

Use and interpretation of signal-model indistinguishability measures for gravitational-wave astronomy

Jonathan E. Thompson^{1,2}, Charlie Hoy³, Edward Fauchon-Jones⁴, and Mark Hannam⁴

¹*Mathematical Sciences and STAG Research Centre,*

University of Southampton, Southampton, SO17 1BJ, United Kingdom

²*Theoretical Astrophysics Group, California Institute of Technology, Pasadena, California 91125, USA*

³*Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth, United Kingdom*

⁴*School of Physics and Astronomy, Cardiff University, Cardiff, CF24 3AA, United Kingdom*



(Received 15 June 2025; accepted 15 August 2025; published 5 September 2025)

The difference (“mismatch”) between two gravitational-wave signals is often used to estimate the signal-to-noise ratio (SNR) at which they will be distinguishable in a measurement or, alternatively, when the errors in a signal model will lead to biased measurements. It is well known that the standard approach to calculate this “indistinguishability SNR” is too conservative: a model may fail the criterion at a given SNR, but not necessarily incur a biased measurement of any individual parameters. This problem can be solved by taking into account errors orthogonal to the model space (which, therefore, do not induce a bias), and calculating indistinguishability SNRs for individual parameters, rather than the full N -dimensional parameter space. We illustrate this approach with the simple example of aligned-spin binary black hole signals, and calculate accurate estimates of the SNR at which each parameter measurement will be biased. In general, biases occur at much higher SNRs than predicted from the standard mismatch calculation. Which parameters are most easily biased depends sensitively on the details of a given waveform model, and the location in parameter space, and in some cases the bias SNR is as high as the conservative estimate. We also illustrate how the parameter bias SNR can be used to robustly specify waveform accuracy requirements for future detectors.

DOI: [10.1103/ddd7-x9zz](https://doi.org/10.1103/ddd7-x9zz)

I. INTRODUCTION

Due to improvements in detector sensitivity, the observation potential of gravitational-wave (GW) astronomy has grown rapidly since the first direct GW detection in 2015 [1–5], and by current forecasts it will continue to do so over the next two decades [6–13]. Detector networks in 2015 could be expected to observe $O(10)$ black hole binaries per year, while current networks should be sensitive to $O(100)$ binaries per year [6]. At the projected sensitivity of next-generation ground-based detectors, we will observe many thousands of binaries per year, and will be sensitive to all black hole mergers in the universe. The additional upcoming space-based GW detector LISA [14,15] will be sensitive to massive black hole signals of similar morphology to those seen in ground-based detectors (but at total masses $> 10^5 M_\odot$).

With increased sensitivity we also observe louder signals, which will allow more accurate measurements; for a given source, measurement accuracy scales roughly linearly with detector sensitivity. However, to realize higher measurement accuracies we also require sufficiently accurate theoretical signal models against which to compare the detector data. The accuracy requirements of our models, in order to make unbiased measurements, become more stringent with increased signal strength.

As such, the question of quantifying model accuracy, and determining future accuracy requirements, is an important one. In the near term we wish to know under what circumstances we can trust the source inference from current models, and when we must beware of systematic biases. In the longer term, as part of the extensive research and development effort to prepare for LISA [14,15], Einstein Telescope [9–12], and Cosmic Explorer [6–8], we also wish to know how accurate our models must be to achieve the observatories’ science goals. These questions are made more urgent by the large resources and many-year timescale required to produce accurate models: large numbers of computationally expensive numerical relativity (NR) simulations, and sophisticated procedures to calibrate

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

semianalytic phenomenological models, or to train surrogate models, e.g., Refs. [16–24].

To date it has been difficult to provide useful waveform accuracy measures. For example, in NR simulations one typically quantifies the signal’s phase accuracy from the beginning of the simulation. In GW data analysis, on the other hand, it is more usual to consider an inner product between waveforms [25], which involves an optimization with respect to an overall phase shift and confuses the nominal phase uncertainty of the waveform. Furthermore, when estimating binary source parameters, we identify the parameters at which our model agrees best with the data—a single NR waveform corresponds to a binary with a single set of parameters, so we also need a way to convert its uncertainty into measurement biases.

A series of previous works have noted that we can define a waveform uncertainty measure based on the inner product above (the mismatch). The mismatch between two waveforms, for example between a true signal waveform and a waveform model, can in turn be related to the signal-to-noise ratio (SNR) at which the two waveforms will be distinguishable in a measurement [26–31]. Unfortunately, in practice this simple mismatch requirement is typically found to be extremely conservative, and so of little use in making realistic estimates of required model accuracy [32,33].

One pragmatic solution to this problem would be to consider our most accurate NR waveforms as proxies for true signals, and to study how well our current models recover true source properties for selected detector configurations and a range of signal strengths, i.e., a range of SNRs. In doing this we could in principle identify the SNR at which a model will lead to a biased measurement in each parameter of interest, and determine how much more accurate the models must be for future high-SNR observations. A first attempt at such an approach was made in Pürrer and Haster’s 2019 study, Ref. [33]. The authors considered two signals and found parameter biases at vastly different SNRs, depending on both the parameter and the signal: some parameters are biased at an SNR of ~ 50 , while others are not biased even at SNRs of ~ 2500 , and some of the parameters that are most susceptible to bias in one signal are not biased at all for the other. In order to draw some general conclusions the authors estimate an approximate “balance SNR” for each signal (SNRs of ~ 50 for both signals) and use this to conclude that the mismatch uncertainties of models for next-generation ground-based detectors must improve on those of ca. 2019 models by 3 orders of magnitude, and the mismatch uncertainty of NR simulations must improve by 1 order of magnitude.

In this work we revisit the indistinguishability mismatch criterion, and show that, if calculated appropriately, it is not conservative, but in fact an accurate measure of the SNR at which a measurement will be biased. The criterion is too conservative in its standard form, partly because it does not account for waveform errors that do not contribute to

measurement bias (i.e., are orthogonal to the signal manifold), but mostly because the criterion applies to an N -dimensional credible interval. The criterion can, however, be calculated in such a way as to accurately predict bias SNRs for individual parameters. In the present work we illustrate each of these features with respect to a simple model for quadrupole-only radiation and spins aligned with the orbital angular momentum (referred to as an “aligned-spin” binary). We will consider state-of-the-art generic binary models in future work. We stress that although we are not aware of this method being applied to binary black hole waveform models with current and future ground-based detectors, the method itself is not new—it is discussed, either implicitly or explicitly, in works from Ref. [26] through to Ref. [31], and the method we use to calculate the 1D parameter bias SNRs is equivalent to that discussed in Ref. [31].

We also note that a more appropriate measure of model uncertainty is not the mismatch but the square root of the mismatch. It is this quantity that scales directly with both the signal SNR and with standard accuracy measures in NR simulations and, under reasonable assumptions, NR computational cost. This seemingly minor change has important implications for future accuracy requirements: a 2 orders of magnitude improvement in mismatch can be achieved with only 1 order of magnitude improvement in NR accuracy, and similarly only 1 order of magnitude increase in computational cost.

Although we consider only a simple proof-of-principle model in this work, we are able to make some broad estimates for the required improvements in model accuracy and NR simulations for next-generation observatories. These sharpen the early estimates from Ref. [33]. In particular, we note that the bias SNRs depend on an individual model’s construction, and in principle a “conservative” estimate of the bias SNR can sometimes be correct. This leads to a far more stringent accuracy requirement than in Ref. [33]: model mismatch uncertainties must be below 10^{-6} to be free of bias in observations with SNRs of ~ 1000 , an improvement of 4 orders of magnitude over some current models. On the other hand, if model construction can be optimized to maximize individual bias SNRs, we may require only modest improvements over the most accurate current models, e.g., NRSUR7DQ4 [21]. This large uncertainty in the required level of improvement highlights the scale of the general problem of accuracy requirement estimates, and we hope that future applications of the method we discuss here to current generic models will provide more refined, and more useful, estimates.

In this work we consider the problem of quantifying the accuracy of waveform models for source measurements, such that those measurements will not be contaminated by systematics. What we *do not* consider is the related (and likely more difficult) problem of identifying when

systematics are present in a measurement. See, for example, Refs. [34,35] for discussions of strategies to identify waveform systematics in observations.

The outline of this paper is as follows. In Sec. II we review the approach to estimating indistinguishability SNR, describe why it is overly conservative, and detail how one may improve upon it. In Sec. III we outline the details of the signal waveforms we use and our parameter estimation injections. Section IV discusses the results for indistinguishability SNR estimation for N -dimensional posterior data, and Sec. V extends this analysis to SNR estimates for individual model parameters. We discuss the impact of SNR estimates on next-generation detectors in Sec. VI and provide concluding thoughts in Sec. VII. Throughout this manuscript we work in units of $G = c = 1$.

II. MODEL INDISTINGUISHABILITY AND BIAS

We consider the data time series d collected by a ground-based GW interferometer. Under the hypothesis that a GW signal exists in the data, we write $d = s + n$, where s is the true signal and n is the stationary and Gaussian-distributed noise of the detector. For most of this work we will consider the “zero-noise case,” where we implicitly replace quantities relating to noise with their expectation values under infinite noise realizations (thereby setting $n \rightarrow \langle n \rangle = 0$) [36].

Consider a gravitational-wave model for a compact binary coalescence, $h(\theta)$, parametrized by a set of intrinsic and extrinsic parameters $\theta \in \Theta$. The set Θ contains intrinsic parameters such as the primary and secondary masses, $m_1 \geq m_2$, the primary and secondary (dimensionless) spin vectors, $\chi_{1,2}$, as well as a number of extrinsic parameters. For the examples in this work we focus on the constrained case of compact binaries with spins strictly aligned with the orbital angular momentum, thereby reducing the spin degrees of freedom to 2, denoted without loss of generality simply as χ_{1z} and χ_{2z} . We emphasize, however, that the general approach outlined in this work does not rely on these simplifying approximations to the signal. We drop explicit parameter dependence wherever convenient for ease of reading.

It is useful to define an inner product between two signals $h_1 = h(\theta_1)$ and $h_2 = h(\theta_2)$ as

$$\langle h_1 | h_2 \rangle = 4\text{Re} \int_{f_{\min}}^{f_{\max}} \frac{\tilde{h}_1(f) \tilde{h}_2^*(f)}{\tilde{S}_n(f)} df, \quad (1)$$

where the tilde denotes the Fourier transform, the signals are written 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 detector’s power spectral density. The SNR of a GW signal h is then given by $\rho^2 = |h|^2 = \langle h | h \rangle$.

We refer to the *indistinguishability SNR* as the SNR below which two GW signals will be indistinguishable in a measurement. This SNR has a natural connection with the

ratio of parameter bias to measurement variance (see Appendix A), as that ratio itself scales directly with the SNR of the signal. When discussing waveform model errors, the indistinguishability SNR indicates the SNR above which a given model will lead to biased parameter inference, and we can then determine, for a given GW observation, whether that model can be trusted to provide unbiased measurements.

There has been extensive discussion of variants of the indistinguishability SNR in the literature [26–31]. The standard calculation provides only a conservative estimate relevant to parameter biases. We give examples in Sec. II B of signals louder than a model’s nominal indistinguishability SNR, for which the model recovers all of the source properties with no bias. In these scenarios the conservative estimate is of little value.

In Sec. II B 2 we will explain the two reasons why the standard estimate is conservative, and how to address them. This will lead us to define quantities that accurately estimate bias SNRs, either for N -D sets of parameters or individual parameters—the N -D qualifier will be explained in due course. In the remainder of the paper we will provide concrete examples to illustrate these points, and demonstrate that we can calculate a reliable bias SNR for all measurable parameters. This approach can then be used to inform waveform accuracy requirements in current and future detectors.

A. Mismatches and distance metrics

One measure of waveform model accuracy is the *match*, defined as the noise-weighted inner product optimized over some subset of parameters $\Theta_{\text{opt}} \subset \Theta$ [37], and normalized with respect to the magnitude of each waveform,

$$M(h_1, h_2) = \max_{\Theta_{\text{opt}}} \frac{\langle h_1 | h_2 \rangle}{|h_1| |h_2|}. \quad (2)$$

The match is unity if the two waveforms are the same, up to an overall amplitude rescaling. To quantify the difference between two waveforms we use the *mismatch*,

$$\mathcal{M} = 1 - M(h_1, h_2). \quad (3)$$

The choice of optimization parameters Θ_{opt} is discussed in more detail later in Sec. III B.

We can identify the mismatch with a measure of normalized difference between two waveforms, $\hat{d} = \sqrt{\mathcal{M}}$ [38]. This is motivated by the usual interpretation of an inner product as the square of a distance, and also a consideration of uncertainties in waveforms. For the former, we can rearrange the inner product between the difference of two waveforms to find [27,28]

$$\begin{aligned}
|h_1 - h_2|^2 &= 2|h_1|^2 \left(1 - \frac{\langle h_1 | h_2 \rangle}{|h_1||h_2|} \right), \\
&= 2\rho^2 \mathcal{M}(h_1, h_2), \\
&= 2\rho^2 \hat{d}^2(h_1, h_2),
\end{aligned} \tag{4}$$

assuming that both waveforms have the same SNR, $\rho^2 = |h_1|^2 = |h_2|^2$. We see then that $\hat{d}$ is proportional to the norm of the difference between the two normalized waveforms. In addition, when written in terms of normalized signals $\hat{h} = h/|h|$ under the same assumptions as above and rearranged, Eq. (4) becomes

$$\hat{d}^2(h_1, h_2) = \frac{1}{2} |\hat{h}_1 - \hat{h}_2|^2, \tag{5}$$

which we discuss further in the context of the linear signal approximation in Appendix A.

To connect the normalized distance to error measures, we note that the mismatch between a waveform and some approximation of it can be related, to leading order in the amplitude and phase uncertainties in the approximate waveform ΔA and $\Delta\phi$, as $\mathcal{M} \sim (\Delta A)^2$ and $\mathcal{M} \sim (\Delta\phi)^2$. (See, for example, the discussion in Sec. IV.C.1 of Ref. [18].) Since we ultimately want to relate mismatch calculations to waveform accuracy requirements, we prefer to use the normalized difference $\hat{d}$, as it is proportional to the standard uncertainty measures of the waveform.

B. Indistinguishability SNR

1. Standard definitions

If we assume that the statistical likelihood behaves as a Gaussian in $|h_1 - h_2|$, which is true in the high-SNR limit [37], then two waveforms will be distinguishable at 1 standard deviation when $|h_1 - h_2| > 1$, or $\mathcal{M} > 1/(2\rho^2)$ [26,39]. More generally, if we optimize the mismatch over all but N parameters in a model, then the signals will be distinguishable with probability p if [29]

$$\mathcal{M} > \frac{\chi_N^2(1-p)}{2\rho^2}, \tag{6}$$

where $\chi_N^2(1-p)$ is the chi-square value at probability p for N degrees of freedom. If we are interested in only 1- σ , then the chi-square value is N , recovering a commonly quoted indistinguishability criterion [30],

$$\mathcal{M} > \frac{N}{2\rho^2}. \tag{7}$$

We see from Eq. (6) that the requirement on a waveform model's mismatch uncertainty scales with $1/\rho^2$. The requirement on the normalized waveform difference, $\hat{d}$, therefore, scales as $1/\rho$, and this reflects the intuitive result

that the accuracy requirements on waveforms (e.g., the accuracy of their amplitude and phase) also scale with $1/\rho$ —if we detect signals twice as loud, we require waveform models twice as accurate.

Equation (6) motivates the standard way to estimate the SNR at which a model will yield biased parameter estimates: we calculate the mismatch between a fiducial signal (e.g., a numerical relativity waveform) and a signal model, keeping the true intrinsic binary parameters (the masses and spins) fixed, and optimizing over the extrinsic parameters (distance, orientation, etc). The number of degrees of freedom is then the number of parameters that we have not optimized over, or which we consider physically meaningful to measure. For example, in an aligned-spin system the intrinsic parameters are the total mass M , the mass ratio q , and the two black hole spins χ_{1z} and χ_{2z} . At low SNRs it is not possible to measure both spins, only a mass-weighted sum (commonly $\chi_{\text{eff}} = (m_1\chi_{1z} + m_2\chi_{2z})/M$ [40]), and so we may consider this system to have 3 rather than 4 degrees of freedom. In many cases the appropriate number of degrees of freedom may be unclear. We will resolve this apparent confusion in the next section.

2. Issues and resolutions

As noted above, the standard application of Eq. (6) leads to a conservative estimate of the minimum SNR at which parameter biases appear. There are two reasons for this.

Uninformative perpendicular SNR. The first reason is well known from discussions of waveform accuracy dating back to Refs. [26,34,37,39], and becomes clear when we consider the discussion above in more detail. Consider the illustration in Fig. 1. Denote the signal waveform s and the source parameters θ_s . Denote the model evaluated at these true source parameters by h_s . The distance between s and h_s , $\hat{d}_s$, is the quantity typically used in calculating the indistinguishability SNR. However, this is not the relevant quantity when considering parameter biases: we are not interested in whether h_s can be distinguished from s , but whether h_s can be distinguished from h_{bf} , the model evaluated at parameters that give the best agreement between the model and the signal, θ_{bf} , which are those that will be measured in a parameter estimation exercise. We wish to know whether the true parameters θ_s will lie within some credible interval (CI) around θ_{bf} .

It is common to treat numerical relativity waveforms as true signals, and calculate matches between these and a given model calculated with the same intrinsic parameters. In general, this will overestimate the mismatch and lead to an indistinguishability SNR that is too conservative. We must instead find the model parameters that maximize the agreement with the NR signal, and then calculate the mismatch between model waveforms at these two sets of parameters.

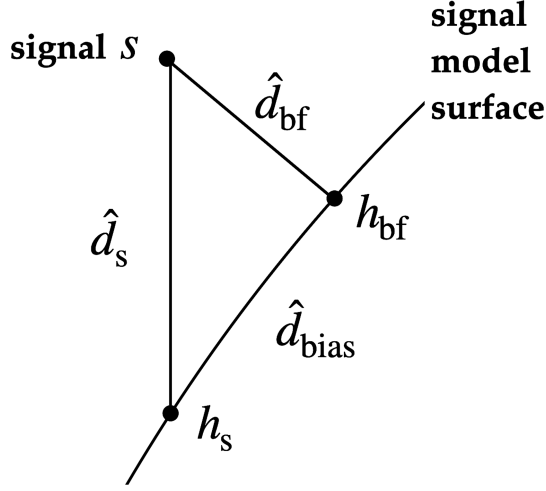


FIG. 1. Illustration of the relationship between the true signal, s , the model signal with the true parameters, h_s , and the model signal with the parameters θ_{bf} that agrees best with the true signal, h_{bf} . We can relate the three waveforms by considering directions parallel and perpendicular to the signal-model manifold, as in Eqs. (10) and (11).

We are, therefore, interested in the distance between the true parameters and the best-fitting model parameters that lie solely within the model manifold, $\hat{d}_{\text{bias}} = \sqrt{\mathcal{M}(h_{bf}, h_s)}$, meaning that we wish to ignore the contribution to the signal waveform that is orthogonal to the manifold of our model. Another way of saying this is that the difference between the true signal waveform and the model at the true parameters is made up of two contributions: one that leads to a bias in the measured parameters, and another that does not introduce a bias, but only reduces the extracted SNR of

the signal [39]. (See also Sec. II.B of Ref. [26].) The appropriate indistinguishability criterion for N degrees of freedom is then

$$\hat{d}_{\text{bias,ND}}^2 = \mathcal{M}(h_{bf}, h_s) > \frac{\chi_N^2(1-p)}{2\rho^2}. \quad (8)$$

It is common to refer to $\hat{d}_s^2 = \mathcal{M}(s, h_s)$ as the *faithfulness* mismatch, because it is a measure of how well the model reproduces the signal when evaluated at the same parameters. The quantity $\hat{d}_{bf}^2 = \mathcal{M}(s, h_{bf})$ is referred to as the *effectualness* mismatch, because it describes how effective the model is at reproducing the signal in total, and is the smallest mismatch that can be achieved in a search or parameter estimation (assuming zero noise).

The result in Eq. (8) has also been recently discussed again in Ref. [31]. Their Eq. (16) takes the place of our Eq. (8) and in our notation would be

$$\mathcal{M}(s, h_s) - \mathcal{M}(s, h_{bf}) > \frac{\chi_N^2(1-p)}{2\rho^2}. \quad (9)$$

This is equivalent to Eq. (8) because, to a good approximation, $\hat{d}_{\text{bias}}^2 = \hat{d}_s^2 - \hat{d}_{bf}^2$, i.e., $\mathcal{M}(h_{bf}, h_s) = \mathcal{M}(s, h_s) - \mathcal{M}(s, h_{bf})$. That the usual Pythagorean relation approximately holds (and we are not just being misled by the notation) can be seen by considering normalized waveforms and writing the signal and model waveforms in terms of the model at the true parameters as

$$\hat{s} = A\hat{h}_{bf} + \sqrt{1-A^2}\hat{h}_\perp, \quad (10)$$

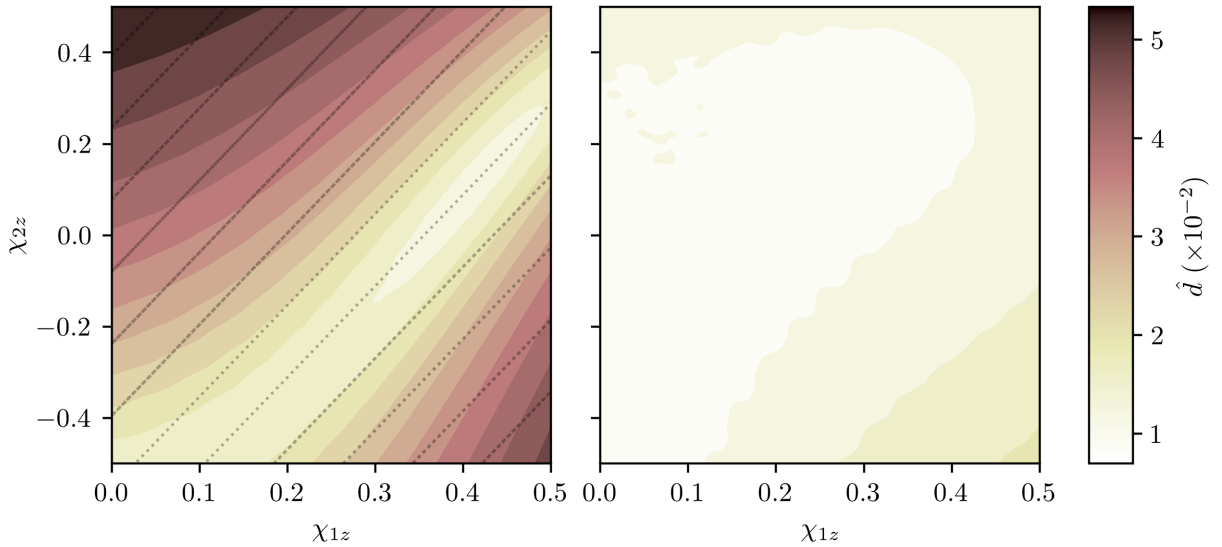


FIG. 2. Contour plots of $\hat{d}$ computed for the models PHENOMD and NRHYBSUR3DQ8, plotted for a range of χ_{1z} and χ_{2z} for fixed values of $(m_1, m_2) = (200, 100)M_\odot$. In the left panel, the faithfulness is used to compute the distance $\hat{d}_s$, while the right panel displays the distance $\hat{d}_{bf}$ arising from the effectualness. The dotted lines show lines of constant χ_{antisym} , indicating that model accuracy varies strongly with changing χ_{eff} .

$$\hat{h}_s = B\hat{h}_{\text{bf}} + \sqrt{1 - B^2}\hat{h}_{\perp}, \quad (11)$$

where $\hat{h}_{\perp}$ and $\hat{h}_{\parallel}$ are both orthogonal to $\hat{h}_{\text{bf}}$ and to each other. We note that $A = 1 - \mathcal{M}(s, h_{\text{bf}})$ and $B = 1 - \mathcal{M}(h_{\text{bf}}, h_s)$, and so $\langle \hat{s} | \hat{h}_s \rangle = AB \approx 1 - \mathcal{M}(s, h_{\text{bf}}) - \mathcal{M}(h_{\text{bf}}, h_s)$ to leading order in the mismatches, which gives us the desired result.

To illustrate the relative importance of $\hat{d}_{\text{bf}}$ to $\hat{d}_s$, we show in Fig. 2 $\hat{d}_s$ in the left panel, compared to $\hat{d}_{\text{bf}}$ in the right panel, between the models PHENOMD and NRHYBSUR3DQ8. These results are plotted for a range of χ_{1z} and χ_{2z} values for fixed masses $(m_1, m_2) = (200, 100)M_{\odot}$. We see that the waveform model error orthogonal to the model surface, $\hat{d}_{\text{bf}}$, varies little over the parameter space, while $\hat{d}_s$ shows a clear trend of variation perpendicular to lines of constant $\chi_{\text{antisym}} = (m_1\chi_{1z} - m_2\chi_{2z})/M$, displayed as dotted lines in the left panel. Seeing these results, one would hypothesize that a parameter estimation study injecting any one of these NRHYBSUR3DQ8 signals and recovering with PHENOMD would find comparable recovered SNRs regardless of the spin values used for the injection. From Eq. (9) one would infer that the difference in any of these injections would be the varying levels of parameter bias seen in the parameter estimation. We also see that $\hat{d}_{\text{bf}}$ is comparable to $\hat{d}_s$ only when $\hat{d}_s$ is small, i.e., the bias distance that we are most interested in, $\hat{d}_{\text{bias}}$, will be well approximated by $\hat{d}_s$ except in cases where the indistinguishability SNR is high. Nonetheless, as we will see, we require $\hat{d}_{\text{bias}}$ to accurately calculate bias SNRs.

We will refer to the simple mismatch-based indistinguishability SNR in Eq. (6) first described in Ref. [29] as the *faithfulness SNR* ρ_{faith} . We call the improved estimate in Eq. (8) the *N-D bias SNR* $\rho_{\text{bias,ND}}$.

Consider Fig. 3, which illustrates parameter measurement for a two-dimensional toy problem where the only parameters in the model are m_1 and m_2 . (See Sec. III A for more details of this configuration.) The figure shows the true parameters, indicated by a black dot, and the 2D 90% CIs for signals at a selection of SNRs. In this example the faithfulness SNR ρ_{faith} from Eq. (6) is 52, and the 2D bias SNR $\rho_{\text{bias,2D}}$ from Eq. (8) is 60. We see that the true parameters lie approximately on the 2D CI boundary for a signal at SNR 60, but are within the CI for a signal with SNR 50, consistent with the discussion above. For this example, these two SNR estimates differ only by about 10%, but we will see cases below for which these SNR estimates may disagree by 150%.

Uninteresting bias directions. The second reason why these estimates of the indistinguishability SNR are too conservative arises from the fact that we are dealing with a multidimensional parameter space [41]. Consider the 1D

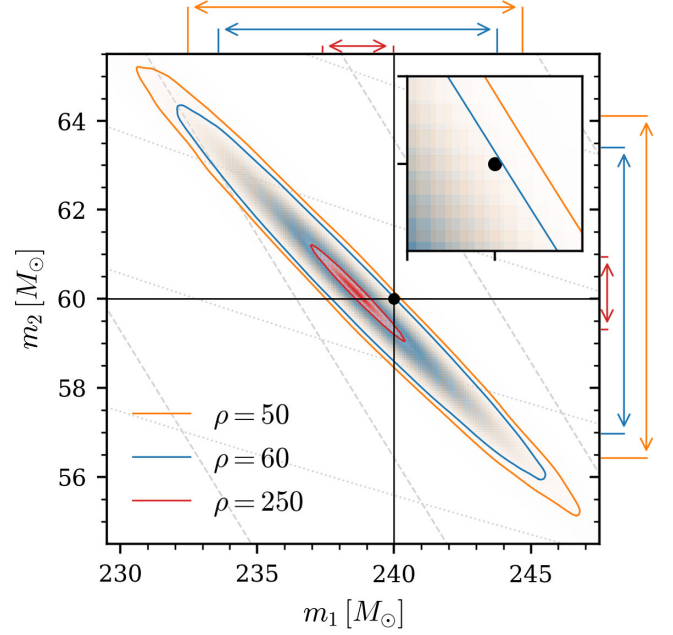


FIG. 3. Measurement of the primary mass, m_1 , and secondary mass, m_2 , for different signal-to-noise ratios ρ for the case BAM-3 described in Table I. The black dot indicates the true parameters, the contours show the 90% credible intervals, and the horizontal/vertical lines above and to the right of the Figure show the 90% symmetric credible intervals for the 1D marginalized posteriors. The inset shows a zoomed in portion of the posterior, focusing on the correlation between the true parameters and the credible interval at which they are biased. We show lines of constant total mass (grey dashed) and constant chirp mass (grey dotted). For this simulation, the faithfulness indistinguishability SNR is 52, the 2D bias SNR is 60, and the primary mass is estimated to be biased at $\rho \approx 250$. We see that the 2D bias SNR correctly identifies the SNR at which the 2D posterior is biased (at the 90% credible interval), and the 1D marginalized posterior for the primary mass remains unbiased until $\rho \approx 250$.

CIs for m_1 and m_2 in Fig. 3 separately (shown as vertical lines above and to the right of the figure—note that the 1D intervals are always narrower than the direct projection of the 2D intervals). Neither m_1 nor m_2 is remotely close to being biased at SNRs of 50 or 60. The faithfulness and 2D bias SNRs tell us nothing about the potential bias of the two parameters we are interested in.

We may expect that the 2D bias SNR estimate may be more accurate if we consider some other parametrization of the masses, e.g., the total mass $M = m_1 + m_2$, or the chirp mass $M_{\text{chirp}} = M\eta^{3/5}$, where $\eta = m_1 m_2 / (m_1 + m_2)^2$ is the symmetric mass ratio. Lines of constant chirp mass and total mass are shown in Fig. 3, and we see that biases will not be incurred in either parameter at SNRs 50 or 60. It may be the case that the 2D bias SNR does apply to the chirp mass in low-mass binaries where the signal is dominated by the inspiral (and, therefore, the leading-order post-Newtonian phasing, from which the chirp mass derives), but in general we do not expect the *N-D* bias SNR to apply to any single

parameter of interest. We provide more detailed examples in Secs. IV and V to illustrate this point.

To accurately estimate the SNR at which individual parameters will be biased we must calculate separately the bias SNR for each parameter (or combination of parameters) that we are interested in. This is straightforward to do. We first calculate the complete set of parameters θ_{bf} at which the model best agrees with the true signal, as before. We then calculate the parameters at which the model best agrees with the true signal, but *keep the one parameter we are interested in fixed to its true value*. In this way, we are considering the distance between the true and best-fit values of that one parameter along the curve connecting these two points within the model manifold that is always closest to the true signal in all other parameters.

We also illustrate this SNR estimate in Fig. 3. To find the SNR at which, say, m_1 is biased, we keep m_1 fixed to its true value and optimize all other parameters to find the parameters that give the best agreement with the true signal under this restriction, $\theta_{\text{bf}|m_1}$. The distance between the model at θ_{bf} and at $\theta_{\text{bf}|m_1}$, $\hat{d}(h_{\text{bf}}, h_{\text{bf}|m_1})$, will tell us the bias SNR for m_1 . In this case it is approximately SNR 250. Since the single-parameter indistinguishability SNRs can be related directly to the SNR at which each parameter will be biased, we call this the “ m_1 bias SNR” and denote it by ρ_{m_1} . We note that this result is presented in a similar fashion in Eq. (26) of Ref. [31], again asserting that the Pythagorean relation holds between the distances, as described in Eq. (9), and our separate maximization over all other parameters at both θ_s and θ_{bf} amounts to their choice of maximum averaged overlap.

We see, then, that the N -D bias SNR *does* reliably predict the SNR at which a measurement will be biased from the true parameters, but only in the N -dimensional credible interval, where N is the number of parameters that were kept fixed in the mismatch optimization. To calculate the SNR at which a particular parameter is biased, we must calculate the appropriate parameter bias SNR.

These statements are based on a small number of assumptions. The derivation of Eq. (6) in Ref. [29] begins with an assumption of sufficiently high SNR to allow for the Gaussian posterior scaling, which is equivalent to assuming the linear signal approximation outlined in Appendix A. This assumption is relaxed somewhat by using the mismatch instead of the Fisher matrix in their calculation, though we still assume that the signal and model have approximately equal SNRs. We further assume that the minimization to find the best-fit parameters has a true minimum (equivalent to the likelihood being unimodal). This is a valid assumption for the comparable-mass black hole binary signals we consider here, though may not always hold for other sources of gravitational waves [42]. For more realistic signal models with higher multipoles, one can find reparametrizations of the extrinsic parameters

to ensure unimodal posteriors [43]. We finally assume that the impact of the (broadly uniform) prior probability is negligible for this analysis, except when the best-fit parameters lie close to a prior boundary, as discussed in Sec. V C.

We have illustrated here that these assumptions hold for one example. In Sec. IV, we will illustrate in more detail with two-dimensional toy models and full four-dimensional examples that Eq. (8) does correctly predict the indistinguishability SNR, so long as we calculate the normalized distance $\hat{d}(h_{\text{bf}}, h_s)$ —the distance used in most applications of Eq. (6), $\hat{d}(s, h_s)$, provides only a conservative estimate. We then show in Sec. V that by optimizing the mismatch over all parameters but one, we can calculate the bias SNR for that parameter.

III. METHODS

We now describe the specifics of the numerical setup used to produce the results in this paper.

A. Signal waveforms and waveform models

In this work we predominantly use numerical relativity waveforms as proxies for our signals s . Numerical relativity solutions of Einstein’s equations for black hole mergers are excellent representations of real astrophysical signals in that the only approximations in the NR calculations are the numerical errors (which can in principle be reduced to any desired level with sufficient numerical resolution) and the calculation of the GW signal at a finite distance from the source. (Recall that gravitational waves are formally defined at null infinity.) With sufficiently careful simulation setup and postprocessing, other uncertainties, e.g., initial data specification and gauge choices, can be reduced to the same problems of numerical resolution and finite-radius wave extraction.

We use NR waveforms produced by the BAM code, which solves the moving-puncture treatment of Baumgate-Shapiro-Shibata-Nakamura formulation with finite-difference methods [44]. The waveforms listed in Table I were previously published in Ref. [45] and used to tune or verify the phenomenological model PHENOMD [46]. We consider only quasicircular aligned-spin binaries, where the black hole spins are aligned with the orbital angular momentum, and the binary’s orbital plane is fixed, i.e., there is no spin precession. We use only the dominant ($\ell = 2, m = 2$) multipole, so that the signal’s orientation and polarization can be absorbed into an overall amplitude factor. To generate the NR signals down to the required starting frequency, we use the hybrids constructed in Ref. [47] (restricting to the $\ell = 2, m = \pm 2$ multipoles).

We also use the NRHYBSUR3DQ8 model [48] to produce proxy signals. This model is calibrated to NR waveforms from binaries with mass ratios between $q = 1$ and $q = 8$, and spins up to $\chi = 0.8$, and allows us to consider signals at

TABLE I. Table of simulation configurations used in this work, listing the mass ratio and aligned dimensionless spins of each black hole, described in Sec. III A. All simulated signals are generated at a total mass of $300M_\odot$ and a starting frequency $f_{22} = 2$ Hz. We also present values for the faithfulness and effectualness mismatches, $\hat{d}_s^2$ and $\hat{d}_{\text{bf}}^2$, respectively, described in Sec. II B 2, between these signals and PHENOMD computed over a frequency range of 5–128 Hz. We finally tabulate the 2D and 4D bias distances $\hat{d}_{\text{bias}}^2$ to be used in Sec. IV. Note that the relation $\hat{d}_s^2 = \hat{d}_{\text{bf}}^2 + \hat{d}_{\text{bias}}^2$ holds to a good approximation, independently of the number of degrees of freedom used in the optimization.

Simulation ID	q	χ_{1z}	χ_{2z}	$\hat{d}_s^2(\times 10^{-3})$	$\hat{d}_{\text{bf},4\text{D}}^2(\times 10^{-3})$	$\hat{d}_{\text{bias},4\text{D}}^2(\times 10^{-3})$	$\hat{d}_{\text{bf},2\text{D}}^2(\times 10^{-3})$	$\hat{d}_{\text{bias},2\text{D}}^2(\times 10^{-3})$
BAM-1	3	-0.5	-0.5	1.22	0.13	1.09	0.23	1.00
BAM-2	4	0.25	0	0.40	0.14	0.26	0.34	0.06
BAM-3	4	-0.75	0	0.86	0.23	0.63	0.25	0.62
BAM-4	10	0	0	1.48	0.16	1.32	0.21	1.26
BAM-5	18	0	0.4	3.68	0.29	3.38	0.51	3.16
SUR-1	2	0.5	-0.5	1.75	0.13	1.62	0.73	1.02
SUR-2	2	0.4	0.1	0.08	0.07	0.01	0.07	0.01
SUR-3	2	0.3	-0.4	0.22	0.09	0.13	0.16	0.05
SUR-4	2	0.05	0.47	2.49	0.07	2.40	0.87	1.59

arbitrary points in this parameter space. As with the BAM NR waveforms, for this study we only consider aligned-spin binaries, and the dominant (2,2) multipole.

As an example waveform model to assess systematics, we chose PHENOMD for three reasons: (1) PHENOMD models only the dominant (2,2) multipole of aligned-spin binaries, which provided a convenient reduced parameter space on which to test our approach, (2) PHENOMD is a relatively old model with larger uncertainties than more recent models, ensuring that our model is less accurate than our proxy signal waveforms, and (3) PHENOMD was calibrated to a subset of the BAM NR waveforms that we use in this study listed in Table I, allowing a consistent test of the performance of the model against waveforms that were treated as true signals in the model's construction.

In any consideration of waveform systematics it is important that the uncertainties in the signal proxy waveforms are far smaller than the uncertainties in the models we are assessing, otherwise the signal uncertainties will contaminate our results. Error estimates for NR waveforms can be expressed as mismatch uncertainties, which allow us to calculate their faithfulness indistinguishability SNR. We estimate the mismatch uncertainties of the BAM NR signals and the NRHYBSUR3DQ8 signals as $\sim 10^{-4}$, i.e., $\hat{d} \sim 10^{-2}$, putting the faithfulness indistinguishability SNR at approximately $1/\hat{d} \sim 100$. We will see in Sec. V that parameter bias SNRs can be much larger, and so we must be cautious in interpreting these results. This is not a serious issue for this proof-of-principle study, where we are considering only the (2,2) multipole of aligned-spin models, since these will not be used to measure properties of loud GW observations, but this will be a crucial point to bear in mind when we assess systematics for state-of-the-art models in future work.

B. Optimal model parameters

The distance measure introduced in Sec. II A requires optimization over a set of parameters Θ_{opt} . When computing the faithfulness between two spin-aligned quadrupolar signals, Θ_{opt} only includes the coalescence time, t_c , and coalescence phase, φ_c , and in this case we express the time- and phase-shift optimization of the match in a computationally efficient manner using an inverse fast Fourier transform (iFFT) [41,49],

$$M_{\text{iq}}(h_1, h_2) = \max_{\{t_c, \varphi_c\}} \frac{\langle h_1 | h_2 \rangle}{|h_1| |h_2|} \\ = \frac{4}{|h_1| |h_2|} \max_{t_c} \left| \text{iFFT} \left[\frac{\tilde{h}_1(f) \tilde{h}_2^*(f)}{\tilde{S}_n(f)} \right] (t_c) \right|. \quad (12)$$

Maximization of φ_c is achieved by taking the norm in Eq. (12), and maximization over t_c is done by taking the maximum component of the output iFFT array. We can increase the resolution of the discrete time step used for the time-shift optimization by padding the frequency domain data before taking the iFFT, which is especially important for signals with only slight differences between the linear-in-frequency contributions to their phases [41,50].

To compute the bias distances we need the appropriate best-fit parameters between the model and signal. We write the best-fit parameters θ_{bf} as the union between a set of optimized parameters $\xi_{\text{bf}} \in \Theta_{\text{opt}}$ and a set of parameters held fixed, $\bar{\theta} \in \Theta \setminus \Theta_{\text{opt}}$, such that $\theta_{\text{bf}} = \xi_{\text{bf}} \cup \bar{\theta}$ is found through the *minimization* of the mismatch,

$$\xi_{\text{bf}}(s; \bar{\theta}) = \arg \min_{\xi \in \Theta_{\text{opt}}} \mathcal{M}(s, h(\xi; \bar{\theta})). \quad (13)$$

For the case of computing the N -D bias SNR in Eq. (8), the best-fit parameters are found by optimizing over all signal parameters, thus $\Theta_{\text{opt}} = \Theta$. To compute the distances for individual parameter biases, for example the bias SNR estimate for the primary mass m_1 , then $\bar{\theta} = \{m_1\}$, and we minimize Eq. (13) over all remaining parameters, in this case $\xi = \{m_2, \chi_{1z}, \chi_{2z}, t_c, \varphi_c\}$ for our quadrupolar, spin-aligned model. This minimization is in practice reduced to a three- or four-dimensional numerical optimization over at most $\{m_1, m_2, \chi_{1z}, \chi_{2z}\}$ using Eq. (12) to compute the time- and phase-optimized mismatch. We then recover $\theta_{\text{bf}|m_1} = \xi_{\text{bf}} \cup \{m_1\}$ introduced in Sec. II B 2.

We choose to use the Nelder-Mead [51] algorithm implemented in the Python library SciPy [52] to perform the numerical minimizations. Nelder-Mead does not rely on numerical derivatives of the objective function and generally requires only a small number of function evaluations to converge sufficiently to a minimum. We run the minimization over a spread of initial values, starting at the true parameters of the signal θ_s and expanding away in quadratically increasing step sizes in each parameter to ensure at least minimal coverage of parameter space regions far from the true parameters. The initial parameter guesses for the Nelder-Mead minimization are precomputed, and then the minimization is performed in parallel, taking the global minimum found across all resulting values. Finally, we ensure a fine resolution for the time step optimization by padding the frequency domain signals with an array of zeros to a length equal to a large power of 2 (2^{22}) [41] and run the Nelder-Mead algorithm with an absolute error tolerance of 10^{-14} .

Regardless of the stated error tolerance, we also check the efficacy of the optimization by finding the 4D best-fit parameters using two different parametrizations, $\{m_1, m_2, \chi_{1z}, \chi_{2z}\}$ and $\{M_{\text{chirp}}, \eta, \chi_{\text{eff}}, \chi_{\text{antisym}}\}$, and compare the effectualness in both sets of parameters. This effectualness typically disagrees with a relative error of 10^{-5} , and we find that using either set of “best-fit” parameters impacts the SNRs computed below when the SNRs reach values above ~ 600 . We, therefore, strongly suggest caution when considering any high-SNR predictions in the tabulated data below—we leave the values in for comparison between methods.

C. Parameter estimation methods

To demonstrate that the effectual indistinguishable SNR correctly corresponds to the biases in our inferred estimates for the true source parameters, we perform Bayesian inference to estimate the *posterior probability density function* for a given signal s . Bayesian inference is the process of estimating the properties of the signal for a given model h and observed data d . A posterior probability distribution for the parameters θ can be obtained through Bayes’ theorem,

$$p(\theta|d, h) = \frac{p(\theta|h)p(d|\theta, h)}{\mathcal{Z}}, \quad (14)$$

where $p(\theta|h)$ is the prior probability of the parameters θ given our model h , otherwise known as the prior, $p(d|\theta, h)$ is the likelihood of the data given the parameters θ and model h , and $\mathcal{Z} = \int_{\Theta} p(\theta|h)p(d|\theta, h)d\theta$. Under the noise assumptions outlined in Sec. II, the Whittle likelihood in gravitational-wave physics is proportional to [25]

$$p(d|\theta, n) \propto \exp \left\{ -\frac{1}{2} \langle d - h(\theta) | d - h(\theta) \rangle \right\}. \quad (15)$$

An aligned-spin quasicircular binary black hole signal is fully characterized by 11 parameters: four intrinsic describing the component masses m_1 and m_2 and the spins aligned with the orbital angular momenta of each black hole χ_{1z} and χ_{2z} , and seven extrinsic parameters describing the source location, inclination angle, merger time, etc. For gravitational-wave astronomy it is difficult to analytically calculate the posterior distribution, as it requires evaluating an eleven-dimensional integral.

To further reduce the dimensionality of the evidence integral, it is possible to analytically marginalize over some parameters [53–57]. In this work we marginalize over the luminosity distance [55,56] and coalescence phase of the binary [53]. We also fix the inclination angle, polarization, and sky location of the binary to their true values. For models that only consider the dominant quadrupole of aligned-spin binaries, the sky location and inclination angle only affect the overall amplitude of the GW and are degenerate with the luminosity distance. In addition, the polarization angle is completely degenerate with the coalescence phase. As such, we only sample over the merger time along with the masses and spins of each black hole. We note that at high SNRs ($\gtrsim 600$) we observed non-negligible differences between the posterior distributions obtained with and without distance and phase marginalization. We, therefore, do not show posterior distributions for SNRs > 600 in subsequent sections.

Given the large parameter space of the evidence integral, stochastic sampling [58–60] is often employed to draw samples from the unknown posterior distribution. Numerous tools are available to perform Bayesian inference for gravitational-wave astronomy [61–68], and many commonly employ the nested sampling algorithm, which iteratively evolves a set of *live points* randomly drawn from the prior to converge to regions of high probability [59,60]. In this work, we perform Bayesian inference using BILBY [63] with the DYNESTY [69] nested sampler.

Since we are interested in confidently identifying the 90% credible region at potentially high SNRs, we use 3000 live points and combine the results from six independent chains to obtain our final posterior distribution. This compares to 1000 live points and four independent chains

commonly used by LIGO-Virgo-KAGRA Collaboration in their production analyses [5]. We employ the `BILBY`-implemented `RWALK` sampling algorithm with an average of 60 steps per Markov chain Monte Carlo, and we also assume wide and agnostic priors for all parameters. Specifically, we employ uniform priors on the component masses with chirp mass and mass ratio constraints. Constraints are chosen to ensure regions of high probability are sufficiently sampled, while also reducing computational costs where possible. We also assume uniform priors on the aligned-spin components of the binary (Eq. (A7) in Ref. [62]).

Although directly translatable to any GW detector network, in this work we focus on next-generation GW detectors. Specifically, we assume a single detector network consisting of Einstein Telescope (ET) [9] and assume a prospective power spectral density [10] when evaluating the inner product. Since the exact configuration of ET is still under discussion, for simplicity we assume that ET is formed of a single L-shaped interferometer [70].

IV. RESULTS: N -D BIAS SNR

We discussed in Sec. II B 2 how to identify the SNR at which the true parameters will be observably biased from the posterior distribution using the appropriate distance measure $\hat{d}_{\text{bias}}$. When this distance is used in Eq. (8), the number of degrees of freedom, N , is not immediately specified. In fact the value of N depends on the dimensionality of the (marginalized) posterior distribution of interest [29] and, therefore, relates to the number of model parameters held fixed to their “true” values during the mismatch optimizations performed in finding θ_{bf} , i.e., the dimensionality of $\bar{\theta}$. In this section we explore the validity of Eq. (8) through direct comparison to parameter estimation results and the scaling of the posterior’s 90% CI.

We begin by using a model with two effective degrees of freedom, m_1 and m_2 , so that we can view samples from the entire posterior in a two-dimensional scatter plot. We construct this effective model from `PHENOMD` by fixing the component spins to the values of the injected signal we compare against, listed in Table I, both when sampling in parameter estimation and when computing optimal mismatch parameters. Of interest to us is the SNR at which the intrinsic masses are biased, and we consider the posterior distribution marginalized over $\{d_L, t_c, \varphi_c\}$, yielding an effective 2D posterior distribution in m_1 and m_2 , such as the one plotted in Fig. 3. The best-fit parameters are found by optimizing the mismatch over all parameters except the spins. Afterwards, we will extend the analysis to the full four-dimensional model.

A. Principal component posteriors and SNR scaling

The N -D bias SNR formula dictates the scaling of the bulk N -dimensional posterior’s credible region under the

assumption that the posterior is approximately a multivariate normal distribution. We can approximate this assumption on the posterior samples by using principal component analysis (PCA). PCA aims to find a linear transformation between the component directions of the signal parameters with minimal covariance by diagonalizing the covariance matrix (i.e., maximizing the variance in each parameter).

In order to perform the PCA on the posterior samples, we first normalize each component of the data, shifting the posterior to have zero mean and unit variance using the `SCIKIT-LEARN` class `StandardScaler`. The resulting posterior of the PCA will then be an approximate multivariate normal distribution with zero mean, up to nonlinear correlations present in the data. In this idealized PCA representation of the data, the 90% CI will approximate an N -dimensional sphere of radius $[\chi_N^2(0.1)]^{1/2}$.

As PCA is a linear transformation that projects the data onto axes constructed from linear combinations of the input parameters, its ability to produce multivariate normal samples depends on the strength of nonlinear correlations between the input parameters. As the Jacobian between different mass parametrizations is nonlinear, we can hope to improve the effectiveness of the PCA by choosing a sample parametrization that reduces the nonlinearities in sample correlations. Ultimately, the choice of input parameters will depend on the structure of the posterior for each of our injection cases. Given that the PCA computation is not expensive, we choose to compute the PCA of our posteriors using all possible combinations of input parametrizations (for the 4D cases, all pairs of mass parameters and all pairs of spin parameters), testing for multivariate normality on the PCA-transformed posteriors using the Henze-Zirkler test with a significance of 0.05 [71], available as the function `multivariate_normality` in the `PINGOUIN` Python package [72].

This test sometimes fails with our chosen significance, so in addition we also compute the Jensen-Shannon (JS) divergence [73] between each one-dimensional marginalized posterior of the PCA data and a zero mean unit variance normal distribution with an equal number of samples as the posterior, taking as representative of non-Gaussianity the maximum JS divergence across all 1D marginal posterior distributions. We finally choose a parametrization that minimizes this maximal JS divergence and passes the Henze-Zirkler test, if available, otherwise we take the parametrization that simply minimizes the representative JS divergence.

Once the PCA is performed, the variance in the posterior distribution increases inversely with the square of the signal SNR. This scaling is robust, and we can verify that it holds by comparing the PCA posteriors from injections at two different SNRs. The results of such a comparison for an injection of the 2D signal BAM-5, first at SNR 250 and again at SNR 20, show that after training the PCA transformation on the SNR 250 posterior data and applying the

same transformation to the SNR 20 posterior, the rescaled 90% CI circle of the SNR 250 posterior captures 89.3% of the posterior SNR 20 samples.

B. SNR comparisons

Given the approximate scaling of the PCA 90% CI, we can estimate the SNR at which the true parameters are biased through a simple rescaling. Define the norm of the true injection parameters, transformed using the same PCA transformations trained on the posterior data, to be r_{inj} . Then the SNR at which r_{inj} will fall outside the 90% CI is computed by rescaling the injected SNR ρ_{inj} by the ratio of the 90% CI sphere radius and the norm of the injection parameters,

$$\rho_{\text{PCA,ND}} = \frac{\sqrt{\chi_N^2(0.1)}}{r_{\text{inj}}} \rho_{\text{inj}}, \quad (16)$$

where, for the examples we consider, $\rho_{\text{inj}} = 250$.

In Table II we present the results of the PCA rescaling alongside the computed faithfulness SNRs and bias SNRs for both the 2D and 4D models. The columns of $\rho_{\text{faith,ND}}$ contain the values arising from computing the faithfulness SNR in Eq. (6). The SNRs $\rho_{\text{bias,ND}}$ use Eq. (8), and $\rho_{\text{PCA,ND}}$ are the SNRs computed by rescaling the injected SNRs of the parameter estimation samples such that the 90% CIs of the PCA samples contain r_{inj} .

The faithfulness SNR is a consistent lower bound for the bias SNRs. This should not be a surprise, as the faithfulness SNR includes contributions in the faithfulness mismatch coming from signal components orthogonal to the model

TABLE II. Table presenting 2D and 4D bias SNRs for the injected cases of study listed in Table I. The faithfulness SNR $\rho_{\text{faith,ND}}$ is computed using the faithfulness mismatch or, equivalently, the $\hat{d}_s^2$ values from Table I, assuming 2 or 4 free degrees of freedom in Eq. (6). The values for $\rho_{\text{bias,ND}}$ arise from Eq. (8) and the values of the bias distances given in Table I, assuming 2 and 4 free degrees of freedom. The SNRs $\rho_{\text{PCA,2D}}$ and $\rho_{\text{PCA,4D}}$ are computed by rescaling the approximate 90% PCA posterior volume of the recovered parameter estimation posterior samples for the 2D and 4D injections, respectively.

Simulation ID	$\rho_{\text{faith,4D}}$	$\rho_{\text{bias,4D}}$	$\rho_{\text{PCA,4D}}$	$\rho_{\text{faith,2D}}$	$\rho_{\text{bias,2D}}$	$\rho_{\text{PCA,2D}}$
BAM-1	57	60	60	44	48	47
BAM-2	98	122	120	76	195	182
BAM-3	67	79	79	52	61	61
BAM-4	51	54	56	39	43	43
BAM-5	33	34	38	25	27	27
SUR-1	47	49	53	36	47	47
SUR-2	221	538	556	170	433	428
SUR-3	134	174	186	103	213	214
SUR-4	40	40	39	30	38	38

manifold that do not affect the systematic bias. For some cases, such as BAM-4, the difference between the faithfulness and bias SNRs is small. This will happen when the true signal sits close to the model manifold near the best-fit parameters compared to the distance between the injection and best-fit parameters, as we see when comparing the values of $\hat{d}_s^2$ and $\hat{d}_{\text{bf}}^2$ in Table I. Under the linear assumption used in the PCA, we see that the bias SNR is consistent in reproducing an estimate for the SNR at which the posterior bulk will no longer contain the true signal parameters within its 90% CI.

For the 2D model, we can fully visualize the rescaling in Fig. 4, where we show the results of performing PCA on the 2D posterior samples for SNR 250 injections of BAM-5 (top) and BAM-2 (bottom), in the left column of the figure. We plot circles with radius r_{inj} as the blue circles. In the right column we show the samples and rescaled contours mapped back to the physical parameter space. For the case of BAM-5 we see that the norm of the injected parameters is considerably larger than the PCA 90% CI radius, indicating that biased recovery occurs at much lower SNRs, roughly a factor of 10 lower than the injected SNR according to the estimates in Table II.

The BAM-2 PCA posterior shows that the biased SNR is much closer to the SNR 250 injection value, estimated to be around SNR 180 from the posterior scaling and SNR 195 from Eq. (8). We find that an SNR 195 injection places the injected values on the 2D 90% CI boundary, as is shown below in Fig. 7. Furthermore, for this case, we see that the true values of the simulation lie far along the semimajor axis of the sample correlation ellipse, meaning that the 1D projections of the 2D posterior onto the m_1 and m_2 axes will still show bias even at the lower SNR required for the 2D posterior to contain the injected values. We explore resolutions to this below in Sec. V.

Finally, we compare the predictions of the PCA rescaled SNR and N -D bias SNR estimates to parameter estimation results using the full 4D model. Shown in Fig. 5 are the marginalized two-dimensional projections of the full 4D PCA posterior for BAM-3, injected at the approximate 4D bias SNR 80. At this SNR our linearized PCA approximation still holds (despite the noticeable railing visible in the PC1–PC3 plane), and the 4D sphere of radius $[\chi_4^2(0.1)]^{1/2}$ contains 90.7% of the samples in the posterior. We also see that the norm of the true signal parameters in the PCA projection matches very closely to this radius value. When looking across all cases of interest, we find that the 4D sphere estimate works well at containing approximately 90% of the posterior samples for all cases when injected at the 4D bias SNR values listed in Table II, and this radius matches the norm of the injected values to a relative error within 8% for the majority of cases, with the notable outlier being SUR-3 with a relative error of 15%.

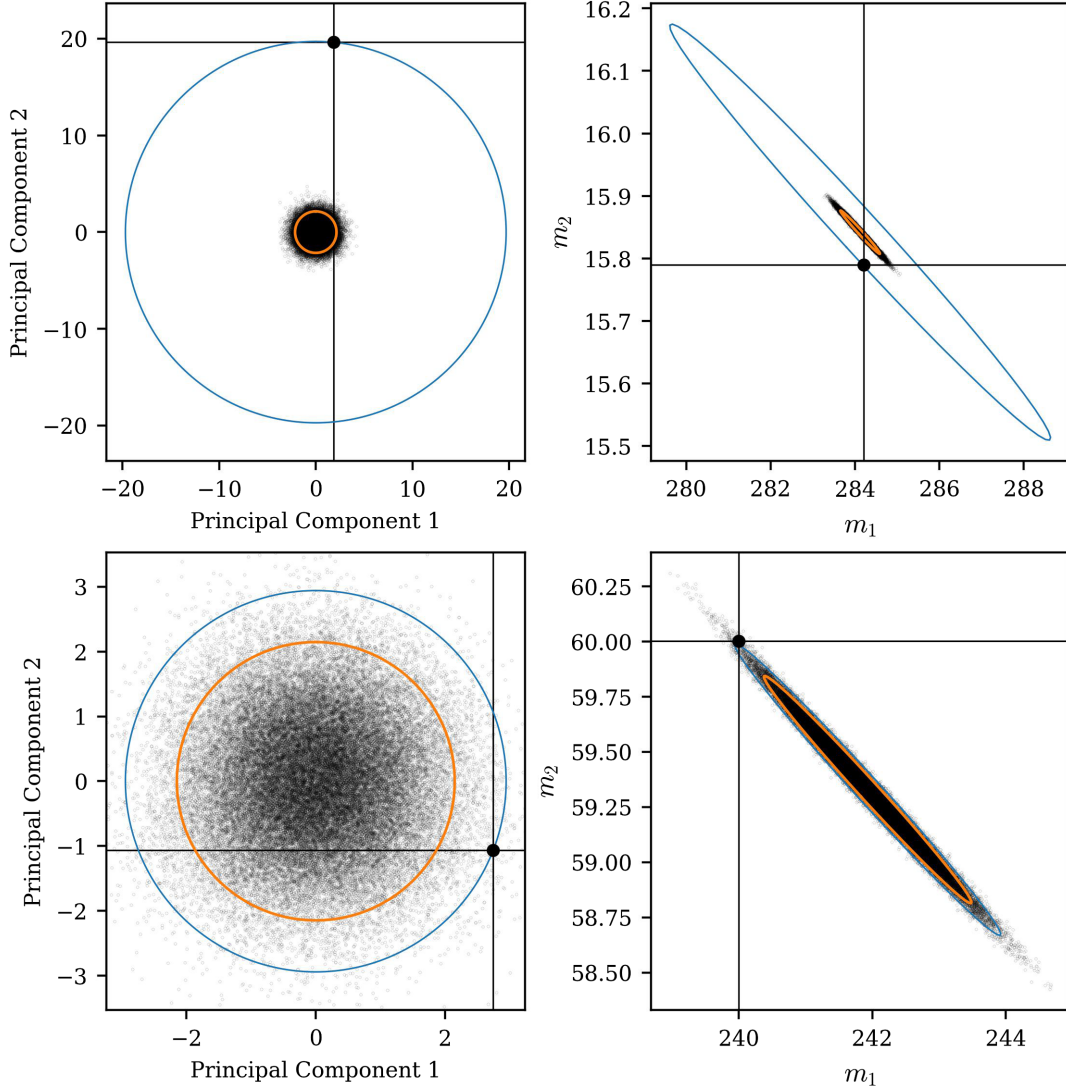


FIG. 4. Posterior probability distributions for the recovery of BAM-5 (top row) and BAM-2 (bottom row) described in Table I using the 2D model restriction of PHENOMD. The left column shows the posteriors after applying principal component analysis, detailed in Sec. IV A, with orange circles showing the approximate 2D 90% credible region for the $\rho = 250$ injection. The blue circles in the left column are generated using the norm of the injected signal parameters (shown as the black dot) as a radius, i.e., representing the SNR at which the true parameters will lie at the edge of the 90% credible region. The right column shows the samples in the physical m_1 – m_2 parametrization, with the blue and orange circles mapped into correlation ellipses using the inverse principal component transformation.

V. RESULTS: PARAMETER BIAS SNRS

In this section we compute the one-dimensional bias SNRs for all cases of interest and present these results in Tables III and IV found in Appendix C, both for the 2D model and the full 4D injections, respectively, and comment on the results below in Sec. V A. We also compare the results of the bias SNR computation to Fisher analysis results in Sec. V B and discuss the impacts of prior railing in Sec. V C.

A. Parameter bias SNRs for cases of interest

The bias SNRs calculated in Sec. IV estimate the SNR required for the N -dimensional injection parameter vector

to lie outside of the 90% CI of the N -dimensional posterior probability density. As discussed in Sec. II B 2, it does not tell us whether any given parameter of interest is biased. We see this fact demonstrated in Fig. 3, which displays the two-dimensional posterior distribution of m_1 and m_2 for BAM-3 along with the 90% CIs for the marginal one-dimensional posteriors of each mass separately as arrows along the plot edge. The SNR at which the 2D 90% CI just contains the injected parameters is comparable to the value of $\rho_{\text{bias},2\text{D}}$ presented in Table II, but each individual mass parameter becomes biased at SNRs much greater than this value (though this may not always be the case, as we discuss below). To investigate the bias SNR for m_1 in this example we instead compute the parameters $\theta_{\text{bf}|m_1}$ and use them in

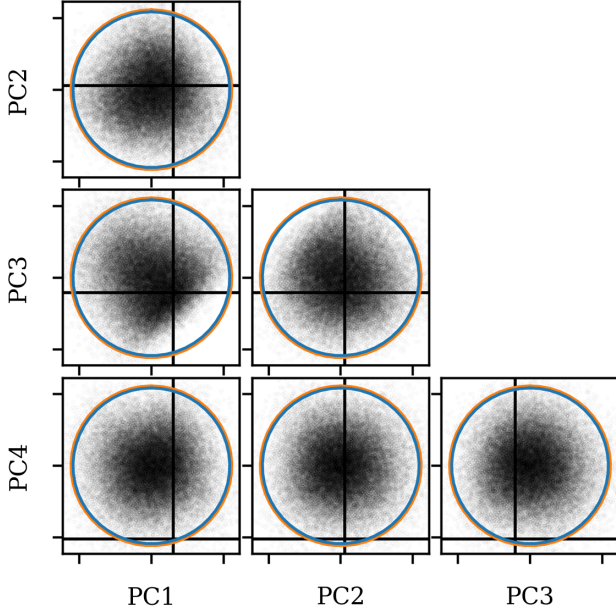


FIG. 5. Two-dimensional marginal posterior projections of the four-dimensional posterior distribution for BAM-3, detailed in Table I, after principal component analysis is applied to the posterior samples, with, e.g., PC1 denoting the first principal component. The signal is injected with a SNR of 80, matching the predicted 4D bias SNR predicted using Eq. (8) in Table II. The orange circles show the 2D projections of the 4D sphere with radius $[\chi_4^2(0.1)]^{1/2}$ that approximates the 90% credible region for the posterior and contains 90.7% of the posterior samples. The blue circles show the 2D projections of the 4D sphere with a radius determined by the norm of the true signal parameters.

Eq. (8) with $N = 1$ in place of the full signal parameters θ_s . In this way, we compute the distance between the true value of m_1 and the effectual value of m_1 along the one-dimensional submanifold described by choosing, at each value of m_1 , the remaining signal parameters from $\Theta_{\text{opt}} = \Theta \setminus \{m_1\}$ utilizing Eq. (13).

The results for computing the one-dimensional bias SNRs for the 2D model cases are presented in Table III, and the 4D results are presented in Table IV. The general trend of these SNRs is that they are higher than the N -D bias SNR for each model (i.e., 2D or 4D), meaning that for many signals of interest calculating either the faithfulness SNR or the N -D bias SNR will provide lower bounds to the parameter bias SNRs, but these lower bounds may sometimes be orders of magnitude too conservative. We plot a selection of one-dimensional marginalized posteriors for the individual component masses and spins of the cases BAM-2 and BAM-5 in Fig. 6, injected at varying SNRs estimated by the predicted parameter bias SNRs. The injected values of these parameters are shown as vertical black lines and the 90% CI boundaries for the different SNR injections are shown above each figure panel. The SNRs predicted from Eq. (8) provide a robust estimate for the SNR at which these individual parameters become

TABLE III. Values of the parameter bias SNRs computed for the 2D restriction of PHENOMD for all four mass parameters considered in this work. We replicate the values of the 2D bias SNR from Table II for comparison. We leave all values as computed for comparison but caution that SNR values above 500 may not be reliable given the numerical accuracy thresholds used in this work.

Simulation ID	$\rho_{\text{bias},2\text{D}}$	m_1	m_2	M_{chirp}	η
BAM-1	48	101	173	371	138
BAM-2	195	164	157	151	159
BAM-3	61	285	2449	523	1079
BAM-4	43	75	112	165	96
BAM-5	27	730	127	66	204
SUR-1	47	271	148	45	197
SUR-2	433	345	333	360	337
SUR-3	213	382	492	323	431
SUR-4	38	81	120	612	99

biased. We show the full 1D comparison results in Figs. 10 and 11 in Appendix C.

The variation between the parameter bias SNRs for different parameters has no obvious correlation, and parameter bias SNRs for a given parameter can vary significantly across parameter space. We also note the occurrence of a parameter bias SNR being *smaller* than the N -D bias SNR, which happens for individual parameters in a few cases but is most prominent in the 2D example for BAM-3, where $\rho_{\text{bias},2\text{D}} = 195$, but all of the parameter bias SNRs for the mass parameters listed in Table III are between 150 and 165. This case is shown in Fig. 7 along with the posterior results for parameter estimation runs performed at SNRs of 76 (the faithfulness SNR estimate), 160, and 195. We verify that indeed the true parameters lie within the 90% CI for the two-dimensional posterior at an SNR of 195 and fall near the boundary of the one-dimensional marginalized posterior 90% CIs for both individual masses at SNR 160. In this case, then, should one wish to produce a truly conservative estimate of the bias SNR, one should compute $\rho_{\text{bias},1\text{D}}$ regardless of the numbers of model degrees of freedom being measured. For the 2D BAM-3 case, we would then arrive at a conservative estimate of the bias SNR to be $\rho_{\text{bias},1\text{D}} = 195 \times \sqrt{\chi_2^2(0.1)/\chi_1^2(0.1)} = 195 \times \sqrt{4.6/2.7} = 149$, which indeed is a lower bound on the parameter bias SNRs computed for this model. This holds true for all cases examined, though again for many of the cases listed in Tables III and IV these 1D estimates are overly conservative.

To further investigate the variation of parameter bias SNRs of a given model as we move across parameter space, we compute the parameter bias SNRs between PHENOMD and NRHYBSUR3DQ8 for fixed values of $(m_1, m_2) = (200, 100)M_\odot$ and ranging over equivalent spin values shown in Fig. 2. We display results for six parameters in Fig. 8. The top row of panels shows the variation of the

TABLE IV. Values of the parameter bias SNRs computed for PHENOMD for all mass and spin parameters considered in this work. We replicate the values of the 4D bias SNR from Table II for comparison. The asterisks denote signals for which the parameter estimation is heavily impacted by the χ_{2z} prior bound, discussed in Sec. V C. We leave all values as computed for comparison but caution that SNR values above 500 may not be reliable given the numerical accuracy thresholds used in this work.

Simulation ID	$\rho_{\text{bias,4D}}$	m_1	m_2	M_{chirp}	η	χ_{1z}	χ_{2z}	χ_{eff}	χ_{antisym}
BAM-1	60	686	216	181	354	1009	557	125	712
BAM-2	122	498	258	704	374	340	256	117	293
BAM-3*	79	503	680	2125	585	451	610	1085	513
BAM-4*	54	787	517	2098	658	449	288	201	352
BAM-5	34	85	82	171	82	103	204	112	131
SUR-1	49	1727	540	115	2372	636	3572	49	1109
SUR-2	538	97725	6423	1636	16303	1716	1618	754	1715
SUR-3	174	844	107964	216	1424	518	631	166	569
SUR-4	40	357	183	197	257	317	238	41	271

parameter bias SNR for $[m_1, m_2, M_{\text{chirp}}, \eta]$ from left to right. The bottom row of panels displays the bias SNR for $[\chi_{1z}, \chi_{2z}, \chi_{\text{eff}}, \chi_{\text{antisym}}]$ from left to right. The structures of the parameter bias SNR contours show some similarity,

especially between the component masses and spins, η and χ_{antisym} . One notes that all parameters have relatively high parameter bias SNRs across most of the spin parameter space except for χ_{eff} . The structure of the parameter bias

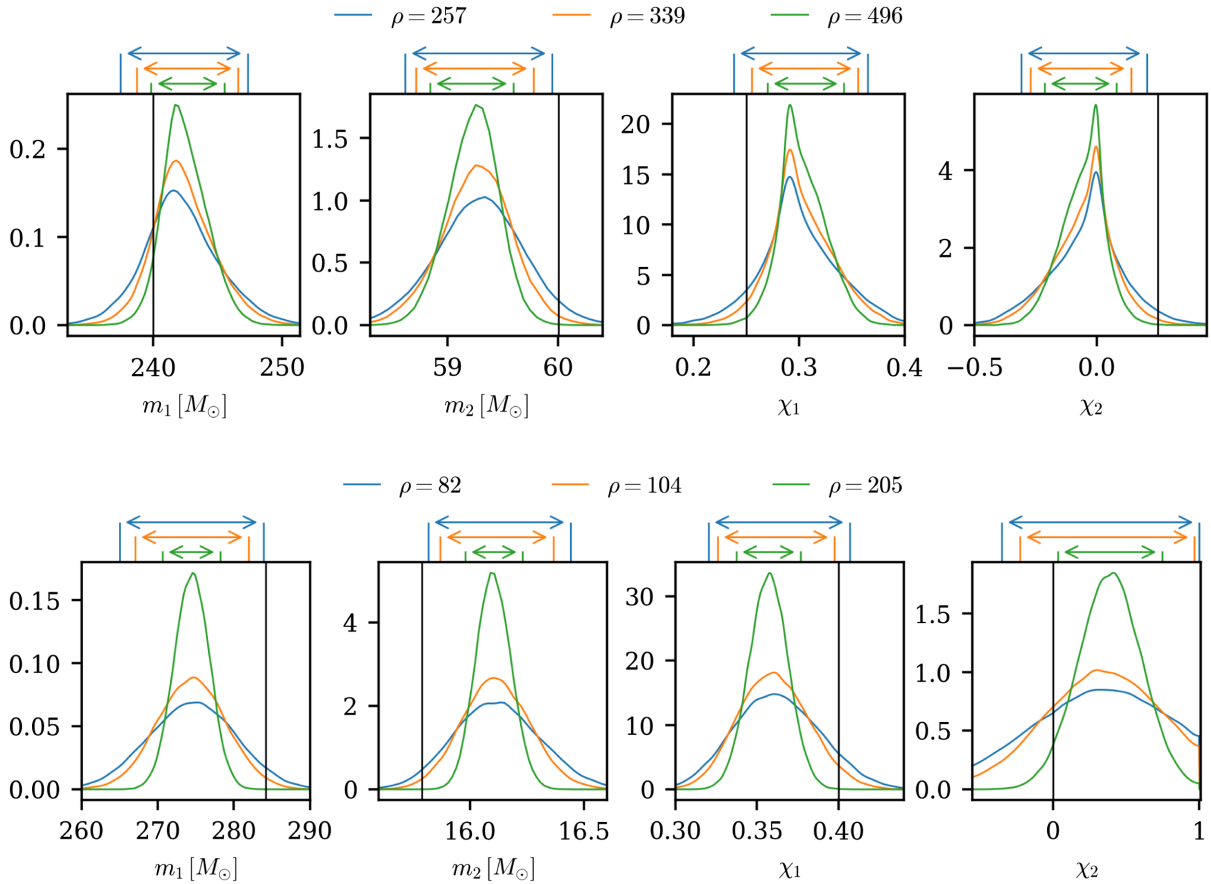


FIG. 6. Measurement of the primary mass m_1 , secondary mass m_2 , primary spin χ_1 , and secondary spin χ_2 for different signal-to-noise ratios ρ . The top row shows our analysis of BAM-2. For this simulation, the 1D effectual SNRs for $[m_1, m_2, \chi_1, \chi_2]$ are $\rho = [496, 257, 339, 256]$, respectively. The bottom row shows our analysis of BAM-5. For this simulation the 1D effectual SNRs for $[m_1, m_2, \chi_1, \chi_2]$ are $\rho = [85, 82, 104, 205]$, respectively. In all cases, the black vertical line indicates the true value, the horizontal arrows and vertical bars display the 90% CIs, and we sample over m_1, m_2, χ_1 , and χ_2 .

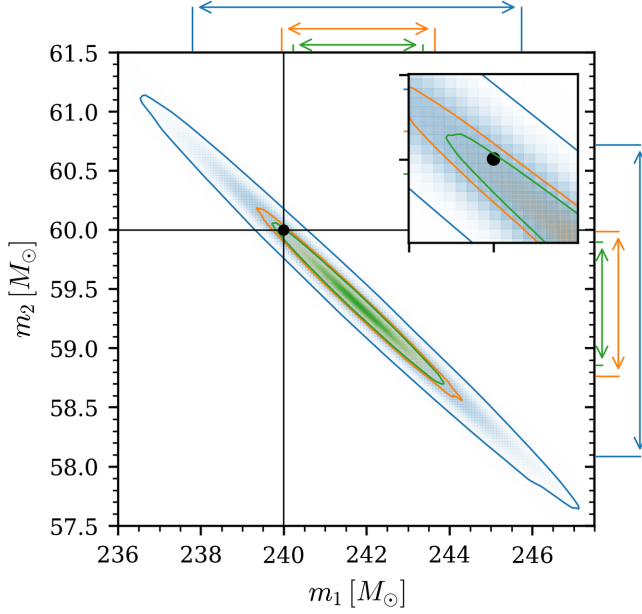


FIG. 7. Measurement of the primary mass, m_1 , and secondary mass, m_2 , for different signal-to-noise ratios ρ . Here we show our analysis of BAM-2 when only sampling over m_1 and m_2 (other parameters are held fixed to their true values). The black dot indicates the true parameters, the contours show the 90% credible intervals, and the horizontal/vertical lines above and to the right of the figure show the 90% symmetric credible intervals for the 1D marginalized posteriors. The inset shows a zoomed in portion of the posterior, focusing on the correlation between the true parameters and the credible interval at which they are biased. For this simulation, the faithfulness indistinguishability SNR is 76, the 2D bias SNR is 195, and both the primary and secondary masses are estimated to be biased at $\rho \approx 160$. We see that the effectual SNR correctly identifies the SNR at which the 2D posterior is biased (at the 90% credible interval) and the 1D marginalized posteriors for the primary and secondary masses remain unbiased until $\rho \approx 160$.

SNR for χ_{eff} mimics closely the structure of $\hat{d}_s$ plotted in Fig. 2. This behavior is also visible in the tabulated bias SNR data in Table IV when comparing $\rho_{\text{bias,4D}}$ to the parameter bias SNR for χ_{eff} in the four SUR cases, hinting that the χ_{eff} modeling bias between PHENOMD and NRHYBSUR3DQ8 is the driving systematic cause of differences between the two waveform models.

B. Comparisons to Fisher biases

One approach to computing bias estimates common in the GW literature is that of Fisher analysis, which we summarize in Appendix A. The main results of interest for this study are the estimate to the shift in measured parameters due to systematic errors, $\Delta\theta_{\text{sys}}$ in Eq. (A5), and the definition of the Fisher matrix Γ given in Eq. (A6). From these two quantities we can estimate an SNR at which the Fisher bias in the parameter θ^i will become larger than the Fisher estimate of the 90% CI by

$$\rho_{\text{Fisher}}^i = 1.645 \rho_{\text{inj}} \frac{\sqrt{\Gamma_{ii}}}{\Delta\theta_{\text{sys}}^i}. \quad (17)$$

Here, $\sqrt{\Gamma_{ii}}$ (i not summed over) is used as the approximate standard deviation in the measurement of parameter θ^i [25], the numerical factor rescales the significance to represent the 90% CI, and ρ_{inj} is the SNR of the injected signal s used in Eq. (A5). For this paper we consistently use $\rho_{\text{inj}} = 250$.

We present the results of ρ_{Fisher}^i for the cases of interest in Table V found in Appendix C. When compared to the bias SNRs in Table IV, we see that the two methods broadly agree to within 10% for most cases except for BAM-3 and BAM-4, which are the two cases impacted by bounded priors and discussed in Sec. V C, or where the bias SNR is particularly large, where results from both analyses may become unreliable due to either limited numerical precision in the minimization for the bias distance calculation or accuracy of numerical derivatives and conditioning of the Fisher matrix in the case of the Fisher estimates. We conclude from this comparison that both approaches are equivalent at estimating the bias SNR for the cases we have considered when the Fisher analysis is done correctly (see discussions in Appendix A about alignment and parameter choices), and one should use whichever method is most convenient to calculate when investigating for potential systematic biases.

C. Impact of bounded priors

One of the assumptions made for this work in Sec. II B 2 is that one may overlook the impact of priors on the posterior probability scaling when estimating the bias SNR. We have seen that this assumption is upheld when comparing our bias SNR values to the posteriors resulting from parameter estimation, even at moderately low SNRs around 40, but have also found two 4D cases for which this is not true: BAM-3 and, to a lesser extent, BAM-4. These two cases are denoted with asterisks in Table IV and in Fig. 10.

In both of these cases the best-fitting values of the secondary spin χ_{2z} for the model PHENOMD lie close to the physically imposed $\chi_{2z} = -1$ boundary, and the railing of the posterior against this boundary produces large shifts in the recovery of the other parameters when considering each parameter's one-dimensional marginalized posterior. The presumption that the posterior is a multivariate normal distribution no longer holds, and instead the full posterior is a *truncated* multivariate normal distribution [74,75], in this case truncated in one dimension, and the severity of the truncation will impact the recovered means and covariances of all one-dimensional marginalized parameter posteriors. We leave to future work further handling of bounded priors on bias SNR estimation, but make a few remarks.

First, the parameter bias SNRs computed for these cases tend to *overestimate* the SNR at which parameters are

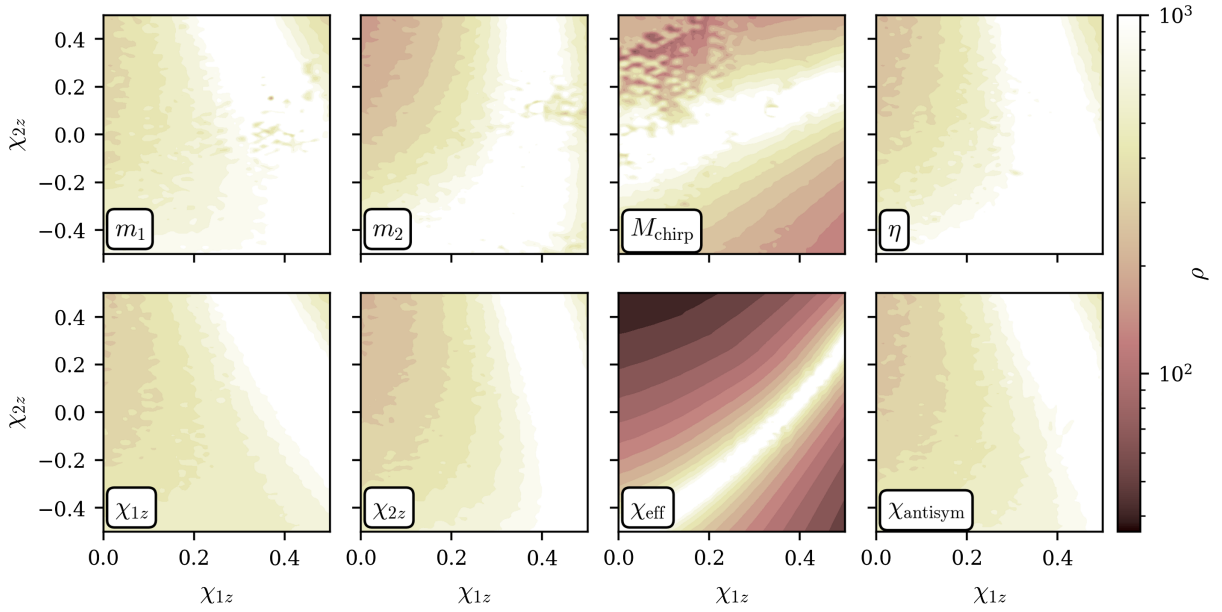


FIG. 8. Contour plots of parameter bias SNRs computed between the models PHENOMD and NRHYBSUR3DQ8, plotted for a range of χ_{1z} and χ_{2z} for fixed values of $(m_1, m_2) = (200, 100)M_\odot$. The top row of panels displays, from left to right, the parameter bias SNR contours for $[m_1, m_2, M_{\text{chirp}}, \eta]$, and the bottom row displays the parameter bias SNR contours for the parameters $[\chi_{1z}, \chi_{2z}, \chi_{\text{eff}}, \chi_{\text{antisym}}]$ from left to right. The cloudy structures visible in the contour plots arise both from the interpolation used to construct the contours and the fluctuations in the lower minimization tolerances used to compute the optimal parameters from Eq. (13). We further cap the color range to a maximum value of 10^3 , as SNR predictions above this value are not reliable from the numerical thresholds used in this work.

correctly recovered, leading to overconfidence in model performance. This is especially true for BAM-3, where the parameter estimation shows biases at SNRs significantly lower than the estimated values from both Eq. (8) and the Fisher analysis. In this instance, the best-fitting value for χ_{2z} lies at the physical lower bound, and the Fisher bias value for $\Delta\chi_{2z}$ is near the unphysical spin value of $\chi_{2z} \approx -1.2$. Both of these facts provide clear indicators that the prior bound is impacting our parameter bias SNR estimates and should be watched for when applying these methods.

The second remark is that, again for these two cases, the 4D bias SNR estimates are *not* greatly impacted by the railing posterior against the χ_{2z} boundary, as seen in Table II and shown in Fig. 5. While the boundary is still clearly visible in the PCA of the posterior samples, its impact on the total posterior scaling is seemingly minor. Assessing how robust of an observation this is we leave to future work.

VI. ACCURACY REQUIREMENTS FOR FUTURE DETECTORS

We expect to observe signals with SNRs of $O(1000)$ with the next-generation detectors Einstein Telescope and Cosmic Explorer. This is 2 orders of magnitude above the $O(10)$ SNRs of LIGO-Virgo-KAGRA observations to date. How do our results translate into waveform accuracy

requirements, both for models and for NR simulations and inspiral approximations? (For other studies on NR and waveform model accuracy needs for next-generation detectors, see Refs. [33,76,77].) The examples in this paper are specific to a quadrupole-only aligned-spin model, and we plan to extend and apply our methods to state-of-the-art generic binary models in future work. But a number of aspects of our study on indistinguishability SNRs—from simple conservative estimates using the faithfulness through to individual-parameter bias SNRs—allow us to make some general statements about accuracy requirements over the next 10–15 years.

Our focus has been on models tuned to NR simulations, so let us first put this discussion in the context of NR simulation uncertainties and computational cost. Phase errors in NR waveforms are dominated by numerical resolution: with a 4th-order accurate scheme (in both time and space discretization), a factor of 2 improvement in resolution leads to errors reduced by a factor of 16. The higher-resolution simulation requires eight times the memory (in a 3D code), and also double the number of time integration steps, so the computational cost also increases by a factor of 16. This tells us that the computational cost scales roughly linearly with the accuracy. If we require an order of magnitude improvement in accuracy, we need an order of magnitude increase in computational resources. For higher-order or pseudospectral codes, the scaling may be better, with a slower increase in

computational costs, but assuming a linear scaling between accuracy and computational cost allows us to make an approximate translation of mismatch requirements to computational resource needs.

In the following, therefore, we recall that the normalized difference between two waveforms relates to the mismatch as $\hat{d} = \sqrt{\mathcal{M}}$, and errors in NR simulations (e.g., the waveform phase and amplitude) scale as $\hat{d}$. For the purposes of this discussion we will, therefore, assume that computational cost scales linearly with $1/\hat{d}$. It is straightforward to adjust our estimates for different computational cost scalings.

Equation (6) provides the most conservative mismatch accuracy requirement if we use $N = 1$ for individual parameter measurements. To guarantee no parameter biases due to model inaccuracies for $\rho > 1000$, this criterion requires $\mathcal{M} \lesssim 10^{-6}$. Current binary black hole (BBH) NR waveforms and waveform models quote mismatch uncertainties of $10^{-4} - 10^{-2}$, for example, see Refs. [18,21,78,79]. This suggests a necessary improvement of between 2 and 4 orders of magnitude in mismatch uncertainty, or 1–2 orders of magnitude improvement in simulation accuracy and computational cost.

As we have seen, the true bias SNRs are typically 5–10 times larger than those predicted by the most conservative estimate, due mostly to parameter correlations over the high-dimensional binary parameter space. The scaling will depend on both the parameter of interest and the point in parameter space, but could be determined by studying the parameter correlations of the model, largely independently of any accuracy analysis. However, we have also seen that in some cases the model error is along the principal direction of signal variation, and the true bias SNR can be comparable to conservative estimates, such as happens in the BAM-2 case in Fig. 4 and Table IV. In general, for any given model, we cannot know *a priori* the distribution of bias SNRs across the parameter space. To properly determine the limits of a model, we must calculate the true bias SNR over a sufficiently dense sampling of the binary parameter space. If a model’s biases are always approximately orthogonal to the principal parameter directions of the signal space, then the mismatch accuracy requirements will be less strict for some parameters. From the above estimate that bias SNRs are 5–10 times higher in these cases, and the mismatch requirement scaling with $1/\rho^2$, we expect mismatch requirements to be 1–2 orders of magnitude less strict. This means that the required improvements in mismatch accuracy of 2–4 orders of magnitude could reduce to only 1–2 orders of magnitude, and a factor of 3–10 increase in computational cost. This suggests that an important diagnostic in the construction of future waveform models will be the direction of parameter biases. It remains to be seen whether it is possible to optimize a model’s construction to ensure that parameter biases are always approximately orthogonal to the principal

parameter directions, though techniques introduced to mitigate waveform modeling errors and applied to extreme mass ratio inspiral signal models may well be suited to this task [80,81]. We note that the analysis of the parameter space variations of the parameter bias SNRs in PHENOMD was possible only because we have access to a much more accurate model from which to construct proxy true signals, NRHYBSUR3DQ8—cutting-edge model development will not have that luxury.

For fiducial “true” signals, the only accuracy measure available to us is the faithfulness SNR, and so for these signals we cannot escape the requirement of mismatch uncertainties of $\sim 10^{-6}$. (Note, however, that NR accuracies at this level are already achievable in principle, as seen in the tail of the mismatch distribution in Fig. 4 of Ref. [21].) How smoothly the bias SNRs vary across the parameter space, and therefore the density of much more accurate NR waveforms required to fully assess a model’s accuracy, will also depend on the details of the model. Another goal of modeling procedures should be to achieve bias SNRs that vary as little as possible, and as slowly as possible, across the binary parameter space.

Our overall conclusion would then be that current NR and model mismatches need to improve by up to 4 orders of magnitude for next-generation detectors, requiring a roughly 2 orders of magnitude increase in computational cost. However, further improvements in modeling techniques, and a more complete understanding of the parameter correlations for generic binaries over the full binary parameter space, may soften these requirements, and only modest improvements may be necessary over the most accurate current NR simulations and waveform models. We should make clear that improved accuracy is not the only factor that affects computational cost. We likely require much longer NR simulations (i.e., including many more inspiral orbits) than at present, a more dense sampling of binary parameter space, and an extension to more extreme parts of parameter space (higher mass ratios, routine simulations of near-extreme-spin black holes, and eccentric orbits). See, for example, Sec. 4.1.5 of Ref. [82] for a discussion on the scaling of NR computational costs.

VII. CONCLUSIONS

We have discussed a common estimate of the indistinguishability SNR of BBH waveforms and waveform models, based on the mismatch of a signal against a model evaluated at the signal source’s parameters, or the mismatch uncertainty of a waveform. We also stress that the square root of the mismatch, $\hat{d} = \sqrt{\mathcal{M}}$, which is the normalized distance between two waveforms, is a more intuitive measure of waveform differences. The standard indistinguishability SNR estimate is known to be conservative, sometimes by as much as an order of magnitude. This is because (a) measurement biases relate instead to the

difference between the model at its true parameters θ_s and the model at the best-fit parameters θ_{bf} that give the best agreement between the model and signal (see Fig. 1), and (b) the correct indistinguishability SNR calculated from the distance $\hat{d}_{\text{bf}}$ is the SNR at which the true parameters lie outside an N -D confidence surface, where N is the number of fixed parameters in the mismatch calculation. It cannot be used to estimate the indistinguishability SNR for single parameters, except as a conservative lower bound, calculated using 1 degree of freedom in χ^2 in Eq. (8). The correct indistinguishability SNR for each parameter, which we call the *parameter bias SNR*, is calculated by optimizing all other parameters in the mismatch calculation (keeping the parameter we are interested in fixed) and using the distance between the model at those parameters and the true parameters to calculate the indistinguishability SNR in Eq. (8) with 1 degree of freedom.

We have illustrated that this approach provides accurate estimates of both the N -D and 1D parameter bias SNRs. For the N -D case we used a PCA to demonstrate that the N -D 90% CI in a parameter estimation analysis agrees well with that predicted from the N -D bias SNR. In the case of parameter bias SNRs, we performed an extensive set of parameter estimation analyses to confirm that the parameter bias SNRs calculated from the appropriate normalized distance (mismatch) correctly predicted the SNR at which the true value of each parameter would lie on the 90% CI in a measurement. We also compared with estimates from Fisher methods and found that both methods were in good agreement for the cases we considered, with the caveat that both methods will fail if the best-fit parameters rail against a parameter boundary.

Previous works have typically used $\hat{d}_s$ (in our notation from Fig. 1), and chosen the number of degrees of freedom in Eq. (6) in either an *ad hoc* manner, or based on the number of intrinsic parameters in the system [19,29,30,33,83]. As we have illustrated, this *does not* in general predict the correct parameter bias SNR, and, although the answer is often lower than the true parameter bias SNRs, it is not necessarily so. Unless one identifies the principal parameter directions for the given point in parameter space, the relationship between common bias SNR estimates, e.g., $N/(2\rho^2)$, and the true parameter SNR biases is unknown.

In this work we restricted examples to the simple test case of the (2,2)-mode from aligned-spin binaries. In future work, we aim to extend these results to state-of-the-art generic models, to provide robust statements on the reliability of these models across the binary parameter space. To do that will require mismatch optimizations over a fifteen- or seventeen-dimensional parameter space (for noneccentric and eccentric binaries, respectively), for which the optimization algorithms used in the four-dimensional models in this paper are not expected to be adequate. However, in addition to being able to optimize

over some parameters analytically [84–86], we expect that calculating optimum mismatches will be computationally cheaper than resolving the full posterior distribution function in the vicinity of the best-fit parameters, as is done in standard parameter estimation exercises. We must also overcome the issues of railing against prior boundaries, as discussed in Sec. VC.

For now, we can nonetheless make broad statements about the required model accuracy, and levels of accuracy improvements, for future GW observatories. We estimate that model accuracy must improve by up to 2 orders of magnitude for next-generation detectors, but, depending on the details of model construction, only modest improvements may be sufficient.

We caution, however, that the methods we have discussed here, and the statements we have made about future accuracy needs, apply only to situations where we can calculate a sufficiently accurate “true” signal against which to evaluate models. This is currently limited to the last orbits and merger of binary black hole systems. We do not have a means to calculate long-duration, fully general relativistic inspiral, and for systems with matter (binary neutron stars or black hole–neutron star binaries) we have neither a full understanding of all physical processes involved, nor as yet sufficiently accurate numerical relativity codes to calculate true waveforms. Quantifying the necessary level of modeling accuracy and physical completeness for future science goals is an important open question in source modeling and requires a great deal of further work over the next decade.

Various plots and analyses in this paper were made using Python software packages LALSuite [87], PyCBC [88], PESummary [89], Matplotlib [90], NumPy [91], and SciPy [52].

ACKNOWLEDGMENTS

The authors would like to thank Alvin Chua, Stephen Fairhurst, and Frank Ohme for enlightening discussions on waveform systematics and bias estimation. We thank Jannik Mielke for comments during the LIGO-Virgo-KAGRA internal review. J.T. acknowledges support from the NASA LISA Preparatory Science Grant 20-LPS20-0005. C.H. thanks the UKRI Future Leaders Fellowship for support through the Grant MR/T01881X/1. E.F.-J and M.H. were supported in part by Science and Technology Facilities Council (STFC) Grant ST/V00154X/1. 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. In part, the computational resources at Cardiff University were also supported by STFC Grant ST/I006285/1. We are also grateful for the Sciama High Performance Compute (HPC) cluster, which is supported by the Institute of

Cosmology and Gravitation (ICG), SEPNet, and the University of Portsmouth. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation grants PHY-0757058 and PHY-0823459. The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available. The data are available from the authors upon reasonable request.

APPENDIX A: FISHER UNCERTAINTY AND BIAS

We briefly review the formalism for approximating modeling bias in the context of GW presented in Refs. [34,39]. When considering the errors introduced in GW parameter inference, it is convenient to expand the GW signal about the parameters θ_{bf} that maximize the likelihood in Eq. (15),

$$\langle \partial_i h(\theta_{\text{bf}}) | d - h(\theta_{\text{bf}}) \rangle = 0, \quad (\text{A1})$$

where $\partial_i h(\theta) \equiv \partial h(\theta) / \partial \theta^i$. Defining $\Delta \theta^i \equiv (\theta_{\text{bf}} - \theta)^i$, the theoretical signal model h may be expanded about θ_{bf} as

$$h(\theta) = h(\theta_{\text{bf}}) + \partial_i h(\theta_{\text{bf}}) \Delta \theta^i + \frac{1}{2} \partial_i \partial_j h(\theta_{\text{bf}}) \Delta \theta^i \Delta \theta^j + \dots \quad (\text{A2})$$

If we assert that the difference between the true and best-fit parameters is small, we are enforcing the *linear signal approximation* by truncating the expansion at $\mathcal{O}(\Delta \theta^2)$,

$$h(\theta_s) \approx h(\theta_{\text{bf}}) + \partial_i h(\theta_{\text{bf}}) \Delta \theta^i + \mathcal{O}(\Delta \theta^2). \quad (\text{A3})$$

It follows then directly from the above approximation that the difference between the data and the model evaluated at the maximum likelihood parameters leads to two distinct biases arising from statistical and systematic errors in θ as, respectively,

$$\Delta \theta_{\text{stat}}^i = (\Gamma^{-1}(\theta_{\text{bf}}))^{ij} \langle \partial_j h(\theta_{\text{bf}}) | n \rangle, \quad (\text{A4})$$

$$\Delta \theta_{\text{sys}}^i = (\Gamma^{-1}(\theta_{\text{bf}}))^{ij} \langle \partial_j h(\theta_{\text{bf}}) | s - h(\theta_s) \rangle, \quad (\text{A5})$$

where Γ is the *Fisher information matrix*,

$$\Gamma_{ij}(\theta) = \langle \partial_i h(\theta) | \partial_j h(\theta) \rangle. \quad (\text{A6})$$

When working with loud signals, the error from Eq. (A4) becomes subdominant to Eq. (A5) and we may write the bias-to-variance ratio condition as $\Delta \theta_{\text{sys}}^i / \sigma^i \leq 1$, having made the usual approximation that the variance of any measured parameter θ^i is $\sigma^i \approx \sqrt{(\Gamma^{-1})^{ii}}$ [25].

The authors of Refs. [34,39] use the fact that $h(\theta_{\text{true}}) - h(\theta_{\text{bf}})$ can be well approximated by its leading term in the Taylor expansion above to recover the Fisher matrix in Eq. (A5), but one can go in the opposite direction, directly substituting into Eq. (A1) that the directional derivative of h along $\Delta \theta^i$ is approximated by the difference of the signals, in which case one arrives at

$$|h(\theta_{\text{bf}})|^2 - \langle h(\theta_s) | h(\theta_{\text{bf}}) \rangle = \langle h(\theta_{\text{bf}}) | s \rangle - \langle h(\theta_s) | s \rangle, \quad (\text{A7})$$

recovering the result used to derive Eq. (9) when the SNRs of the signals are all comparable and thus contribute to an overall scaling of both sides, which can be removed. This expression also assumes [as was done for Eqs. (A4) and (A5)] that higher-order terms (e.g., $\Delta \theta^i \Delta \theta^j \partial_i \partial_j h$) can be sufficiently ignored, which is a statement about the curvature effects in the model manifold.

Under the linear signal approximation between $\hat{h}_1$ and $\hat{h}_2$ from Eq. (A3), we find that the distance formula in Eq. (5) simplifies to

$$\hat{d}^2(\hat{h}_1, \hat{h}_2) \approx \frac{1}{2} \hat{\Gamma}_{ij}(\theta_1) \Delta \theta^i \Delta \theta^j, \quad (\text{A8})$$

which is the (local) half squared geodesic distance between the two signals on the signal manifold, known as Synge's world function [38,92], thereby further justifying our interpretation of $\hat{d}$ as a distance. This distance is also related to the Mahalanobis distance [93] away from θ_{bf} .

The estimate for $\Delta \theta_{\text{sys}}^i$ in Eq. (A5) is commonly referred to as the Cutler-Vallisneri (CV) criterion for the systematic bias, based on the authors of Ref. [34], and in that work it is discussed how the validity of Eq. (A5) depends heavily on the phase difference between s and $h(\theta_s)$. Recent work [94] has shown that the CV criterion can be improved through the use of an alignment procedure that performs a time and phase shift separately between s and both $h(\theta_{\text{bf}})$ and $h(\theta_s)$ in Eq. (A5), thereby helping to ensure that any potential phase differences between the signal and model evaluations are minimized.

The impact of this alignment is to bring the Fisher bias estimate close to the true bias values we might see in parameter estimation. The time shift and phase shift both rotate the Fisher bias about the true values of the signal by shifting the mean, as visualized in Fig. 9. Here, the posterior probability distribution for SUR-3, injected at an SNR of 250, is plotted in green alongside a multivariate normal distribution in blue produced with mean $\theta_s + \Delta \theta_{\text{sys}}$ and covariance Γ^{-1} computed using Eqs. (A5) and (A6), after

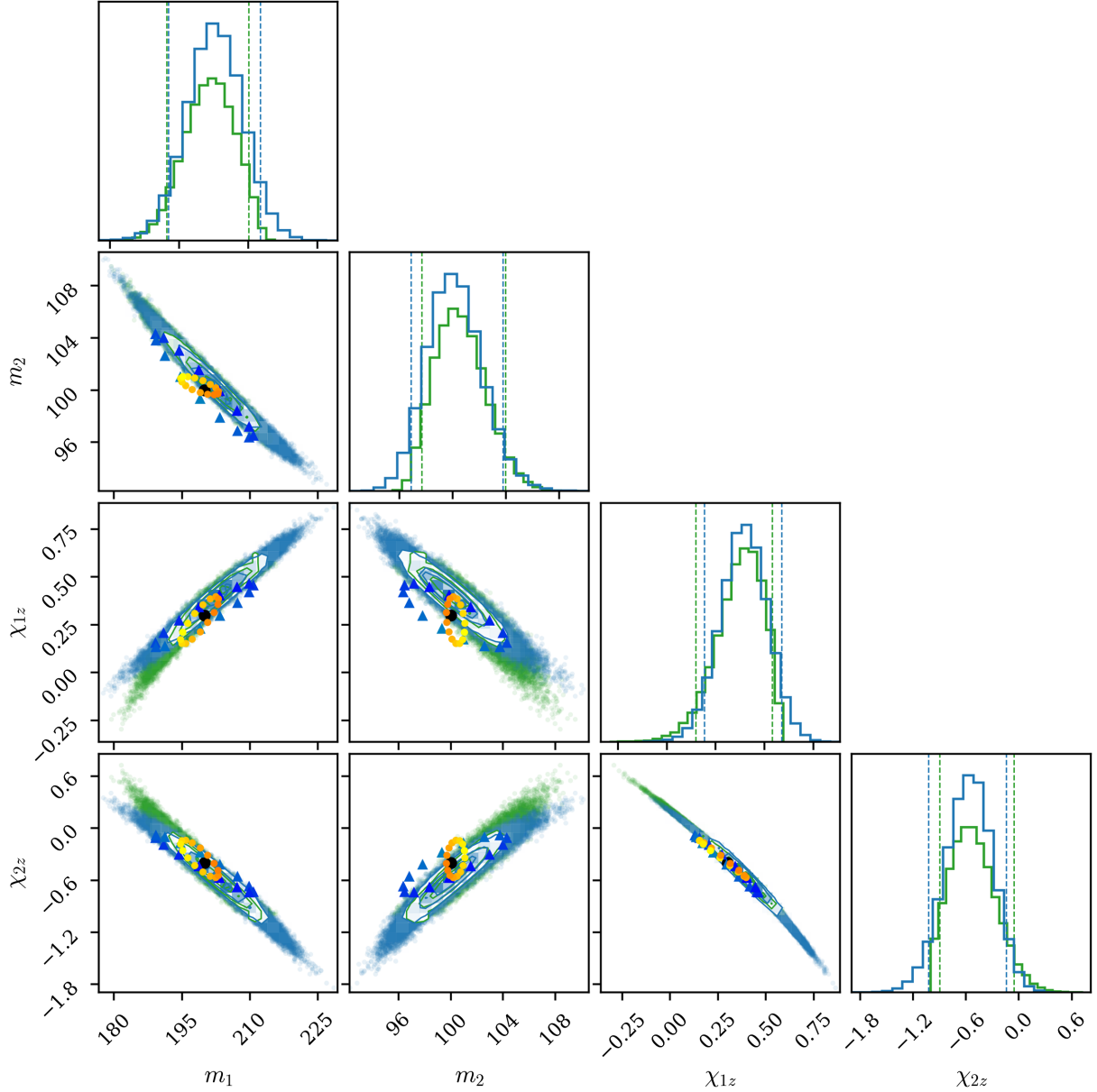


FIG. 9. The posterior distribution for SUR-3 shown in green alongside a multivariate normal distribution with mean and covariance computed using Fisher analysis, as outlined in Appendix A, having applied a time- and phase-shift alignment. The blue-green triangles show the impact of a phase shift ranging between $[0, 2\pi]$ applied to $h(\theta_{\text{bf}})$ in Eq. (A5), while the orange-yellow dots show the impact of the same phase shift applied to $h(\theta_s)$ in the same equation.

employing the alignment procedure. Waveform derivatives are calculated using the finite-difference approximations summarized in Appendix B. We see that the Fisher approximation works well in reproducing the posterior except near some of the posterior tails, where nonlinear correlations begin to appear. The black dots denote the location of the true parameters.

The variation in the mean of the Fisher samples under different phase shifts is shown in Fig. 9. After aligning the signals in time, we apply an arbitrary phase shift ranging between $[0, 2\pi]$, shown as dots when applied to $h(\theta_s)$ and as triangles when applied to $h(\theta_{\text{bf}})$. We can see that the

phase shift rotates the Fisher bias about the true parameters. Finally, the importance of this alignment procedure is made clear in Table VI, where we have computed the same estimates of ρ_{Fisher} as in Table V expect without using the alignment procedure. One notices that the bias SNRs in this case are dramatically lower, implying that the larger phase difference in $s - h(\theta_s)$ dramatically decreases our estimation of model accuracy.

We also remark on the choice of parameters used in Eq. (A5). As derived, this equation requires us to evaluate the model h at both the best-fitting point in parameter space, θ_{bf} , when computing the waveform derivatives and

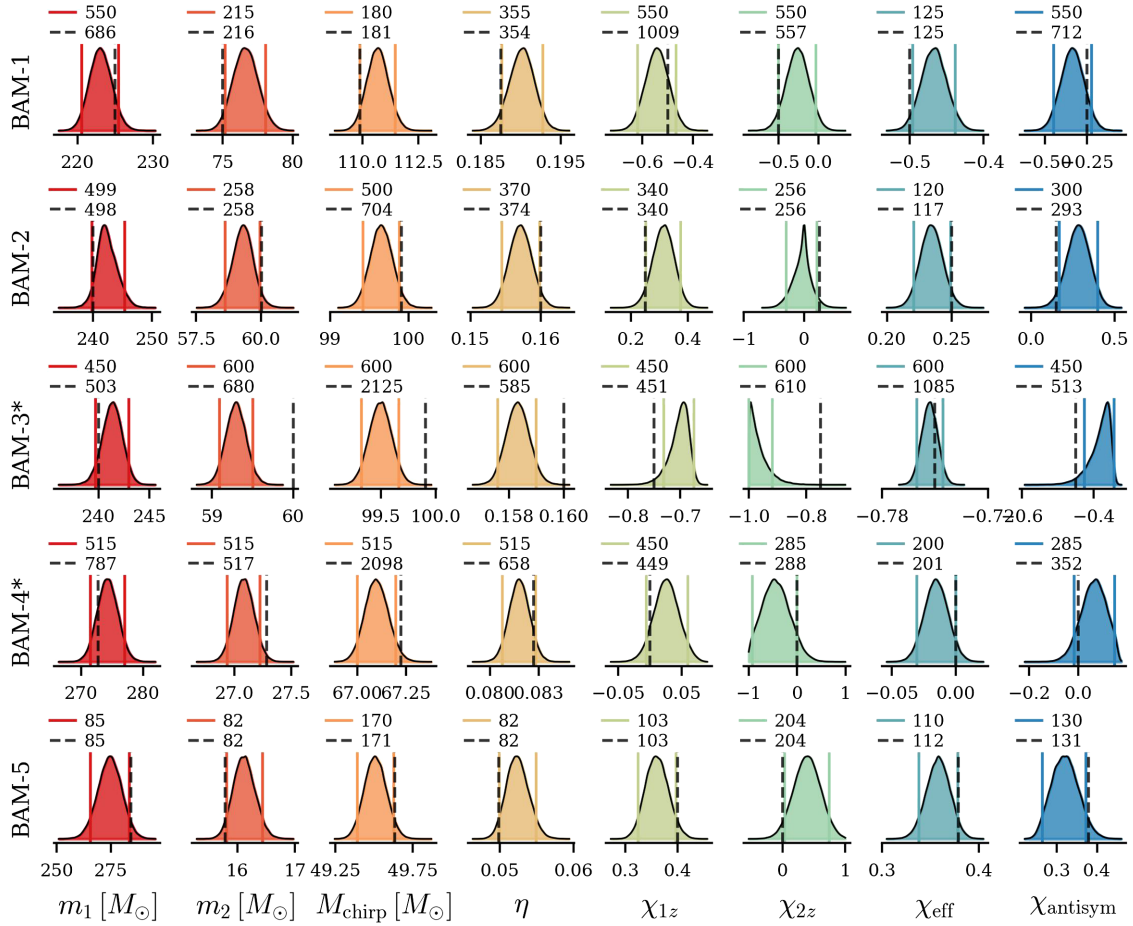


FIG. 10. One-dimensional marginalized posteriors for parameter estimation of the five BAM cases, one in each row, listed in Table I. Each column corresponds to a separate parameter listed at the bottom of the figure. We plot the true injected value as a dashed black line and the 90% CI as solid vertical lines. The legend for each plot shows the injected SNR next to the solid line and the predicted parameter bias SNR next to the dashed line, as discussed in Sec. VA. The asterisks denote cases where railing in χ_{2z} influences the SNR prediction, as discussed in Sec. VC.

Fisher matrix, and at the true injection values θ_s when computing the signal difference $s - h(\theta_s)$. Often times in the literature one sees this fact overlooked or, at least, not

clearly distinguished, and the impact of using one set of parameters rather than both is something we wish to clarify. We compute the Fisher SNRs in Eq. (17) using either only

TABLE V. Fisher bias SNRs computed from Eq. (17) and applying the time- and phase-shift alignment procedure outlined in Appendix A. The asterisks denote signals for which the parameter estimation is heavily impacted by the χ_{2z} prior bound, discussed in Sec. VC. We leave all values as computed for comparison but caution that SNR values above 500 may not be reliable given the numerical accuracy thresholds used in this work.

Simulation ID	m_1	m_2	M_{chirp}	η	χ_{1z}	χ_{2z}	χ_{eff}	χ_{antisym}
BAM-1	623	212	198	336	969	567	134	732
BAM-2	475	250	747	360	330	249	114	284
BAM-3*	310	407	25184	326	251	282	1221	265
BAM-4*	1293	733	1570	1011	532	303	180	383
BAM-5	80	75	174	78	100	221	101	128
SUR-1	24164	422	115	974	874	3958	51	2569
SUR-2	23602	8316	1578	31018	1589	1526	764	1620
SUR-3	900	11074	218	1609	517	675	168	585
SUR-4	455	208	186	307	377	255	40	303

TABLE VI. Fisher bias SNRs computed from Eq. (17) but without applying the time- and phase-shift alignment procedure outlined in Appendix A. The asterisks denote signals for which the parameter estimation is heavily impacted by the χ_{2z} prior bound, discussed in Sec. V C. We leave all values as computed for comparison but caution that SNR values above 500 may not be reliable given the numerical accuracy thresholds used in this work.

Simulation ID	m_1	m_2	M_{chirp}	η	χ_{1z}	χ_{2z}	χ_{eff}	χ_{antisym}
BAM-1	302	19	11	40	180	62	10	94
BAM-2	7	8	6	7	6	6	17	6
BAM-3*	27	19	30	21	149	271	97	322
BAM-4*	32	21	99	27	14	9	7	11
BAM-5	7	7	11	7	11	43	9	16
SUR-1	20	26	12	22	19	20	20	20
SUR-2	152	103	15	1321	98	422	7	182
SUR-3	6	6	5	6	5	5	6	5
SUR-4	3	2	4	3	3	3	475	3

the parameters θ_s or θ_{bf} and the alignment procedure, with results presented in Table VII for θ_s and in Table VIII for θ_{bf} .

When only using θ_s in Eq. (A5), the bias estimates produced increase, thereby moderately lowering the estimated SNR at which the model will show bias. The results compared to those in Table V show larger differences between the two Fisher calculations than between the

TABLE VII. Fisher bias SNRs computed from Eq. (17) using only the parameters θ_s and applying the time- and phase-shift alignment procedure outlined in Appendix A. The asterisks denote signals for which the parameter estimation is heavily impacted by the χ_{2z} prior bound, discussed in Sec. V C. We leave all values as computed for comparison but caution that SNR values above 500 may not be reliable given the numerical accuracy thresholds used in this work.

Simulation ID	m_1	m_2	M_{chirp}	η	χ_{1z}	χ_{2z}	χ_{eff}	χ_{antisym}
BAM-1	601	197	180	316	798	475	121	607
BAM-2	420	236	901	328	298	231	97	262
BAM-3*	336	573	2589	398	272	323	2619	295
BAM-4*	806	519	1485	667	472	295	195	361
BAM-5	79	80	132	78	87	146	115	105
SUR-1	751	999	107	2788	479	1214	49	705
SUR-2	36274	7355	1611	24298	1655	1561	737	1660
SUR-3	747	7216	212	1176	485	585	165	530
SUR-4	220	139	266	177	210	172	39	189

Fisher analysis and the bias distance estimates discussed in Sec. V B, with an average relative difference of 25%. The results of using only θ_{bf} are expectedly worse, where the improved difference between signals $s - h(\theta_{\text{bf}})$ greatly underestimates the bias, causing the bias SNR to greatly overestimate the accuracy of the model. The only case for which this does not hold is BAM-3, which is impacted severely by the prior bound on χ_{2z} and discussed in

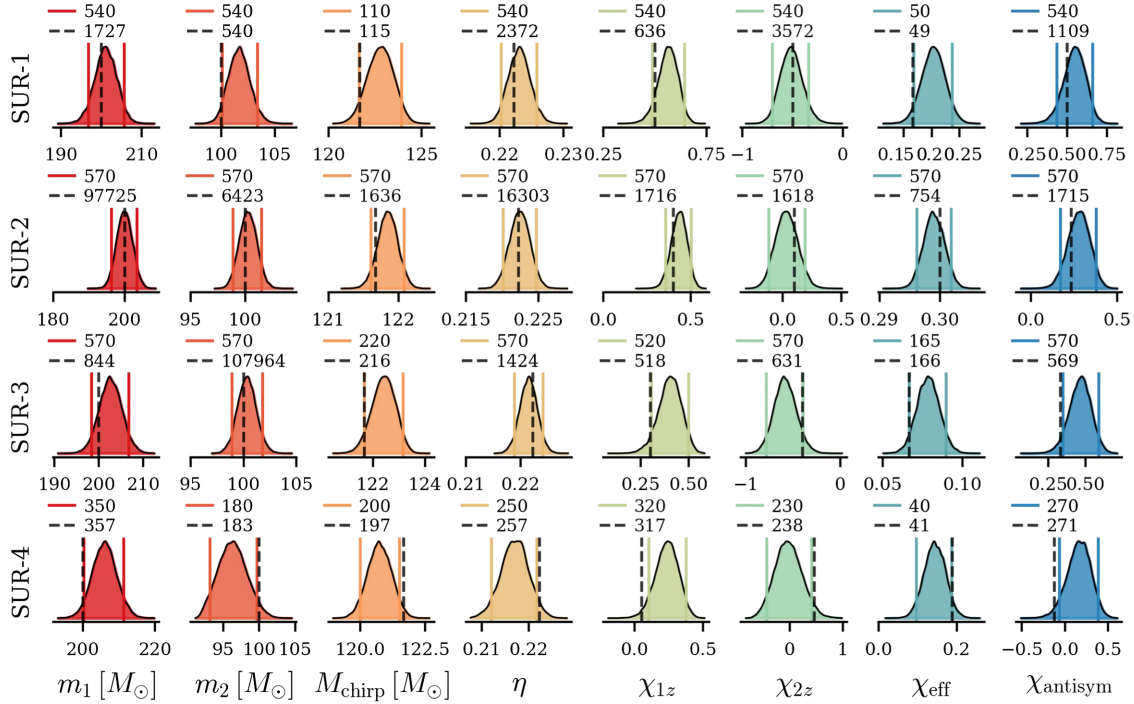


FIG. 11. One-dimensional marginalized posteriors for parameter estimation of the four SUR cases, one in each row, listed in Table I. Each column corresponds to a separate parameter listed at the bottom of the figure. We plot the true injected value as a dashed black line and the 90% CI as solid vertical lines. The legend for each plot shows the injected SNR next to the solid line and the predicted parameter bias SNR next to the dashed line, as discussed in Sec. V A.

TABLE VIII. Fisher bias SNRs computed from Eq. (17) using only the parameters θ_{bf} and applying the time- and phase-shift alignment procedure outlined in Appendix A. The asterisks denote signals for which the parameter estimation is heavily impacted by the χ_{2z} prior bound, discussed in Sec. V C. We leave all values as computed for comparison but caution that SNR values above 500 may not be reliable given the numerical accuracy thresholds used in this work.

Simulation ID	m_1	m_2	M_{chirp}	η	χ_{1z}	χ_{2z}	χ_{eff}	χ_{antisym}
BAM-1	22504	17353	5114	417209	6053	5335	3596	5537
BAM-2	10129	7786	30028	9102	14947	15737	11788	15214
BAM-3*	843	466	568	551	660	559	490	603
BAM-4*	64329	9900	6250	21901	230639	174300	27077	9696014
BAM-5	14638	17164	15499	15271	28389	76884	11574	77380
SUR-1	9066	8823	13757	8935	9632	9482	42356	9542
SUR-2	14469	12252	86069	12387	9163	8487	4433	8432
SUR-3	2162	2602	1641	2314	1985	2077	2186	2029
SUR-4	3375	3683	3286	3476	3590	3700	4524	3641

Sec. V C. The condition numbers of the Fisher matrices computed in this analysis are large, but we have verified the robustness of the results in the tables provided in Appendix C by comparing the results of the bias SNR estimates after adding uniform random noise several orders of magnitude larger than the inverse condition number to each Fisher matrix before inversion [95]. Signal-to-noise ratio estimates above 1000 are more sensitive to this added noise, but the leading-order results hold in these cases.

Our overall conclusion from this exercise is that care should be taken when computing bias estimates with Fisher analyses to apply the alignment procedure and use both appropriate sets of parameters in Eq. (A5).

APPENDIX B: WAVEFORM MODEL DERIVATIVES

The Fisher analysis outlined in Appendix A requires differentiating the waveform model h , and we discuss our approach to waveform differentiation in this appendix. The waveform model we use in this analysis, PHENOMD, is written in C-code inside of the LALSuite software library [87] and is not readily amenable to modern approaches to function differentiation, like autodifferentiation [96]. While Python libraries exist to compute derivatives of PHENOMD, such as RIPPLE [97], we decided to implement a simpler framework for waveform derivatives.

Certain parameters in θ are straightforward to differentiate with respect to in PHENOMD due to the simple functional dependence of h on these parameters. This fact

is (implicitly) outlined in Appendix A of Ref. [94]. For the luminosity distance d_L , coalescence time t_c , and coalescence phase φ_c , the partial derivative of h can be analytically written as

$$\frac{\partial h}{\partial d_L} = -\frac{h}{d_L}, \quad (\text{B1})$$

$$\frac{\partial h}{\partial t_c} = -2\pi i f h, \quad (\text{B2})$$

$$\frac{\partial h}{\partial \varphi_c} = -i h. \quad (\text{B3})$$

For all other parameters no simple functional dependence exists, so we compute these derivatives numerically using fourth-order finite difference stencils. For a function $f(x)$ and some finite step size Δx , the centered fourth-order finite difference stencil is

$$\frac{df}{dx} \approx \frac{f(x-2\Delta x) - 8f(x-\Delta x) + 8f(x+\Delta x) - f(x+2\Delta x)}{12\Delta x}. \quad (\text{B4})$$

On rare occasions, in particular near the boundaries of parameter priors, we may need to use a forward or backward directed stencil instead of the centered stencil. These expressions are given by

$$\left. \frac{df}{dx} \right|_{\text{forward}} \approx -\frac{3f(x+4\Delta x) - 16f(x+3\Delta x) + 36f(x+2\Delta x) - 48f(x+\Delta x) + 25f(x)}{12\Delta x}, \quad (\text{B5})$$

$$\left. \frac{df}{dx} \right|_{\text{backward}} \approx \frac{3f(x-4\Delta x) - 16f(x-3\Delta x) + 36f(x-2\Delta x) - 48f(x-\Delta x) + 25f(x)}{12\Delta x}. \quad (\text{B6})$$

The stencils all require specification of a step size Δx . For the differentiation of PHENOMD with respect to various parameters, we do not know *a priori* what appropriate step size to choose at any given point in parameter space. Instead we guess an initial step size, $\Delta\theta^i = 2^{-11}$, for the chosen parameter θ^i and compute derivatives at this chosen $\Delta\theta^i$ as well as at a coarser resolution $2\Delta\theta^i$ and a finer resolution $\Delta\theta^i/2$. After computing the numerical derivative at these three initial step sizes, we inspect the relative difference of the overlap, Eq. (1), between each increasingly finer resolution. If the relative difference in the

overlap is below 10^{-8} , we choose the middle step size resolution $\Delta\theta^i$. If not, we decrease all step sizes by 2 and repeat until convergence or after six iterations.

APPENDIX C: BIAS SNR RESULTS

We present the tabulated parameter bias SNRs, which are discussed in Sec. V and Appendix A, along with complete 1D posterior plots for the parameter bias estimates given in Sec. V.

-
- [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. X* **6**, 041015 (2016); **8**, 039903 (E) (2018).
 - [3] R. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Phys. Rev. X* **11**, 021053 (2021).
 - [4] R. Abbott *et al.* (LIGO Scientific Collaboration and VIRGO Collaboration), *Phys. Rev. D* **109**, 022001 (2024).
 - [5] R. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration), *Phys. Rev. X* **13**, 041039 (2023).
 - [6] R. Abbott, T. Abbott, K. Ackley, C. Adams, V. Adya, C. Affeldt, M. Agathos *et al.*, *Living Rev. Relativity* **23**, 1 (2020).
 - [7] D. Reitze *et al.*, *Bull. Am. Astron. Soc.* **51**, 035 (2019), <https://inspirehep.net/literature/1743201>.
 - [8] M. Evans *et al.*, [arXiv:2306.13745](https://arxiv.org/abs/2306.13745).
 - [9] M. Punturo *et al.*, *Classical Quantum Gravity* **27**, 084007 (2010).
 - [10] S. Hild *et al.*, *Classical Quantum Gravity* **28**, 094013 (2011).
 - [11] M. Maggiore *et al.* (ET Collaboration), *J. Cosmol. Astropart. Phys.* **03** (2020) 050.
 - [12] A. Abac *et al.*, [arXiv:2503.12263](https://arxiv.org/abs/2503.12263).
 - [13] P. Amaro-Seoane *et al.* (LISA Collaboration), [arXiv:1702.00786](https://arxiv.org/abs/1702.00786).
 - [14] S. Babak, A. Petiteau, and M. Hewitson, [arXiv:2108.01167](https://arxiv.org/abs/2108.01167).
 - [15] M. Colpi *et al.* (LISA Collaboration), [arXiv:2402.07571](https://arxiv.org/abs/2402.07571).
 - [16] J. Healy and C. O. Lousto, *Phys. Rev. D* **102**, 104018 (2020).
 - [17] D. Ferguson *et al.*, [arXiv:2309.00262](https://arxiv.org/abs/2309.00262).
 - [18] E. Hamilton *et al.*, *Phys. Rev. D* **109**, 044032 (2024).
 - [19] M. A. Scheel *et al.*, [arXiv:2505.13378](https://arxiv.org/abs/2505.13378).
 - [20] A. Nagar *et al.*, *Phys. Rev. D* **98**, 104052 (2018).
 - [21] 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).
 - [22] G. Pratten, S. Husa, C. Garcia-Quiros, M. Colleoni, A. Ramos-Buades, H. Estelles, and R. Jaume, *Phys. Rev. D* **102**, 064001 (2020).
 - [23] 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).
 - [24] L. Pompili *et al.*, *Phys. Rev. D* **108**, 124035 (2023).
 - [25] L. S. Finn, *Phys. Rev. D* **46**, 5236 (1992).
 - [26] L. Lindblom, B. J. Owen, and D. A. Brown, *Phys. Rev. D* **78**, 124020 (2008).
 - [27] S. T. McWilliams, B. J. Kelly, and J. G. Baker, *Phys. Rev. D* **82**, 024014 (2010).
 - [28] M. Hannam, S. Husa, F. Ohme, and P. Ajith, *Phys. Rev. D* **82**, 124052 (2010).
 - [29] E. Baird, S. Fairhurst, M. Hannam, and P. Murphy, *Phys. Rev. D* **87**, 024035 (2013).
 - [30] K. Chatziioannou, A. Klein, N. Yunes, and N. Cornish, *Phys. Rev. D* **95**, 104004 (2017).
 - [31] A. Toubiana and J. R. Gair, [arXiv:2401.06845](https://arxiv.org/abs/2401.06845).
 - [32] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Classical Quantum Gravity* **34**, 104002 (2017).
 - [33] M. Pürrer and C.-J. Haster, *Phys. Rev. Res.* **2**, 023151 (2020).
 - [34] C. Cutler and M. Vallisneri, *Phys. Rev. D* **76**, 104018 (2007).
 - [35] Q. Hu and J. Veitch, *Phys. Rev. D* **106**, 044042 (2022).
 - [36] D. Markovic, *Phys. Rev. D* **48**, 4738 (1993).
 - [37] C. Cutler and E. E. Flanagan, *Phys. Rev. D* **49**, 2658 (1994).
 - [38] B. J. Owen, *Phys. Rev. D* **53**, 6749 (1996).
 - [39] E. E. Flanagan and S. A. Hughes, *Phys. Rev. D* **57**, 4566 (1998).
 - [40] P. Ajith *et al.*, *Phys. Rev. Lett.* **106**, 241101 (2011).
 - [41] F. Ohme, Bridging the gap between post-Newtonian theory and numerical relativity in gravitational-wave data analysis, Ph.D. thesis, Potsdam University, 2012.
 - [42] A. J. K. Chua and C. J. Cutler, *Phys. Rev. D* **106**, 124046 (2022).
 - [43] J. Roulet, S. Olsen, J. Mushkin, T. Islam, T. Venumadhav, B. Zackay, and M. Zaldarriaga, *Phys. Rev. D* **106**, 123015 (2022).
 - [44] B. Bruegmann, J. A. Gonzalez, M. Hannam, S. Husa, U. Sperhake, and W. Tichy, *Phys. Rev. D* **77**, 024027 (2008).

- [45] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. Jiménez Forteza, and A. Bohé, *Phys. Rev. D* **93**, 044006 (2016).
- [46] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez Forteza, and A. Bohé, *Phys. Rev. D* **93**, 044007 (2016).
- [47] C. Kalaghatgi, M. Hannam, and V. Raymond, *Phys. Rev. D* **101**, 103004 (2020).
- [48] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, L. E. Kidder, and H. P. Pfeiffer, *Phys. Rev. D* **99**, 064045 (2019).
- [49] R. Balasubramanian, B. S. Sathyaprakash, and S. V. Dhurandhar, *Phys. Rev. D* **53**, 3033 (1996); , **54**, 1860(E) (1996).
- [50] P. Ajith *et al.*, *Classical Quantum Gravity* **29**, 124001 (2012); **30**, 199401(E) (2013).
- [51] J. A. Nelder and R. Mead, *Computer Journal (UK)* **7**, 308 (1965).
- [52] P. Virtanen *et al.* (SciPy 1.0 Contributors), *Nat. Methods* **17**, 261 (2020).
- [53] J. Veitch and W. Del Pozzo, Analytic marginalisation of phase parameter, Tech. Report No. T1300326 (Nikhef, Amsterdam, The Netherlands, 2013).
- [54] W. M. Farr, Marginalization of the time and phase parameters in CBC parameter estimation, Tech. Report.No T14 00460 (University of Birmingham, 2014).
- [55] L. P. Singer and L. R. Price, *Phys. Rev. D* **93**, 024013 (2016).
- [56] L. P. Singer *et al.*, *Astrophys. J. Lett.* **829**, L15 (2016).
- [57] E. Thrane and C. Talbot, *Pub. Astron. Soc. Aust.* **36**, e010 (2019); **37**, e036(E) (2020).
- [58] N. Metropolis and S. Ulam, *J. Am. Stat. Assoc.* **44**, 335 (1949).
- [59] J. Skilling, in *AIP Conference Proceedings* (AIP, 2004).
- [60] J. Skilling, *Bayesian Anal.* **1**, 833 (2006).
- [61] J. Veitch *et al.*, *Phys. Rev. D* **91**, 042003 (2015).
- [62] J. Lange, R. O’Shaughnessy, and M. Rizzo, *arXiv:1805.10457*.
- [63] G. Ashton *et al.*, *Astrophys. J. Suppl. Ser.* **241**, 27 (2019).
- [64] C. M. Biwer, C. D. Capano, S. De, M. Cabero, D. A. Brown, A. H. Nitz, and V. Raymond, *Publ. Astron. Soc. Pac.* **131**, 024503 (2019).
- [65] R. J. E. Smith, G. Ashton, A. Vajpeyi, and C. Talbot, *Mon. Not. R. Astron. Soc.* **498**, 4492 (2020).
- [66] G. Ashton and C. Talbot, *Mon. Not. R. Astron. Soc.* **507**, 2037 (2021).
- [67] M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno, and B. Schölkopf, *Phys. Rev. Lett.* **127**, 241103 (2021).
- [68] V. Tiwari, C. Hoy, S. Fairhurst, and D. MacLeod, *Phys. Rev. D* **108**, 023001 (2023).
- [69] J. S. Speagle, *Mon. Not. R. Astron. Soc.* **493**, 3132 (2020).
- [70] M. Branchesi *et al.*, *J. Cosmol. Astropart. Phys.* **07** (2023) 068.
- [71] N. Henze and B. Zirkler, *Commun. Stat.-Theory Methods* **19**, 3595 (1990).
- [72] R. Vallat, *J. Open Source Software* **3**, 1026 (2018).
- [73] J. Lin, *IEEE Trans. Inf. Theory* **37**, 145 (1991).
- [74] Z. W. Birnbaum and P. L. Meyer, *On the Effect of Truncation in Some or All Coordinates of a Multinormal Population* (University of Washington, 1951).
- [75] G. M. Tallis, *J. R. Stat. Soc. Ser. B* **23**, 223 (1961).
- [76] D. Ferguson, K. Jani, P. Laguna, and D. Shoemaker, *Phys. Rev. D* **104**, 044037 (2021).
- [77] A. Jan, D. Ferguson, J. Lange, D. Shoemaker, and A. Zimmerman, *Phys. Rev. D* **110**, 024023 (2024).
- [78] J. E. Thompson, E. Hamilton, L. London, S. Ghosh, P. Kolitsidou, C. Hoy, and M. Hannam, *Phys. Rev. D* **109**, 063012 (2024).
- [79] 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).
- [80] A. J. K. Chua, N. Korsakova, C. J. Moore, J. R. Gair, and S. Babak, *Phys. Rev. D* **101**, 044027 (2020).
- [81] M. Liu, X.-D. Li, and A. J. K. Chua, *Phys. Rev. D* **108**, 103027 (2023).
- [82] N. Afshordi *et al.* (LISA Consortium Waveform Working Group), *arXiv:2311.01300*.
- [83] M. Hannam *et al.*, *Nature (London)* **610**, 652 (2022).
- [84] C. Capano, Y. Pan, and A. Buonanno, *Phys. Rev. D* **89**, 102003 (2014).
- [85] P. Schmidt, F. Ohme, and M. Hannam, *Phys. Rev. D* **91**, 024043 (2015).
- [86] I. Harry, S. Privitera, A. Bohé, and A. Buonanno, *Phys. Rev. D* **94**, 024012 (2016).
- [87] LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration, LVK Algorithm Library—LALSuite, Free software (GPL) (2018).
- [88] A. Nitz *et al.*, gwastro/pycbc: v2.3.3 release of pycbc (2024).
- [89] C. Hoy and V. Raymond, *SoftwareX* **15**, 100765 (2021).
- [90] J. D. Hunter, *Comput. Sci. Eng.* **9**, 90 (2007).
- [91] C. R. Harris *et al.*, *Nature (London)* **585**, 357 (2020).
- [92] J. L. Synge, ed., *Relativity: The General Theory* (Elsevier Science Publishing, Amsterdam, Netherlands, 1960).
- [93] P. Mahalanobis, *Sankhya A* **80**, 1 (2018).
- [94] A. Dhani, S. Völkel, A. Buonanno, H. Estelles, J. Gair, H. P. Pfeiffer, L. Pompili, and A. Toubiana, *Phys. Rev. X* **15**, 031036 (2025).
- [95] M. Vallisneri, *Phys. Rev. D* **77**, 042001 (2008).
- [96] R. D. Neidinger, *SIAM Rev.* **52**, 545 (2010).
- [97] T. D. P. Edwards, K. W. K. Wong, K. K. H. Lam, A. Coogan, D. Foreman-Mackey, M. Isi, and A. Zimmerman, *Phys. Rev. D* **110**, 064028 (2024).