Waveform accuracy and systematic uncertainties in current gravitational wave observations

Caroline B. Owen,[1,*] Carl-Johan Haster,[2,3] Scott Perkins,[1,†] Neil J. Cornish,[4] and Nicolás Yunes[1]

[1] *Illinois Center for Advanced Studies of the Universe, Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*
[2] *Department of Physics and Astronomy, University of Nevada, Las Vegas, 4505 South Maryland Parkway, Las Vegas, NV 89154, USA*
[3] *Nevada Center for Astrophysics, University of Nevada, Las Vegas, NV 89154, USA*
[4] *eXtreme Gravity Institute, Department of Physics, Montana State University, Bozeman, MT 59717, USA*

(Dated: January 31, 2023)

The post-Newtonian formalism plays an integral role in the models used to extract information from gravitational wave data, but models that incorporate this formalism are inherently approximations. Disagreement between an approximate model and nature will produce mismodeling biases in the parameters inferred from data, introducing systematic error. We here carry out a *proof-of-principle* study of such systematic error by considering signals produced by quasi-circular, inspiraling black hole binaries through an injection and recovery campaign. In particular, we study how unknown, but calibrated, higher-order post-Newtonian corrections to the gravitational wave phase impact systematic error in recovered parameters. As a first study, we produce injected data of non-spinning binaries as detected by a current, second-generation network of ground-based observatories and recover them with models of varying PN order in the phase. We find that the truncation of higher order (>3.5) post-Newtonian corrections to the phase can produce significant systematic error even at signal-to-noise ratios of current detector networks. We propose a method to mitigate systematic error by marginalizing over our ignorance in the waveform through the inclusion of higher-order post-Newtonian coefficients as new model parameters. We show that this method can reduce systematic error greatly at the cost of increasing statistical error.

I. INTRODUCTION

Gravitational waves emitted during the coalescence of compact-object binaries offer a unique way to directly observe strong-gravity systems. Models of the signals produced by coalescence events, derived from general relativity or modified theories of gravity, are used to extract information about the astrophysical sources from the gravitational wave data. However, due to the complexity of the theories and computational time constraints, these models are necessarily approximations (either numerical, analytical, or some hybrid of the two). Mismatch between an approximate model and nature will always result in biases in the parameters inferred from data. Such bias is known as mismodeling systematic error.

In this study, we investigate mismodelling error in the context of the post-Newtonian (PN) approximation to binary inspiral signals. While contemporary analyses use more sophisticated models that include the merger and ringdown phase of the signal, we use the simpler PN inspiral regime to illustrate the effects of mismodeling and how these effects can be mitigated. Similar conclusions will apply to the full inspiral-merger-ringdown waveforms.

The approximate nature of waveform models means that some amount of systematic error is unavoidable. This systematic error is tolerable so long as it is significantly smaller than the statistical error caused by the finite signal-to-noise ratio (SNR) of the observations. Figure 1 shows a schematic diagram to illustrate why this is the case. Both panels in the figure represent a one-dimensional posterior probability distribution on a hypothetical parameter obtained by analyzing some synthetic noise-free data using some model. The locations of the injected (true) and recovered values for that parameter are indicated, where the recovered value is taken from the maximum posterior point. In this context, the systematic error is given by the difference between the recovered and injected values of the parameter. In data with noise, the offset will be a combination of systematic and statistical error. The statistical error is reflected by the width of the posterior probability distribution, and is due to the noise weighting in the likelihood.

In Case A, the systematic error is smaller than the statistical error, and therefore, the injected parameter is within the relevant credible interval of the recovered parameter. In Case B, however, the systematic error is dominant and the true value lies outside of the credible interval.

One strives to create models that are accurate enough such that the posterior probability distributions on every parameter are reliable (i.e. they resemble Case A in Fig. 1). However, statistical error scales inversely with the SNR while the systematic error is independent of the SNR. Therefore, as instruments are upgraded and improved, the systematic error can become dominant.

* cbo4@illinois.edu
† perkins35@llnl.gov

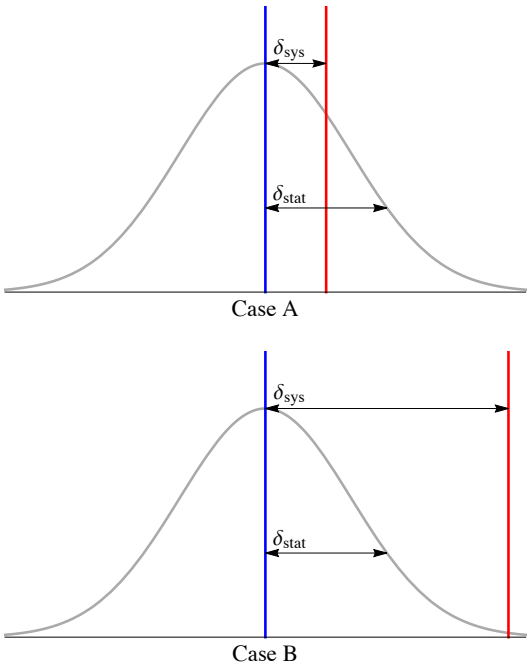

FIG. 1. Schematic diagram to illustrate the relationship between systematic error $\delta_{\rm sys}$ and statistical error $\delta_{\rm stat}$. The gray curve in each panel depict a one-dimensional, inferred posterior probability distribution on a hypothetical parameter obtained by analyzing a synthetic noise-free data set and recovering it with some model that has some amount of inherent inaccuracy with respect to the true signal. The injected (true) value of the parameter is indicated with a red line, while the recovered value, taken in this case from the maximum posterior point, is indicated with a blue line. In Case A, the systematic error is smaller than the statistical error and the value of the injected parameter lies within the credible intervals of the recovered best-fit value of the parameter. In Case B, however, the systematic error is dominant and the injected value lies outside of the credible interval.

As we progress in our capacity to observe gravitational waves, understanding how mismodeling impacts parameter estimation becomes ever more important.

The importance of bias from mismodeling depends on the accuracy of the models used, but the construction of models is a complicated matter due to the intrinsic nonlinearity of the theory and the broad parameter space of interest to us. On solar system scales, Newtonian gravity is adequate for most calculations. However, in the extreme environments that produce the gravitational waves observable by ground-based detectors, Newtonian gravity does not suffice. Instead, one must account for post-Newtonian corrections to Newtonian gravity. These corrections are derived through a formal expansion of the field equations (Einstein's or otherwise) in powers of $(v/c)^2$ [1], where v is the characteristic speed of the system and c is the speed of light and gravity.

During the quasi-circular inspiral of compact objects, a waveform based on the PN approximation has been shown to be highly accurate when enough terms are kept in the small-velocity expansion [2], but this is not the case for sufficiently massive binaries. As the total mass of the system increases, the merger and post-merger parts of the coalescence signal become more dominant, and the classic pre-merger inspiral PN expansion is no longer sufficient. This fact has has led to the creation of two classes of inspiral-merger-ringdown (IMR) models: effective-one-body (EOB) waveforms [3–6] and phenomenological waveforms [7–10]. The phenomenological framework defines the waveform as a piecewise function, with one piece modeling the inspiral, one piece the late inspiral and merger, and one piece the post-merger. The inspiral piece is constructed from the classic 3.5PN order approximation[1], enhanced with 4, 4.5, 5 and 5.5PN terms that are calibrated to a suite of EOB and numerical relativity simulations. The late inspiral and merger piece and the post-merger piece are modeled in a similar way, by fitting coefficients in an ansatz function to numerical relativity simulations when needed. Both the EOB and the phenomenological waveforms have been validated by showing high matches against a set of numerical relativity simulations [11] and are used today for gravitational wave parameter estimation by the LIGO/Virgo collaboration [12–15].

The post-Newtonian approximation plays a critical role in the creation of waveforms [2] either because it is later resummed as it is in EOB waveforms or because it is enhanced with fitting coefficients as it is in both the phenomenological and EOB models. However, the inherent approximation will necessarily introduce systematic uncertainties, and therefore, systematic bias has been studied in depth over the last three decades [18–21].

Early studies of mismodeling relied mostly on the Fisher information, an approximation to the full likelihood that is only accurate for sufficiently loud signals in Gaussian noise [22]. Cutler and Vallisneri [18] developed a Fisher information formalism to estimate mismodeling bias and applied it to a 3.5PN inspiral signal recovered with a 3PN inspiral model. Focusing on binary black hole signals detected by LISA [23] with SNR of 1000, they concluded that it is possible for systematic error to dominate the statistical error by several orders of magnitude for intrinsic parameters like masses and spins. This paper was preceded by the work of Canitrot [19], who used match filtering techniques to show that recovering a 2.5PN binary black hole signal detected by Virgo [13] with a 2PN model can produce systematic error of order $\sim 1\%$ in the chirp mass and order $\sim 50\%$ in the symmetric mass ratio.

Fisher studies, however, are not necessarily accurate for low SNR events, like those detected by current

[1] A term of order $(v/c)^{2n}$ relative to the leading-order term is considered to be of nPN order.

[2] One exception to this is the family of Numerical Relativity Surrogate waveform models, constructed directly from catalogues of numerical relativity waveforms [16, 17].

ground-based detectors, and thus, a Bayesian approach is preferred. To this end, recent studies have employed injection and recovery campaigns: synthetic data (with a set of known injection parameters) is generated using some model, and then the data is analyzed with another model. Through this approach, one can then determine whether systematic bias is a concern by comparing the injected parameters to the posterior probability distributions obtained.

Littenberg et al. performed such an analysis, injecting numerical relativity waveforms and recovering with EOB models [20]. Focusing on stellar-mass binary black holes observed by advanced LIGO [12] and Virgo[13], they found that at SNRs < 50 systematic error was smaller than or comparable to statistical error for a wide range of mass ratios. However, for SNRs > 100, as is expected from 3G detectors, systematic error can dominate. More recently, Pürrer and Haster also injected numerical relativity waveforms and recovered with various semi-analytic frequency-domain models [21]. Focusing on 3G events at SNRs $\in [466, 2598]$ that spend much longer in the 3G sensitivity band than those studied in [20], the authors concluded that systematic error induced by these semi-analytic models needs to be reduced by three orders of magnitude.

In spite of all of this work in the study of systematic errors, no work has yet been done to employ Bayesian methods to assess the inaccuracies of the latest waveform models with respect to our ignorance of (unknown) high-order PN order terms in the era of current ground-based detectors. This is a problem because evidence already exists from observations during the O3 LIGO/Virgo campaign that systematic error may be influencing posterior probability distributions [14, 15]. Indeed, for certain O3 events, the LIGO/Virgo collaboration found somewhat different posterior probability distributions for various source parameters when analyzing the data with the `IMRPhenomXPHM` [10] or `SEOBNRv4PHM` [6] models. The analyses of GW191109_163120 with these two models produced differences in the inferred spins and mass ratios, analyses of GW191219_010717 produced differences in the inferred inclination angle, total mass and distance, analyses of GW200129_065458 produced differences in the evidence for precession and the inferred mass ratio, and analyses of GW200208_222617 produced a multimodal mass posterior probability distribution with different waveforms preferring difference modes [15].

Given the above, in this paper we embark on a *proof-of-principle* study to determine whether the unknown, yet numerically-calibrated, PN terms in the `IMRPhenomD` [7, 8] waveform model can systematically bias the extraction of intrinsic source parameters (like the masses) at SNRs consistent with current and near-future observations with LIGO/Virgo/KAGRA [24]. We focus on inspiraling black hole binaries of mass ratios that are detected by a current-generation gravitational wave detector network at SNRs of $\mathcal{O}(10)$. We consider an *inspiral-only* `IMRPhenomD` model created by stopping the `IMRPhenomD` waveform at the transition frequency between the inspiral and intermediate regions of the waveform. We produce injected data from this model and recover the injected data using models inspired by the same model but truncated at 5, 4.5, 4 or 3.5PN order in the inspiral phase. Using standard Bayesian inference packages, we then explore the parameter space of interest to construct posterior probability distributions. From these distributions we can then estimate the statistical and systematic errors in the inferred parameters.

Overall, we find that the truncation of the `IMRPhenom` family even at 5PN order can lead to systematic biases larger than statistical uncertainties in the chirp mass, mass ratio and individual masses already at the SNRs expected in the fourth observing run of the LIGO/Virgo/KAGRA detectors. Of course, this does not necessarily imply that the `IMRPhenomD` family contains these intrinsic errors, since the 4, 4.5, 5 and 5.5PN terms are never set to zero in any of the `IMRPhenomD` models used in the analysis of real astrophysical observations. Our results, however, do imply that fitting errors in the 4, 4.5, 5, and 5.5PN coefficients as well as the absence of higher order, currently-unknown PN terms could introduce systematic errors that are not currently being accounted for. These fitting errors arise due to both the finite and discrete nature of the numerical relativity simulation sets used in the fit, as well as possibly the choice of the fitting functions.

In order to deal with these potential systematic uncertainties, we propose a method to ameliorate them: to include our ignorance of the missing terms directly in the model and then to marginalize over them. More specifically, we propose that uncertainties be included in the waveform model as higher PN order parameters, with physically uninformed (uniform) priors that encapsulate the residual from the fits. We then propose to marginalize over these systematic-uncertainty parameters simultaneously while exploring the astrophysically informative parameter space. We show that the effect of the inclusion of these new systematic parameters is to significantly reduce the bias (by pushing the peak of the posterior probability distribution back to the injected value), at the cost of increasing the statistical error (by widening the distribution).

The remainder of this paper is structured as follows: in Sec. II we review the waveform models used in our analyses, followed by Sec. III where we lay out the details of the experimental design of our injection and recovery campaign. The results of this campaign are presented in Sec. IV. In Sec. V, we explore a way to account for our ignorance of the true waveforms by marginalizing over missing terms in our model. And in Sec. VI we provide discussion on the results obtained in this study. In what follows, we will work in geometric units where $G = 1 = c$.

II. A TRUNCATED IMRPHENOMD MODEL TO REPRESENT MISMODELING

In this section, we provide a summary of the IMRPhenomD waveform approximant, which is a phenomenological frequency domain waveform model for spin-aligned black hole binaries [7, 8]. In doing so, we generally follow the notation of [7, 8]. The frequency domain strain is a complex function, expressed in polar form as

$$\tilde{h}(f,\boldsymbol{\theta}) = A(f,\boldsymbol{\theta})e^{-i\phi(f,\boldsymbol{\theta})}, \tag{1}$$

where f is the frequency of the gravitational wave and $\boldsymbol{\theta}$ is a vector of waveform parameters. This approximant includes only the dominant quadrupolar radiation mode. The phenomenological modelling of the strain is broken up into three regions: the inspiral at low frequencies, the post-merger and ringdown at high frequencies, and an intermediate region for the plunge and merger.

In this proof-of-principle study, we will make some simplifications to render the problem tractable. First, we focus on the inspiral region alone, as this is where the PN formalism is valid. This simplification limits our analysis to relatively low (total) mass sources, such that the total SNR from the whole coalescence is dominated by the inspiral region. Second, we focus only on the accuracy of the phase of the inspiral frequency-domain strain $\phi(f,\boldsymbol{\theta})$ and leave the amplitude as is. We do so because the phase of the gravitational wave signal provides the most constraining information for the inference of binary parameters. In order to minimize the effects of both amplitude modulations and subdominant waveform harmonics, we limit our study to quasi-circular binary inspirals, with black hole spins aligned or anti-aligned with the orbital angular momentum and with comparable mass ratios. In spite of these approximations, the study we carry out here will still allow conclusions that should apply generally to a large number of gravitational wave sources.

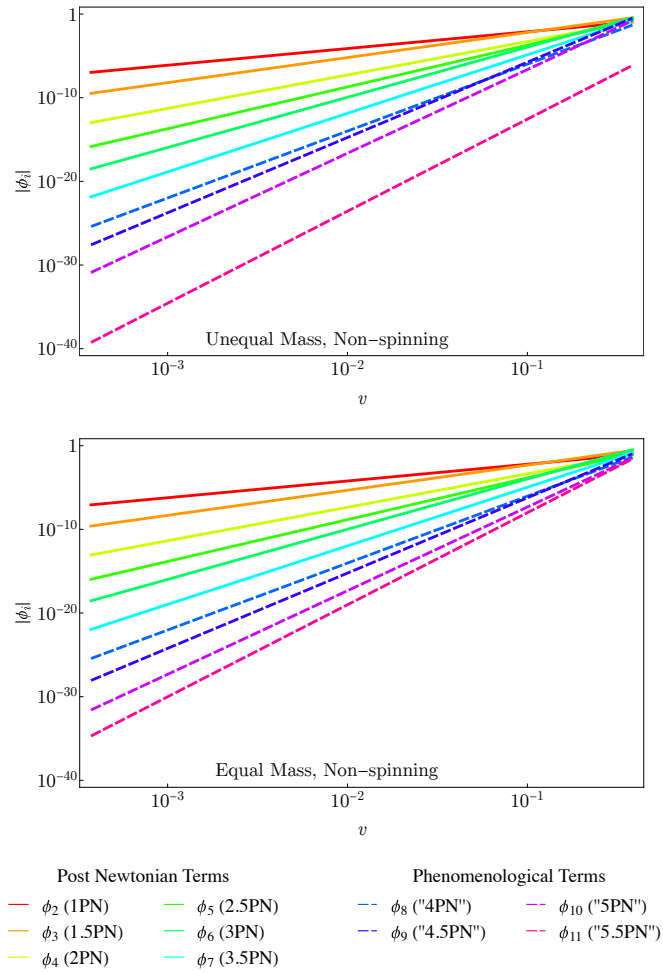

FIG. 2. Absolute values of series functions $\phi_i = \phi_i(f,\boldsymbol{\theta})$ of the IMRPhenomD inspiral phase, as given in Eqs. (2-6), plotted as a function of velocity $v = (\pi M f)^{1/3}$ for the systems presented in Table I. Observe that the terms in the phase have a hierarchy, with the lowest PN order terms larger than the higher PN order terms.

The IMRPhenomD phase during the inspiral

$$\phi_{\rm Ins}(f,\boldsymbol{\theta}) = \phi_{\rm TF2}(f,\boldsymbol{\theta}) + \phi_{\rm phenom}(f,\boldsymbol{\theta}) \tag{2}$$

is a hybrid model. The first term in the above equation comes from the TaylorF2 model, which derives from the stationary phase approximation of the waveform constructed in a standard Taylor expansion within PN theory [25]. This phase is known to 3.5PN order and it can be expressed as

$$\begin{aligned}\phi_{\rm TF2}(f,\boldsymbol{\theta}) = \; & 2\pi f t_c - \varphi_c - \pi/4 \\ & + (\pi f M)^{-5/3}\sum_{i=0}^{7}\phi_i(f,\boldsymbol{\theta})\,,\end{aligned} \tag{3}$$

where the TaylorF2 series functions are

$$\phi_i(f,\boldsymbol{\theta}) = \frac{3}{128\eta}(\pi f M)^{(i/3)}\varphi_i\,, \qquad 0 \le i \le 7, \tag{4}$$

and the PN coefficients φ_i are functions of the masses and the magnitude of the spin angular momenta [8]. Above, $M = m_1 + m_2$ is the total mass and $\eta = m_1 m_2/M^2$ is the symmetric mass ratio, determined by the component masses m_1 and m_2 where by convention we assume $m_1 \geq m_2$. The time t_c and phase φ_c of coalescence are, for the purpose of data analysis, just a reference time and phase offset. Since the orbital velocity of the system is $v = (\pi M f)^{1/3}$, we see that $\phi_{\rm TF2}(f,\boldsymbol{\theta})$ is a Frobenius series in powers of v, with a controlling factor that goes as v^{-5}. In a PN series of this type, a term proportional to $v^{2n} \propto f^{2n/3}$ relative to the controlling factor is said to be of nPN order.

The second term in Eq. (2) contains phenomenological corrections that are expected to enter at 4PN order and above within PN theory. More specifically, the IMRPhenomD model postulates the 5.5PN-accurate phe-

nomenological corrections to the phase [3]

$$\phi_{\rm phenom}(f,\boldsymbol{\theta}) = (\pi f M)^{-5/3} \sum_{i=8}^{11} \phi_i(f,\boldsymbol{\theta}) \,, \qquad (5)$$

where the phenomenological series functions are

$$\phi_i(f,\boldsymbol{\theta}) = \frac{3\pi^{5/3}}{(i-5)\eta}(Mf)^{i/3}\sigma_{i-7} \,, \quad 8 \le i \le 11, \qquad (6)$$

The σ_i coefficients are unknown functions of the masses and spin angular momenta. The `IMRPhenomD` model makes use of an *ansatz* where the σ_i are represented as bi-variate polynomials in η and $(\chi_{\rm PN} - 1)$. Here, $\chi_{\rm PN}$ is a certain function of the dimensionless spins χ_1 and χ_2, η and M. More precisely, the `IMRPhenomD` model represents the σ_i coefficients via

$$\sigma_i = \sum_{j=0}^{2}\sum_{k=0}^{2} \lambda^i_{jk}\, \eta^j \, \left(\chi_{\rm PN} - 1\right)^k \,. \qquad (7)$$

The fitting coefficients λ^i_{jk} are determined by fitting the waveform phase in the frequency domain to the phase of the Fourier transform of a finite set of numerical relativity simulations and EOB waveforms [8]. The fits will be susceptible to statistical fitting error, fitting error due to the finite nature and finite accuracy of the numerical relativity waveforms used to do the fits, and fitting error due to the functional form of the ansatz from Eq. (7). Nonetheless, in the `IMRPhenomD` model, one picks the best-fit values for these λ^i_{jk} coefficients and disregards these fitting uncertainties.

The phenomenological terms of the waveform phase present a similar convergence behaviour to the `TaylorF2` terms, as shown in Fig. 2. That is, the individual contributions of the 4, 4.5, 5 and 5.5PN terms are smaller than that of proceedings terms, presenting a PN hierarchy up to the upper bound of the inspiral region, which in the `IMRPhenomD` model is defined at $v_{\rm max}/c = (.018\pi)^{1/3} \sim 0.38$. This suggests that we could consider the phenomenological terms as behaving as higher PN order terms, even though they formally do not need to. In order to make this distinction clear, we will put quotation marks around the PN order of phenomenological terms, for example referring to the σ_4 term as a "5.5PN" term. We use this different notation to remind ourselves that the σ_i coefficients in the phase are not derived directly from the underlying theory (GR), but rather, are obtained from fitting, which in itself carries both systematic and statistical uncertainties.

[3] Note that the controlling (leading) factors of the `TaylorF2` and phenomenological Frobenius series as defined above differ by a factor of $\pi^{-5/3}/128$. In principle, from the basics of PN theory, the controlling factors should be the same, but we will not rescale them here to stay close to historical conventions.

III. INJECTION AND RECOVERY CAMPAIGN

In this paper, we set out to understand how PN corrections to the inspiral phase impact systematic error in intrinsic parameters estimated from binary black hole inspirals. The systematic error present in a given signal will, as detectors improve, eventually become dominant over statistical error, as the latter scales inversely with the SNR of the signal. We therefore ask the following question: As a function of the PN order of the recovery model and the SNR present in an inspiral signal, at what point does the systematic error become larger than the statistical error? In order to answer this question, we will perform a synthetic injection and recovery campaign. In this section, we present the details of the injection and the recovery models.

Given an advanced LIGO/Virgo observation, one does not know the exact waveform generated by Nature or the exact binary parameters that produced the signal. Therefore, the synthetic injected data will be created with the full `IMRPhenomD` inspiral model, for a set of systems detailed in Table I. The recovery model will also be the `IMRPhenomD` inspiral model, but truncated at a given PN order (as listed in Table II), since we have seen that the phenomenological terms represent PN corrections.

Our choice of the `IMRPhenomD` model will not affect the conclusions of this work because both the injection and recovery models are built from this base. Therefore, our analysis will be subject to the same inherent model inaccuracies in both the injection and the recovery. In turn, this implies that our analysis will only be sensitive to the systematic errors we are introducing ourselves through truncation of the PN terms in the phenomenological Fourier phase. A similar analysis could be carried out with the updated `IMRPhenomXAS` model[9], or extended to include modifications to the merger and ringdown, but we leave this for future work.

The details of the injection parameters that define the systems we study are the following. Our analysis will focus only on the inspiral regime, and therefore, although we consider two separate binary configurations, we choose a total mass of $20 M_\odot$ for both. This guarantees that the SNR is dominated by the inspiral part of coalescence. One injected configuration will have two bodies of equal mass, while the other will have one body that is three times more massive than the other. In both cases, we choose to set the spin to zero. The SNR scales inversely with the luminosity distance D_L of the binary, so for each configuration, we choose injected luminosity distances to obtain the SNRs listed in Table I. We do so to roughly fix the statistical error (which predominantly scales as 1/SNR) across injected configurations. We select SNRs to encompass the sensitivities of our current (second-generation, 2G) detectors, as well as that of near future detectors. Other details of the system parameters used in the injections can be found in Table I.

The SNR and the results of our parameter estimation studies will depend on the spectral noise density of the

TABLE I. Properties of injected binary system configurations. Here, m_i is the mass of the i-th black hole and D_L is the luminosity distance between Earth and the source that ensures the listed SNR as measured by a current detector network. In addition to the parameters listed in the table below, we set the aligned dimensionless spins $\chi_{1,2} = 0$, the inclination angle $\iota = 0.4$ rad, the polarization angle $\psi = 2.659$ rad, the right ascension ra $= 1.375$ rad, the declination dec $= -1.2108$ rad, the coalescence phase $\phi_c = 1.3$ rad, and the geocentric coalescence time $t_c = 1126259642.413$ sec for each injection.

	m_1 ($\mathrm{M}_\odot$)	m_2 ($\mathrm{M}_\odot$)	D_L(Mpc)	SNR
Equal Mass	10	10	629.853	20
	10	10	314.926	40
	10	10	157.463	80
Unequal Mass	15	5	534.780	20
	15	5	267.390	40
	15	5	133.695	80

TABLE II. Injection and recovery models used in our analyses. The coefficients σ_i are the phenomenological coefficients contained in Eq. 5.

	Label	Coefficients set to zero
Injection model:	"5.5PN"	none
Recovery models:	"5.5PN"	none
	"5PN"	$\sigma_4 = 0$
	"4.5PN"	$\sigma_3 = \sigma_4 = 0$
	"4PN"	$\sigma_2 = \sigma_3 = \sigma_4 = 0$
	3.5PN	$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 0$

detectors assumed to have measured the synthetic injected signals. To represent a current detector network, we focus on a set of ground-based detectors, comprised of LIGO Hanford [12], LIGO Livingston [12], and Virgo [13]. In particular, we use analytic approximations for the design sensitivities of advanced LIGO and advanced Virgo [26], which should approximately correspond to that of the fourth-observing run.

To perform the parameter estimation analyses, we use the Bayesian inference library BILBY [27, 28]. This library obtains its waveform approximants from the LALSimulation [29] package and provides access to several Bayesian inference packages, including `dynesty` [30], the nesting sampling package that we use for this project. In each parameter estimation study, we vary over all model parameters of the `IMRPhenomD` model except for the dimensionless spins, namely chirp mass, $\mathcal{M}$, mass ratio q, luminosity distance D_L, right ascension ra, declination dec, inclination angle ι, polarization angle ψ, coalescence phase ϕ_c, and coalescence time t_c. All analyses are performed using 1024 live points for each of the three duplicate analyses done for each model variant. Otherwise, our analyses settings (both sampling configuration, and prior choices) match those from recent LIGO/Virgo/KAGRA publications [14, 15, 31, 32]. After exploring the 9-dimensional likelihood surface, we will present a subset of corner plots (to avoid crowding the presentation), focusing only on some astrophysically relevant parameters.

With these details in hand, we perform parameter estimation analyses on each injection configuration with each recovery model, leaving us with 30 analyses to complete. When injecting the data and exploring the likelihood surface, we set an upper bound on the frequency at the end of the inspiral region ($M f_{\max} = 0.018$) to isolate the impact of modifying the inspiral model. For the binaries considered here with a total mass of $M = 20\mathrm{M}_\odot$, this corresponds to a frequency of $f_{\max} = 182.7$Hz.

The truncation of the frequency range used is applied only to the evaluation of the likelihood itself without any additional termination conditions applied to the underlying waveforms used, this in order to avoid parameter biases introduced by such unphysical sharp waveform features as explored in [33, 34]. We will comment later on how our results are impacted by a different choice of maximum frequency.

IV. SYSTEMATIC AND STATISTICAL UNCERTAINTIES THROUGH BAYESIAN PARAMETER ESTIMATION

In this section, we present the results of the Bayesian parameter estimation analyses laid out in Sec. III. In particular, we will focus on the systematic errors introduced into the chirp mass $\mathcal{M} = \eta^{3/5} M$ and the mass ratio $q = m_2/m_1 < 1$ due to truncation of the inspiral `IMRPhenomD` model. As we will see, strong correlations with the reference time t_c will also force us to include this parameter in the partial corner plots we present to explain our results. We emphasize again that even though we present corner plots for only a subset of the parameters of the `IMRPhenomD` model, we do vary over all the parameters of the model (except the spins), as discussed in the previous section.

Let us first consider the unequal mass injected signal with SNR = 80. The left panel of Fig. 3 presents the partial corner plot obtained from our Bayesian parameter estimation analyses, using each of the recovery models detailed in Table II. The figure depicts one-dimensional marginalized posterior probability distributions and the 90% credible region contours of the marginalized two-dimensional posterior distribution. While our focus is on $\mathcal{M}$ and q, we also include the reference (geocentric) coalescence time t_c, because its correlation with mass parameters helps to explain the results. Observe that the posterior probability distributions obtained using the "5.5PN" model are centered on the injected values, indicated with black vertical lines. This is as expected because the recovery model is identical to the one used to produce the signal injection. Therefore, there is no systematic error introduced by the model and any small bias present is due

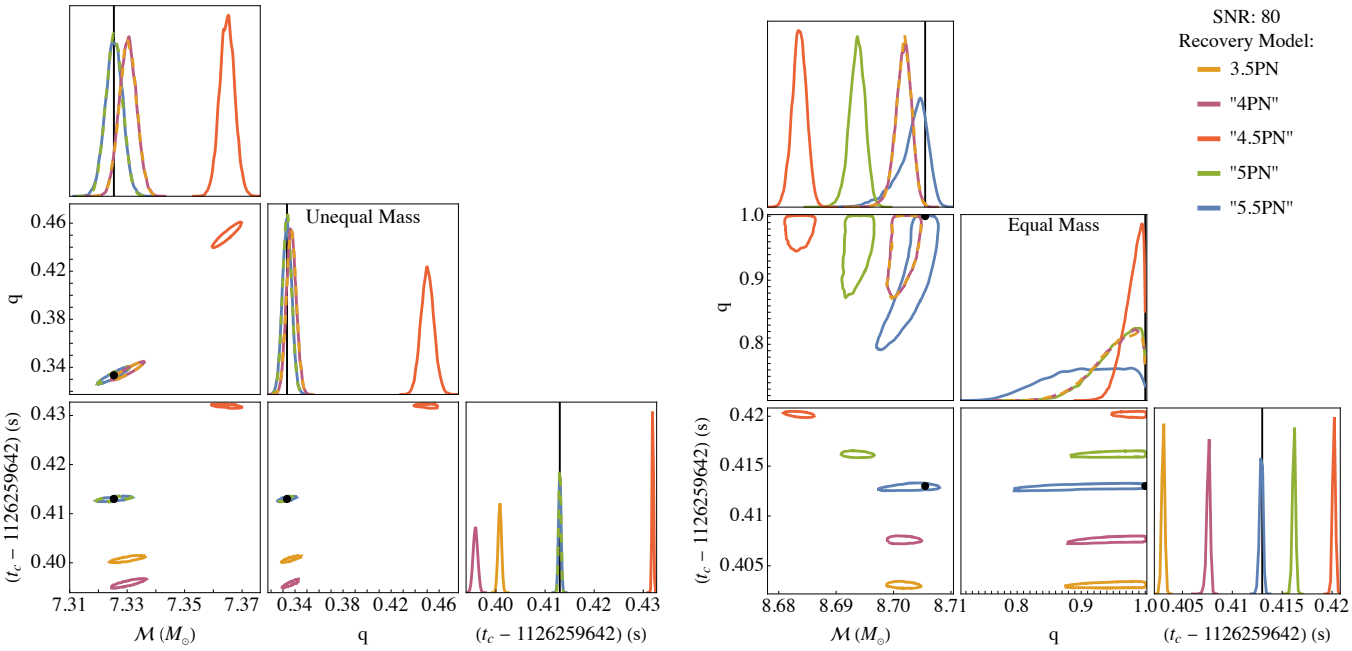

FIG. 3. Partial corner plot produced after a Bayesian parameter estimation study, using each of the recovery models detailed in Table II to infer parameters from the unequal mass (left) and equal mass (right), SNR 80, injected signals, described in Table I. The corner plots show the marginalized one-dimensional posterior probability distributions and the 90% credible region contours of the two-dimensional posterior probability distributions on chirp mass $\mathcal{M}$, mass ratio q, and (geocentric) reference time t_c. The injected "true" values are shown as black dots and vertical lines. Observe that the bias is very small for the "5.5PN" model in both cases, because the the recovery model is the same as the injection model. The bias is also suppressed for the "5PN" model in the unequal mass case where the removed "5.5PN" is very small. However, in the equal mass case where the "5.5PN" term has a larger contribution, the bias from the "5PN" model is significant. The bias of the "4.5PN" model is very large in both cases, because of the alternating structure of the phenomenological coefficients. The bias becomes smaller again, but is still not negligible, for the 3.5 and "4PN" models. This suggests that the phenomenological coefficients of the `IMRPhenomD` model are important for parameter estimation and their inaccurate determination could lead to systematic bias.

to sampling error. Observe also that the "5PN" model preforms similarly well, this behavior can be explained from Fig. 2. For the unequal mass injection, the magnitude of the "5.5PN" term is significantly smaller than that of any other term. Therefore, the systematic error introduced by removing this term from the model is negligible and again the small bias is dominated by sampling error.

The 3.5PN, "4PN" and "4.5" models, however, show a different behavior in the left panel of Fig. 3. Both the 3.5PN and "4PN" models produce a slight bias from the injected values in the mass parameters, with very similar posterior probability distributions. On the other hand, the posterior probability distributions obtained with the "4.5PN" model indicate a large bias with negligible support at the injected values. At first glance, these results are surprising, because one would naïvely expect the "4.5PN" model to be more accurate, and thus produce a smaller bias, than the 3.5 and "4PN" models. However, some of the phenomenological coefficients have alternating signs: while the "4PN", "5PN" and "5.5PN" terms are all positive on the relevant range of parameters, the"4.5PN" term is negative (Fig 2 plots absolute values) but of similar magnitude to the "4PN" term. This means that the "4.5PN" term approximately cancels the "4PN" term, and so truncating the model at this order produces a less accurate model. Once this term is removed to produce the "4PN" model we are able to more accurately recover the mass parameters.

This alternating behavior of the series is not just a feature of the phenomenological coefficients but also a known feature of the classic PN expansion. Indeed, it has been known for a while that parameter estimation with a 2.5PN model is less accurate than with a 2PN model, even though the former is formally more accurate [35]. This is precisely because of the alternating structure of the series. In the case of the `IMRPhenomD` model, however, the higher than 3.5PN order coefficients are all phenomenological, so it is not clear that this alternating sequence should continue, or that the magnitude of the terms in all of parameter space have been accurately determined through fitting. This can only be determined by either calculating the higher PN order term, or re-doing the fits with a denser set of numerical relativity simulations and quantify the statistical uncertainty of the fits.

The similarity between the one-dimensional posterior

probability distributions in the left panel of Fig. 3 found using the 3.5PN and "4PN" models can be explained by studying the Fourier phase model of Eqs. (2)-(5). The total phase includes the term $2\pi f t_c$, which has the same linear dependence on frequency as the "4PN" term when accounting for the control factor of the PN expansion. The presence of this term compensates for the removal of the "4PN" term and allows the 3.5PN model to recover similar mass parameters at the cost of biasing t_c. We see this in the left panel of Fig. 3: while the posteriors in the mass parameters coincide, the posteriors on t_c show a large bias between the two models.

The same qualitative conclusions hold for other mass ratios, although for equal-mass injections one can encounter boundary effects from the prior, as shown in the right panel of Fig. 3. The right panel is analogous to the left panel of Fig. 3, but for the equal-mass binary with the same SNR. The posterior probability distributions for $\mathcal{M}$, for example, are very similar to that of the unequal mass case. However, because there is not such a large difference between the "5.5PN" and "5PN" terms, as shown in Fig. 2, we see that the systematic bias in $\mathcal{M}$ grows with the "5PN" and "4.5PN" models, and then decreases for the 3.5PN and "4PN" models.

The right panel of Fig. 3 also shows that the posterior probability distribution for q is affected by the boundary of the q prior (at $q = 1$), and thus it provides less information about systematic bias. This boundary is largely artificial, as it is set by the convention of $m_1 \geq m_2$, but nonetheless, it reduces the accessible parameter space. While, for the unequal-mass binary, a mismatch between the full and truncated `IMRPhenomD` waveforms could be "corrected" by introducing a bias that either increases or decreases the mass ratio around the true value, the equal-mass case is only provided with freedom in one direction.

Let us now study how the above conclusions are affected by the SNR of the signal. Figure 4 illustrates the impact of the SNR on the relationship between systematic and statistical error, focusing on only three recoveries, all with the "4PN" model but each with a signal of different SNR. As expected, at the lower SNRs, the injected values are contained within the credible contours of the recovered posterior probability distribution. However, at the highest SNR, this is no longer the case and the posterior probability distributions have less support at the injected values. Clearly then, as the SNR of the observed signals increases, models that were once accurate enough to recover parameters reliably are no longer sufficient, as anticipated.

Let us now collect all of our results and study how the systematic and statistical uncertainties behave with SNR for different recovery models, as shown in Fig. 5. To do so, we take the statistical error $\delta_{\rm stat}\theta$ on a given parameter θ to be half of the 90% credible interval on the one-dimensional marginalized posterior probability distribution on that parameter. We then take the bias in the recovered parameter to be $\Delta\theta = |\theta_{\rm rec} - \theta_{\rm inj}|$,

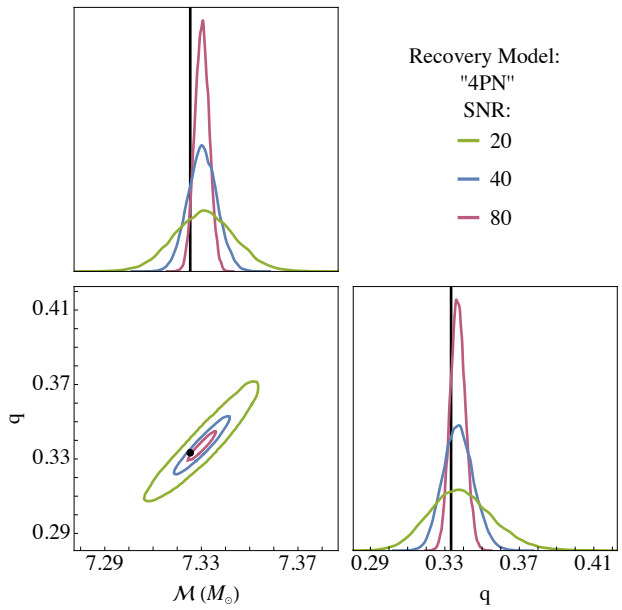

FIG. 4. Partial corner plot produced using the "4PN" recovery model detailed in Table II to estimate parameters from each of the unequal mass injected signals described in Table I. We depict the marginalized one-dimensional posterior probability distributions and 90% credible region contours of the two-dimensional posterior probability distributions on chirp mass $\mathcal{M}$ and mass ratio q. The injection values are indicated in black. Observe that, as expected, as the SNR increases, the width of the posterior probability distribution shrinks, reducing the support of the posterior on the injected values. This figure indicates clearly that models once thought to be accurate enough at small SNR can rapidly become insufficiently accurate at higher SNRs.

where $\theta_{\rm inj}$ is the injected value and the recovered value $\theta_{\rm rec}$ is determined by the maximum of the full nine-dimensional posterior probability distribution [4]. Figure 5 shows $\log_{10}(\Delta\theta/\delta_{\rm stat}\theta)$ as a function of SNR of the injection, for each recovery model and for both the unequal mass (left) and equal mass (right) systems, with $\theta = \mathcal{M}$ (top) and $\theta = q$ (bottom). In the case of the equal mass configuration, we use a one-sided credible interval as the statistical error when producing plots in the right panels of Fig. 5 because the injected value $q = 1$ is at the boundary of the allowed range.

[4] It is essential that the set of recovered parameters $\boldsymbol{\theta}_{\rm rec}$ is taken from the maximum of the full N-dimensional posterior probability distribution (where N is the number of parameters varied in the Bayesian analysis) rather than the maximum of marginalized one- or two-dimensional posterior probability distributions such as those depicted in Fig 3. While we have verified that the peaks depicted in the corner plots in this paper correspond closely to the peaks of the full posterior probability distribution, this need not always be the case. Especially in the case of highly correlated parameters, it is possible that the maximum of a marginalized posterior probability distribution does not line up with the maximum of the full posterior probability distribution.

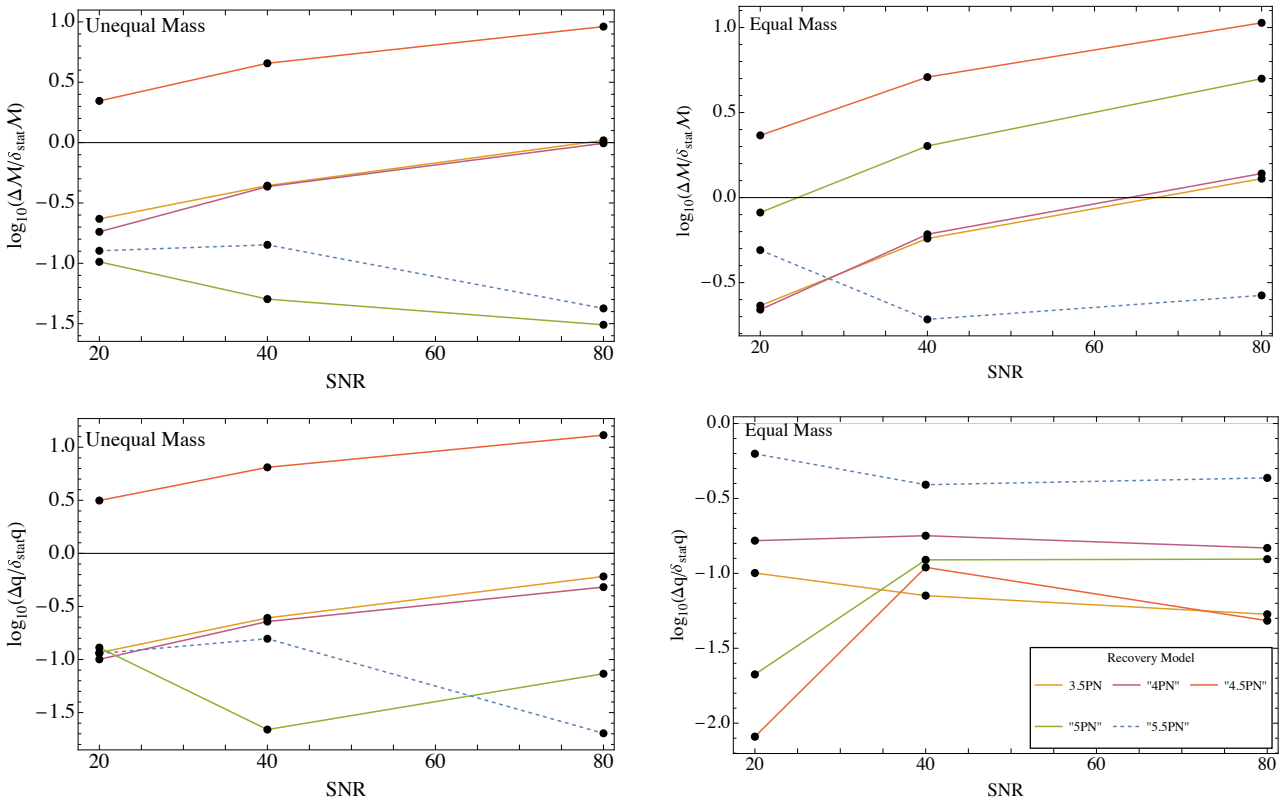

FIG. 5. Ratio of bias to statistical error in the chirp mass (top) and mass ratio (bottom) for various recovery models (Table II) and injected parameter (Table I), including the unequal mass cases (right) and the equal mass cases (left). The black dots indicate the configurations of the parameter estimation analyses we performed, which we join through a nearest-neighbour linear interpolation. When the recovery model matches the injection model (i.e. in the "5.5PN" case), the bias is due entirely to sampling, and it is thus distinguished by plotting it with dashed lines. Observe that for the 3.5PN, "4PN," and "4.5PN" recovery models in the unequal mass case and 3.5PN, "4PN," "4.5PN," and "5PN" recovery models in the equal mass case, where the bias is dominated by systematic error, the bias grows with SNR relative to the statistical error. Observe that this trend is violated for the mass ratio in the equal-mass case due to prior boundary effects. Observe also that systematic bias grows as "PN" terms are removed, until the "4.5PN" term is removed. Most significantly, note the instances where $\log_{10}(\Delta\theta/\delta_{\rm stat}\theta) > 0$. In these cases the recovery model was not accurate enough to reliably recover the injection parameters.

This figure can be read and interpreted as follows. The bias for a given parameter and recovery model is identical to the statistical error when $\log_{10}(\Delta\theta/\delta_{\rm stat}\theta) = 0$. When $\log_{10}(\Delta\theta/\delta_{\rm stat}\theta) < 0$, the injected value of the parameter falls within the confidence intervals of the recovered posterior probability distribution, which one interprets as a sign of trust in the inferred point-estimate value. However, when $\log_{10}(\Delta\theta/\delta_{\rm stat}\theta) > 0$, the injected parameter is outside of the recovered confidence intervals. When this is the case, one interprets that the recovery models is not accurate enough to reliably recover the injection parameters.

Figure 5 allows us to make several observations. When the recovery model matches the injection model, as is the case for the "5.5PN" recovery model, there is no mismodeling error and $\Delta\theta$ is entirely produced due to sampling error. We include the recoveries done with this model in our results to demonstrate the bias that can be expected even in the absence of systematic error (but we represent it with dashed lines to emphasize its difference from the recoveries done with the other models). As shown in the figure, the sampling error of the "5.5PN" model is comparable to the systematic error of the "5PN" model in the unequal mass case, because, as stated before, the "5.5PN" term in this case is very small. In the equal-mass case, however, the systematic error of the "5PN" model is much larger than the sampling error, and grows with SNR. Indeed, the ratio of the systematic error to the statistical error tends to grow with SNR for all recovery models, except for those referring to the mass ratio in the equal mass case (due to prior boundary effects). Another trend we observe is that, as expected from Fig. 3, systematic bias grows as "PN" terms are removed, until the "4.5PN" term is removed, which leads to the most systematic error for the reasons explained earlier. Additionally, because of correlations with the t_c term, which as explained before enters at an effective 4PN order, the 3.5PN and "4PN" recovery models produce similar biases.

To verify that the biases seen here are in fact due to

the truncation of the higher PN corrections in our recovery models, we preformed a set of analyses where we lowered the maximum frequency of our injected signal from $Mf_{\max} = 0.018$ to $Mf_{\max} = 0.01$. We focused on our SNR=80 unequal mass injection and preformed two analyses with a lowered maximum frequency: one where all injection parameters are kept the same, resulting in an SNR < 80 signal of shorter duration, and one where the luminosity distance D_L is rescaled to maintain an SNR of 80 while leaving all other parameters unchanged. We found that, in lowering the maximum frequency, the contribution to the signal from the higher order terms is reduced, as expected. Moreover, the biases in inferred parameters become significantly smaller for the shorter signal, indicating that the biases in our original analyses are a result of the truncated PN approximation. However, we also find that the width of posterior probability distributions increases for the shorter signals, consequently increasing the statistical error. This increase is seen in both the analyses where D_L is left unchanged and where it is rescaled to maintain an SNR of 80. The reason for this is that statistical error scales inversely with SNR only approximately; in reality, the likelihood surface can have multiple valleys and peaks, none of which need to be perfectly Gaussian (as assumed when deriving the 1/SNR scaling of the statistical error with a Fisher analysis). The structure of the likelihood, in turn, depends on the duration of the signal, since shorter signals will allow for stronger correlations and more structure.

The biases introduced during the inspiral by truncating PN corrections to the phase do, of course, depend on the total mass M of the injected signal. If the power spectral density were flat, then the biases would be total mass insensitive because the inspiral frequency cutoff $f_{\max} \propto 1/M$ and so $v_{\max} = (0.018\pi)^{(1/3)}$ is independent of total mass. The power spectral density, however, is very much not flat, and it rises at lower frequencies. This implies that higher total mass signals spend less of their inspiral in band, until eventually, at a sufficiently high total mass, only the merger and ringdown are observed. Shorter in-band inspirals will lead to a less accurate posterior recovery and larger correlations with other parameters, as the signal will contain less information. This is why in our analysis we selected a total mass of $M = 20\mathrm{M}_\odot$ for all injections, so as to ensure that the SNR of the observed signal would be dominated by the inspiral, allowing us to disregard the merger and ringdown.

A useful way to quantify the disagreement between two waveforms h_1 and h_2 is the mismatch

$$MM(h_1, h_2) = \max_{t_c,\phi_c} \left[1 - \frac{(h_1|h_2)}{\sqrt{(h_1|h_1)(h_2|h_2)}} \right], \qquad (8)$$

where the inner product is defined as

$$(h_1|h_2) = 4\mathrm{Re} \int_0^\infty \frac{\tilde{h_1^*}\tilde{h_2}}{S_n(f)} df, \qquad (9)$$

with $S_n(f)$ the noise power spectral density of the detector and the asterisks denoting complex conjugation. The mismatch takes values between zero and unity, with $MM = 0$ corresponding to perfect agreement between h_1 and h_2 up to time and phase offsets (see [36] for a useful discussion on computing mismatch for quasi-circular and eccentric waveform models). Even with perfect waveform models the expected value of the mismatch between injected and recovered waveforms is non-zero due to noise. For signals in stationary, Gaussian noise, the expected mismatch is $\mathrm{E}[MM] = (D-1)/(2\,\mathrm{SNR}^2)$, where D are the number of parameters that describe the waveform model [37]. Note that this statistical error in the match scales inversely with the square of the signal-to-noise ratio, so as detectors become more sensitive, the statistical error drops very rapidly, potentially exposing various sources of systematic error from waveform mis-modeling. Bias in inferred parameters is produced as a result of disagreement between the injection and recovery models. We can therefore use mismatch as a tool to predict the relative magnitude of the bias a given recovery model will cause and in doing so analytically corroborate the Bayesian results presented in this section.

We compute the mismatch between our injected signal $h_{\text{“5.5PN”}}(\boldsymbol{\theta}_{\mathrm{inj}})$ and our recovery models injected with with our injection parameters $h_{\mathrm{x}}(\boldsymbol{\theta}_{\mathrm{inj}})$, x = {3.5PN, “4PN”,“4.5PN”,“5PN”}. When x = “5.5PN”, the mismatch is identically zero. We expect that the relative sizes of these mismatches should correlate to the relative sizes of the biases in recovered parameters produced by these models.

We consider $\boldsymbol{\theta}_{\mathrm{inj}}$ as given by the unequal and equal mass injections at SNR = 80 detailed Tab. I. In Fig. 6, we plot $MM[h_{\text{“5.5PN”}}(\boldsymbol{\theta}_{\mathrm{inj}}), h_{\mathrm{x}}(\boldsymbol{\theta}_{\mathrm{inj}})]$. Indeed, we see that the relative magnitudes of the mismatches closely correspond to the relative magnitudes of the biases in mass ratio and chirp mass depicted in Fig 3. Notably, we see that the the “5PN” recovery model in the unequal mass case most closely matches the injected signal, while the largest mismatch in both cases is produced by the “4.5PN” model. Biases in the reference time are not captured by the mismatch because it is maximized over this parameter. Through this analysis, we illustrate the connection between bias in the recovered parameters and mismatch between the injection and recovery models and we lend additional credence to the conclusions drawn in this paper.

V. MARGINALIZING OVER UNCERTAINTIES

The most striking aspect of the results presented in the previous section is the magnitude of the systematic error that we found. As we discussed in that section, we considered relatively low total mass systems (for which the total SNR from the whole coalescence is dominated by the inspiral region) and we truncated higher PN-like corrections that we have shown to be appropriately small.

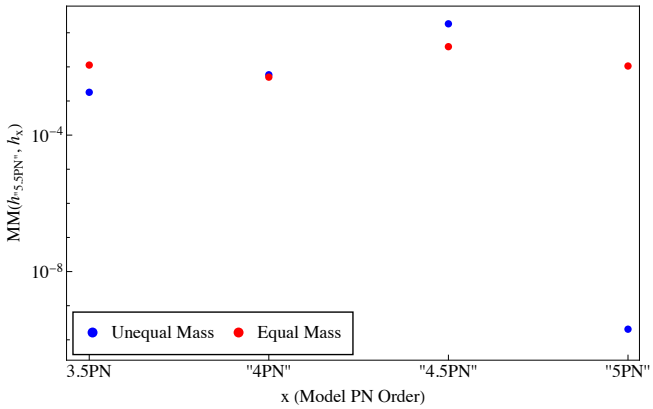

FIG. 6. Mismatch, as defined in Eq. 8, between the injected signal $h_{\text{“5.5PN”}} = h_{\text{“5.5PN”}}(\boldsymbol{\theta}_{\rm inj})$ and signals produced by injecting the recovery models with the injection parameters $h_{\rm x} = h_{\rm x}(\boldsymbol{\theta}_{\rm inj})$ We consider the SNR = 80 unequal and equal mass injection detailed in Table I. Observe that the relative magnitudes of the mismatch values are comparable to the relative magnitudes of the biases in chirp mass and mass ratio depicted in the corner plots in Fig. 3

Despite this, the posterior distributions obtained indicate significant biases in the recovered mass parameters, often with very little support at the injected values. These biases in chirp mass and mass ratio will translate to similarly large biases in the inferred component masses. Most notably, these biases occur even at SNRs expected in the next observing runs with current detectors, and will certainly become more significant at higher SNRs with for third-generation detectors. Of course, systematic bias is not necessarily going to be this large with `IMRPhenomD`-type models where one never sets the terms in the phenomenological phase to zero. However, the phase is indeed truncated at "5.5PN" order and the phenomenological coefficients do contain systematic fitting errors (due to the reasons discussed in Sec. II). In this section, therefore, we search for, propose and develop a method to ameliorate systematic uncertainties due to unknown higher PN order terms.

Previous attempts at mitigating waveform inaccuracies have shown great promise, but are often limited in scope and applicability [38–42]. Such attempts either only account for and correct model uncertainties in a subset of the full parameter space covered by the full analyses, or provide diagnostics without suggested improvements for the full parameter space. A method which is capable of both full coverage and general directive corrections is yet to emerge, but an important first step is presented in [43] basing the model for the waveform corrections around the noise-dependent indistinguishability between the waveform models under comparison.

Our proposed method will differ from these previous attempts because our philosophy will be to "parameterize our ignorance and then marginalize over it." In essence, what we propose is that, for any waveform that contains systematic inaccuracies, one should do the following:

(i) introduce a model for these inaccuracies, characterized by a set of parameters $\boldsymbol{\lambda}$,

(ii) explore the likelihood by varying over all parameters $(\boldsymbol{\theta} \cup \boldsymbol{\lambda})$,

(iii) marginalize over $\boldsymbol{\lambda}$ to produce reduced corner plots that range over $\boldsymbol{\theta}$ only.

The expected end result will be to ameliorate the systematic uncertainty at the cost of broadening the posterior probability distributions and thus increasing the statistical error.

A key element in this method is to properly model our ignorance. For the case considered in this study, our ignorance is entirely encapsulated by unknown higher PN order terms. For the `IMRPhenomD` model, this means inaccuracies (a) in the "4PN" to "5.5PN" order terms of the phenomenological phase (due to fitting inaccuracies), and (b) inaccuracies in the higher than "5.5PN" order terms (due to truncation). One way to address (a) is to promote the λ^i_{jk} constants as new parameters of the `IMRPhenomD` model, with priors taken to be the posterior probability distributions of the λ^i_{jk} fits (as opposed to simply picking numbers for these λ^i_{jk}, which would correspond to delta-function priors). This method is similar in spirit to how the equation-of-state sensitivity of approximately universal relations [44, 45] is taken into account when extracting the mass and radius of neutron stars from binary inspirals [46, 47].

Let us now focus on how to address inaccuracies due to truncation (case (b) above). The best way to model this ignorance is to use our analytic knowledge of the terms that are being ignored. For example, if we have a model that is accurate to 5.5PN order, then the next term in the series will scale as $[1/\eta(Mf)^{-5/3}]\bar{\sigma}_5(Mf)^4$, where the term in square brackets is the controlling factor of the phenomenological approximation. Technically, the 6PN term could also scale as $[1/\eta(Mf)^{-5/3}]\bar{\sigma}_5(Mf)^4 \log Mf$ (due to certain tail terms that arise in the PN approximation); however, the log-correction is mild and similar enough to the non-log term that, for the purpose of marginalizing over our ignorance, either model will suffice. One could of course include not just the 6PN term, but also the 6.5PN term simultaneously, and maybe also higher PN order terms. How many terms one must keep will depend on the accuracy of the base model and the SNR of the signal.

An important question now presents itself: how does one set the priors on these new nuisance parameters $\bar{\sigma}_i$? One option would be to use completely uninformative (i.e. flat) priors with infinite (or very large) boundaries, but this is not smart. If one were to do this, then some region of the prior would allow sufficiently large values of $\bar{\sigma}_i$, which would render the ignorance term $[1/\eta(Mf)^{-5/3}]\bar{\sigma}_5(Mf)^4$ much larger than terms at lower (and known) PN order. In fact, for sufficiently large $\bar{\sigma}_i$, the ignorance term would dominate over the rest of the

known PN series, rendering the entire approximation invalid. A better choice of prior is to use our analytic knowledge of the structure of our ignorance. Even though we do not know the $\bar{\sigma}_i$ precisely, we do know that they are functions of the system parameters $\boldsymbol{\theta}$ and we know the values of σ_i at lower (known) PN orders. We can then infer the range that $\bar{\sigma}_i$ can have based on these lower PN terms, and use this as the boundary of a flat prior. Such a methodology was recently used successfully to study tests of general relativity with the LVK implementation of the parameterized post-Einsteinian model [48].

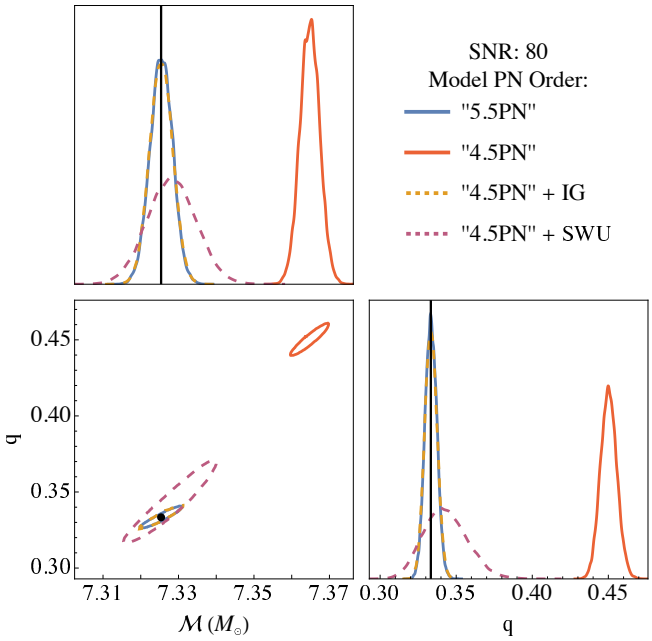

FIG. 7. Partial corner plot using the "5.5PN" and "4.5PN" recovery models, as well as the enhanced model with a "4.5PN" order base and the two priors (IG and SWU). All partial corner plots are marginalized over the parameters not shown in the figure. In particular, the enhanced "4.5PN" model are marginalized over the higher PN order terms introduced. The upper left and lower right panels contain the one-dimensional PDFs for chirp mass $\mathcal{M}$ and mass ratio q, respectively, and the lower left panel depicts the 90% credible region contours, with the injection values are indicated in black. Observe that the enhanced models remove the systematic bias introduced by the plain "4.5PN" model, at the cost of enlarging the width of the posterior probability distribution when the range of $\bar{\sigma}_{3,4}$ is assumed unknown (the SWU prior).

With all of this in mind, let us now consider a particular example. Let us focus on the "4.5PN" recovery model, because the analyses of the previous section demonstrated that this model leads to the largest systematic biases relative to statistical error for both the equal and unequal mass injections. Let us then enhance this "4.5PN" recovery model with two ignorance terms, one at "5PN" order (controlled by a new parameter $\bar{\sigma}_3$) and one at "5.5PN" order (controlled by a new parameter $\bar{\sigma}_4$). For the priors on $\bar{\sigma}_3$ and $\bar{\sigma}_4$, let us consider two cases:

- **Informed Gaussian (IG):** We collect samples on q produced during the recovery done with the "5.5PN" model. We then substitute these q samples into the $\sigma_3(q)$ and $\sigma_4(q)$ functions (defined in the `IMRPhenomD` model through the fitted polynomial ansatz) to produce two-dimensional distributions in (σ_3, σ_4). We use these distributions to fit a two-dimensional correlated Gaussian, and we use these Gaussian distributions as the priors on $\bar{\sigma}_3$ and $\bar{\sigma}_4$.

- **Super Wide Uniform (SWU):** With the prior bounds on $q \in [1/8, 1]$, we infer upper and lower prior boundaries on $\bar{\sigma}_3$ and $\bar{\sigma}_4$ via $\bar{\sigma}_{3,4}^{\max/\min} = \sigma_{3,4}(q_{\max/\min})$. With this in hand, we then assign uniform and uncorrelated priors inside that range.

While the IG prior is fully informed by the knowledge of the higher PN order terms, the SWU is much more relaxed. The way we have constructed the SWU prior does use knowledge of the higher PN order terms, but this is only a matter of convenience here. We could have, instead, used a completely agnostic approach by requiring that each $\bar{\sigma}_i$ term be smaller than the one preceding it. Such an assumption is applicable when considering inspirals, which can be modeled as PN series. We already demonstrated in Fig. 2 that the terms in the phenomenological part of the `IMRPhenomD` phase obey this requirement well. Therefore, the SWU prior ends up producing a prior that is very similar to what one would find with the agnostic approach just described.

With this enhanced recovery model defined, we then carry out a Bayesian parameter estimation analysis for the unequal mass signal injection at SNR=80, the results of which are summarized in the partial corner plot of Fig. 7. The upper left and lower right panels represent the one-dimensional marginalized posterior probability distributions on $\mathcal{M}$ and q respectively and the lower left panel depicts the 90% credible region contours of the marginalized two-dimensional posterior probability distributions. Let us compare the posterior probability distributions from the analyses done with the "5.5PN" recovery model, the standard "4.5PN" recovery model, and the enhanced "4.5PN" recovery model with the IG and SWU priors described above. Observe that the enhanced recovery model removes the systematic bias of the "4.5 PN" model. Observe also that the use of the IG prior preserves the width of the posterior probability distribution, while the SWU increases this width, therefore increasing the amount of statistical uncertainty. This is as expected because the IG prior uses knowledge of the higher PN order terms that is not truly accessible, while the SWU prior models this ignorance at the cost of a wider posterior probability distribution. This analysis provides a proof of principle that the method proposed to marginalize over our ignorance in our models can produce much more accurate recovered parameters at the cost of more statistical error.

VI. DISCUSSION

In this paper, we have presented an injection and recovery campaign performed to understand how systematic error is impacted by PN corrections to the phase of the gravitational waves emitted in the quasi-circular inspiral portion of black hole binary coalescence. We have considered injected data of equal and unequal mass, non-spinning black hole binaries for a selection of SNRs and recovered with models truncated at different PN orders in the `IMRPhenomD` phase. We have shown that even truncation at "5PN" order can lead to systematic error that dominates over statistical error at SNRs expected in the next observing runs of current detectors. This does not necessarily indicate that the `IMRPhenomD` model contains these errors, because the fits that produced the higher PN order terms of this model were done assuming the model would never be truncated at a given PN order. However, our results demonstrate the importance of these higher order corrections for accurate parameter estimation.

An important reminder that arises as a consequence of our work is that both the IMRPhenom and the EOB waveform models used in gravitational wave data analysis to date carry uncertainties even in the inspiral phase due to fits of certain (unknown high PN order) parameters. These fits are not perfect and carry both statistical error (due to the finite size of the "data" the models are fit to) and systematic error (due to the finite accuracy of the "data"). To date, these errors are not accounted for in parameter estimation, and could, in principle, affect parameter inferences for sufficiently high SNR events. Our work suggests that this systematic error need not only be relevant to observations with third-generation detectors, but they could already influence parameter inferences in the next observing runs with current detectors.

In an initial attempt to tackle this issue, we here propose a method to ameliorate the impact of systematic uncertainties in parameter estimation. Our proposal relies on the idea of creating a model to characterize our ignorance in the waveform and then to marginalize over it. In the inspiral phase, this can be done by promoting the fitting coefficients of IMRPhenom and EOB models to new waveforms parameters (with priors set by the posterior probability distributions obtained in the fit) and by introducing higher PN order terms that capture terms not included in the waveform model. We have shown that this method can all but remove the systematic error introduced by truncating the inspiral phase at a fixed PN order, at the cost of widening the posterior probability distribution, and thus, increasing the statistical uncertainties.

The work presented here opens several opportunities for further work. One such opportunity would be to investigate the inclusion of the fitting parameters of IMRPhenom and EOB models in parameter estimation, as described above. This would probably require re-doing the fits to numerical relativity simulations to account for (i) a larger set of such simulations and (ii) the finite accuracy of the simulations (with error estimates taken from the simulations themselves). Once these fits are re-done, the posterior probability distributions on the fitting parameters could be modeled through a multi-dimensional Gaussian or a kernel density estimator to prescribe priors for the fitting parameters in future parameter estimation studies. Another opportunity for future work would be to consider the impact of different (higher PN order) phenomenological functions when fitting against numerical relativity simulations in the inspiral. Ideally, the degree of the polynomial to be fitted would be determined by the information content of the numerical relativity signal, and the inherent numerical error in the simulations. The promising results of our proposed method to ameliorate systematic error presented here suggests that our ignorance of higher PN order terms (which will always exist, at any given finite PN order) is likely to not hinder future parameter inferences and the potential of gravitational waves to learn about astrophysics and fundamental physics.

ACKNOWLEDGMENTS

The authors are grateful to Katerina Chatziioannou, Jon Gair, Michael Pürrer, Leo Stein, Helvi Witek, and Simone Mezzasoma for useful comments and discussions. CBO and NY acknowledge support from NSF Grants PHY-1759615 and PHY-1949838. NJC acknowledges support from NSF Grants PHY-1912053 and PHY-2207970. The authors are also grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants PHY-0757058 and PHY-0823459. This is LIGO Document Number LIGO-P2300015.

[1] L. Blanchet, Living Rev. Rel. **5**, 3 (2002), arXiv:gr-qc/0202016.
[2] C. M. Will, Proc. Nat. Acad. Sci. **108**, 5938 (2011), arXiv:1102.5192 [gr-qc].
[3] A. Buonanno and T. Damour, Phys. Rev. D **59**, 084006 (1999), arXiv:gr-qc/9811091.
[4] A. Bohé et al., Phys. Rev. D **95**, 044028 (2017), arXiv:1611.03703 [gr-qc].
[5] A. Nagar et al., Phys. Rev. D **98**, 104052 (2018), arXiv:1806.01772 [gr-qc].
[6] S. Ossokine et al., Phys. Rev. D **102**, 044055 (2020), arXiv:2004.09442 [gr-qc].
[7] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. Jiménez Forteza, and A. Bohé, Phys. Rev. D **93**, 044006 (2016), arXiv:1508.07250 [gr-qc].

[8] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez Forteza, and A. Bohé, Phys. Rev. D **93**, 044007 (2016), arXiv:1508.07253 [gr-qc].

[9] G. Pratten, S. Husa, C. Garcia-Quiros, M. Colleoni, A. Ramos-Buades, H. Estelles, and R. Jaume, Phys. Rev. D **102**, 064001 (2020), arXiv:2001.11412 [gr-qc].

[10] G. Pratten et al., Phys. Rev. D **103**, 104056 (2021), arXiv:2004.06503 [gr-qc].

[11] I. Hinder et al., Class. Quant. Grav. **31**, 025012 (2014), arXiv:1307.5307 [gr-qc].

[12] J. Aasi et al. (LIGO Scientific), Class. Quant. Grav. **32**, 074001 (2015), arXiv:1411.4547 [gr-qc].

[13] F. Acernese et al. (VIRGO), Class. Quant. Grav. **32**, 024001 (2015), arXiv:1408.3978 [gr-qc].

[14] R. Abbott et al. (LIGO Scientific, VIRGO), (2021), arXiv:2108.01045 [gr-qc].

[15] R. Abbott et al. (LIGO Scientific, VIRGO, KAGRA), (2021), arXiv:2111.03606 [gr-qc].

[16] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, D. Gerosa, L. C. Stein, L. E. Kidder, and H. P. Pfeiffer, Phys. Rev. Research. **1**, 033015 (2019), arXiv:1905.09300 [gr-qc].

[17] T. Islam, S. E. Field, S. A. Hughes, G. Khanna, V. Varma, M. Giesler, M. A. Scheel, L. E. Kidder, and H. P. Pfeiffer, Phys. Rev. D **106**, 104025 (2022), arXiv:2204.01972 [gr-qc].

[18] C. Cutler and M. Vallisneri, Phys. Rev. D **76**, 104018 (2007), arXiv:0707.2982 [gr-qc].

[19] P. Canitrot, Phys. Rev. D **63**, 082005 (2001).

[20] T. B. Littenberg, J. G. Baker, A. Buonanno, and B. J. Kelly, Phys. Rev. D **87**, 104003 (2013), arXiv:1210.0893 [gr-qc].

[21] M. Pürrer and C.-J. Haster, Phys. Rev. Res. **2**, 023151 (2020), arXiv:1912.10055 [gr-qc].

[22] M. Vallisneri, Phys. Rev. D **77**, 042001 (2008), arXiv:gr-qc/0703086.

[23] A. R. Kaiser and S. T. McWilliams, Class. Quant. Grav. **38**, 055009 (2021), arXiv:2010.02135 [gr-qc].

[24] B. P. Abbott et al. (KAGRA, LIGO Scientific, Virgo, VIRGO), Living Rev. Rel. **21**, 3 (2018), arXiv:1304.0670 [gr-qc].

[25] T. Damour, B. R. Iyer, and B. S. Sathyaprakash, Phys. Rev. D **66**, 027502 (2002), arXiv:gr-qc/0207021.

[26] B. O'Reilly, M. Branchesi, and G. Gemme, LIGO Technical Document T2000012-v2, Tech. Rep. (2022).

[27] G. Ashton et al., Astrophys. J. Suppl. **241**, 27 (2019), arXiv:1811.02042 [astro-ph.IM].

[28] I. M. Romero-Shaw et al., Mon. Not. Roy. Astron. Soc. **499**, 3295 (2020), arXiv:2006.00714 [astro-ph.IM].

[29] LIGO Scientific Collaboration, "LIGO Algorithm Library - LALSuite," free software (GPL) (2018).

[30] J. S. Speagle, Monthly Notices of the Royal Astronomical Society (2020), arXiv:1904.02180 [astro-ph].

[31] L. S. Collaboration and V. Collaboration, (2022).

[32] L. S. Collaboration, V. Collaboration, and K. Collaboration, (2021).

[33] C. L. Rodriguez, B. Farr, W. M. Farr, and I. Mandel, Phys. Rev. D **88**, 084013 (2013), arXiv:1308.1397 [astro-ph.IM].

[34] I. Mandel, C. P. L. Berry, F. Ohme, S. Fairhurst, and W. M. Farr, Class. Quant. Grav. **31**, 155005 (2014), arXiv:1404.2382 [gr-qc].

[35] E. Poisson, Phys. Rev. D **52**, 5719 (1995), [Addendum: Phys.Rev.D 55, 7980–7981 (1997)], arXiv:gr-qc/9505030.

[36] B. Moore, T. Robson, N. Loutrel, and N. Yunes, Class. Quant. Grav. **35**, 235006 (2018), arXiv:1807.07163 [gr-qc].

[37] K. Chatziioannou, A. Klein, N. Yunes, and N. Cornish, Phys. Rev. D **95**, 104004 (2017), arXiv:1703.03967 [gr-qc].

[38] C. J. Moore, C. P. L. Berry, A. J. K. Chua, and J. R. Gair, Phys. Rev. D **93**, 064001 (2016), arXiv:1509.04066 [gr-qc].

[39] Z. Doctor, B. Farr, D. E. Holz, and M. Pürrer, Phys. Rev. D **96**, 123011 (2017), arXiv:1706.05408 [astro-ph.HE].

[40] B. Edelman et al., Phys. Rev. D **103**, 042004 (2021), arXiv:2008.06436 [gr-qc].

[41] A. Z. Jan, A. B. Yelikar, J. Lange, and R. O'Shaughnessy, Phys. Rev. D **102**, 124069 (2020), arXiv:2011.03571 [gr-qc].

[42] Q. Hu and J. Veitch, Phys. Rev. D **106**, 044042 (2022), arXiv:2205.08448 [gr-qc].

[43] J. S. Read, (2023), arXiv:2301.06630 [gr-qc].

[44] K. Yagi and N. Yunes, Science **341**, 365 (2013), arXiv:1302.4499 [gr-qc].

[45] K. Yagi and N. Yunes, Phys. Rev. D **88**, 023009 (2013), arXiv:1303.1528 [gr-qc].

[46] K. Chatziioannou, Phys. Rev. D **105**, 084021 (2022), arXiv:2108.12368 [gr-qc].

[47] B. P. Abbott et al. (LIGO Scientific, Virgo), Phys. Rev. Lett. **121**, 161101 (2018), arXiv:1805.11581 [gr-qc].

[48] S. Perkins and N. Yunes, Phys. Rev. D **105**, 124047 (2022), arXiv:2201.02542 [gr-qc].