta'_{\rm net}\|$ is greater than 2 at a sampling point, it implies the difference between `IMRPhenomXPHM` and `SEOBNRv4PHM` is so large at this point that they could not both be accurate enough. Furthermore, the difference in waveform will induce a difference in likelihood, and therefore has the potential to affect the results of a parameter estimation algorithm. This yields a systematic difference in parameter estimates, and so the results from different waveform models may not coincide. Therefore, in addition to $\|\Delta'_{\rm net}\|$, we also calculate Jensen–Shannon (J-S) divergence between `IMRPhenomXPHM` samples and `SEOBNRv4PHM` samples (if available). The J-S divergence is a measurement of the similarity between two probability distributions and is used in GWTC-2 [2]. The greater it is, the greater the difference between the two distributions and there may be potential systematic errors in the

data analysis. Since the J-S divergence for samples from a distribution is easiest to evaluate in one dimension, we choose the greatest J-S divergence among samples for the following parameters: mass ratio q, chirp mass $\mathcal{M}$, effective spin $\chi_{\rm eff}$ and effective precession spin $\chi_{\rm p}$ as a measurement of inconsistency of posterior samples, for they are the major physical parameters to be studied. We use `gaussian_kde` in `SciPy` to estimate probability density functions. The base of J-S divergence is chosen to be 2, so that the divergence ranges between 0 and 1.

The full results of the 89 BBH and NSBH events are showed splitly in Tab. I and Tab. II in App. A for reference. We list the basic information of each event, including some source parameters and network SNR, and statistics we construct, including mean value of $\|\Delta'_{\rm net}\|$, normalized $\|\Delta'_{\rm net}\|$ (which equals to $\|\Delta'_{\rm net}\|/{\rm SNR}$), fraction of $\|\Delta'_{\rm net}\| < 2$ samples, and the J-S divergence. We highlight some points in the rest of this subsection.

1. *Overall accuracy*

We show the relations between the waveform difference $\|\Delta'_{\rm net}\|$ of different events and SNR in Fig. 2, in which each point represents a GW event. We find every event has samples that can not meet the $\|\Delta'_{\rm net}\| < 2$ requirement (left panel), but most events have mean $\|\Delta'_{\rm net}\|$ around 2 (right panel). This means some waveform pairs used in data analysis can pass (and are near the edge of) our accuracy standard, but violations exist. We could not identify whether one or both waveform models is inaccurate. Later in Sec. III A 3 we will show it is the samples with large spin or small mass ratio or edge-on inclination that contribute to $\|\Delta'_{\rm net}\| < 2$ fraction. Overall speaking, considering that the violations are generally not strong, we conclude that the current waveform accuracy is around the edge of our assessment standard for the current detector sensitivity which makes detections of SNRs ranging from 8 to ~ 30.

Although the properties of GW sources differ, there is a tendency that large SNR events are more likely to have greater waveform difference (as expected by Eq. 16), and have fewer samples that meet the $\|\Delta'_{\rm net}\| < 2$ requirement. This emphasizes the importance of waveform modeling for future GW detections, in which the SNR can reach hundreds to thousands. We can also make a rough estimation of waveform accuracy requirements for future detectors. The mismatch $\mathcal{M}$ with the "true" waveform is widely-used to assess the waveform accuracy, and the relation between $\|\Delta\|$ and $\mathcal{M}$ can be derived with Eqs. 16 and 18:

$$\mathcal{M}(h_1, h_2) \approx \frac{1}{2\rho_0^2}\|\Delta_{{\rm SNR}=\rho_0}(h_1, h_2)\|^2 \qquad (19)$$

Eq. 19 gives the mismatch between two waveform models, but limited by the triangle inequality, the mismatch between models $\mathcal{M}(h_1, h_2)$ should be at the same order of magnitude as the mismatch between a model and the real waveform $\mathcal{M}(h_1, h_0)$, under the assumption that both models are well-calibrated by high precision waveforms like NR simulation. From our previous discussion we know the $\|\Delta\|$ is around the edge of its upper limit under the current detector sensitivity. If we assume $\|\Delta\|$ is of the same range for future detectors, and SNR is roughly 30–100 times higher, we can determine that the mismatch should decrease 3–4 orders of magnitude. This is consistent with the results reported in Ref. [48].

2. *Impact on parameter estimation*

From previous discussions, the waveforms generated by posterior samples of GWTC-3 and GTWC-2.1 are mostly within the waveform difference bound, yet there are some exceptions. Seven GW events have more than 60% posterior samples violating the standard, which means the difference of two waveform models might be too large to ensure their accuracies. The difference of waveforms may result in difference in parameter estimation, indicating systematic errors [48, 53, 65].

We show the relation between waveform difference and posterior sample consistency (maximum J-S divergence) in Fig. 3, where we can see a weak tendency that events with large waveform difference are more likely to have large J-S divergence, i.e., difference in waveform models may lead to inconsistency in parameter estimation. Particularly, when the fraction of $\|\Delta'_{\rm net}\| < 2$ samples is below 40%, the maximum J-S divergence would be greater than the majority of the GW events. This coincides with our expectation.

However, the inverse statement is not necessarily true. When most posterior samples meet our accuracy standard, it is also possible that two waveform models give inconsistent results. In fact, the waveform error is not the only factor that causes two sets of posterior samples to differ. The behavior of samplers or packages (`bilby` [19] vs `RIFT` [18]) and the prior choice (such as high-spin and low-spin prior for neutron stars [66]) can both influence the consistency between the two posterior samples, although the latter one is not involved in our analysis. Even if we exclude these factors in a full Bayesian analysis, theoretically, it is the combination of waveform gradients, covariance matrices and waveform difference that contributes to systematic errors in parameter estimation [65], not just waveform difference. Besides, we use the maximum J-S divergence as the measurement of posterior difference, which might be influenced when the parameter estimation does not work efficiently on some specific parameters. The last row in Fig. 3 shows such cases. This makes the correlation between posterior sample consistency and waveform difference more statistically dispersed.

In the last three rows of Fig. 3, we give some examples of inconsistent posterior samples. GW200129_065458 has the largest J-S divergence among GWTC-3 events, and GW190412_053044 has the largest J-S divergence among

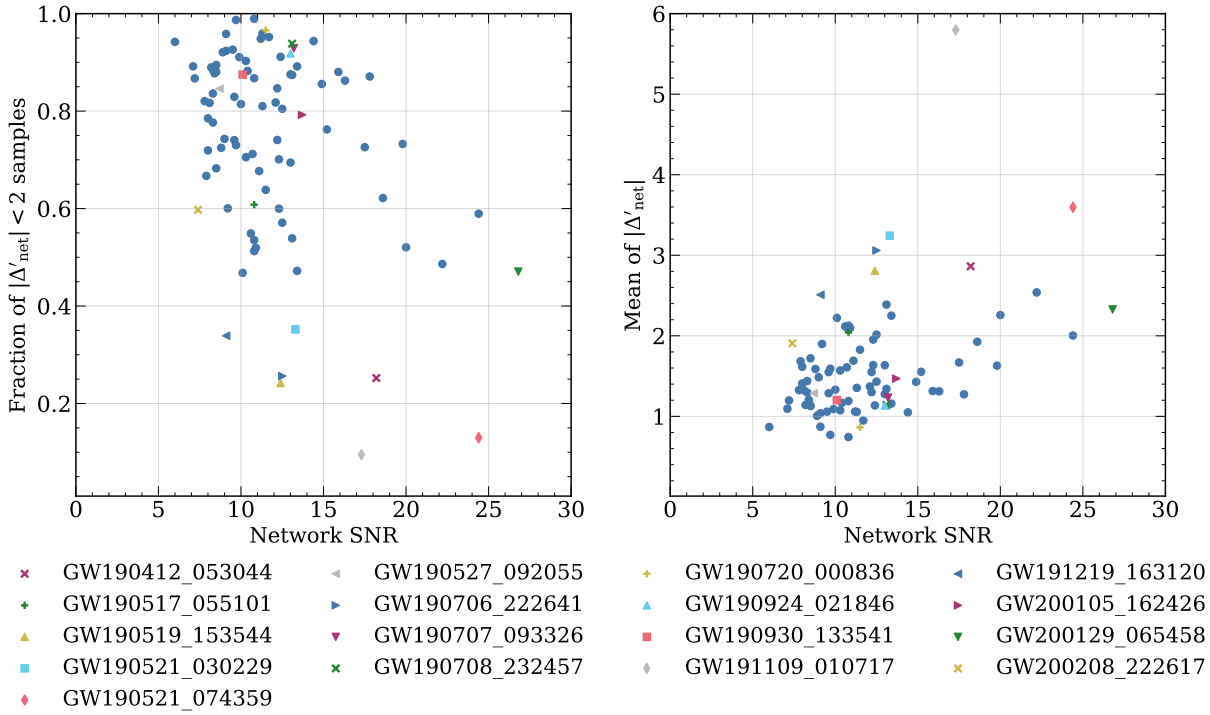

FIG. 2. The left panel shows the relation between fraction of samples that meet our accuracy standard ($\|\Delta'_{\rm net}\| < 2$) and network SNR, and right panel shows mean value of all samples' $\|\Delta'_{\rm net}\|$ and network SNR. We highlight the events whose $\|\Delta'_{\rm net}\| < 2$ samples fraction is less than 0.4 and whose maximum J-S divergence is greater than 0.1. These two plots show waveforms of higher SNR events are more likely to violate our waveform accuracy standard, and given the current detector sensitivity, we are already observing some events which violate our assessment criterion. Note the normalized $\|\Delta'_{\rm net}\|$ can also be read out from the right panel: it is the slope of the line that connects the origin and each point. We can compare the waveform difference of these events when they have the same SNR by comparing the slope. The numerical values are given in the 7th-10th columns of Tab. I and II.

GWTC-2.1 events. The posterior sample inconsistencies of the two events are also reported in GWTC-3 [3] and GWTC-2.1 [4]. In both events, the result with `IMRPhenomXPHM` suggests the possibility of a low mass ratio binary, while that with `SEOBNRv4PHM` does not. GW191219_163120 is the lowest mass ratio detection to-date. Its estimated mass ratio is out of the calibration range of waveform models, so potential systematic error may lie in its data analysis [3]. In our analysis, GW191219_163120 does have less posterior samples that pass our assessment than most other events, but it is not the worst one. Besides, its high SNR (26.8) also contributes to waveform difference: its waveform difference becomes small after normalization. This might be caused by the small spins indicated by parameter estimation. Therefore, we suppose the waveform modelling is not that problematic in the low mass ratio and small spin region, but its high SNR reduces model waveform accuracy. We show its estimation of effective precession spin in Fig. 3, in which we see result of `IMRPhenomXPHM` supports high precession effects in this binary system, while result of `SEOBNRv4PHM` prefers lower precession effects. GW191109_010717 produces the largest waveform difference in our analysis. In a later section III A 3 we will illustrate it might be caused by its special spin effects and higher modes. We show its estimation of effective spin in Fig. 3: results from two waveform models show different multimodality. We also give examples which do not significantly violate our accuracy standard but have inconsistent posterior samples: GW190930_133541 and GW190708_232457. Their results from `IMRPhenomXPHM` seem unable to find the most probable mass ratio. There are six events having this behaviour in GWTC-2.1, as labeled by red circles in Fig. 3. Further investigation is needed, but this is beyond the scope of this work.

Since most posterior samples in this analysis satisfy or just slightly violate our accuracy standard, and samples from two waveform models, generated by different samplers and packages, are mixed as the final results to counterbalance systematic errors, we suppose the waveform modelling error will not induce significant systematic error in data analysis for current detector sensitivity at the population level, while some special events still need further investigation.

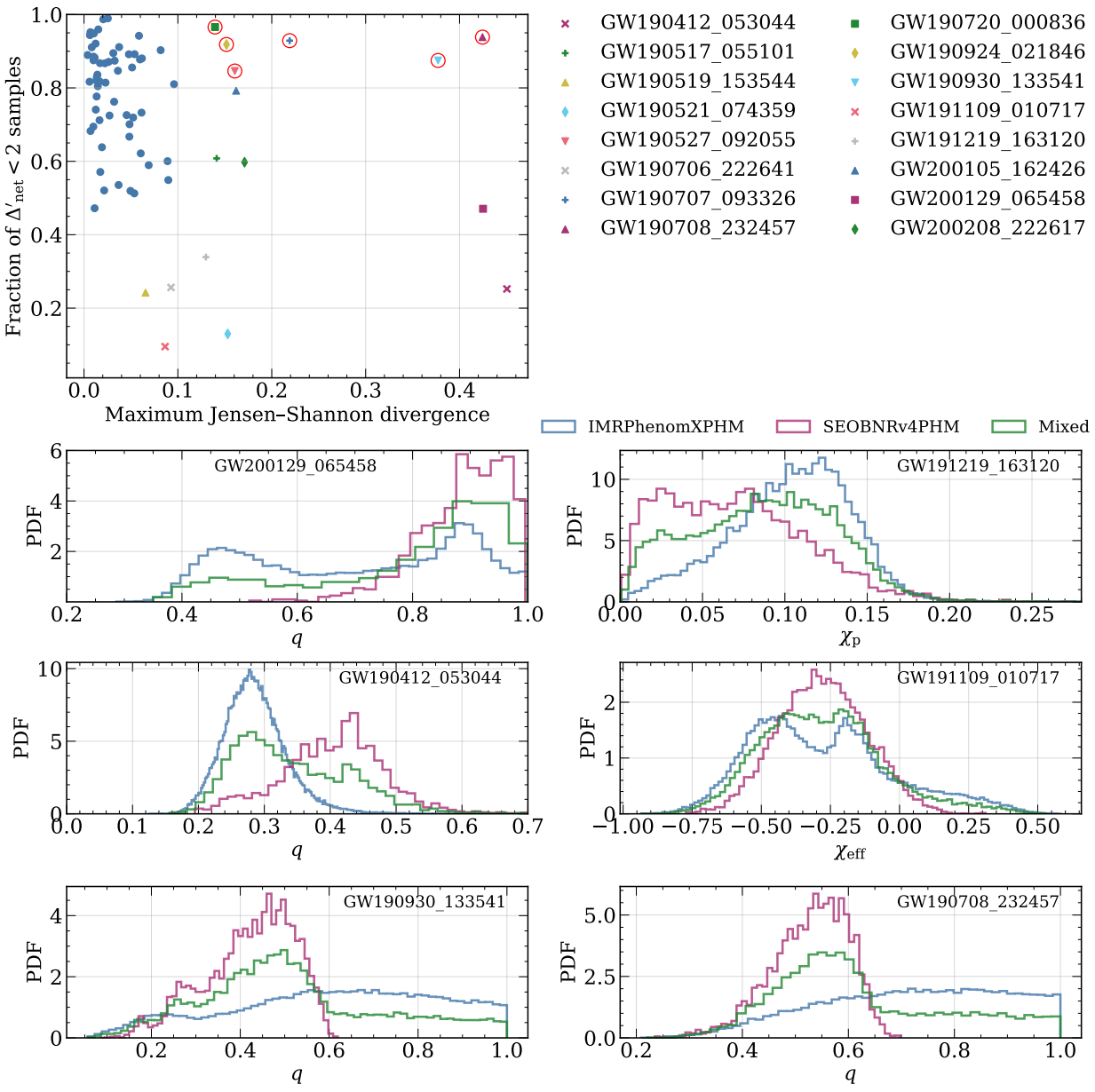

FIG. 3. First row: We visualize the fraction of samples that meet our accuracy standard and maximum J-S divergence in $\{q, \mathcal{M}, \chi_{\rm eff}, \chi_{\rm p}\}$ of the two samples (10th and 11th columns of Tab. I and II). We highlight the events whose fraction of $\|\Delta'_{\rm net}\| < 2$ samples is less than 40%, and the events whose maximum J-S divergence is greater than 0.1. Some GWTC-2.1 events have nearly flat `IMRPhenomXPHM` posteriors for mass ratio (as showed in the undermost row), which caused large J-S divergence despite the small waveform difference. We use red circles to label these events.
Bottom three rows: We show some examples of inconsistent posterior samples; the parameter name and event name are shown in the plots.

3. *Waveform difference in different parameter regions*

In Sec III A, from the angle of data analysis, we discussed SNR's impact on waveform accuracy. What is more physically interesting is how the waveform accuracy varies with intrinsic properties of the GW source. It is plausible that model accuracy decreases when the system includes some complex processes, such as a highly asymmetric mass ratio, high spin effects, high eccentricity and so forth. Accuracy may also drop when the contributions from higher modes increase, which usually happens to edge-on binaries [67–69].

We plot posterior samples of selected events and highlight the samples whose waveform difference is greater than 2 in Fig. 4. In the nearly equal mass region and small spin region, `IMRPhenomXPHM` and `SEOBNRv4PHM` agree with each other and have waveform difference less than 2. However, when mass ratio deviates from 1, or when spin parameters deviates from 0, the waveform difference grows and the waveform pair fails to pass the accuracy standard. For extrinsic parameters, we find that the waveform difference is largest when inclination angle $\theta_{\rm JN}$ for precessing systems is close to $\pi/2$. We attribute this to two causes: the contribution of higher modes increases when the source is egde-on, and the amplitude modulations caused by precession become increasingly visible, magnifying differences in the way precession is modelled [67–72]. This is the reason why events like GW191109_010717 have a small fraction of posterior samples that pass the accuracy standard: estimations of their parameters mainly lead to low mass ratio, high spin regions or edge-on regions.

We then perform simulations of BBH events on the design sensitivity of Advanced LIGO [73]. We set the primary mass at $30M_\odot$, and mass ratio at $1, 0.8, 0.5$ and 0.2. The spin of each component is randomly generated: spin magnitude is uniformly distributed between 0 and 1, and spin direction is isotropic. Inclination angle is isotropic as well. We neglect detector response functions and only include plus polarization here, which will not change our qualitative conclusions. For each mass ratio we simulate 6000 BBH events and calculate the waveform difference between `IMRPhenomXPHM` and `SEOBNRv4PHM`. Since waveform difference $\|\Delta\|$ is proportional to SNR, we introduce a SNR threshold above which $\|\Delta\|$ will be greater than 2. In Fig. 5, we plot the distributions of simulation parameters in the style of corner plot for different mass ratios, and the corresponding SNR thresholds in colors. We find the SNR threshold can reach 30 in the low spin and face-on region, but gradually drops below 10 as the spin parameters increase or $\theta_{\rm JN}$ tends to $\pi/2$. The change in mass ratio has the same effect, $\|\Delta\|$ can reach 2 at a smaller spin if the mass ratio is low. However, we find the $q = 0.2$ simulations can achieve high SNR threshold for low spin face-on sources, while high-spin or edge-on simulations are more likely to produce low SNR thresholds regardless of the mass ratio. Therefore, for current waveform modeling, spin effects and higher modes may need more improvements than low mass ratio cases. This coincides with our calculation on the asymmetric mass ratio but small spin event GW191219_163120. The disagreement in high-spin CBC waveforms and its impact on parameter estimation is also reported in Ref. [53].

Our simulation is consistent with the calculation for real events. Given GW events with SNRs ranging from 8 to 30 (for current detector sensitivity), those generated by nearly equal mass systems or low spin systems would have more $\|\Delta'_{\rm net}\| < 2$ samples, while the other events' posterior samples mostly fail our test, like GW191109_010717. Using the same method in Sec. III A 1 and comparing current SNR threshold with expected SNR of 3rd generation GW detectors, we can also conclude that, in general, the waveform accuracy should be improved for 3 to 4 orders of magnitude. However, for high spin and low mass ratio regions, as well as higher modes, the current waveform models may need more improvements. To calibrate waveform models, these regions might be where NR simulations are most needed for future waveform modelling.

B. NSBH and BNS waveforms

NSBH and BNS events are much less frequent than BBH events so far - only three events are generally considered as NSBH candidates: GW191219_163120, GW200105_162426, and GW200115_042309, and two are considered as BNS events: GW170817 and GW190425_081805. Due to the complexity of these systems (e.g., highly asymmetric mass ratio, eccentricity for NSBH binaries, and matter effects for both types), some physical effects are yet to be included in their waveform models. Current NSBH waveform models of `IMRPhenom` and `SEOBNR` familes, `IMRPhenomNSBH` and `SEOBNRv4_ROM_NRTidalv2_NSBH` [74], are calibrated by non-spinning neutron star simulations and only allow aligned spins. For current BNS models, `IMRPhenomPv2_NRTidalv2` supports precessing spins while `SEOBNRv4T_surrogate` only supports aligned spins. Recent works have made `TEOBResumS` able to generate waveforms for precessing BNS systems with higher modes [75] , as well as waveforms for eccentric BNS systems [76], but they have not been applied to the GWTC-2.1 and -3.

For the three NSBH events, we calculate the $\|\Delta'_{\rm net}\|$ of their posterior samples generated by `IMRPhenomNSBH` and `SEOBNRv4_ROM_NRTidalv2_NSBH`. The fraction of $\|\Delta'_{\rm net}\| < 2$ samples are $99.4\%, 99.6\%$ and 100% for GW191219_163120, GW200105_162426, and GW200115_042309, respectively. Low SNR of these three events may contribute to the small waveform differences, but compared with the BBH events, lacking of precession effects and higher modes should be the decisive factors that make the waveform pair coincide, and it does not necessarily mean these models can describe general NSBH systems with high accuracy. As for BNS events, `IMRPhenomPv2_NRTidalv2` is the only model used in GWTC-2.1 and -3 that includes precession effects, it is not feasible to compare waveform difference of its posterior samples with others. Hence we do not include calculation of BNS waveforms for real events in this work.

We perform simulations for NSBH and BNS systems respectively. For NSBH waveform models, we assume zero spin and secondary mass of 1.4 $M_\odot$. We change mass ratio between 0.02 and 0.25, and tidal deformability parameter between 0 and 2000. For BNS, we assume the two neutron stars are exactly the same: same mass

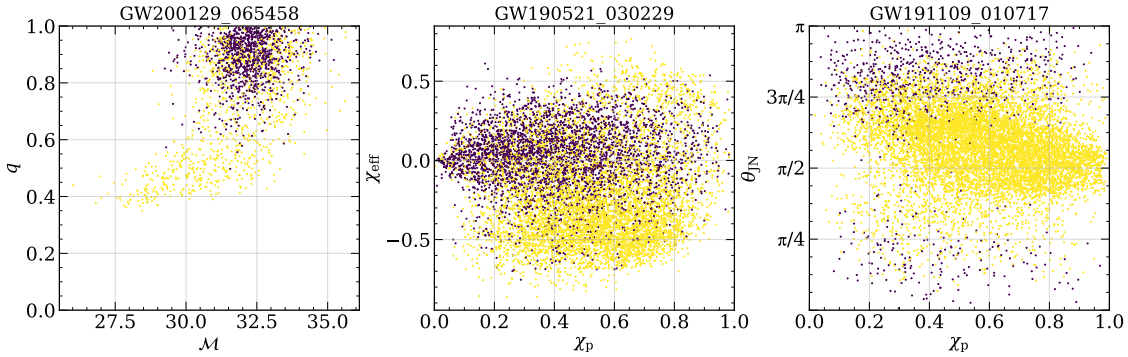

FIG. 4. Posterior scatter of selected events. Yellow points represents samples with $\|\Delta'_{\rm net}\| > 2$, purple points are samples with $\|\Delta'_{\rm net}\| < 2$. We show three representative events with in two-dimension parameter plane $(\mathcal{M}, q)$, $(\chi_{\rm p}, \chi_{\rm eff})$ and $(\chi_{\rm p}, \theta_{\rm JN})$, respectively. It shows the inaccuracies mainly come from high spin region, low mass ratio region, and edge-on region.

1.4 $M_\odot$, same tidal deformability parameter and spin. Then we change spin magnitude between -0.2 and 0.2, and tidal deformability parameter between 0 and 2000. We assume zero inclination for both systems. The results are shown in Fig. 6.

We find the main disagreement for NSBH waveform models lies in mass ratio, as Fig. 6 shows the waveform difference drops with q but is insensitive to the tidal deformability parameter Λ_2 of the neutron star. The latter is because both approximants use the `NRTidalv2` [77] phase description to model the matter effects. SNR threshold can drop below 5 when q is small, but all the three NSBH candidates have SNRs lower than the thresholds indicated in the corresponding regions in Fig. 6. Note we assume zero spin in this simulation, but non-zero spin samples exist in the three NSBH candidates and would make extra contributions to waveform difference. Therefore, they still have a small fraction of $\|\Delta'_{\rm net}\| > 2$ samples. Given the SNR threshold in this simulation, NSBH waveform model accuracies (in terms of the mismatch from real waveform) also need an improvement of 3–4 orders of magnitude for future detection, leaving aside the unincluded physical effects.

As for BNS waveforms `IMRPhenomPv2_NRTidalv2` and `SEOBNRv4T_surrogate`, we change values of Λ and spin magnitude. We assume both components have the same aligned spin and mass, so the individual spin magnitude is equal to the effective spin. We find two waveform models agree with each other quite well in the $\Lambda < 500$, $\chi_{\rm eff} < 0.05$ region, with SNR thresholds up to 100. This is the region that coincides with our current understanding of neutron stars. However, when spin increases, the SNR threshold can drop below 20. This also implies accuracy of future waveform models should be improved by several orders of magnitude.

We do not discuss further about NSBH and BNS waveform models, for we suppose the number of real events is not enough for us to perform analysis on the population level, and further work on precession, higher modes, and even eccentricity should be done for more NSBH and BNS waveform models.

IV. CONCLUSIONS

In this work, we developed a diagnostic test for the presence of waveform mismodelling. This extends the work of Ref. [20] to realistic analyses. While Ref. [20] suggests a waveform model should have error (as a vector) shorter than 1 to be accurate enough, we introduce two waveform models and find their difference should be shorter than 2 if they are both accurate enough. This method frees accuracy evaluation from the unknown true waveform, and it enables the evaluation to be performed in larger, continuous regions in the parameter space: the regions where waveform models can work, rather than where NR simulations are done. We should note that our method can only tell the existence of inaccurate waveform models. It can not tell which one (or both) is (are) inaccurate if the pair fail, or guarantee any accuracy when the pair do not fail. The key idea is: If two models have significant difference, they can not be both accurate enough, but when the difference is small, we can not rule out the possibility that two models are making similar mistakes.

For BBH waveform models, we choose the statr-of-art models from `IMRPhenom` and `SEOBNR` family, `IMRPhenomXPHM` and `SEOBNRv4PHM` for illustration. We have applied our test to existing parameter estimates from the GWTC-3 and GWTC-2.1 (which used the waveform models mentioned above), and found differences in the results of data analysis from different waveform models. The samples that fail our test are mostly located in the low mass ratio, high spin or egde-on regions in the parameter space, which means waveform models become less reliable in these regions. Our simulations agree with this: the waveform difference between `IMRPhenomXPHM` and `SEOBNRv4PHM` can reach the threshold 2 when SNR is less than 10 in high spin regions and edge-on regions; waveform difference increases in low mass ratio region the as well. We also note that spin effects and inclina-

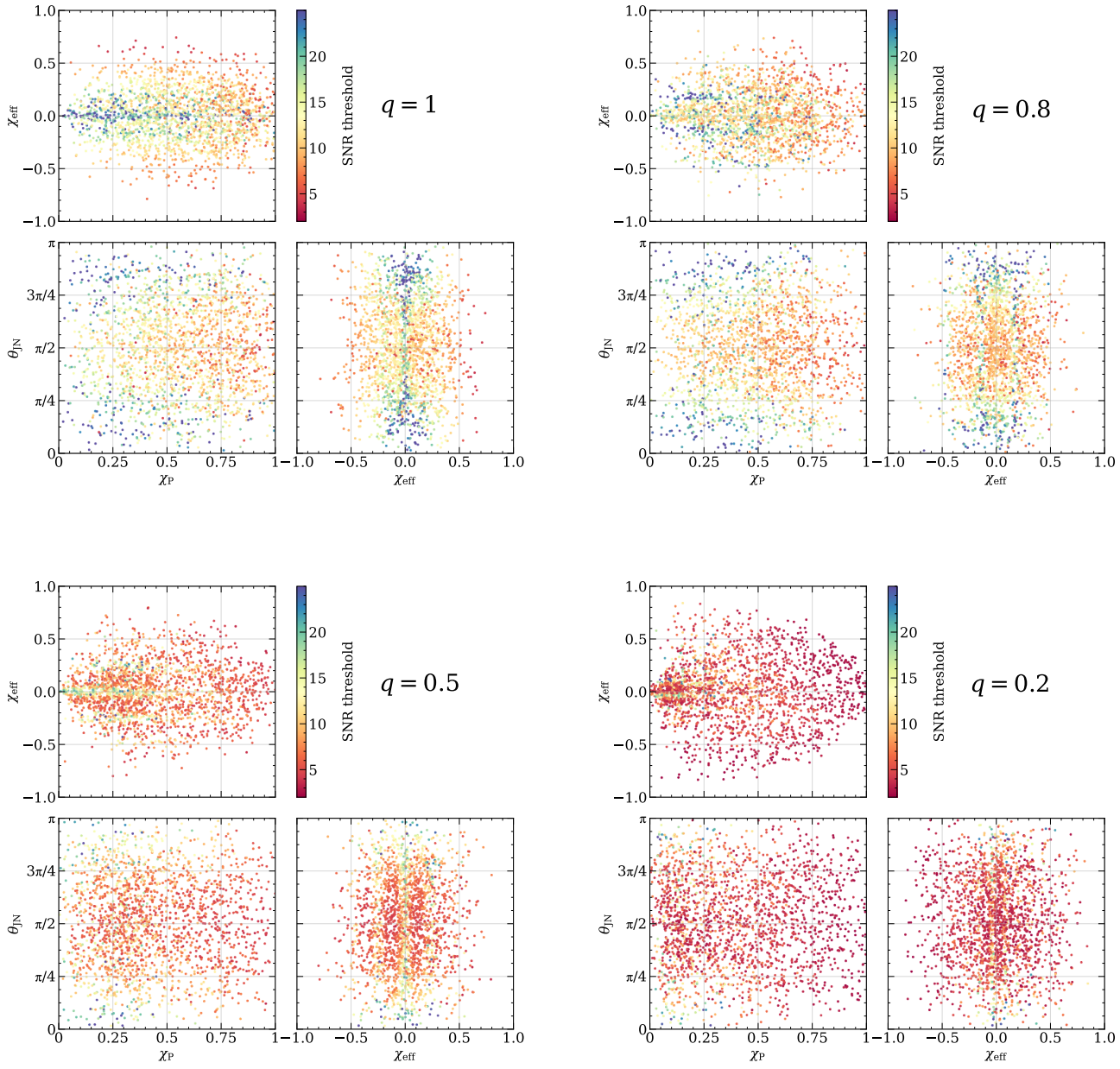

FIG. 5. Simulations of random spin and inclination BBHs under LIGO design sensitivity. The primary mass is fixed at $30M_\odot$ and mass ratio varies from 1 to 0.2, as showed in the top right corner of each figure. We calculate waveform difference $\|\Delta\|$ between `IMRPhenomXPHM` and `SEOBNRv4PHM` for each simulation and the SNR when $\|\Delta\|$ reaches 2. The SNR threshold is shown in different colors. Face-on events with smaller spins and equal masses tend to have a higher SNR threshold.

tions (higher modes) are more problematic for waveform modelling than mass ratio. This points out where NR simulations are needed most for future waveform calibration.

We have investigated the correlation between waveform difference and inconsistency of parameter estimation samples given by different waveform models. The latter is measured by the J-S divergence. For the GWTC-3 and GWTC-2.1 posterior samples, we find when the fraction of $\Delta < 2$ samples is less than 40%, it is more likely to obtain a J-S divergence larger than most other events, which is a sign of underlying systematic errors caused by waveform error. We also note that the inverse is not necessarily true, as the waveform model is not the only factor that can influence the generation of posterior samples, but nonetheless it is always helpful to have one of the factors checked. Since multi-waveform analysis is becoming a standard way of reducing systematic errors

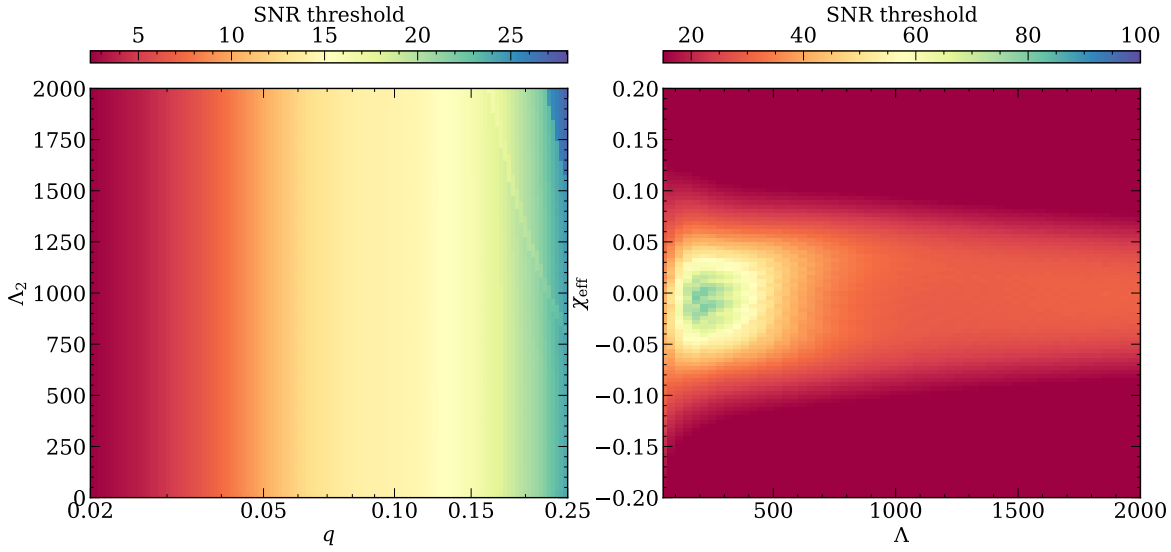

FIG. 6. Left panel: Simulations of NSBH binaries in the mass ratio q - tidal demformability Λ_2 plane. The mass of the neutron star is fixed at 1.4 $M_\odot$, and we assume both components have zero spin. The colors in the plane represent the SNR threshold for the waveform difference between `IMRPhenomNSBH` and `SEOBNRv4_ROM_NRTidalv2_NSBH`, defined in the same way as before.
Right panel: Simulations of BNS binaries in the tidal deformability Λ - effective spin $\chi_{\rm eff}$ plane. We assume both neutron stars are 1.4 $M_\odot$ and they have the same spin and tidal deformability parameter. Colors represent SNR threshold for `IMRPhenomPv2_NRTidalv2` and `SEOBNRv4T_surrogate`.

in parameter estimation of GW sources, we suggest that waveform difference analysis can be used as a real-time quantitative check in the parameter estimation workflow.

For NSBH waveforms, we select `IMRPhenomNSBH` and `SEOBNRv4_ROM_NRTidalv2_NSBH`, the two models used in GWTC-3 and GWTC-2.1 parameter estimation. The posterior samples of the 3 NSBH candidates have small NSBH waveform difference compared to BBH waveforms. We credit this to the fact that these waveform models do not include non-aligned spins or higher modes as BBH waveforms. As expected, we find waveform difference increase when mass ratio decreases in our simulation. The SNR threshold drops below 10 when mass ratio is less than 0.05, indicating that more calibrations are needed for this region, leaving aside the lack of some other physical effects.

For BNS waveforms, we have not applied our test on real events samples, for only `IMRPhenomPv2_NRTidalv2` is used in GWTC-2.1 and -3 data analysis, and we could not find another comparable model to be paired with it. We simulate aligned spin BNSs for `IMRPhenomPv2_NRTidalv2` and `SEOBNRv4T_surrogate` instead. We find the systematic differences between the approximants we examined are small in the region where $\Lambda < 500$, and $|\chi_{\rm eff}| < 0.05$, which should be the case for our current understanding of neutron stars. However, there are some differences when $\Lambda < 50$ where the waveforms appear to diverge again. In the high spin regions, the SNR threshold drops below 20, which can not meet future high SNR detections.

The waveform difference is related to the widely used overlap (or mismatch) through Eqs. 18 and 19. If we assume two models are well calibrated and have comparable errors, we can give a rough estimate of future waveform accuracy requirement. This complements previous works on waveform accuracy [48]. Looking at the SNR thresholds for the three types of waveforms, we know the current waveform accuracy is not enough for future high SNR detections where SNR can reaches up to 1000. The mismatch from the real waveform needed to be reduced by at least 3 orders of magnitude. This is consistent with previous work.

Finally, this method can be extended to more complex GW waveform models for future GW detection, such as waveforms including eccentricity. We can perform this analysis as long as there are at least two waveform models with the similar accuracy and which include the same physical parameters. Our method can work beyond the NR calibration range, thus it can be efficient way to study the waveform models' extrapolation performance. This may also be a guide of where NR simulations are most needed in the parameter space.

ACKNOWLEDGMENTS

The authors would like to thank Daniel Williams, Michael Pürrer, Christopher Berry, Ik Siong Heng, Rossella Gamba and Jacob Lange for helpful discussions and suggestions. Daniel also helped us use the GWTC-3 and GWTC-2.1 PE data release. The authors are grateful for computational resources provided by the LIGO Lab at Caltech which is supported by National Science Foundation Grants PHY-0757058 and PHY-0823459. QH is supported by CSC. JV is supported by STFC grant ST/V005634/1.

Appendix A: Full results of BBH waveform in GWTC-3 and GWTC-2.1

[1] LIGO Scientific Collaboration and Virgo Collaboration, GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs, Physical Review X **9**, 031040 (2019).

[2] R. Abbott *et al.*, GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo during the First Half of the Third Observing Run, Physical Review X **11**, 1–54 (2021).

[3] R. Abbott *et al.*, GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run (2021), arXiv:2111.03606.

[4] The LIGO Scientific Collaboration and The Virgo Collaboration, GWTC-2.1: Deep Extended Catalog of Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run (2021), arXiv:2108.01045 [gr-qc].

[5] A. H. Nitz, C. Capano, A. B. Nielsen, S. Reyes, R. White, D. A. Brown, and B. Krishnan, 1-OGC: The First Open Gravitational-wave Catalog of Binary Mergers from Analysis of Public Advanced LIGO Data, The Astrophysical Journal **872**, 195 (2019).

[6] A. H. Nitz, T. Dent, G. S. Davies, S. Kumar, C. D. Capano, I. Harry, S. Mozzon, L. Nuttall, A. Lundgren, and M. Tápai, 2-OGC: Open Gravitational-wave Catalog of Binary Mergers from Analysis of Public Advanced LIGO and Virgo Data, The Astrophysical Journal **891**, 123 (2020).

[7] A. H. Nitz, C. D. Capano, S. Kumar, Y.-F. Wang, S. Kastha, M. Schäfer, R. Dhurkunde, and M. Cabero, 3-OGC: Catalog of gravitational waves from compact-binary mergers, The Astrophysical Journal **922**, 76 (2021), arXiv:2105.09151.

[8] A. H. Nitz, S. Kumar, Y.-F. Wang, S. Kastha, S. Wu, M. Schäfer, R. Dhurkunde, and C. D. Capano, 4-OGC: Catalog of gravitational waves from compact-binary mergers, arXiv (2021), arXiv:2112.06878.

[9] B. Zackay, L. Dai, T. Venumadhav, J. Roulet, and M. Zaldarriaga, Detecting gravitational waves with disparate detector responses: Two new binary black hole mergers, Physical Review D **104**, 10.1103/physrevd.104.063030 (2021).

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

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

[12] B. Zackay, T. Venumadhav, L. Dai, J. Roulet, and M. Zaldarriaga, Highly spinning and aligned binary black hole merger in the advanced ligo first observing run, Phys. Rev. D **100**, 023007 (2019).

[13] J. Aasi *et al.*, Advanced LIGO, Classical and Quantum Gravity **32**, 074001 (2015), arXiv:1411.4547.

[14] F. Acernese *et al.*, Advanced Virgo: A 2nd generation interferometric gravitational wave detector, Classical and Quantum Gravity **32**, 024001 (2015), arXiv:1408.3978.

[15] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, FINDCHIRP: An algorithm for detection of gravitational waves from inspiraling compact binaries, Physical Review D **85**, 122006 (2012).

[16] J. Veitch *et al.*, Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library, Physical Review D **91**, 1–25 (2015).

[17] C. M. Biwer, C. D. Capano, S. De, M. Cabero, D. A. Brown, A. H. Nitz, and V. Raymond, PyCBC inference: A python-based parameter estimation toolkit for compact binary coalescence signals, Publications of the Astronomical Society of the Pacific **131**, 10.1088/1538-3873/aaef0b (2019).

[18] J. Lange, R. O'Shaughnessy, and M. Rizzo, Rapid and accurate parameter inference for coalescing, precessing compact binaries (2018), arXiv:1805.10457.

[19] G. Ashton *et al.*, BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy, Astrophys. J. Suppl. **241**, 27 (2019), arXiv:1811.02042 [astro-ph.IM].

[20] L. Lindblom, B. J. Owen, and D. A. Brown, Model waveform accuracy standards for gravitational wave data analysis, Physical Review D **78**, 1–12 (2008).

[21] L. Blanchet, Gravitational radiation from post-newtonian sources and inspiralling compact binaries, Living Reviews in Relativity **17**, 1–185 (2014).

[22] E. Berti, V. Cardoso, and A. O. Starinets, Quasinormal modes of black holes and black branes, Classical and Quantum Gravity **26**, 163001 (2009).

[23] K. Jani, J. Healy, J. A. Clark, L. London, P. Laguna, and D. Shoemaker, Georgia tech catalog of gravitational waveforms, Classical and Quantum Gravity **33**, 1–8 (2016).

[24] A. H. Mroué *et al.*, Catalog of 174 Binary Black Hole Simulations for Gravitational Wave Astronomy, Physical Review Letters **111**, 241104 (2013).

[25] M. Boyle *et al.*, The SXS collaboration catalog of binary

Event name	m_1	m_2	$\chi_{\rm eff}$	$\chi_{\rm p}$	$\theta_{\rm JN}$	SNR	Mean	$\rm Mean_{Norm}$	Fraction	Max J-S Div.
GW150914_095045	$37.99^{+5.11}_{-2.89}$	$32.96^{+3.18}_{-5.03}$	$-0.04^{+0.12}_{-0.14}$	$0.51^{+0.35}_{-0.38}$	$2.70^{+0.32}_{-0.71}$	24.4	2.0	0.08	0.59	0.069
GW151012_095443	$29.62^{+17.08}_{-6.87}$	$16.30^{+5.73}_{-6.02}$	$0.12^{+0.28}_{-0.21}$	$0.36^{+0.43}_{-0.27}$	$1.70^{+1.13}_{-1.40}$	10	1.33	0.13	0.81	0.023
GW151226_033853	$15.50^{+12.19}_{-4.00}$	$8.18^{+2.62}_{-3.02}$	$0.20^{+0.23}_{-0.08}$	$0.52^{+0.36}_{-0.35}$	$0.88^{+2.00}_{-0.66}$	13.1	2.39	0.18	0.54	N/A
GW170104_101158	$34.78^{+7.73}_{-4.85}$	$25.25^{+4.43}_{-5.60}$	$-0.04^{+0.15}_{-0.19}$	$0.40^{+0.40}_{-0.31}$	$1.10^{+1.79}_{-0.86}$	13	1.28	0.1	0.88	0.01
GW170608_020116	$11.41^{+4.36}_{-1.49}$	$8.38^{+1.21}_{-2.06}$	$0.05^{+0.13}_{-0.05}$	$0.32^{+0.41}_{-0.24}$	$2.37^{+0.58}_{-2.06}$	14.9	1.43	0.1	0.86	0.051
GW170729_185629	$77.88^{+17.89}_{-14.48}$	$44.13^{+18.11}_{-17.46}$	$0.29^{+0.25}_{-0.33}$	$0.39^{+0.40}_{-0.29}$	$1.35^{+1.44}_{-1.03}$	10.8	2.08	0.19	0.54	0.037
GW170809_082821	$41.19^{+9.63}_{-6.27}$	$29.25^{+5.63}_{-6.46}$	$0.07^{+0.17}_{-0.17}$	$0.39^{+0.44}_{-0.30}$	$2.61^{+0.39}_{-0.59}$	12.4	1.14	0.09	0.91	0.037
GW170814_103043	$34.70^{+6.01}_{-3.52}$	$28.03^{+3.22}_{-4.59}$	$0.08^{+0.13}_{-0.12}$	$0.48^{+0.38}_{-0.37}$	$0.69^{+1.92}_{-0.48}$	15.9	1.31	0.08	0.88	0.062
GW170818_022509	$42.14^{+7.99}_{-5.04}$	$33.42^{+5.07}_{-6.14}$	$-0.06^{+0.19}_{-0.22}$	$0.56^{+0.34}_{-0.41}$	$2.46^{+0.47}_{-0.50}$	11.3	1.06	0.09	0.96	0.015
GW170823_131358	$51.67^{+11.92}_{-7.85}$	$39.42^{+8.11}_{-11.03}$	$0.05^{+0.21}_{-0.22}$	$0.47^{+0.41}_{-0.35}$	$1.73^{+1.16}_{-1.48}$	11.5	1.83	0.16	0.64	0.019
GW190408_181802	$31.70^{+6.96}_{-3.88}$	$23.83^{+3.56}_{-5.01}$	$-0.03^{+0.13}_{-0.17}$	$0.37^{+0.41}_{-0.29}$	$1.01^{+1.85}_{-0.79}$	14.4	1.05	0.07	0.94	0.006
GW190412_053044	$31.76^{+6.81}_{-6.60}$	$10.34^{+2.19}_{-1.55}$	$0.21^{+0.12}_{-0.13}$	$0.19^{+0.22}_{-0.12}$	$0.92^{+1.69}_{-0.40}$	18.2	2.86	0.16	0.25	0.45
GW190413_052954	$52.79^{+15.00}_{-10.02}$	$37.90^{+9.95}_{-11.57}$	$-0.04^{+0.27}_{-0.32}$	$0.44^{+0.42}_{-0.33}$	$0.79^{+2.01}_{-0.58}$	8.5	1.13	0.13	0.88	N/A
GW190413_134308	$83.12^{+19.30}_{-15.64}$	$50.15^{+19.47}_{-23.80}$	$-0.01^{+0.28}_{-0.38}$	$0.55^{+0.36}_{-0.41}$	$1.85^{+1.01}_{-1.52}$	10.3	1.57	0.15	0.71	N/A
GW190421_213856	$60.79^{+13.06}_{-8.71}$	$46.94^{+9.04}_{-14.55}$	$-0.10^{+0.21}_{-0.27}$	$0.45^{+0.41}_{-0.34}$	$2.03^{+0.85}_{-1.70}$	9.7	1.59	0.16	0.73	N/A
GW190503_185404	$53.32^{+12.17}_{-10.28}$	$36.45^{+10.17}_{-12.62}$	$-0.05^{+0.23}_{-0.30}$	$0.43^{+0.40}_{-0.33}$	$2.50^{+0.46}_{-0.62}$	12.2	1.55	0.13	0.74	0.013
GW190512_180714	$29.34^{+6.90}_{-6.72}$	$15.83^{+4.64}_{-2.97}$	$0.02^{+0.13}_{-0.14}$	$0.26^{+0.41}_{-0.20}$	$1.79^{+1.07}_{-1.50}$	12.2	1.3	0.11	0.85	0.036
GW190513_205428	$49.99^{+14.75}_{-12.78}$	$25.60^{+10.88}_{-6.98}$	$0.16^{+0.29}_{-0.22}$	$0.35^{+0.43}_{-0.26}$	$0.79^{+2.01}_{-0.58}$	12.3	1.64	0.13	0.7	0.048
GW190514_065416	$66.96^{+21.39}_{-12.57}$	$47.33^{+12.11}_{-18.23}$	$-0.08^{+0.29}_{-0.35}$	$0.45^{+0.42}_{-0.33}$	$1.47^{+1.38}_{-1.20}$	8.3	1.44	0.17	0.78	0.013
GW190517_055101	$52.92^{+14.76}_{-9.99}$	$32.75^{+9.55}_{-12.25}$	$0.49^{+0.21}_{-0.28}$	$0.55^{+0.31}_{-0.32}$	$2.12^{+0.70}_{-1.18}$	10.8	2.04	0.19	0.61	0.141
GW190519_153544	$94.80^{+15.84}_{-12.40}$	$59.88^{+16.88}_{-18.85}$	$0.33^{+0.20}_{-0.24}$	$0.45^{+0.36}_{-0.28}$	$1.61^{+0.95}_{-1.02}$	12.4	2.81	0.23	0.24	0.066
GW190521_030229	$152.36^{+31.65}_{-17.62}$	$89.65^{+48.96}_{-52.17}$	$-0.14^{+0.50}_{-0.45}$	$0.49^{+0.33}_{-0.35}$	$1.38^{+1.40}_{-1.07}$	13.3	3.24	0.24	0.35	N/A
GW190521_074359	$52.19^{+7.65}_{-5.39}$	$40.36^{+5.87}_{-7.16}$	$0.10^{+0.13}_{-0.13}$	$0.39^{+0.37}_{-0.30}$	$1.48^{+1.27}_{-1.10}$	24.4	3.6	0.15	0.13	0.153
GW190527_092055	$51.05^{+29.88}_{-9.27}$	$32.36^{+11.95}_{-13.36}$	$0.10^{+0.22}_{-0.22}$	$0.36^{+0.47}_{-0.28}$	$1.15^{+1.65}_{-0.87}$	8.7	1.29	0.15	0.85	0.161
GW190602_175927	$106.81^{+24.90}_{-17.99}$	$67.63^{+23.03}_{-31.56}$	$0.12^{+0.25}_{-0.28}$	$0.45^{+0.43}_{-0.34}$	$2.13^{+0.79}_{-1.87}$	12.3	1.95	0.16	0.6	N/A
GW190620_030421	$87.13^{+23.93}_{-17.28}$	$53.35^{+17.64}_{-24.30}$	$0.34^{+0.22}_{-0.29}$	$0.48^{+0.38}_{-0.32}$	$2.07^{+0.82}_{-1.74}$	10.9	2.1	0.19	0.52	0.049
GW190630_185205	$41.40^{+8.12}_{-6.50}$	$28.24^{+5.79}_{-5.63}$	$0.10^{+0.14}_{-0.13}$	$0.33^{+0.36}_{-0.24}$	$1.41^{+1.43}_{-1.16}$	15.2	1.55	0.1	0.76	0.032
GW190701_203306	$74.56^{+15.98}_{-10.85}$	$56.08^{+12.08}_{-17.93}$	$-0.08^{+0.23}_{-0.31}$	$0.44^{+0.41}_{-0.33}$	$0.58^{+0.55}_{-0.42}$	11.7	0.95	0.08	0.95	0.007
GW190706_222641	$117.53^{+22.57}_{-18.55}$	$63.74^{+27.20}_{-27.42}$	$0.28^{+0.25}_{-0.31}$	$0.47^{+0.38}_{-0.33}$	$1.86^{+0.96}_{-1.50}$	12.5	3.06	0.24	0.26	0.093
GW190707_093326	$14.07^{+3.01}_{-2.33}$	$9.28^{+1.69}_{-1.48}$	$-0.04^{+0.10}_{-0.09}$	$0.28^{+0.39}_{-0.22}$	$2.12^{+0.81}_{-1.89}$	13.2	1.23	0.09	0.93	0.219
GW190708_232457	$23.41^{+4.97}_{-5.07}$	$13.70^{+3.57}_{-2.17}$	$0.05^{+0.10}_{-0.10}$	$0.26^{+0.44}_{-0.20}$	$1.39^{+1.54}_{-1.18}$	13.1	1.14	0.09	0.94	0.424
GW190719_215514	$58.83^{+75.12}_{-16.34}$	$32.78^{+14.38}_{-16.38}$	$0.25^{+0.33}_{-0.32}$	$0.45^{+0.39}_{-0.32}$	$1.61^{+1.26}_{-1.33}$	8	1.41	0.18	0.79	N/A
GW190720_000836	$16.57^{+6.36}_{-3.94}$	$8.76^{+2.44}_{-2.10}$	$0.19^{+0.14}_{-0.11}$	$0.29^{+0.39}_{-0.20}$	$2.59^{+0.41}_{-1.99}$	11.5	0.86	0.08	0.97	0.14
GW190727_060333	$58.87^{+13.07}_{-8.01}$	$46.18^{+8.22}_{-13.24}$	$0.09^{+0.25}_{-0.27}$	$0.50^{+0.38}_{-0.37}$	$1.54^{+1.35}_{-1.27}$	12.1	1.37	0.11	0.82	0.017
GW190728_064510	$14.56^{+8.19}_{-2.60}$	$9.39^{+1.92}_{-2.91}$	$0.13^{+0.19}_{-0.07}$	$0.29^{+0.39}_{-0.20}$	$1.11^{+1.77}_{-0.90}$	13.4	1.16	0.09	0.89	0.048
GW190731_140936	$64.91^{+15.63}_{-10.80}$	$46.27^{+12.67}_{-17.54}$	$0.07^{+0.28}_{-0.25}$	$0.41^{+0.43}_{-0.32}$	$1.23^{+1.63}_{-0.98}$	8.5	1.13	0.13	0.89	0.031
GW190803_022701	$57.71^{+13.23}_{-8.92}$	$42.96^{+9.20}_{-13.81}$	$-0.01^{+0.23}_{-0.28}$	$0.44^{+0.42}_{-0.34}$	$0.91^{+1.93}_{-0.70}$	9.1	0.87	0.1	0.96	N/A
GW190814_211039	$24.48^{+1.55}_{-1.43}$	$2.72^{+0.11}_{-0.11}$	$0.00^{+0.07}_{-0.07}$	$0.04^{+0.04}_{-0.03}$	$0.90^{+1.40}_{-0.24}$	22.2	2.54	0.11	0.49	N/A
GW190828_063405	$43.27^{+7.36}_{-4.65}$	$35.40^{+4.69}_{-6.98}$	$0.15^{+0.15}_{-0.16}$	$0.43^{+0.41}_{-0.32}$	$2.38^{+0.57}_{-2.02}$	16.3	1.31	0.08	0.86	N/A
GW190828_065509	$30.42^{+8.31}_{-8.29}$	$13.44^{+4.82}_{-2.71}$	$0.05^{+0.16}_{-0.17}$	$0.26^{+0.42}_{-0.20}$	$1.86^{+0.96}_{-1.52}$	11.1	1.69	0.15	0.68	N/A
GW190910_112807	$56.54^{+8.69}_{-6.47}$	$44.50^{+7.47}_{-9.24}$	$-0.00^{+0.17}_{-0.20}$	$0.38^{+0.43}_{-0.30}$	$1.62^{+1.16}_{-1.25}$	13.4	2.25	0.17	0.47	0.011
GW190915_235702	$43.05^{+11.11}_{-6.00}$	$32.50^{+5.62}_{-7.95}$	$-0.03^{+0.19}_{-0.24}$	$0.56^{+0.34}_{-0.40}$	$1.84^{+0.99}_{-1.48}$	13	1.63	0.13	0.69	0.01
GW190924_021846	$9.78^{+4.83}_{-2.00}$	$5.67^{+1.35}_{-1.61}$	$0.03^{+0.20}_{-0.08}$	$0.25^{+0.41}_{-0.19}$	$0.84^{+1.95}_{-0.64}$	13	1.13	0.09	0.92	0.152
GW190929_012149	$101.70^{+25.99}_{-18.32}$	$41.64^{+25.15}_{-18.10}$	$-0.03^{+0.23}_{-0.28}$	$0.31^{+0.51}_{-0.25}$	$1.45^{+1.18}_{-0.97}$	10.1	2.22	0.22	0.47	N/A
GW190930_133541	$16.36^{+9.27}_{-4.57}$	$7.98^{+2.78}_{-2.34}$	$0.19^{+0.22}_{-0.16}$	$0.30^{+0.42}_{-0.21}$	$0.72^{+2.06}_{-0.55}$	10.1	1.2	0.12	0.87	0.377

TABLE I. The first half of our BBH analysis results, including 10 GWTC-1 events and 35 GWTC2.1 events that are included in GWTC-2. First seven columns are basic information of the events: event names in YYMMDD_HHMMSS form, component masses $m_{1,2}$ in detector frame (which have a difference of factor $1+z$ from Ref. [3, 4], z is the cosmological redshift), effective spin $\chi_{\rm eff}$, effective precession spin $\chi_{\rm p}$, inclination angle $\theta_{\rm JN}$ and network SNR. The parameters are showed by 50% percentile and 90% confidence error bar, but note $\theta_{\rm JN}$ usually has a bimodal distribution, the 50% percentile might be misleading. Last four columns are statistics we construct: mean value and normalized mean value of $\|\Delta'_{\rm net}\|$ among all the samples (latter one is simply mean value divided by network SNR), fraction of $\|\Delta'_{\rm net}\| < 2$ samples, and the maximum J-S divergence between samples of q, $\mathcal{M}$, $\chi_{\rm eff}$ and $\chi_{\rm p}$. It is labeled as "N/A" if result from one of the waveforms is not available in GWTC-2.1 data release.

Event name	m_1	m_2	$\chi_{\rm eff}$	$\chi_{\rm p}$	$\theta_{\rm JN}$	SNR	Mean	$\text{Mean}_{\rm Norm}$	Fraction	Max J-S Div.
GW190403_051519	$185.98^{+37.48}_{-58.13}$	$44.64^{+61.34}_{-24.08}$	$0.68^{+0.16}_{-0.43}$	$0.32^{+0.38}_{-0.22}$	$1.84^{+1.12}_{-1.66}$	8	1.62	0.2	0.72	0.052
GW190426_190642	$178.48^{+87.83}_{-31.54}$	$132.65^{+32.32}_{-63.94}$	$0.23^{+0.42}_{-0.41}$	$0.51^{+0.37}_{-0.36}$	$2.08^{+0.80}_{-1.72}$	9.6	1.55	0.16	0.74	N/A
GW190725_174728	$14.23^{+12.21}_{-3.62}$	$7.59^{+2.40}_{-3.00}$	$-0.04^{+0.36}_{-0.16}$	$0.37^{+0.46}_{-0.28}$	$1.00^{+1.79}_{-0.74}$	9.1	1.04	0.11	0.92	N/A
GW190805_211137	$87.19^{+24.88}_{-15.28}$	$59.60^{+16.40}_{-23.31}$	$0.37^{+0.29}_{-0.39}$	$0.50^{+0.34}_{-0.32}$	$1.00^{+1.75}_{-0.74}$	8.3	1.3	0.16	0.84	0.014
GW190916_200658	$78.32^{+32.30}_{-21.54}$	$42.37^{+24.07}_{-21.18}$	$0.20^{+0.33}_{-0.31}$	$0.37^{+0.43}_{-0.28}$	$1.61^{+1.27}_{-1.35}$	8.2	1.14	0.14	0.89	0.004
GW190917_114630	$11.15^{+3.72}_{-4.48}$	$2.35^{+1.21}_{-0.48}$	$-0.08^{+0.21}_{-0.43}$	$0.17^{+0.42}_{-0.13}$	$1.35^{+1.60}_{-1.16}$	9.5	1.06	0.11	0.93	N/A
GW190925_232845	$24.69^{+7.70}_{-3.17}$	$18.46^{+2.68}_{-4.16}$	$0.09^{+0.16}_{-0.15}$	$0.39^{+0.43}_{-0.30}$	$0.77^{+2.07}_{-0.58}$	9.9	1.09	0.11	0.91	0.008
GW190926_050336	$63.55^{+31.78}_{-14.11}$	$31.96^{+21.57}_{-14.17}$	$-0.02^{+0.25}_{-0.33}$	$0.37^{+0.48}_{-0.29}$	$1.67^{+1.03}_{-1.19}$	9	1.49	0.17	0.74	N/A
GW191103_012549	$14.03^{+7.42}_{-2.34}$	$9.42^{+1.79}_{-2.85}$	$0.21^{+0.16}_{-0.10}$	$0.40^{+0.41}_{-0.26}$	$1.38^{+1.52}_{-1.14}$	8.9	1.01	0.11	0.92	0.012
GW191105_143521	$13.00^{+4.54}_{-1.78}$	$9.36^{+1.45}_{-2.19}$	$-0.02^{+0.13}_{-0.09}$	$0.30^{+0.45}_{-0.24}$	$1.07^{+1.82}_{-0.85}$	9.7	0.77	0.08	0.99	0.02
GW191109_010717	$81.16^{+12.89}_{-8.89}$	$59.72^{+15.58}_{-17.43}$	$-0.29^{+0.42}_{-0.31}$	$0.63^{+0.29}_{-0.37}$	$1.91^{+0.87}_{-1.18}$	17.3	5.8	0.34	0.1	0.086
GW191113_071753	$36.10^{+14.71}_{-16.18}$	$7.31^{+6.49}_{-1.57}$	$0.00^{+0.37}_{-0.29}$	$0.20^{+0.54}_{-0.16}$	$1.70^{+1.08}_{-1.32}$	7.9	1.68	0.21	0.67	0.048
GW191126_115259	$15.71^{+7.24}_{-2.51}$	$10.75^{+1.94}_{-2.98}$	$0.21^{+0.15}_{-0.11}$	$0.39^{+0.40}_{-0.26}$	$1.71^{+1.20}_{-1.48}$	8.3	1.14	0.14	0.89	0.01
GW191127_050227	$86.41^{+60.12}_{-37.31}$	$38.45^{+31.09}_{-25.26}$	$0.18^{+0.34}_{-0.36}$	$0.52^{+0.41}_{-0.41}$	$1.46^{+1.40}_{-1.16}$	9.2	1.9	0.21	0.6	0.089
GW191129_134029	$12.29^{+4.87}_{-2.26}$	$7.80^{+1.67}_{-1.94}$	$0.06^{+0.16}_{-0.08}$	$0.26^{+0.36}_{-0.19}$	$1.73^{+1.16}_{-1.46}$	13.1	1.34	0.1	0.87	0.033
GW191204_110529	$36.20^{+15.49}_{-5.66}$	$26.21^{+5.17}_{-7.54}$	$0.05^{+0.26}_{-0.27}$	$0.52^{+0.38}_{-0.39}$	$1.57^{+1.24}_{-1.24}$	8.8	1.59	0.18	0.72	0.027
GW191204_171526	$13.44^{+3.77}_{-1.98}$	$9.29^{+1.54}_{-1.84}$	$0.16^{+0.08}_{-0.05}$	$0.39^{+0.35}_{-0.26}$	$2.26^{+0.66}_{-2.00}$	17.5	1.67	0.1	0.73	0.045
GW191215_223052	$33.48^{+9.29}_{-4.68}$	$24.46^{+4.07}_{-5.17}$	$-0.04^{+0.17}_{-0.21}$	$0.50^{+0.37}_{-0.38}$	$1.20^{+1.50}_{-0.85}$	11.2	1.06	0.09	0.95	0.01
GW191216_213338	$12.95^{+4.92}_{-2.38}$	$8.23^{+1.73}_{-1.99}$	$0.11^{+0.13}_{-0.06}$	$0.23^{+0.35}_{-0.16}$	$2.50^{+0.44}_{-0.81}$	18.6	1.93	0.1	0.62	0.06
GW191219_163120	$34.73^{+2.27}_{-2.68}$	$1.30^{+0.08}_{-0.05}$	$-0.00^{+0.07}_{-0.09}$	$0.09^{+0.07}_{-0.07}$	$1.76^{+1.13}_{-1.49}$	9.1	2.51	0.28	0.34	0.13
GW191222_033537	$67.15^{+14.73}_{-9.54}$	$52.41^{+10.67}_{-15.40}$	$-0.04^{+0.20}_{-0.25}$	$0.41^{+0.41}_{-0.32}$	$1.62^{+1.24}_{-1.33}$	12.5	2.02	0.16	0.57	0.017
GW191230_180458	$82.70^{+19.46}_{-13.10}$	$62.85^{+13.88}_{-21.42}$	$-0.05^{+0.26}_{-0.31}$	$0.52^{+0.38}_{-0.39}$	$2.03^{+0.85}_{-1.67}$	10.4	1.17	0.11	0.88	0.011
GW200105_162426	$9.57^{+1.85}_{-1.82}$	$2.02^{+0.35}_{-0.25}$	$0.00^{+0.13}_{-0.18}$	$0.09^{+0.17}_{-0.07}$	$1.54^{+1.28}_{-1.22}$	13.7	1.47	0.11	0.79	0.162
GW200112_155838	$44.01^{+8.22}_{-5.16}$	$35.18^{+5.11}_{-7.42}$	$0.06^{+0.15}_{-0.15}$	$0.39^{+0.39}_{-0.30}$	$0.88^{+2.04}_{-0.68}$	19.8	1.63	0.08	0.73	0.061
GW200115_042309	$6.30^{+2.15}_{-2.69}$	$1.53^{+0.91}_{-0.30}$	$-0.15^{+0.24}_{-0.42}$	$0.20^{+0.34}_{-0.16}$	$0.62^{+1.94}_{-0.43}$	11.3	1.35	0.12	0.81	0.096
GW200128_022011	$65.05^{+15.99}_{-9.41}$	$51.16^{+10.32}_{-13.35}$	$0.12^{+0.24}_{-0.25}$	$0.57^{+0.34}_{-0.40}$	$1.38^{+1.46}_{-1.06}$	10.6	2.12	0.2	0.55	0.09
GW200129_065458	$40.25^{+12.23}_{-3.33}$	$34.06^{+3.32}_{-10.82}$	$0.11^{+0.11}_{-0.16}$	$0.52^{+0.42}_{-0.37}$	$0.66^{+0.59}_{-0.41}$	26.8	2.33	0.09	0.47	0.425
GW200202_154313	$11.02^{+3.83}_{-1.51}$	$7.99^{+1.21}_{-1.85}$	$0.04^{+0.13}_{-0.06}$	$0.28^{+0.40}_{-0.22}$	$2.57^{+0.42}_{-0.59}$	10.8	0.74	0.07	0.99	0.025
GW200208_130117	$52.95^{+12.23}_{-8.40}$	$38.51^{+8.64}_{-11.18}$	$-0.07^{+0.22}_{-0.27}$	$0.38^{+0.41}_{-0.29}$	$2.53^{+0.44}_{-0.58}$	10.8	1.19	0.11	0.87	0.017
GW200208_222617	$83.38^{+171.77}_{-48.68}$	$21.91^{+13.04}_{-11.81}$	$0.45^{+0.43}_{-0.44}$	$0.41^{+0.37}_{-0.30}$	$1.54^{+1.30}_{-1.22}$	7.4	1.91	0.26	0.6	0.171
GW200209_085452	$55.70^{+14.96}_{-9.88}$	$42.95^{+11.00}_{-13.60}$	$-0.12^{+0.24}_{-0.30}$	$0.51^{+0.39}_{-0.37}$	$1.74^{+1.12}_{-1.44}$	9.6	1.29	0.13	0.83	0.014
GW200210_092254	$28.68^{+8.45}_{-5.24}$	$3.38^{+0.52}_{-0.52}$	$0.02^{+0.22}_{-0.21}$	$0.15^{+0.22}_{-0.12}$	$2.31^{+0.60}_{-1.97}$	8.4	1.2	0.14	0.88	0.06
GW200216_220804	$84.36^{+28.39}_{-21.07}$	$50.74^{+22.50}_{-30.85}$	$0.10^{+0.34}_{-0.36}$	$0.45^{+0.42}_{-0.35}$	$0.89^{+1.87}_{-0.69}$	8.1	1.32	0.16	0.82	0.006
GW200219_094415	$58.57^{+13.46}_{-8.95}$	$44.41^{+9.26}_{-14.14}$	$-0.08^{+0.23}_{-0.29}$	$0.48^{+0.40}_{-0.35}$	$1.19^{+1.59}_{-0.92}$	10.7	1.61	0.15	0.71	0.017
GW200220_061928	$165.56^{+62.24}_{-27.93}$	$120.51^{+29.47}_{-55.27}$	$0.06^{+0.40}_{-0.38}$	$0.50^{+0.37}_{-0.37}$	$1.24^{+1.55}_{-0.96}$	7.2	1.2	0.17	0.87	0.022
GW200220_124850	$64.13^{+16.51}_{-10.34}$	$46.89^{+11.54}_{-16.61}$	$-0.07^{+0.27}_{-0.33}$	$0.49^{+0.39}_{-0.37}$	$1.66^{+1.17}_{-1.34}$	8.5	1.72	0.2	0.68	0.007
GW200224_222234	$52.30^{+9.08}_{-5.37}$	$42.82^{+5.83}_{-9.89}$	$0.10^{+0.15}_{-0.15}$	$0.49^{+0.37}_{-0.36}$	$0.62^{+0.55}_{-0.45}$	20	2.26	0.11	0.52	0.022
GW200225_060421	$23.58^{+5.65}_{-3.25}$	$17.20^{+3.05}_{-4.55}$	$-0.12^{+0.17}_{-0.28}$	$0.53^{+0.34}_{-0.38}$	$1.31^{+1.47}_{-1.00}$	12.5	1.43	0.11	0.8	0.015
GW200302_015811	$48.45^{+10.33}_{-9.58}$	$25.50^{+12.02}_{-7.63}$	$0.01^{+0.25}_{-0.26}$	$0.37^{+0.45}_{-0.28}$	$1.34^{+1.43}_{-1.01}$	10.8	2.13	0.2	0.51	0.054
GW200306_093714	$39.52^{+21.47}_{-10.80}$	$20.98^{+8.86}_{-9.91}$	$0.32^{+0.28}_{-0.46}$	$0.43^{+0.39}_{-0.31}$	$1.12^{+1.74}_{-0.87}$	7.8	1.33	0.17	0.82	0.013
GW200308_173609	$143.98^{+299.28}_{-85.60}$	$52.82^{+83.32}_{-33.18}$	$0.16^{+0.58}_{-0.49}$	$0.41^{+0.42}_{-0.30}$	$1.55^{+1.19}_{-1.27}$	7.1	1.1	0.15	0.89	0.054
GW200311_115853	$41.83^{+8.21}_{-4.59}$	$34.00^{+4.68}_{-7.26}$	$-0.02^{+0.16}_{-0.20}$	$0.45^{+0.40}_{-0.35}$	$0.55^{+0.52}_{-0.40}$	17.8	1.27	0.07	0.87	0.026
GW200316_215756	$15.98^{+12.23}_{-3.33}$	$9.52^{+2.32}_{-3.49}$	$0.13^{+0.27}_{-0.10}$	$0.29^{+0.38}_{-0.20}$	$2.32^{+0.58}_{-1.85}$	10.3	1.08	0.1	0.9	0.082
GW200322_091133	$105.87^{+392.13}_{-84.00}$	$26.23^{+58.87}_{-18.59}$	$0.08^{+0.51}_{-0.47}$	$0.50^{+0.36}_{-0.41}$	$1.66^{+1.04}_{-1.12}$	6	0.87	0.14	0.94	0.059

TABLE II. The second half of our BBH analysis results, including 8 new events in GWTC-2.1 (compared to GWTC-2) and 36 events in GWTC-3. Columns have the same meaning as Tab. I.

black hole simulations, Classical and Quantum Gravity **36**, 195006 (2019).
[26] B. Szilágyi, J. Blackman, A. Buonanno, A. Taracchini, H. P. Pfeiffer, M. A. Scheel, T. Chu, L. E. Kidder, and Y. Pan, Approaching the post-newtonian regime with numerical relativity: A compact-object binary simulation spanning 350 gravitational-wave cycles, Physical Review Letters **115**, 10.1103/physrevlett.115.031102 (2015).
[27] P. Ajith *et al.*, A phenomenological template family for black-hole coalescence waveforms, Classical and Quantum Gravity **24**, S689 (2007).
[28] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez Forteza, and A. Bohé, Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era, Phys. Rev. D **93**, 044007 (2016), arXiv:1508.07253 [gr-qc].
[29] G. Pratten, S. Husa, C. Garcia-Quiros, M. Colleoni, A. Ramos-Buades, H. Estelles, and R. Jaume, Setting the cornerstone for a family of models for gravitational waves from compact binaries: The dominant harmonic for nonprecessing quasicircular black holes, Phys. Rev. D **102**, 064001 (2020), arXiv:2001.11412 [gr-qc].
[30] C. García-Quirós, M. Colleoni, S. Husa, H. Estellés, G. Pratten, A. Ramos-Buades, M. Mateu-Lucena, and R. Jaume, Multimode frequency-domain model for the gravitational wave signal from nonprecessing black-hole binaries, Phys. Rev. D **102**, 064002 (2020), arXiv:2001.10914 [gr-qc].
[31] G. Pratten *et al.*, Computationally efficient models for the dominant and subdominant harmonic modes of precessing binary black holes, Physical Review D **103**, 104056 (2021).
[32] A. Buonanno and T. Damour, Effective one-body approach to general relativistic two-body dynamics, Physical Review D **59**, 084006 (1999).
[33] T. Damour, Coalescence of two spinning black holes: an effective one-body approach, Phys. Rev. D **64**, 124013 (2001), arXiv:gr-qc/0103018.
[34] A. Bohé *et al.*, Improved effective-one-body model of spinning, nonprecessing binary black holes for the era of gravitational-wave astrophysics with advanced detectors, Phys. Rev. D **95**, 044028 (2017), arXiv:1611.03703 [gr-qc].
[35] R. Cotesta, A. Buonanno, A. Bohé, A. Taracchini, I. Hinder, and S. Ossokine, Enriching the Symphony of Gravitational Waves from Binary Black Holes by Tuning Higher Harmonics, Phys. Rev. D **98**, 084028 (2018), arXiv:1803.10701 [gr-qc].
[36] Y. Pan, A. Buonanno, A. Taracchini, L. E. Kidder, A. H. Mroué, H. P. Pfeiffer, M. A. Scheel, and B. Szilágyi, Inspiral-merger-ringdown waveforms of spinning, precessing black-hole binaries in the effective-one-body formalism, Phys. Rev. D **89**, 084006 (2014), arXiv:1307.6232 [gr-qc].
[37] S. Ossokine *et al.*, Multipolar effective-one-body waveforms for precessing binary black holes: Construction and validation, Physical Review D **102**, 044055 (2020).
[38] A. Nagar *et al.*, Time-domain effective-one-body gravitational waveforms for coalescing compact binaries with nonprecessing spins, tides and self-spin effects, Phys. Rev. D **98**, 104052 (2018), arXiv:1806.01772 [gr-qc].
[39] A. Nagar, G. Riemenschneider, G. Pratten, P. Rettegno, and F. Messina, Multipolar effective one body waveform model for spin-aligned black hole binaries, Phys. Rev. D **102**, 024077 (2020), arXiv:2001.09082 [gr-qc].
[40] A. Nagar, G. Pratten, G. Riemenschneider, and R. Gamba, Multipolar effective one body model for nonspinning black hole binaries, Phys. Rev. D **101**, 024041 (2020), arXiv:1904.09550 [gr-qc].
[41] J. Blackman, S. E. Field, C. R. Galley, B. Szilágyi, M. A. Scheel, M. Tiglio, and D. A. Hemberger, Fast and Accurate Prediction of Numerical Relativity Waveforms from Binary Black Hole Coalescences Using Surrogate Models, Physical Review Letters **115**, 121102 (2015).
[42] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, D. Gerosa, L. C. Stein, L. E. Kidder, and H. P. Pfeiffer, Surrogate models for precessing binary black hole simulations with unequal masses, Phys. Rev. Research. **1**, 033015 (2019), arXiv:1905.09300 [gr-qc].
[43] D. Williams, I. S. Heng, J. Gair, J. A. Clark, and B. Khamesra, Precessing numerical relativity waveform surrogate model for binary black holes: A Gaussian process regression approach, Physical Review D **101**, 10.1103/PhysRevD.101.063011 (2020).
[44] S. E. Field, C. R. Galley, J. S. Hesthaven, J. Kaye, and M. Tiglio, Fast prediction and evaluation of gravitational waveforms using surrogate models, Phys. Rev. X **4**, 031006 (2014), arXiv:1308.3565 [gr-qc].
[45] M. Pürrer, Frequency domain reduced order models for gravitational waves from aligned-spin compact binaries, Class. Quant. Grav. **31**, 195010 (2014), arXiv:1402.4146 [gr-qc].
[46] R. Abbott *et al.*, GW190521: A Binary Black Hole Merger with a Total Mass of 150 M, Physical Review Letters **125**, 1–17 (2020).
[47] R. Abbott *et al.*, GW190412: Observation of a binary-black-hole coalescence with asymmetric masses, Physical Review D **102**, 043015 (2020).
[48] M. Pürrer and C. J. Haster, Gravitational waveform accuracy requirements for future ground-based detectors, Physical Review Research **2**, 1–30 (2020).
[49] P. Kumar, T. Chu, H. Fong, H. P. Pfeiffer, M. Boyle, D. A. Hemberger, L. E. Kidder, M. A. Scheel, and B. Szilagyi, Accuracy of binary black hole waveform models for aligned-spin binaries, Physical Review D **93**, 1–25 (2016).
[50] R. Abbott *et al.* (LIGO Scientific Collaboration, VIRGO Collaboration, KAGRA Collaboration), The population of merging compact binaries inferred using gravitational waves through GWTC-3 (2021), arXiv:2111.03634 [astro-ph.HE].
[51] M. Hannam, C. Hoy, J. E. Thompson, S. Fairhurst, and V. Raymond (VIRGO), Measurement of general-relativistic precession in a black-hole binary (2021), arXiv:2112.11300 [gr-qc].
[52] A. Z. Jan, A. B. Yelikar, J. Lange, and R. O'Shaughnessy, Assessing and marginalizing over compact binary coalescence waveform systematics with RIFT, Phys. Rev. D **102**, 124069 (2020), arXiv:2011.03571 [gr-qc].
[53] A. R. Williamson, J. Lange, R. O'Shaughnessy, J. A. Clark, P. Kumar, J. Calderón Bustillo, and J. Veitch, Systematic challenges for future gravitational wave measurements of precessing binary black holes, Phys. Rev. D **96**, 124041 (2017), arXiv:1709.03095 [gr-qc].
[54] D. Ferguson, K. Jani, P. Laguna, and D. Shoemaker, Assessing the readiness of numerical relativity for LISA and 3G detectors, Physical Review D **104**, 044037 (2021).
[55] N. Kunert, P. T. H. Pang, I. Tews, M. W. Coughlin, and

T. Dietrich, Quantifying modeling uncertainties when combining multiple gravitational-wave detections from binary neutron star sources, Phys. Rev. D **105**, L061301 (2022), arXiv:2110.11835 [astro-ph.HE].
[56] R. Gamba, M. Breschi, S. Bernuzzi, M. Agathos, and A. Nagar, Waveform systematics in the gravitational-wave inference of tidal parameters and equation of state from binary neutron-star signals, Phys. Rev. D **103**, 124015 (2021).
[57] L. S. Finn, Detection, measurement, and gravitational radiation, Physical Review D **46**, 5236 (1992).
[58] C. Cutler and I. E. Flanagan, Gravitational waves from merging compact binaries: How accurately can one extract the binary's parameters from the inspiral waveform?, Physical Review D **49**, 2658 (1994).
[59] T. Damour, A. Nagar, and M. Trias, Accuracy and effectualness of closed-form, frequency-domain waveforms for non-spinning black hole binaries, Phys. Rev. D **83**, 024006 (2011), arXiv:1009.5998 [gr-qc].
[60] P. Jaranowski, A. Królak, and B. F. Schutz, Data analysis of gravitational-wave signals from spinning neutron stars: The signal and its detection, Physical Review D **58**, 063001 (1998).
[61] L. Santamaria *et al.*, Matching post-Newtonian and numerical relativity waveforms: systematic errors and a new phenomenological model for non-precessing black hole binaries, Phys. Rev. D **82**, 064016 (2010), arXiv:1005.3306 [gr-qc].
[62] S. T. McWilliams, B. J. Kelly, and J. G. Baker, Observing mergers of non-spinning black-hole binaries, Phys. Rev. D **82**, 024014 (2010), arXiv:1004.0961 [gr-qc].
[63] L. Sun *et al.*, Characterization of systematic error in Advanced LIGO calibration, Class. Quant. Grav. **37**, 225008 (2020), arXiv:2005.02531 [astro-ph.IM].
[64] R. Abbott *et al.* (LIGO Scientific Collaboration, KAGRA Collaboration, Virgo Collaboration), Observation of Gravitational Waves from Two Neutron Star–Black Hole Coalescences, Astrophys. J. Lett. **915**, L5 (2021), arXiv:2106.15163 [astro-ph.HE].
[65] C. Cutler and M. Vallisneri, LISA detections of massive black hole inspirals: Parameter extraction errors due to inaccurate template waveforms, Physical Review D **76**, 104018 (2007), arXiv:0707.2982.
[66] B. P. Abbott *et al.*, Properties of the Binary Neutron Star Merger GW170817, Physical Review X **9**, 011001 (2019).
[67] S. Biscoveanu, M. Isi, V. Varma, and S. Vitale, Measuring the spins of heavy binary black holes, Phys. Rev. D **104**, 103018 (2021), arXiv:2106.06492 [gr-qc].
[68] V. Varma and P. Ajith, Effects of nonquadrupole modes in the detection and parameter estimation of black hole binaries with nonprecessing spins, Phys. Rev. D **96**, 124024 (2017), arXiv:1612.05608 [gr-qc].
[69] M. Colleoni, M. Mateu-Lucena, H. Estellés, C. García-Quirós, D. Keitel, G. Pratten, A. Ramos-Buades, and S. Husa, Towards the routine use of subdominant harmonics in gravitational-wave inference: Reanalysis of GW190412 with generation X waveform models, Phys. Rev. D **103**, 024029 (2021), arXiv:2010.05830 [gr-qc].
[70] S. Vitale, R. Lynch, J. Veitch, V. Raymond, and R. Sturani, Measuring the spin of black holes in binary systems using gravitational waves, Phys. Rev. Lett. **112**, 251101 (2014), arXiv:1403.0129 [gr-qc].
[71] S. Fairhurst, R. Green, C. Hoy, M. Hannam, and A. Muir, Two-harmonic approximation for gravitational waveforms from precessing binaries, Phys. Rev. D **102**, 024055 (2020), arXiv:1908.05707 [gr-qc].
[72] N. V. Krishnendu and F. Ohme, Interplay of spin-precession and higher harmonics in the parameter estimation of binary black holes, Phys. Rev. D **105**, 064012 (2022), arXiv:2110.00766 [gr-qc].
[73] B. P. Abbott and et al, Prospects for observing and localizing gravitational-wave transients with Advanced LIGO, Advanced Virgo and KAGRA, Living Reviews in Relativity **23**, 3 (2020).
[74] F. Pannarale, E. Berti, K. Kyutoku, B. D. Lackey, and M. Shibata, Aligned spin neutron star-black hole mergers: a gravitational waveform amplitude model, Phys. Rev. D **92**, 084050 (2015), arXiv:1509.00512 [gr-qc].
[75] R. Gamba, S. Akçay, S. Bernuzzi, and J. Williams, Effective-one-body waveforms for precessing coalescing compact binaries with post-Newtonian Twist (2021), arXiv:2111.03675 [gr-qc].
[76] D. Chiaramello and A. Nagar, Faithful analytical effective-one-body waveform model for spin-aligned, moderately eccentric, coalescing black hole binaries, Phys. Rev. D **101**, 101501 (2020), arXiv:2001.11736 [gr-qc].
[77] T. Dietrich, A. Samajdar, S. Khan, N. K. Johnson-McDaniel, R. Dudi, and W. Tichy, Improving the NRTidal model for binary neutron star systems, Phys. Rev. D **100**, 044003 (2019), arXiv:1905.06011 [gr-qc].