A Parameter Estimation Method that Directly Compares Gravitational Wave Observations to Numerical Relativity

J. Lange,[1] R. O'Shaughnessy,[1] M. Boyle,[2] J. Calderón Bustillo,[3] M. Campanelli,[1] T. Chu,[4,5] J. A. Clark,[3] N. Demos,[6] H. Fong,[5,7] J. Healy,[1] D. A. Hemberger,[8] I. Hinder,[9] K. Jani,[3] B. Khamesra,[3] L. E. Kidder,[2] P. Kumar,[5] P. Laguna,[3] C. O. Lousto,[1] G. Lovelace,[6] S. Ossokine,[9] H. Pfeiffer,[5,9,10] M. A. Scheel,[8] D. M. Shoemaker,[3] B. Szilagyi,[8,11] S. Teukolsky,[2,8] and Y. Zlochower[1]

[1]*Center for Computational Relativity and Gravitation, Rochester Institute of Technology, 85 Lomb Memorial Drive, Rochester, NY 14623, USA*
[2]*Center for Astrophysics and Planetary Science, Cornell University, Ithaca, New York 14853, USA*
[3]*Center for Relativistic Astrophysics and School of Physics, Georgia Institute of Technology, Atlanta, GA 30332, USA*
[4]*Department of Physics, Princeton University, Jadwin Hall, Princeton, NJ 08544, USA*
[5]*Canadian Institute for Theoretical Astrophysics, University of Toronto, Toronto M5S 3H8, Canada*
[6]*Gravitational Wave Physics and Astronomy Center, California State University Fullerton, Fullerton, California 92834, USA*
[7]*Department of Physics, University of Toronto, Toronto M5S 3H8, Canada*
[8]*Theoretical Astrophysics 350-17, California Institute of Technology, Pasadena, CA 91125, USA*
[9]*Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Am Mühlenberg 1, 14476 Potsdam-Golm, Germany*
[10]*Canadian Institute for Advanced Research, 180 Dundas St. West, Toronto, ON M5G 1Z8, Canada*
[11]*Caltech JPL, Pasadena, California 91109, USA*

We present and assess a Bayesian method to interpret gravitational wave signals from binary black holes. Our method directly compares gravitational wave data to numerical relativity simulations. This procedure bypasses approximations used in semi-analytical models for compact binary coalescence. In this work, we use only the full posterior parameter distribution for generic nonprecessing binaries, drawing inferences away from the set of NR simulations used, via interpolation of a single scalar quantity (the marginalized log-likelihood, $\ln \mathcal{L}$) evaluated by comparing data to nonprecessing binary black hole simulations. We also compare the data to generic simulations, and discuss the effectiveness of this procedure for generic sources. We specifically assess the impact of higher order modes, repeating our interpretation with both $l \leq 2$ as well as $l \leq 3$ harmonic modes. Using the $l \leq 3$ higher modes, we gain more information from the signal and can better constrain the parameters of the gravitational wave signal. We assess and quantify several sources of systematic error that our procedure could introduce, including simulation resolution and duration; most are negligible. We show through examples that our method can recover the parameters for equal mass, zero spin; GW150914-like; and unequal mass, precessing spin sources. Our study of this new parameter estimation method demonstrates we can quantify and understand the systematic and statistical error. This method allows us to use higher order modes from numerical relativity simulations to better constrain the black hole binary parameters.

Contents

I. INTRODUCTION

On September 14, 2015 gravitational waves (GW) were detected for the first time at the Laser Interferometer Gravitational Wave Observatory (LIGO) in both Hanford, Washington and Livingston, Louisiana [1]. The LIGO Scientific Collaboration and Virgo Collaboration (LVC) concluded that the source of the GW signal was a binary black hole (BBH) system with masses $m_1 = 26.2^{+5.2}_{-3.8}$ and $m_2 = 29.1^{3.7}_{-4.4}$ that merged into a more massive black hole (BH) with mass $m_f = 62.3^{+3.7}_{-3.1}$ [2]. These parameters were estimated by comparing the signal to state-of-the-art semi-analytic models [3–5]. However, in this mass regime, LIGO is sensitive to the last few cycles of coalescence, characterized by a strongly nonlinear phase not comprehensively modeled by analytic inspiral or ringdown models. In [6], the LVC reanalyzed GW150914 with an alternative method that compares the data directly to numerical relativity (NR), which include aspects of the gravitational radiation omitted by the aforementioned models. This additional information led to a shift in some inferred parameters (e.g., the mass ratio) of the coalescing binary.

In this work, we assess the reliability and utility of this novel parameter estimation method in greater detail. For clarity and relevance, we apply this method to synthetic data derived from black hole binaries qualitatively similar to GW150914. Previous work [6] demonstrated by example that this method could access information about GW sources using higher order modes that was not presently accessible by other means. In this work, we demonstrate the utility of this method with a larger set of examples, showing we recover (known) parameters of a synthetic source more reliably when higher order modes are included. More critically, we present a detailed study of the systematic and statistical parameter estimation errors of this method. This analysis demonstrates that these sources of error are under control allowing us to identify source parameters and conduct detailed investigations into subtle systematic issues, such as the impact of higher order modes on parameter estimation. For simplicity and to best leverage the most exhaustively explored region of binary parameters, our analysis emphasizes simulations of nonprecessing black hole binaries as in [6], particularly simulations with mass ratios and spins that are highly consistent with GW150914.

The paper is outlined as follows. Section II lists the simulations used in the study (both for our template bank and synthetic sources), describes our method of choice with regards to waveform extraction, and briefly describes the method (see Section III in [6]). Section III describes the diagnostics used in our assessment of the systematics, illustrating each with concrete examples. Section IV describes several sources of error and their relative impact on our results. Section V presents 3 end-to-end runs, $q = 1$ zero spin; $q = 1.22$ anti-aligned (GW150914-like); and $q = 1.23$ short precessing, including both $l \leq 2$ and $l \leq 3$ (for the GW150914-like) results. Section VI summarizes our findings. Appendix A includes more end-to-end studies that use intrinsically different sources to explore more of the parameter space using our method. For context, the same method used to analyze GW150914 has also been applied to synthetic data using numerical relativity simulations [7].

II. METHODS AND INPUTS

A. Numerical relativity simulations

A numerical relativity (NR) simulation of a coalescing compact binary can be completely characterized by its intrinsic parameters, namely its individual masses and spins. We parameterize the binary using the mass ratio $q = m_1/m_2$ with the convention $q \geq 1$ ($m_1 \geq m_2$) and the dimensionless spin

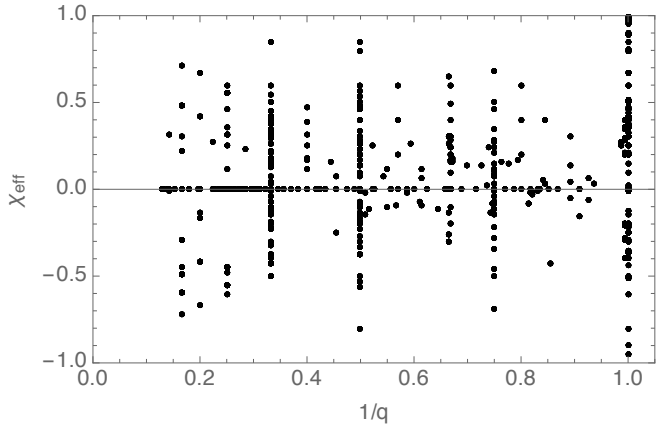

FIG. 1: **NR template bank**: An illustration of all the simulations used in this study in the 2D space of $1/q$ and $\chi_{\rm eff}$ [Eq. (2)]. Combined with our interpolation methods, the wide range of mass ratios and spins represented in this illustration allow us to reproduce binary parameters for much of the parameter space.

parameters

$$\boldsymbol{\chi_i} = \boldsymbol{S_i}/m_i^2. \tag{1}$$

where $i = 1, 2$ indexes the component black holes in the binary. With regard to spin, we define another dimensionless parameter that is a combination of the spins [8–10]:

$$\chi_{\rm eff} = (\boldsymbol{S_1}/m_1 + \boldsymbol{S_2}/m_2) \cdot \hat{L}/M. \tag{2}$$

Figure 1 illustrates our NR template bank, with each simulation represented as a point in the $\chi_{\rm eff}, q$ plane. Finally we quantify the duration of each simulation signal by a dimensionless parameter $M\omega_0$, corresponding to the dimensionless starting binary frequency measured at infinity.

For a given simulation, the GW strain $h(t, r, \hat{n})$ can be characterized by a spin-weighted spherical harmonic decomposition at large enough distance: $h(t, r, \hat{n}) = \Sigma_{l\geq 2}\Sigma_{m=-l}^{l} h_{lm}(t, r)\,{}_{-2}Y_{lm}(\hat{n})$. In this expression, $\hat{n}$ is characterized by polar angles $\iota, -\phi_{\rm ref}$; see [11]. For the majority of sources, the $(2, \pm 2)$ mode dominates the summation and can adequately characterize the observationally-accessible radiation in any direction to a relative good approximation; however, other higher modes can often contribute in a significant way to the overall signal [12]. More exotic sources (i.e. high mass ratio and/or precessing, high spins) have significant power in higher modes [13–17].

B. Simulations used

In this work, we use a wide parameter-range of NR simulations similar to the set used in [6]. We use all of the 300 public and 13 non-public SXS simulations for a total of 313 [18]. From the RIT group, we use all 126 public and 281 non-public simulations to bring the total contribution up to 407 [19]. We also use a total of 282 simulations provided from the GT group [20]. Including all the contributions from these three groups, we have a total NR template bank of 1002 simulations. Figure 1 shows all the NR simulations in the 2D parameter space of $\chi_{\rm eff}$, as defined in Eq. (2), vs $1/q$ i.e. the mass ratio. All these simulations have already been published and were produced by one of three familiar procedures, see Appendix A in [6] for more details for each particular group.

From these simulations, we selected 12 simulations to focus on as candidate synthetic sources. Table I shows the specific simulations used, specifying the mass ratio ($q > 1$), component spins of each BH, and total mass. To simplify the process of referring to these heterogeneous simulations, in the last column we assign a shorthand label to each one. These candidates have a variety of mass ratios and spins including zero, aligned, and precessing systems from different NR groups. The first three simulations (RIT-1a,-1b, and -1c) have identical initial conditions/parameters, carried out with different simulation numerical resolution. In many of the validation studies, RIT-1a is used; this is a GW150914-like simulation with comparable masses and anti-aligned spins. We use this simulation for its relative simplicity (higher order modes start to become important at the total mass we'll scale the simulation to, namely $70M_\odot$) and to relate it to our similar work done on the real event GW150914.

In this paper, we present 3 end-to-end studies of our parameter estimation method using data from synthetic sources. We use: a zero spin q=1.0 NR simulation (SXS-1) to show that the method recovers the parameters for the most basic source, an aligned spin GW150914-like simulation (SXS-0233) to show that higher order modes and therefore NR is needed to optimally recover the parameters even with aligned spin cases, and a precessing source (SXS-0234v2) to show our method arrives at reasonable conclusions for any heavy, comparable-mass binary system with generic spins.

C. Extracting asymptotic strain from $\psi_4(r, t)$

From our large and heterogeneous set of simulations, we need to consistently and reproducibly estimate $rh_{lm}(t)$. Many general methods for strain estimation exist; see the review in [21]. The method adopted here must be robust, using the minimal subset of all groups' output; function with all simulations, precessing or not; and rely on only knowledge of asymptotic properties, not (gauge-dependent) information about dynamics. For these reasons, we implemented our own strain reconstruction and extrapolation algorithm, which as input requires only $\psi_{4,lm}(t)$ on some (known) code extraction radius. This method combines two standard tools – perturbative extrapolation [22] and the fixed-frequency integration method [23] – into a single step.

Group	Param	$M/M_\odot$	q	$s_{1,x}$	$s_{1,y}$	$s_{1,z}$	$s_{2,x}$	$s_{2,y}$	$s_{2,z}$	ι	Label
Sequence-RIT-Generic	D12.25_q0.82_a-0.44_0.33_n120	70	1.22	-	-	0.330	-	-	-0.440	$0, \frac{\pi}{6}, \frac{\pi}{4}, \frac{\pi}{3}, \frac{\pi}{2}, \frac{3\pi}{4}$	RIT-1a
Sequence-RIT-Generic	D12.25_q0.82_a-0.44_0.33_n110	70	1.22	-	-	0.330	-	-	-0.440	$\frac{\pi}{4}$	RIT-1b
Sequence-RIT-Generic	D12.25_q0.82_a-0.44_0.33_n100	70	1.22	-	-	0.330	-	-	-0.440	$\frac{\pi}{4}$	RIT-1c
Sequence-RIT-Generic	DD_D10.99_q2.00_a-0.8_n100	70	2.0	-	-	-0.801	-	-	-0.801	$\frac{\pi}{4}$	RIT-2
Sequence-RIT-Generic	U0_D9.53_q1.00_a0.0_n100	70	1.0	-	-	-	-	-	-	$0, \frac{\pi}{6}, \frac{\pi}{4}, \frac{\pi}{3}, \frac{\pi}{2}, \frac{3\pi}{4}$	RIT-3
Sequence-RIT-Generic	D21.5_q1_a0.2_0.8_th104.4775_n100	70	1.0	-	-	0.200	0.775	0	-0.200	$0, \frac{\pi}{6}, \frac{\pi}{4}, \frac{\pi}{3}, \frac{\pi}{2}, \frac{3\pi}{4}$	RIT-4
Sequence-RIT-Generic	D11_q0.50_a0.0_0.0_n100	70	2.0	-	-	-	-	-	-	$\frac{\pi}{4}$	RIT-5
Sequence-SXS-All	1	70	1.0	-	-	-	-	-	-	$0, \frac{\pi}{6}, \frac{\pi}{4}, \frac{\pi}{3}, \frac{\pi}{2}, \frac{3\pi}{4}$	SXS-1
Sequence-SXS-All	Ossokine_0233	70	1.23	-	-	0.320	-	-	-0.580	$0, \frac{\pi}{6}, \frac{\pi}{4}, \frac{\pi}{3}, \frac{\pi}{2}, \frac{3\pi}{4}$	SXS-0233
Sequence-SXS-All	Ossokine_0234v2	70	1.23	0.0943	0.0564	0.322	0.266	0.213	-0.576	$0, \frac{\pi}{6}, \frac{\pi}{4}, \frac{\pi}{3}, \frac{\pi}{2}, \frac{3\pi}{4}$	SXS-0234v2
Sequence-SXS-All	BBH_SKS_d14.4_q1.19_sA_0_0_0.420_sB_0_0_0.380	70	1.19	-	-	0.420	-	-	0.380	$\frac{\pi}{4}$	SXS-$\chi_{\rm eff}$0.4
Sequence-SXS-All	BBH_SKS_d12.8_q1.31_sA_0_0_0.962_sB_0_0_-0.900	70	1.31	-	-	0.962	-	-	-0.900	$\frac{\pi}{4}$	SXS-high-antispin

TABLE I: **Synthetic sources**: A list of the synthetic sources used in our mismatch studies and end-to-end runs. These are done at different inclinations and with higher order modes. All synthetic sources are performed using the same SNR (20) and the same extrinsic parameters: GPS time 10^9s; RA=0; DEC=3.1; and line of sight relative to the NR simulation characterized by Euler angles ι, ϕ, ψ with ι provided in the table and $\phi = \psi = 0$.

Specifically, we extract $rh(t)$ at infinity from $\psi_4(r,t)$ at finite radius using a perturbative extrapolation technique based on Eq. (29) in [22], implemented in the fourier domain and using a low-frequency cutoff [23]. Specifically, if $f_{\rm min}$ is identified as the minimum frequency content for the mode, we construct the gravitational wave strain from ψ_4 at a single finite radius from

$$\begin{aligned} r\tilde{h}_{lm}(f) = \frac{\tilde{\psi}_{4,lm}}{(i\omega)^2}(1-2M/r)[1 - \frac{(\ell-1)(\ell+2)}{2r}\frac{1}{i\omega} + \frac{(\ell-1)(\ell+2)(\ell^2+\ell-4)}{8r^2}\frac{1}{(i\omega)^2}] \\ + \frac{\tilde{\psi}_{4,l+1,m}}{(i\omega)^2}\frac{2ia}{(\ell+1)^2}\sqrt{\frac{(\ell+3)(\ell-1)(\ell+m+1)(\ell-m+1)}{(2l+1)(2l+3)}}[i\omega - \frac{\ell(\ell+3)}{r}] \\ - \frac{\tilde{\psi}_{4,l-1,m}}{(i\omega)^2}\frac{2ia}{(\ell)^2}\sqrt{\frac{(\ell+1)(\ell-2)(\ell+m)(\ell-m)}{(2l-1)(2l+1)}}[i\omega - \frac{(\ell-2)(\ell+1)}{r}] \end{aligned} \tag{3}$$

where the effective frequency is implemented as

$$i\omega = i2\pi \text{sign}(f)\text{max}(|f|, f_{\rm min}) \tag{4}$$

and where a is an estimate for the final black hole spin. This method nominally introduces an obvious obstacle to practical calculation: the last two terms manifestly require an estimate of a and are tied to a frame in which the final black hole spin is aligned with our coordinate axis. In practice, the two spin-dependent terms are small and can be safely omitted in most practical calculations; moreover, each group provides a suitable estimate for the final state. We will clearly indicate when these terms are incorporated into our analysis in subsequent discussion.

When implementing this procedure numerically, we first clean $\psi_{4,lm}$ using pre-identified simulation-specific criteria to eliminate junk radiation at early and late times, tapering the start and end of the signal to avoid introducing discontinuities. For example, for many simulations and for all modes, any content in $\psi_{4,lm}$ prior to $t \leq r + t_0$ was set to zero, for some suitable t_0 (fixed for all modes); subsequently, to eliminate the discontinuity this choice introduces, each mode was multiplied by a Tukey window chosen to cover 5% of the remaining waveform duration. Similarly, all data after a mode-dependent time t_e was set to zero, where the time t_e was identified via the first time (after the time where $|\psi_{4,22}|$ is largest) where $r|\psi_{4,lm}|$ fell below a fixed, mode-independent threshold. To smooth discontinuity, a cosine taper was applied at the end, with duration the larger of either 15 M or 10% of the remaining post-coalescence duration, whichever is larger.

The Fourier transform implementation includes additional interpolation/resampling and padding. First, particularly to enable non-uniform time-sampling, each mode is interpolated and resampled to a uniform grid, with spacing set by the time-sampling rate of the underlying simulation. In carrying out this resampling, the waveform is padded to cover a duration $2T + 100M$, where T is the remaining duration of the (2,2) mode after the truncation steps identified above. To simplify subsequent visual interpretation and investigation, the padding is aligned such that the peak of the (2,2) mode occurs near the center of the interval ($t = 0$).

Finally, the characteristic frequency $Mf_{\rm min,(l,m)}$ is identified from the starting frequency of each $\psi_{4,lm}$. In cases where the starting frequency cannot be reliably identified (e.g., due to lack of resolution), the frequency is estimated from the minimum frequency of the 22 mode as $|m|f_{min,(2,2)}/2$.[1] In Section IV B we will demonstrate the reliability of this procedure to extract

[1] This fallback approximation is not always appropriate for strongly precessing systems. However, for strongly precessing systems, the relevant starting frequency can be easily identified.

$h(t)$ from ψ_4.

D. Framework for directly comparing simulations to observations I: Single simulations

In this section, we briefly review the methods introduced in [11] and [6] to infer compact binary parameters from GW data. All analyses of the data begin with the likelihood of the data given noise, which always has the form (up to normalization)

$$\ln \mathcal{L}(\lambda;\theta) = -\frac{1}{2}\sum_k \langle h_k(\lambda,\theta) - d_k | h_k(\lambda,\theta) - d_k\rangle_k - \langle d_k|d_k\rangle_k, \tag{5}$$

where h_k are the predicted response of the kth detector due to a source with parameters (λ, θ) and d_k are the detector data in each instrument k; λ denotes the combination of redshifted mass M_z and the numerical relativity simulation parameters needed to uniquely specify the binary's dynamics; θ represents the seven extrinsic parameters (4 spacetime coordinates for the coalescence event and 3 Euler angles for the binary's orientation relative to the Earth); and $\langle a|b\rangle_k \equiv \int_{-\infty}^{\infty} 2df \tilde{a}(f)^* \tilde{b}(f)/S_{h,k}(|f|)$ is an inner product implied by the kth detector's noise power spectrum $S_{h,k}(f)$. In all calculations, we adopt the fiducial O1 noise power spectra associated with data near GW150914 [1]. In practice we adopt a low-frequency cutoff $\mathrm{f_{min}}$ so all inner products are modified to

$$\langle a|b\rangle_k \equiv 2\int_{|f|>f_{\min}} df \frac{\tilde{a}(f)^*\tilde{b}(f)}{S_{h,k}(|f|)}. \tag{6}$$

The joint posterior probability of λ, θ follows from Bayes' theorem:

$$p_{\rm post}(\lambda,\theta) = \frac{\mathcal{L}(\lambda,\theta)p(\theta)p(\lambda)}{\int d\lambda d\theta \mathcal{L}(\lambda,\theta)p(\lambda)p(\theta)}, \tag{7}$$

where $p(\theta)$ and $p(\lambda)$ are priors on the (independent) variables θ, λ. For each λ, we evaluate the marginalized likelihood

$$\mathcal{L}_{\rm marg} \equiv \int \mathcal{L}(\lambda,\theta)p(\theta)d\theta \tag{8}$$

via direct Monte Carlo integration, where $p(\theta)$ is uniform in 4-volume and source orientation. To evaluate the likelihood in regions of high importance, we use an adaptive Monte Carlo as described in [11]. We will henceforth refer to the algorithm to "integrate over extrinsic parameters" as ILE. The marginalized likelihood is a way to quantify the similarity of the data and template. If we integrate out all the parameters except total mass, we get a curve that looks like Figure 2. Having $\ln \mathcal{L}$ in this form is the most useful for our purposes, and plots involving $\ln \mathcal{L}$ will be as a function of total mass.

E. Framework for directly comparing simulations to observations II: Multidimensional fits and posterior distribution

The posterior distribution for intrinsic parameters, in terms of the marginalized likelihood and assumed prior $p(\lambda)$ on intrinsic parameters like mass and spin, is

$$p_{\rm post} = \frac{\mathcal{L}_{marg}(\lambda)p(\lambda)}{\int d\lambda \mathcal{L}_{\rm marg}(\lambda)p(\lambda)}. \tag{9}$$

As we demonstrate by concrete examples in this work, using a sufficiently dense grid of intrinsic parameters, Eq. (9) indicates that we can reconstruct the full posterior parameter distribution via interpolation or other local approximations. The reconstruction only needs to be accurate near the peak. If the marginalized likelihood $\mathcal{L}_{\rm marg}$ can be approximated by a d-dimensional Gaussian, with (estimated) maximum value $\mathcal{L}_{\rm max}$, then we anticipate only configurations λ with

$$\ln \mathcal{L}_{\rm max}/\mathcal{L}_{\rm marg}(\lambda) > \chi^2_{d,\epsilon}/2 \tag{10}$$

contribute to the posterior distribution at the 1-ϵ creditable interval, where $\chi^2_{d,\epsilon}$ is the inverse-χ^2 distribution. [The practical significance of this threshold will be more apparent in Section III B, which implicitly illustrates it using one dimension.] Since the mass of the system can be trivially rescaled to any value, each NR simulation is represented by particular values for the seven intrinsic parameters (mass ratio and the three components of the spin vectors) and is represented by a one-parameter family of points in the 8-dimensional parameter space of all possible values of λ. Given our NR archive, we evaluate the natural log of the marginalized likelihood as a function of the redshifted mass $\ln \mathcal{L}_{\rm marg}(M_z)$. As in [6], our first-stage result is this function, explored almost continuously in mass and discretely as our fixed simulations permit. This information alone is sufficient to estimate what parameters are consistent with the data: for example, using a cutoff such as Eq. (10), we identify the masses that are most consistent for each simulation.

As demonstrated first in [6] and explored more systematically here, this likelihood is smooth and broad extending over many NR simulations' parameters. As a result, even though our function exploration is a restricted to a discrete grid of NR simulation values, we can interpolate between simulations to reconstruct the entire likelihood and hence entire posterior. We can do this because of the simplicity of the signal, which for the most massive binaries involves only a few cycles. More broadly, our method works because many NR simulations produce very similar radiation, up to an overall mass scale; as a result, as has been described previously in other contexts [24], surprisingly few simulations have been needed to explore the model space (e.g., for nonprecessing binaries).

Finally, as we demonstrate repeatedly below by example, $\ln \mathcal{L}_{\rm marg}$ is often well approximated by a simple low-order series, typically just a quadratic. Moreover, for the short GW150914-like signals here, many nonprecessing simulations fit both observations and even precessing simulations fairly well. As a result, we employ a quadratic approximation to $\ln \mathcal{L}_{\rm marg}$ near the peak under the restrictive approximation that all angular momenta are parallel using information from only nonprecessing binaries. Using this fit, we can estimate

$\ln \mathcal{L}_{\rm marg}$ for all masses and aligned spins and therefore estimate the full posterior distribution. Section IV B in [6] gives the results of this method based on the LIGO data containing GW150914. In this work, we apply this method to a larger set of examples.

III. DIAGNOSTICS

Many steps in our procedure to compare NR simulations to GW observations can introduce systematic error into our inferred posterior distribution. Sources of error include the numerical simulations' resolution; waveform extraction; finite duration; Monte Carlo integration error; the finite, discrete, and sparsely spaced simulation grid; and our fit to said grid. In the following sections, we describe tools to characterize the magnitude and effect of these systematic errors. First and foremost, we introduce the broadly-used *match*, a complex-valued inner product which arises naturally in data analysis and parameter inference applications. Following many previous studies [25], we review how systematic error shows up as a mismatch and parameter bias. Second, we describe an analogy to the match which uses our full multimodal infrastructure and is more directly connected to our final posterior distribution: the marginalized likelihood versus mass $\ln \mathcal{L}_{\rm marg}(M)$, or equivalently (one-dimensional) posterior distribution implied by assuming the data must be drawn from a specific simulation up to overall unknown mass and orientation. Due to systematic error, the inferred one-dimensional distribution (or match versus mass) may change, both globally and through any concrete confidence interval (CI) derived from it. To appropriately quantify the magnitude of these effects, we introduce two measures to compare similar distributions. On the one hand, any change in the 90% CI provides a simple and easily-explained measure of how much an error changes our conclusions. On the one hand, the KL divergence ($D_{\rm KL}$) gives a simple, well-studied, theoretically appropriate, and numerical measure of the difference between two neighboring distributions. In this section we describe these diagnostics and illustrate them using concrete and extreme examples to illustrate how a significant error propagates into our interpretation.

A. Inner products between waveforms: the mismatch

The match is a well-used and data-analysis-driven tool to compare two candidate GW signals in an idealized setting. Unlike most discussions of the match, which derive them from the response of a single idealized instrument, we follow [26] and work with the response of an idealized *two*-detector instrument, with both co-located identical interferometers oriented at 45^o relative to one another, and the source located directly overhead this network.[2] As is well-known, the match arises naturally in the likelihood of a candidate signal, given known and noise-free data – or, in the notation of this work, from Eq. (5) restricted to this idealized network, setting d to $h_0 = h(t, \lambda_0)$ and $h(\lambda, \theta) = h$:

$$\begin{aligned} \ln \mathcal{L} &= -\frac{1}{2}\{\langle h - h_0 | h - h_0\rangle - \langle h_0 | h_0\rangle\} \\ &= -\frac{1}{2}\{\langle h|h\rangle - 2\Re\langle h_0|h\rangle\}, \end{aligned} \tag{11}$$

where $\Re$ is the real part. Again $\langle a|b\rangle$ is the complex overlap (inner product) between two waveforms for a single detector as shown in Eq. (6); the GW strain $h = h_+ - ih_\times$ contains two polarizations, and is assumed to propagate from directly overhead the network; the likelihood reflects the response of both detectors' antenna response and noise. Eq. (11) is slightly different than the the likelihood obtained in Eq. (17) of [26] by an overall constant. What we use, described in [27], is the likelihood ratio (divided by the likelihood of zero signal). If we add this constant back into the equation, we recover Eq. (17) from [26]:

$$\ln \mathcal{L}_{\rm single} = -\frac{1}{2}\{\langle h_0|h_0\rangle + \langle h|h\rangle - 2\Re\langle h_0|h\rangle\}. \tag{12}$$

This single-detector likelihood depends on the parameters λ, θ of h and λ_o, θ_0 of h_0. For the purposes of our discussion, we will include "systematic error" parameters that enhance or change the model space in λ (e.g., changes in simulation resolution).

The parameters which maximize the likelihood identify the configuration of parameters that make h most similar to h_0. For a fixed emission direction from the source, three key parameters in θ dominate how h can be changed to maximize the likelihood: the event time $t_{\rm event}$; the source luminosity distance D_L; and the polarization angle ψ, characterizing rotations of the source (or detector) about the line of sight connecting the source and instrument. In terms of these parameters,

$$h = e^{-2i\psi}\frac{D_{\rm L,ref}}{D_{\rm L}} h_{\rm ref}(t - t_{\rm event}|\lambda, \theta_{\rm rest}) \tag{13}$$

where $h_{\rm ref}$ is the value of h at $D_{\rm L} = D_{\rm L,ref}, t_{\rm event} = 0$, and $\psi = 0$ and $\theta_{\rm rest}$ denotes the four remaining extrinsic parameters besides these three. As noted in [26], a change of the polarization angle ψ corresponds to a rotation of the argument of the complex strain function, $h(\psi) = e^{-2i\psi}h(\psi = 0)$. As a result, maximizing the likelihood versus ψ corresponds to choosing a phase angle so $\langle h|h_0\rangle$ is purely real:

$$\max_\psi \langle h_0|h\rangle = |\langle h_0|h\rangle|. \tag{14}$$

Similarly maximizing the likelihood versus distance, the likelihood becomes

$$\max_{\psi, D_L} \ln \mathcal{L}_{\rm single} = -\rho^2(1 - P_*). \tag{15}$$

[2] Equivalently, we work in the limit of many identical detectors, such that the network has equal sensitivity to both polarizations for all source propagation directions.

where in this expression $\rho^2 = \langle h_0|h_0\rangle = \langle h|h\rangle$ and the function P is

$$P_*(h_0, h) \equiv \max_\psi \frac{|\langle h_0|h\rangle|}{\sqrt{\langle h_0|h_0\rangle\langle h|h\rangle}}, \quad (16)$$

This partially-maximized likelihood depends strongly on the event time. If we furthermore maximize over event time, we find the final and important relationships

$$\ln \mathcal{L}_{\rm single,max} = \max_{\psi, D_L, t_{\rm event}} \ln \mathcal{L}_{\rm single} = -\rho^2(1-P), \quad (17)$$

$$P(h_0, h) \equiv \max_{\psi, t_{\rm event}} \frac{|\langle h_0|h\rangle|}{\sqrt{\langle h_0|h_0\rangle\langle h|h\rangle}}. \quad (18)$$

In the rest of this paper, we will use the mismatch $\mathcal{M}$ between two signals:

$$\mathcal{M}(h_0, h) = 1 - P(h_0, h). \quad (19)$$

Because of its form – an inner product – the mismatch identifies differences between the two candidate signals; substituting this expression into the maximized ideal-detector likelihood [Eq. (17)] yields:

$$\ln \mathcal{L}_{\rm single,max} = -\rho^2 \mathcal{M}. \quad (20)$$

As the above relationships make apparent, a candidate signal h which has a significant mismatch cannot be scaled to resemble h_0 and therefore must be unlikely. This relationship has been used to motivate simple criteria to characterize when two signals h, h_0 are indistinguishable (or, conversely, distinguishable); working to order of magnitude [cf. Eq. (10)], two signals are indistinguishable if [28–31]

$$\mathcal{M} \leq \frac{1}{\rho^2}. \quad (21)$$

In this work, we apply the match criteria to assess when two simulations of the same or similar parameters (or the same simulation at a different mass) can be distinguished from a reference configuration.

As a concrete example, discussed at greater length in Section III E, the top-right panel in Figure 3 shows two plots of mismatch versus total mass. In the black curve, we calculate the match of two identical waveforms from the RIT-1a simulation: one set at a fixed total mass $M = 70M_\odot$ while the other changes over a given mass range. At the true total mass, the mismatch goes to zero. For comparison, the red curve in that figure shows the mismatch between another simulation h and a fixed RIT-1a (h_0), versus total mass for h. As illustrated in the top-left panel of Figure 3, the two simulations are not identical; hence, the mismatch in the top-right panel between h and h_0 never reaches zero. Moreover, due to differences in the source h_o and template family h, the location of the minimum mismatch and hence best fit occurs at a different, offset total mass, close to $50M_\odot$.

As the reader will see in subsequent sections, we can also calculate the mismatch as a function of particular properties of NR simulations to see how much error is introduced, see Section IV.

B. Marginalized likelihood versus mass

Another simple diagnostic is the result $\ln \mathcal{L}_{\rm marg}(M)$ for a single simulation on some reference data (e.g., the simulation itself, or a signal with comparable physical origin). This function enters naturally into our full parameter estimation calculation; therefore, it allows us to test all of the quantities that influence our principal result directly including NR resolution, extraction radius, etc. as described below. For simplicity, as computed for the purposes of this test, this function depends on part (only $l \leq 2$ modes) of the NR radiation and the data. Figure 2 shows a null example run with RIT-1a, a GW150914-like simulation, as a source compared against itself. As previous work from both real LIGO and synthetic data has suggested, $\ln \mathcal{L}(M)$ can be well-approximated by a locally quadratic fit (see Section III D for a more in-depth discussion of this example).

C. Probability Density Function/KL Divergence

To quantitatively assess whether two given versions of $\ln \mathcal{L}(M)$ are demonstrably different, we employ an observationally-motivated diagnostic to prioritize agreement in regions with significant posterior support. Motivated by the applications we perform when comparing results of this kind, we translate $\ln \mathcal{L}(M)$ into a probability distribution (i.e., assuming all other parameters are fixed):

$$p_c(M) = \frac{1}{\int dM e^{\ln \mathcal{L}}} e^{\ln \mathcal{L}}. \quad (22)$$

In practice, this distribution is always extremely well approximated by a gaussian, so we can further simplify by characterizing any 1d distribution by its mean M_* and variance $1/\Gamma_{MM} = \sigma_*^2$. Using this ansatz, we can therefore define a quantity to assess the difference between any pair of results for $\ln \mathcal{L}(M)$. In this work, we use the KL divergence between these two approximately-normal distributions:

$$\begin{aligned} D_{KL}(p_*|p) &= \int dx p(x) \ln p(x)/p_*(x) \\ &= \ln \frac{\sigma}{\sigma_*} - \frac{1}{2} + \frac{(\bar{x} - \bar{x}_*)^2 + \sigma_*^2}{2\sigma^2}. \end{aligned} \quad (23)$$

We also will plot the derived PDF $p_c(M)$ and evaluate the implied 1D 90% CI derived from it.

The implications of a significant disagreement for this diagnostic – already illustrated via high mismatch in Figure 3 – can be clearly seen in the 1D posterior distributions derived from the fit of $\ln \mathcal{L}_{\rm marg}(M)$ as shown in Figure 3 and Figure 4. Loosely following the work in [25] for estimating parameter errors due to mismatch, we expect the parameter error will be a significant fraction of the statistical error. Using the notation above and approximating $P \simeq 1 - \frac{1}{2}\bar{\Gamma}_{xx}\delta x^2$ for some nominal perturbed parameter x, we estimate the statistical error to be $\sigma_{x,stat} \simeq 1/\rho\sqrt{\Gamma_{xx}}$. Conversely, balancing mismatch and

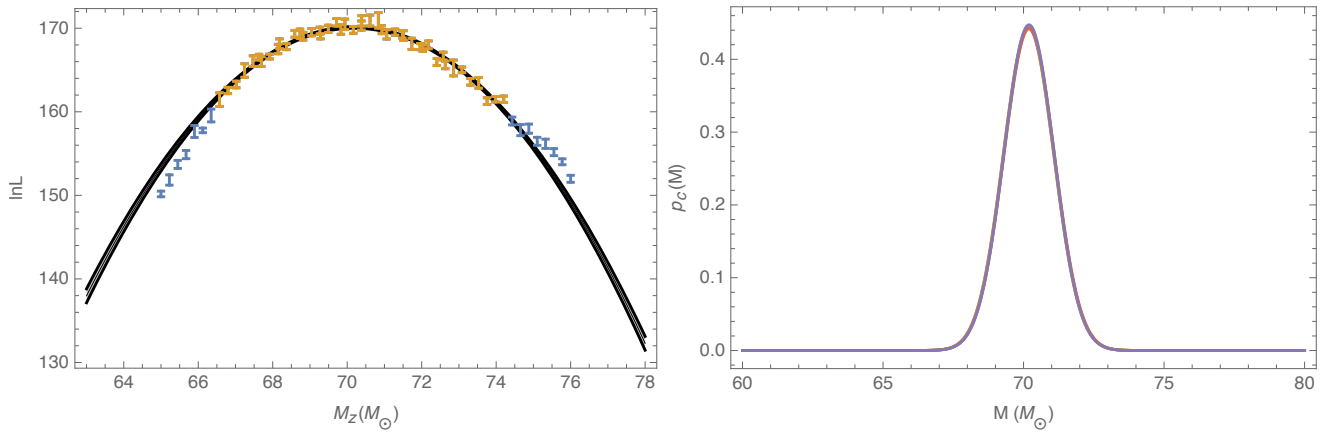

FIG. 2: **Example of $\ln \mathcal{L}_{\rm marg}(M)$: comparing a simulation to itself**: *Left panel*: Blue and yellow points (with error bars) show results of evaluating $\ln \mathcal{L}_{\rm marg}(M)$ with RIT-1a as a source compared to itself. The shaded region is derived by fitting a quadratic to these data via least-squares [Eq. (25)], providing a mean and confidence interval (shown). The reference source has total mass $M = 70 M_\odot$ and an inclination $\imath = 0.785$; all calculations are carried out using $f_{\rm min} = 30$Hz. This curve will be duplicated as a black curve in the right panel from Figure 3 and left panel from Figure 4. *Right panel*: Nominal one-dimensional posterior distributions [Eq. (22)] derived from the fit to left. This figure shows five examples, randomly drawn from the fit coefficient distribution derived by least squares, drawn to exemplify the propagated systematic uncertainties due to Monte Carlo integration error. For studies similar to this one (i.e., high-mass investigations where direct comparison to numerical relativity is most appropriate), this figure suggests that Monte Carlo error is much smaller than the posterior width (i.e., has little relevance given the substantial statistical uncertainty introduced by the limited number of GW cycles available for comparison from short NR simulations).

parameter biases, similar changes in likelihood occur when

$$\delta x \simeq \frac{1}{\bar{\Gamma}^{1/2}_{xx}} \mathcal{M}^{1/2}; \tag{24}$$

however, much more detailed calculations is presented in [25]. The above relationship illustrates how a high mismatch causes a deviation in the $\ln \mathcal{L}_{\rm marg}(M)$ curve as well as its corresponding posterior distribution. Figure 3 show a comparison between two waveforms from RIT-1a and RIT-2 (red curve). With significantly different parameters (see Table I), the mismatch is significantly high. This causes a radical shift in the $\ln \mathcal{L}_{\rm marg}(M)$ result as well as its corresponding PDF compared to to it's true value. This example will be described in greater detail in Section III E.

D. Example 0: Null test/Impact of Monte Carlo Error

To illustrate the use of these diagnostics, we first apply them to the special case where the data contains the response due to a known source. In this case, by construction, the match will be unity when using the same parameters. Following a similar procedure to that we would apply if we didn't know the source mass, we can also plot the mismatch $\langle h_A(M)|h_A(M_*)\rangle / ||h_A(M)||||h_A(M_*)||$. Referring to the notation in Eq. (16), we assign the RIT-1a waveform to $h_0 = h_{RIT-1a}$(source) and again the RIT-1a waveform to $h = h_{RIT-1a}$ (template). This plot can be seen in any of the following examples as the black curve (top-right panels from Figure 3 and Figure 4). It has a peak value of unity (not plotted) and rapidly falls as one moves away from the mass corresponding to the peak match value. The left panel of Figure 2 shows the log likelihood $\ln \mathcal{L}_{\rm marg}$ provided by *ILE* as a function of mass. From here we fit a local quadratic to the $\ln \mathcal{L}_{\rm marg}$ close to the peak. Using the fit, we generate five random samples and use them for subsequent calculations (i.e. 1D distributions). We derived a 1D distribution using Eq. (22).

First and foremost, these figures illustrate the relationships between the three diagnostics. As suggested by Eq. (20), the match and log likelihood $\ln \mathcal{L}_{\rm marg}$ are nearly proportional up to an overall constant. Second, as required by Eq. (22), the one-dimensional posterior is proportional to $\mathcal{L}_{\rm marg}$. This visual illustration corroborates our earlier claim implicit in the left panel of Figure 2: only the part of $\ln \mathcal{L}_{\rm marg}$ within a few of its the peak value contributes in any way to the posterior distribution and to any conclusions drawn from it (e.g., the 90% CI).

Each evaluation of the Monte Carlo integral has limited accuracy, as indicated in Figure 2. By taking advantage of many evaluations of this integral, we dramatically reduce the overall error in the fit. To estimate the impact of this uncertainty, we use standard frequentist polynomial fitting techniques [32] to estimate the best fit parameters and their uncertainties (i.e., of a quadratic approximation to $\ln \mathcal{L}$ near the peak): if $\ln \mathcal{L}_{\rm marg} = \sum_\alpha \lambda_\alpha F_\alpha(M_z)$ and $\gamma_{kk} = 1/\sigma_k^2$ is an inverse covariance matrix characterizing our measurement errors, then the best-fit estimate for $\ln \mathcal{L}_{\rm marg}$ and its variance is

$$\ln \mathcal{L}_{\rm marg,est} = F(F^T \gamma F)^{-1} \gamma y \tag{25a}$$

$$\Sigma(x) = F_\alpha(x)[(F^T \gamma F)^{-1}]_{\alpha\beta} F_\beta(x) \tag{25b}$$

where y is an array representing the $\ln \mathcal{L}_{\rm marg}$ estimates at the

sample	D_{KL}	CI (90%)
1	0	(68.71 - 71.66)
2	2.5e-4	(68.71 - 71.68
3	1.2e-4	(68.71 - 71.68)
4	7.2e-4	(68.71 - 71.67)
5	2.3e-4	(68.70 - 71.68)

TABLE II: **KL Divergence and 90% CI between different samples from the null test fit**: This table shows the D_{KL} and 90% CI for five different sample PDFs. The D_{KL} was calculated comparing the 1D distributions to the first sample (i.e. D_{KL} for sample 1 is zero). The CI are also given to show the change between them. Both diagnostics suggest the distributions are nearly indistinguishable.

data points and F is a matrix representing the values of the basis functions on the data points: $F_\alpha(x_k)$. The left panel of Figure 2 shows the 90% CI derived from this fit, assuming gaussian errors.

To translate these uncertainties into changes in the one-dimensional posterior distribution p_c, we generate random draws from the corresponding approximately multinomial distribution for fit parameters; and thereby generate random samples and hence one-dimensional distributions for $p_c(M)$ consistent with different realizations of the Monte Carlo errors. The right panel of Figure 2 shows five random samples from the fit in the left panel. This figure demonstrates this level of Monte Carlo error, by design, has negligible impact on the posterior distribution. To quantify the impact of Monte Carlo error on the posterior, we calculate the KL Divergence from Eq. (23). In all cases, the KL divergence was small, of order 10^{-4}, see Table II for more details on D_{KL} and the 90% CI. In Section IV A, we further verify this conclusion by repeating our analysis many times.

E. Example 1: Two NR simulations with different parameters/Illustrating how sensitively parameters can be measured

In this example we compare two NR simulations with significantly different parameters to demonstrate how our diagnostics handle waveforms of extreme contrast. The two NR simulations used are RIT-1a and RIT-2. As shown in Table I, these simulations are both aligned spin with different magnitudes with $q = 1.22$ and $q = 2.0$ respectively. To illustrate the extreme differences between the radiation from these two systems, the top-left panel of Figure 3 shows the two simulations' $rh(t)$.

Our three diagnostics equally reveal the substantial differences between these two signals. To be concrete, since these diagnostics treat data and models asymmetrically, we operate on synthetic data containing RIT-1a with inclination $\imath = \pi/4$ in these applications. First, the top-right panel of Figure 3 shows the results of our mismatch calculations. The black curve is the same null test mismatch calculation as in the top-right panel of Figure 4: it has a narrow minimum (of zero) at the true binary mass ($70M_\odot$). For the red curve, we calculate the mismatch while holding RIT-2 at a fixed mass and changing the mass of RIT-1a. Using the notation in Eq. (16), we assign the RIT-2 waveform to $h_0 = h_{\rm RIT-2}$(fixed mass at $M = 70M_\odot$) and the RIT-1a waveform to $h = h_{\rm RIT-1a}$(changing mass). In this case, the match does not reach unity, differing by a few percent, while the peak value occurs at significantly offset parameters (here, in total mass). Second, the bottom-left panel of Figure 3 shows the results for $\ln \mathcal{L}_{\rm marg}(M)$, using these two NR simulations to look at the same stretch of synthetic data including our local quadratic fit to them. Third, the bottom-right panel of Figure 3 shows the implied one-dimensional posterior distribution derived from our fits. There is a clear shift in total mass with the null test again peaking around $70M_\odot$ and this example's peak around $50M_\odot$. There are also orders of magnitude difference between the $\ln \mathcal{L}_{\rm marg}$ of the two cases. These diagnostics show something that could be seen just by looking at the waveforms; however, we now have some idea on how major differences propagate through our diagnostics and how the error in each diagnostic relate to each other. For completeness, we also include the D_{KL} and CI for these two waveforms in Table III. The D_{KL} as well as the CI are both considerably offset, as expected given the two significantly different simulations involved.

ILE run (source/template)	D_{KL}	CI (90%)
RIT-1a/RIT-1a	0.0	(68.8 - 71.4)
RIT-2/RIT-1a	288.8	(49.3 - 52.0)

TABLE III: **KL Divergence and 90% CI between two NR simulations with different parameters**: This table shows the D_{KL} and 90% CI between: RIT-1a/RIT-1a and RIT-1a/RIT-2. The D_{KL} was calculated comparing the 1D distributions to RIT-1a/RIT-1a distribution (notice its D_{KL} is zero i.e. they're identical). The CI are also given to show the difference between these two distributions.

Finally, the parameter shift seen above is roughly consistent in magnitude with what we would expect for such an extreme mismatch error, given the SNR and match: we expect using Eq. (24) $\delta M \simeq \sigma_M \rho \mathcal{M}^{1/2} \simeq 5\sigma_M \simeq 5M_\odot$ (using $\mathcal{M} = 6 \times 10^{-2}, \rho = 20$ and $\sigma_M = 1.1M_\odot$), or a shift in best fit of several standard deviations and many solar masses. While noticeably smaller than our actual best-fit shift, our result from Eq. (24) provides a valuable sense of the order-of-magnitude biases incurred by specific level of mismatch in general. Moreover, this example is a concrete illustration of the critical need to have $\mathcal{M} \leq 1/\rho^2$ to insure that any systematic parameter biases are small and under control.

F. Example 2: Different physics: SEOB vs NR/Illustrating the value of numerical relativity

Several studies have previously demonstrated the critical need for numerical relativity, since even the best models do not yet capture all available physics [33, 34]. For example, these models generally omit higher-order modes, whose omission will impact inferences about the source [35–37].

To illustrate the value of NR in the context of this work, we compare parameter estimation with NR and with an analytic model. In this particular example, we use NR simulation RIT-1a including the $l \leq 2$ modes (see Table I) evalu-

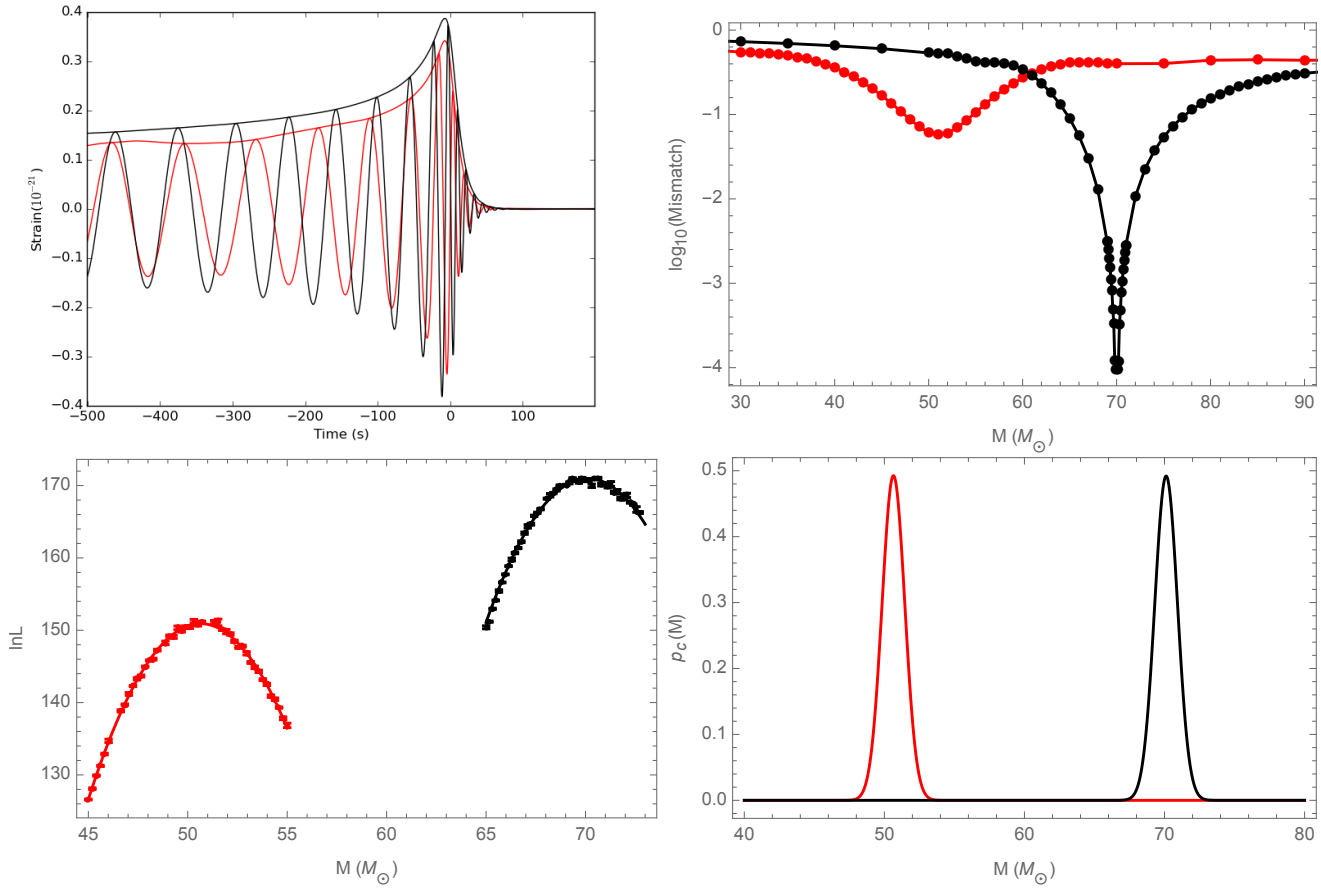

FIG. 3: **Example 1-Assessing differences between two NR simulations with different parameters**: Two representations of the different predictions of RIT-1a and RIT-2, which are aligned spin binaries with mass ratios $q = 1.22$ and $q = 2.0$ respectively, illustrating how dramatic differences propagate into our diagnostics. *Top-left panel*: The strain along a line of sight inclined at $\iota = 0.785$ and evaluated for a total mass $M = 70M_\odot$, with RIT-1a in black and RIT-2 in red. *Top-right panel*: The mismatch between synthetic data and candidate templates as a function of the template's mass. In both cases, the RIT-1a simulation is used as the template (i.e., as h in Eq. (16)). For the black curve, RIT-1a for a $70M_\odot$ binary is also used as the source (i.e., $h_0 = h_{\rm RIT-1a}$). For the red curve, the source is RIT-2 set at M=70 $M_\odot$. while RIT-1a has a changing mass *Bottom-left panel*: Points show the marginalized likelihood versus total mass calculated by applying the same template simulation (RIT-1a) to two different sources: RIT-1a in black and RIT-2 in red. Each source has fixed mass $M = 70M_\odot$ and inclination $\iota = 0.785$; as in Figure 2, we evaluate $\mathcal{L}$ using a low-frequency cutoff $f_{\rm min} = 30$Hz. For context, red and black solid curves show a corresponding quadratic least-squares fit to these data. *Bottom-right panel*: The corresponding one-dimensional posteriors $p_c(M)$ [Eq. (22)]. Both bottom panels illustrate how an ill-suited simulation with large mismatch (i.e., the red curve) correlates with a drastic shift in parameters (here, total mass) relative to the true best-fit solution (here, the black curve), [see Eq. (24)]. Also, the ill-matched simulation cannot recover all the information available to the true solution, so the peak $\ln \mathcal{L}_{\rm marg}$ for the red curve is substantially lower ($\simeq 20$) than the peak of the black curve.

ated along an inclination $\iota = \pi/4$. Using this line of sight and our fiducial mass ($M = 70M_\odot$), higher harmonics play a nontrivial role. For our analytical model, we use an Effective-One-Body model with spin (SEOBNRv2), described in [38], which was one of the models used in the parameter estimation of GW150914 [39] and which was recently compared to this simulation [33]. The top-left panel of Figure 4 shows the time-domain strains from the NR simulation and SEOBNRv2 with the same parameters. To better quantify the small but visually apparent difference in the two waveforms, we use the diagnostics described earlier on these two waveforms.

One way to characterize the differences in these waveforms is the mismatch [Eq. (16)]. In the top-right panel of Figure 4, we calculate the mismatch by holding the SEOBNRv2 waveform at a fixed mass while changing the mass of the NR waveform shown in blue. Referring to the notation in Eq. (16), we assign the SEOBNRv2 waveform to $h_0 = h_{\rm SEOBNRv2}$ and the RIT-1a waveform to $h = h_{\rm RIT-1a}$. For comparison, a mismatch calculation was done with the null test from Section III D (RIT-1a compared to itself) shown here in black. Two differences between the two curves are immediately apparent. First, the blue curve does not go to zero; the mismatch is a few times 10^{-3}, significantly in excess of the typical accuracy threshold [Eq. (21), evaluated at $\rho = 25$]. Second,

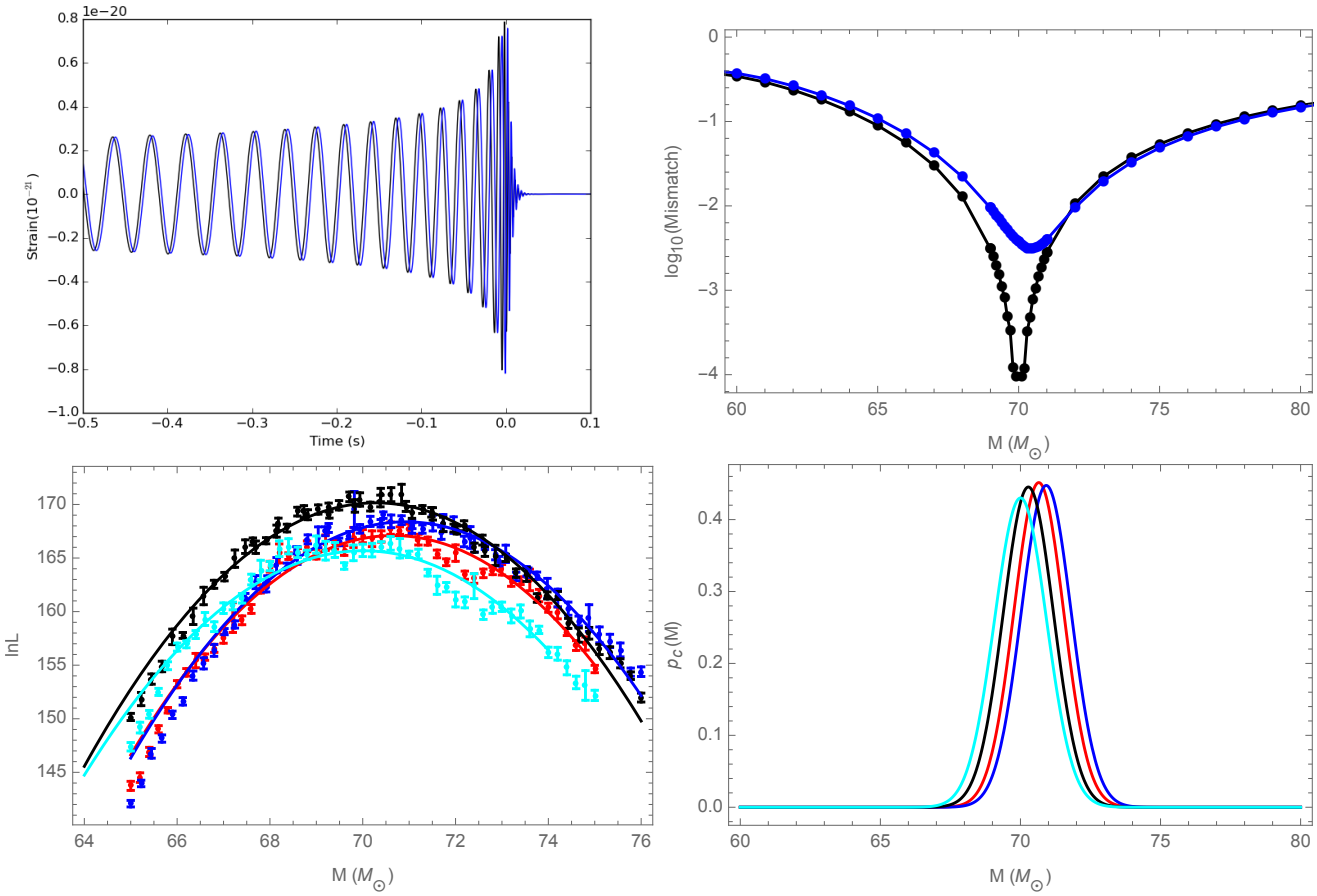

FIG. 4: **Example 2-Assessing differences in SEOB and NR waveforms that have the same parameters**: This figure shows how subtle differences between an NR solution and an approximation to GR (here, EOB) can propagate into mismatch and parameter estimation. These two companion figures follow the pattern of Figures 3. *Top-left panel*: The black and blue curves show the strain evaluated from RIT-a and SEOBNRv2, respectively, for a source with identical parameters. Source parameters and strain results for the black curve are identical to Figure 3 (e.g., $\iota = \pi/4$). *Top right-panel*: Following the top-right panel of Figure 3, this figure shows the match between the two waveforms on the top-left with the corresponding template from RIT-1a. *Bottom-left*: The marginalized likelihood $\ln \mathcal{L}_{\rm marg}$ for the two waveforms shown above, evaluated using *both* RIT-1a and SEOBNRv2 as templates: NR source compared to same NR template in black; the SEOBNRv2 source to a SEOBNRv2 template in red; the SEOBNRv2 source to a NR (RIT-1a) template in blue; and the NR (RIT-1a) source to a SEOBNRv2 template in cyan. *Bottom-right*: The one-dimensional posterior distributions $p_c(M)$ derived from the quadratic fits shown in the bottom-left. Both bottom panels show a clear change along the total mass for SEOBNRv2 sources. The NR/NR comparison has the highest $\ln \mathcal{L}_{\rm marg}$ with with a corresponding total mass $\sim 70 M_\odot$. The NR/SEOBNRv2 template curve correctly finds the total mass $\sim 70 M_\odot$; however, the $\ln \mathcal{L}_{\rm marg}$ is orders of magnitudes different than the null example. The differences between NR simulations and the SEOBNRv2 model is significant for parameter estimation.

the minimum occurs at offset parameters. The best-fit offset and mismatch are qualitatively consistent with the naive estimate presented earlier: a high mismatch yields a high change in total mass [see Eq. (24)]. This simple calculation illustrates how mismatch could propagate directly into significant biases in parameter estimation.

Another and more observationally relevant way to characterize the differences between these two waveforms is by carrying out a full *ILE* based parameter estimation calculation. We carry out four comparisons: the null test (a NR source compared to same NR template (black)); the SEOBNRv2 source compared to a SEOBNRv2 template (red); the NR source compared to a SEOBNRv2 template (cyan); and an SEOBNRv2 source compared to a NR template (blue). The bottom panels of Figure 4 shows both the underlying $\ln \mathcal{L}_{\rm marg}(M)$ results; our quadratic approximations to the data; and our implied one-dimensional posterior distributions [Eq. (22)]. All *ILE* calculations were carried out with $f_{\rm min} = 30$Hz. All four likelihoods $\ln \mathcal{L}_{\rm marg}$ and posterior distributions p_c are manifestly different, with generally different peak locations and widths. Table IV quantifies the differences between the possible four configurations, using D_{KL} and 90% CI. The D_{KL} was always calculated by comparing one of them to the NR/NR case. These systematic differences exist even without higher modes, whose neglect will only exacerbate the biases seen here.

ILE configuration (source/template)	D_{KL}	CI (90%)
SEOB/SEOB	0.086	(69.2 - 72.1)
SEOB/RIT-1a	0.25	(69.4 - 72.4)
RIT-1a/RIT-1a	0	(68.8 - 71.8)
RIT-1a/SEOB	0.050	(68.5 - 71.5)

TABLE IV: **KL Divergence and 90% CI between SEOB and NR**: This table shows the D_{KL} and 90% CI for the four different configurations using SEOBNRv2 and NR as sources and templates. The D_{KL} was calculated comparing the 1D distributions to the NR/NR case (notice its D_{KL} is zero i.e. they're identical). The CI also given to show the change between them. Based on the D_{KL} results, the 1D posteriors are similar but not exactly the same distribution. These nontrivial differences affect our parameter estimation results and also change our astrophysical conclusions about the source.

$f_{\rm min}$ for *ILE* run (Hz)	D_{KL}	CI (90%)
10	0.0	(69.2 - 71.1)
20	1.3e-3	(69.2 - 71.1)
30	0.62	(69.2 - 72.1)
40	7.1	(69.2 - 74.6)

TABLE V: **KL Divergence and 90% CI of PDFs derived from SEOB sources with different low frequency cutoffs**: This table shows the D_{KL} and 90% CI for the four different configurations using SEOBNRv2 source with a set duration of 5Hz and compared against SEOBNRv2 templates with different low frequency cutoffs. The D_{KL} was calculated comparing the 1D distributions to the $f_{\rm min} = 10$Hz case (notice its D_{KL} is zero i.e. they're identical). The CI also given to show the change between them. Based on the D_{KL} results, the 1D posteriors of $f_{\rm min} = 10, 20$Hz seem to be the same distribution; however, they differ significantly to $f_{\rm min} = 30, 40$Hz.

Keeping in mind the two figures adopt a comparable color scheme, the shift in peak value and location between the black and blue curves seen in the bottom panels of Figure 4 can be traced back to the top-right of Figure 4: to a first approximation, systematic errors identified by the mismatch ($\mathcal{M}$) show up in the marginalized likelihood ($\ln \mathcal{L}_{\rm marg}$). Again, based on calculations using Eq. (24), we expect the change in mass location of order unity holding all other things equal, comparable to the observed offset.

In many ways, one-dimensional biases shown in the bottom-right panel *understate* the differences between these signals: that comparison explicitly omits the peak value of $\ln \mathcal{L}_{\rm marg}$, which occurs not only at a different location but also with a different value for all four cases. As we would expect, the NR/NR case has the highest $\ln \mathcal{L}_{\rm marg}$ with a peak near the true total mass $70M_\odot$. The NR/SEOB case can also produce a peak near $70M_\odot$; however, the $\ln \mathcal{L}_{\rm marg}$ is orders of magnitude lower, which translates to a lower likelihood that this was in fact the correct template. When performing a full multidimensional fit, template-dependent biases in the peak value of $\ln \mathcal{L}_{\rm marg}$ can also impact our conclusions.

To summarize, we have shown that using SEOBNRv2 in place of a more precise solution of Einstein's equations introduces non-negligible systematic errors, of a magnitude comparable to the statistical error for plausible sources, and that it can impact astrophysical conclusions.

G. Example 3: Signal duration and cutoff frequency/Illustrating the impact of simulation duration with SEOB

Numerical relativity simulations have finite duration. Until hybrids [40–43] are ubiquitously available, these finite duration cutoffs will impair the utility of direct comparison between data and multimodal NR simulations. To assess this impact of finite simulation duration, we adopt a contrived but easily-controlled approach, using an analytic model where we can freely adjust signal duration. While our specific numerical conclusions depend on the noise power spectrum adopted, as it sets the required low-frequency cutoff, the general principles remain true for advanced instruments.

In this example, we plot $\ln \mathcal{L}_{\rm marg}$ for a fiducial SEOBNRv2 source versus itself using different choices for the low-frequency cutoff (and, equivalently, different initial orbital frequencies for the binary). The left panel of Figure 5 shows $\ln \mathcal{L}_{\rm marg}$ versus M. In this figure, the $\ln \mathcal{L}_{\rm marg}$ curves for $f_{\rm min} = 10$Hz and 20Hz (brown and green) are significantly narrower and higher compared to the $\ln \mathcal{L}_{\rm marg}$ curves for $f_{\rm min} = 30$Hz or 40Hz(red and magenta). As described in [6], even though very little signal power is associated with very low frequencies for this combination of detector and source, a significant amount of information about the total mass is available there with all other parameters of the system perfectly known. These differences are immediately apparent in our one-dimensional diagnostics $\ln \mathcal{L}_{\rm marg}(M)$ and $p_c(M)$, which are both narrower and more informative when more information is included (i.e., for lower $f_{\rm min}$). That said, our PSD does not provide access to arbitrarily low frequencies, and the lowest two frequencies have nearly identical posterior distributions, as measured by KL divergence, see Table V. This investigation strongly suggests our analysis could be sharper with longer simulations or hybrids. That said, [6] demonstrated this procedure will, for GW150914-like data and noise, arrive at similar results to an analysis which includes these lower frequencies. As noted in [6], this virtue leverages a fortuitous degeneracy in astrophysically relevant observables: the limitations of our high-frequency analysis are mostly washed out due to strong degeneracies between mass, mass ratio, and spin.

IV. VALIDATION STUDIES

In this section we self-consistently assess our errors in $h(t)$ and $\ln \mathcal{L}$. Using the diagnostics described above, via targeted one-dimensional studies, we systematically assess the impact of Monte Carlo error; waveform extraction error; simulation resolution; and limited access to low frequency content. We will show via our diagnostics that the effects from these potential sources of error can be either ignored or mitigated (e.g., by a suitable choice of operating point for our analysis procedure, such as a high enough extraction radius). For each potential source of error, we use the KL divergence D_{KL} [Eq. (23)] to quantify small differences in one-dimensional poste-

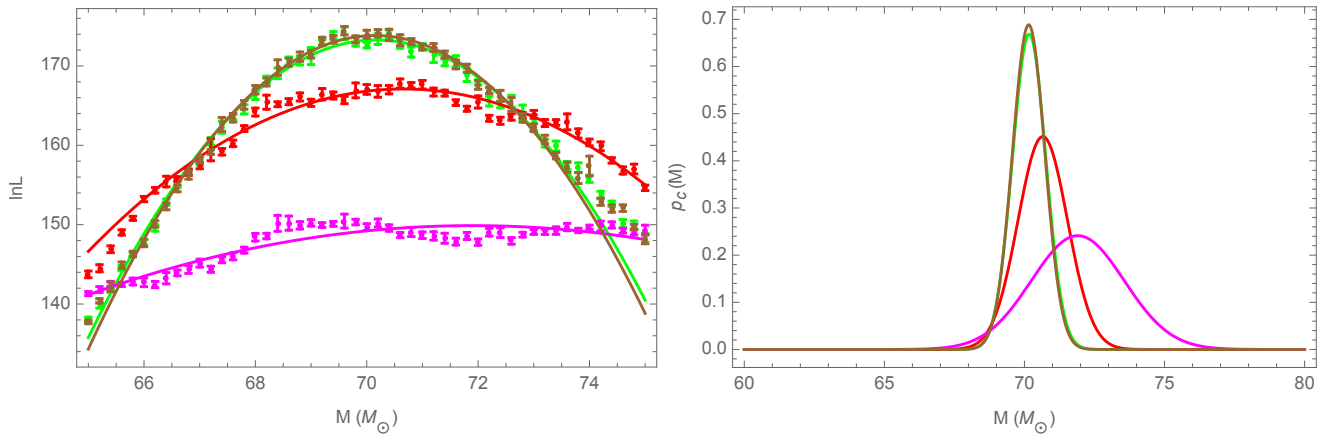

FIG. 5: **Example 3-Quantifying the impact of the low-frequency cutoff**: Using analytic SEOBNRv2 templates with user-specified starting frequency and length, this figure quantifies the impact of our choice of low-frequency cutoff on parameter estimation. *Left panel*: Plot of $\ln \mathcal{L}_{\rm marg}$ versus total mass evaluated using SEOBNRv2 templates with different starting frequencies with $f_{\rm min} = 10$Hz (brown), $f_{\rm min} = 20$Hz (green), $f_{\rm min} = 30$Hz (red), and $f_{\rm min} = 40$Hz (magenta). In all cases, the source signal is also SEOBNRv2 using the same parameters as RIT-1a, but starting frequency $f_{\rm min} = 5$Hz. *Right panel*: The one-dimensional posteriors $p_c(M)$ [Eq. (22)] implied by the results to left. As you increase the low frequency cutoff, the $\ln \mathcal{L}_{\rm marg}$ decreases significantly, and both the posterior and $\ln \mathcal{L}_{\rm marg}$ are wider and offset from the true parameters.

Trial	D_{KL}	CI (90%)
v1	0	(68.9 - 71.9)
v2	4.8e-5	(68.9 - 71.9)
v3	5.6e-5	(68.9 - 71.9)

TABLE VI: **KL Divergence and 90% CI between different runs of the same null test.**: This table shows the D_{KL}, calculated using Eq. (23) and 90% CI for three different runs of the same configuration as described in Section III D. The D_{KL} was calculated comparing the 1D distributions to Trial v1 (notice its D_{KL} is zero i.e. they're identical). The CI also given to show the change between them. Based on the D_{KL} results, the 1D posteriors of these different trials are identical.

rior distributions $p_c(M)$ [Eq. (22)] derived from $\ln \mathcal{L}_{\rm marg}$. We will relate our results to familiar mismatch-based measures of error. To be concrete, we will employ a target signal amplitude (SNR) $\rho = 25$, similar to GW150914. For similarly-loud sources, the mismatch criteria [Eq. (21)] suggests any parameters with mismatch below $\log_{10}(\mathcal{M}) = -2.8$ will lead to "statistical errors" (associated with the width of the posterior) will be smaller than systematic biases.

A. Impact of Monte Carlo error

We have already assessed the error from our Monte Carlo integration in Section III D, directly propagating the (assumed correct) Monte Carlo integration error into our fit. To comprehensively demonstrate the impact of Monte Carlo integration error, we repeat our entire analysis reported in Figure 2 multiple times. Figure 6 shows our directly comparable results; Table VI reports quantitative measures of how these distributions change. Based on these quantities, we conclude the error introduced by our Mont Carlo is negligible. Our results are consistent with Section III D.

B. Error budget for waveform extraction

While gravitational waves are defined at null infinity, the finite size of typical NR computational domains implies a computational technique must identify the appropriate asymptotic radiation from the simulation [44]. This method generally has error, often associated with systematic neglect of near-field physics in the asymptotic expansion used to extract the wave (i.e., truncation error). Our perturbative extrapolation method shares this limitation. As a result, if we decrease the radius at which we extract the asymptotic strain, we increase the error in our approximation. In other words, the mismatch between the waveform extracted at r and some large radius generally decreases with r; the trend of match versus r provides clues into the reliability of our results.

Figure 7 shows an example of a mismatch between two estimates of the strain: one evaluated at finite, largest possible radius and one at smaller (and variable) radius. For context, we show the nominal accuracy requirement corresponding to a SNR=25 [see Eq. (21)] as a black dotted line. First and foremost, this figure shows that, at sufficiently high extraction radius, the error introduced by mismatch errors is substantially below our fiducial threshold for all choices of: cutoff frequency, waveform extraction location, and waveform extraction technique; see also [7]. Second, the second panel shows our perturbative extraction method is reasonably consistent with an entirely independent approach to waveform extraction. Agreement is far from perfect: our study also indicates a noticeable discrepancy between the results of our perturbative extraction technique and the SXS strain extraction method. Due to the good agreement reported elsewhere [33], we suspect these residual disagreements arise from coordinate

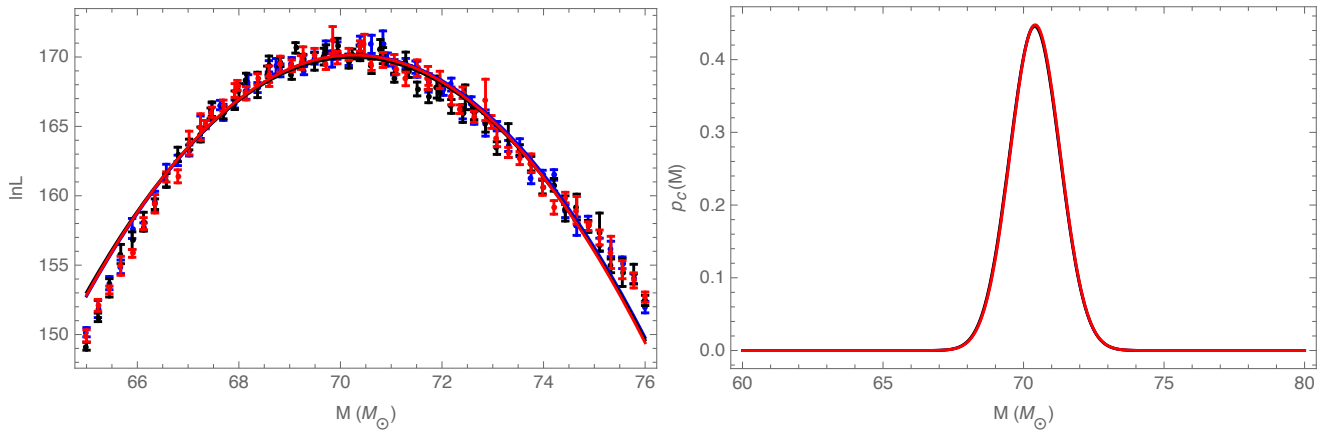

FIG. 6: **Monte Carlo error revisited: Repeating the fitting process multiple times**: This figure shows several repeated, independent end-to-end calculations of $\ln \mathcal{L}_{\rm marg}$ (left panel) and $p_c(M)$ (right panel), shown in different colors. The calculation performed is identical to the calculation described for Figure 2. This figure demonstrates we understand and have control over our Monte Carlo errors.

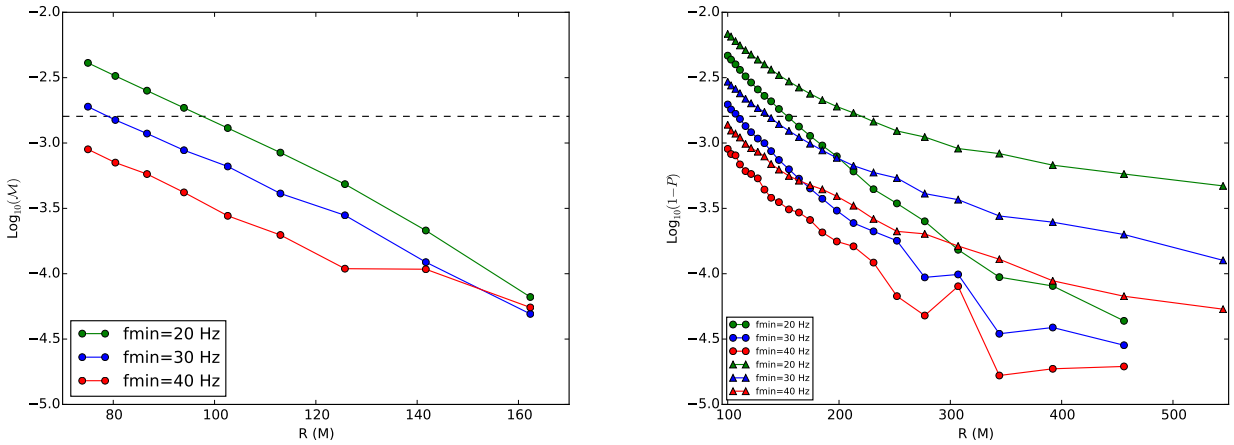

FIG. 7: **Mismatch between waveforms at different extraction radii using different NR groups and extraction techniques**: Both panels show the mismatch between the radiation extracted from RIT-1a (left panel) and SXS-0233 (right panel) as a function of the extraction radius r. All calculations are performed using the same configurations as Figures 3 and 4: a total mass of $70M_\odot$ and an inclination $\iota = 0.785$. In both panels, the green, blue, and red colors represent different choices of low frequency cutoff: $f_{\rm min} = 20, 30, 40$Hz respectively. For context and motivated by Eq. (21), the dashed line denotes the mismatch threshold implied by $\rho = 25$ (i.e., $\log_{10}(1/25^2)$). *Left panel*: Mismatch calculations comparing a waveform perturbatively extracted at $r = 190M$ with a waveform that is perturbatively extracted at other extraction radii, [see Eq. (3)]. *Right panel*: Circles correspond to results using a reference waveform extracted at $r = 545M$ via perturbative extraction from their ψ_4 data; triangles denote calculations using a reference waveform evaluated using the strain provided by SXS (i.e., using a polynomial extrapolation with $N = 2$). In both cases, the reference waveform is compared to a waveform constructed via perturbative extraction using ψ_4 data at the specified radius.

effects unique to our interpretation of SXS data; we will assess this issue at greater depth in subsequent work. Third and finally, as expected, comparisons that employ more of the NR signals are more discriminating: calculations with a smaller $f_{\rm min}$ generally find a higher (i.e., worse) mismatch. Nonetheless, our mismatch calculations significantly improve at large extraction radius, when perturbative extrapolation is carried out well outside the near zone.

To assess the observational impact of waveform extraction systematics, we evaluate $\ln \mathcal{L}_{\rm marg}(M)$ and $p_c(M)$ using waveform estimates produced using different extraction radii.

Extraction Radius (M)	D_{KL}	CI (90%)
190M/190M	0	(68.8 - 71.5)
162.34/190M	9.3e-3	(68.9 - 71.5)
141.71/190M	3.6e-2	(69.0 - 71.8)

TABLE VII: **KL Divergence and 90% CI between PDFs with different extraction radii**: This table shows the D_{KL}, calculated using Eq. (23) and 90% CI for PDFs with three different extraction radii. The D_{KL} was calculated comparing the 1D distributions to the PDF with $r = 190M$ (notice its D_{KL} is zero i.e. they're identical). The CI also given to show the change between them. Based on the D_{KL} results, the 1D posteriors show some differences but are very similar.

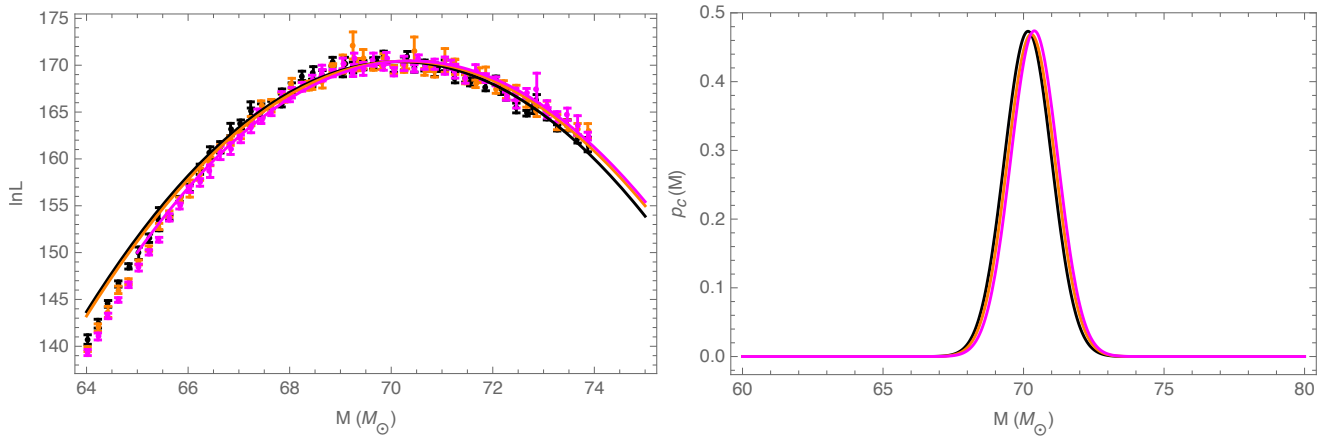

FIG. 8: **Propagating systematic error from finite extraction radius into posterior distributions** : This figure shows how small systematic errors from finite NR extraction radius propagate into parameter estimation posterior distributions, by concrete example. *Left panel*: A plot of $\ln \mathcal{L}_{\rm marg}$ versus total mass. In all cases, the source is RIT-1a at $r = 190M$; the templates are also RIT-1a, using different extraction radii as templates. Here, magenta is $r = 141.71M$, orange is $r = 162.34M$, and black is $r = 190M$. We focus our search on only the last few extraction radii to avoid clutter. The error is relatively small but bigger than what our match study naively suggests (i.e., changes in $\ln \mathcal{L}$ of order $10^{-4}\rho^2/2 \simeq 2 \times 10^{-2}$, though this result only applies to the change in the peak value, which is indeed changes by less than than amount). *Right panel*: One-dimensional posterior distributions $p_c(M)$ of each individual fit derived from the three plots [see Eq. (22)]. Even though there are small differences, these PDFs are virtually identical.

NR Label	Resolution	Mismatch
RIT-1a/RIT-1a	n120/n120	0.0
RIT-1b/RIT-1a	n110/n120	3.90e-5
RIT-1c/RIT-1a	n100/n120	5.27e-5

TABLE VIII: **Mismatch between waveforms with different numerical resolutions**: Here is a mismatch study between the different resolutions for one NR simulation. Specifically RIT-1a vs RIT-1a, RIT-1a vs RIT-1b, and RIT-1a vs RIT-1c. The results were evaluated at $M = 70M_\odot$ and $\imath = 0.785$. The mismatch between the different resolution is very small and is much smaller than our accuracy requirement. We therefore expect the error introduced to be negligible.

Specifically, we take a simulation; use its large-radius perturbative estimate as a source; and follow the procedures used in Figures 3 and 4 to produce $\ln \mathcal{L}_{\rm marg}(M)$ and $p_c(M)$. Figure 8 shows our results; for clarity, we include only the last three extraction radii ($r = 190M, 162M, 141M$). The errors here are relatively small but bigger than expected from our match study; however, the error shown in the match only applies to changes in the peak value $\ln \mathcal{L}_{\rm marg}$, which can be seen in the left panel. To again quantify these small differences, we use D_{KL} and CI, as reported in Table VII. As this table shows, the error introduced is insignificant as long as we pick a relative large extraction radius. This is almost always the case for the current simulations available. Some of the GT simulations require us to chose a lower extraction radius due to an increase in the error as the extraction radius increases beyond a certain point, but this does not affect our overall results.

C. Impact of simulation resolution

Here we analyze errors introduced by different numerical resolutions. Higher resolutions simulations take longer to run and computationally cost more than lower resolution ones. If the effects of different resolutions are insignificant, numerical relativist will be able to run at a lower resolution while not introducing any systematic errors. Table VIII shows a match comparison between the highest resolution RIT-1a and the two lower ones, RIT-1b and RIT-1c. The mismatches are orders of magnitudes better than our accuracy requirement ($\sim 10^{-2.8}$), and therefore introduce errors that are negligible.

Using $\ln \mathcal{L}_{\rm marg}$ as our diagnostic to compare these three simulations, we draw similar conclusions; see Figure 9. We again see a error so small that changes between the three curves are almost impossible to see, even far from the peak. Table IX quantifies these extremely small differences. In short, different resolutions have no noticeable impact on our conclusions. While this resolution study was only done for a aligned RIT simulation, similar conclusions are expected when a wider range of simulations are used.

Even though in this case the mismatch and *ILE* studies show conclusively the minimal impact the numerical resolution has on the waveform, we generate 1D distributions from the fits for completeness. It is not surprising to see in the right panel of Figure 9 the posteriors from the three fits match almost exactly. To quantify this similarity, we calculate D_{KL} as well as the CI for the corresponding PDFs. Based on the D_{KL}, these distributions are clearly identical and using different resolutions does not effect the waveform in any significant way. This resolution study was only done for an aligned RIT simulation; while extraction radius studies have been performed for SXS for other extraction procedures [45], a similar

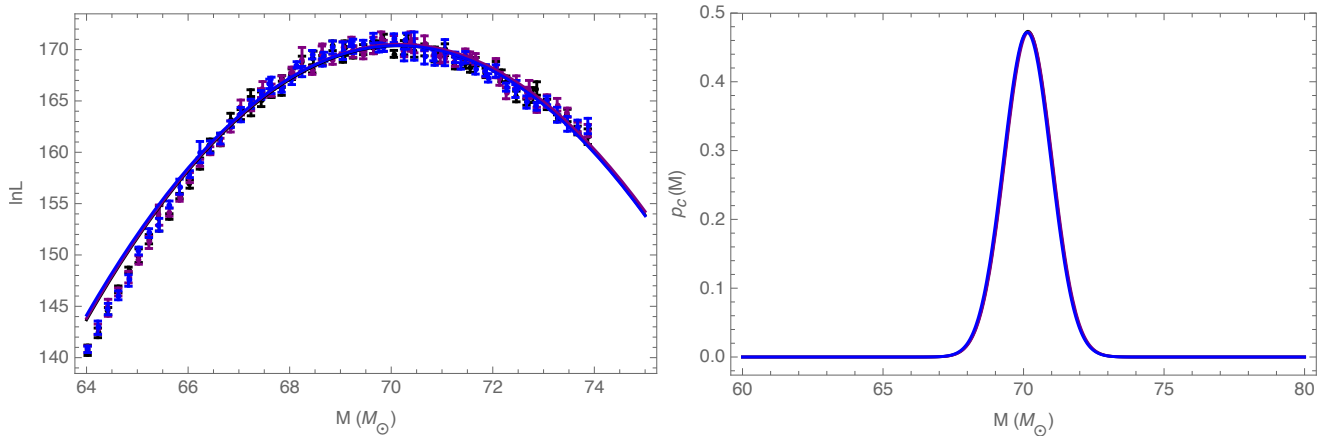

FIG. 9: **Single runs of *ILE* with changing resolution and their corresponding PDFs**: The left panel consists of $\ln \mathcal{L}$ vs total mass curves with different numerical resolution. Here we use RIT-1a as the source and compare it to simulations with the same parameters at different resolutions, specifically RIT-1b and RIT-1c. The results were evaluated with $f_{\rm min} = 30$Hz at a total mass $M = 70 M_\odot$ with a inclination $\imath = 0.785$. Here black is n120, purple is n110, and blue is n100. Even though the error is clearly minuscule, we convert the fits to a PDFs for completeness. The right panel shows the PDFs for the three different resolutions [see Eq. (22)]. It is clear that these are all the same PDFs, and the error introduced by different resolutions is irrelevant.

Resolution (M)	D_{KL}	CI (90%)
n120/n120	0	(68.8 - 71.5)
n110/n120	2.0e-4	(68.8 - 71.6)
n100/n120	6.5e-4	(68.7 - 71.5)

TABLE IX: **KL Divergence and 90% CI between PDFs with different numerical resolution**: This table shows the D_{KL}, calculated using Eq. (23), and 90% CI for PDFs with the three different resolutions for RIT-1a. The D_{KL} was calculated comparing the 1D distributions to the PDF with n120 (notice its D_{KL} is zero i.e. they're identical). The confidence intervals also given to show the change between them. Based on the D_{KL} results, the 1D posteriors are identical.

$f_{\rm min}$ for *ILE* run (Hz)	D_{KL}	CI (90%)
10/10	0.0	(69.2 - 71.2)
20/10	9.2e-3	(69.2 - 71.3)
30/10	0.34	(69.0 - 72.0)
40/10	1.9	(67.8 - 73.0)

TABLE X: **KL Divergence and 90% CI of PDFs derived from RIT-4 sources with different low frequency cutoffs**: This table shows the D_{KL} and 90% CI for the four different configurations using a RIT-4 source with a set duration of 5Hz and compared against RIT-4 templates with different low frequency cutoffs. The D_{KL} was calculated comparing the 1D distributions to the $f_{\rm min} = 10$Hz case (notice its D_{KL} is zero i.e. they're identical). The CI also given to show the change between them. Based on the D_{KL} results, the 1D posteriors of $f_{\rm min} = 10, 20$Hz seem to be the same distribution; however, they differ significantly to $f_{\rm min} = 30, 40$Hz.

resolution investigation needs to be done for SXS simulations for this extraction method. We hypothesize that this effect will also be minimal.

D. Impact of low frequency content and simulation duration

As demonstrated by Example 3 in Section III G above, the available frequency content provided by each simulation and used to the interpret the data can significantly impact our interpretation of results. In this section, we perform a more systematic analysis of simulation duration and frequency content, again using the semi-analytic SEOBNRv2 model as a concrete waveform available at all necessary durations. Before we begin, we first carefully distinguish between two unrelated "minimum frequencies" that naturally show up in our analysis. It is easy to get confused between the low frequency cutoff (in this work called $f_{\rm min}$) and simulation duration (or initial orbital frequency $M\omega_0$). The simulation duration is the true duration of the simulation, which is a property of the binary and can be drastically different over many NR simulations. The low frequency cutoff is an artificial cut to the signal that allows us to normalize the signal duration of all our waveforms. As a result, with a lower $f_{\rm min}$, more of the NR simulation enters into our analysis.

The top panels of Figure 10 shows the result of compare a RIT-4 source with a duration of 5.0 Hz to itself with changing $f_{\rm min}$. As $f_{\rm min}$ increases, a smaller portion of the simulation waveform is being used to analyze the data. When $f_{\rm min}$ is high, we end up cutting off more of the waveform. This results in a sharp decline in $\ln \mathcal{L}_{\rm marg}$ since one is now comparing less of the waveform to itself. In this panel it is clear that $f_{\rm min} \sim 10-20$Hz seems to not significantly affect $\ln \mathcal{L}_{\rm marg}$; however, the curve changes drastically when $f_{\rm min} = 30 - 40$Hz. For completeness Table X shows the corresponding D_{KL} and CI for different $f_{\rm min}$, again showing the similarities between the $f_{\rm min} = 10, 20$Hz frequencies and the differences of the higher frequencies. Hybrid NR waveforms will nullify this source of error by allowing us to compare more of the waveform while at the same time allowing us to standardize durations.

To investigate the shift in mass seen in Figure 5 further,

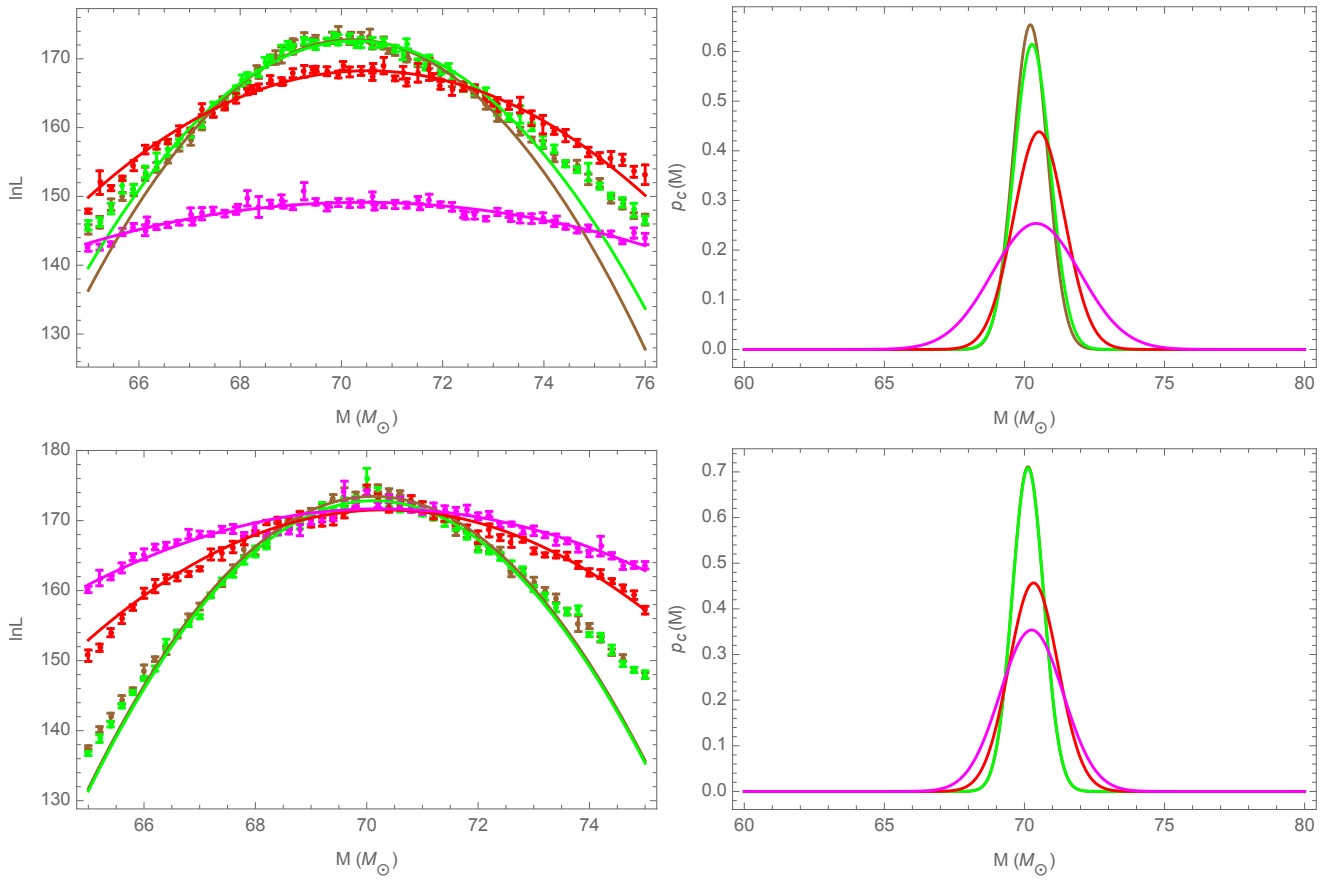

FIG. 10: **Assessing the impact of low frequency cutoff 2: Consistent cutoff choices**: Revisiting the investigations shown in Figure 5, this figure uses (in the top panels) comparisons of an NR simulation to itself as a method to isolate the impact of $f_{\rm min}$ (the low-frequency cutoff appearing in the likelihood). The bottom panels repeat a comparable analysis using SEOB. *Top left panel*: Plots different $\ln \mathcal{L}_{\rm marg}$ vs total mass curves with different $f_{\rm min}$. Here we compared a RIT-4 source with a duration of 5.0 Hz source compared to itself at different $f_{\rm min}$ values. Specifically brown has a $f_{\rm min} = 10$, green has a $f_{\rm min} = 20$, red has a $f_{\rm min} = 30$, and magenta has a $f_{\rm min} = 40$. These results are similar to the SEOBNRv2 case in Figure 5. As the cutoff increases, our $\ln \mathcal{L}$ curve becomes wider, and the peak value $\ln \mathcal{L}_{\rm marg}$ is lower. *Top right panel*: One-dimensional posteriors $p_c(M)$ [Eq. (22)]. This figure qualitatively resembles Figure 5; however, unlike the previous analysis, while the posterior is wider (i.e., less informative), no significant bias is introduced by the low-frequency cutoff. *Bottom left panel*: Similar to prior figures, a plot of $\ln \mathcal{L}_{\rm marg}(M)$, evaluated using SEOBNRv2. In this comparison, the SEOBNRv2 source with a certain duration was compared to a SEOBNRv2 template with the same $f_{\rm min}$. Specifically brown has a $f_{\rm min} = 10$, green has a $f_{\rm min} = 20$, red has a $f_{\rm min} = 30$, and magenta has a $f_{\rm min} = 40$. As the cutoff increases, our $\ln \mathcal{L}$ curve becomes wider. *Bottom right panel*: The corresponding PDFs to the fits [see Eq. (22)]. We again see similarities between this case and Figure 5 minus the shift in total mass with increasing $f_{\rm min}$.

$f_{\rm min}$ for *ILE* run (Hz)	D_{KL}	CI (90%)
10/10	0.0	(69.2 - 71.0)
20/10	1.7e-5	(69.2 - 71.1)
30/10	0.33	(68.9 - 71.8)
40/10	0.85	(68.4 - 72.1)

TABLE XI: **KL Divergence and 90% CI of PDFs derived from SEOB sources**: This table shows the D_{KL} and 90% CI for the four different configurations using a SEOB source compared against SEOB templates with the same duration/$f_{\rm min}$ (i.e. if the source has a duration of 10 Hz, the template has a $f_{\rm min} = 10$Hz). The D_{KL} was calculated comparing the 1D distributions to the $f_{\rm min} = 10$Hz case (notice its D_{KL} is zero i.e. they're identical). The CI also given to show the change between them. Based on the D_{KL} results, the 1D posteriors of $f_{\rm min} = 10, 20$Hz seem to be the same distribution; however, they differ significantly to $f_{\rm min} = 30, 40$Hz.

we compare a SEOBNRv2 source to a SEOBNRv2 template with the same duration/$f_{\rm min}$ (i.e. the source has a duration of 10 Hz therefore the template has a $f_{\rm min} = 10$Hz). This was done to investigate the shift in total mass seen in Figure 5 for a SEOBNRv2 source with a fixed duration compared to a SEOBNRv2 template with different low frequency cutoffs. As the bottom panels of Figure 10 now show, this shift was a product of comparing a source and templates with different signal lengths. When we now set the same duration for the source and $f_{\rm min}$ for the template, the *ILE* results and their corresponding PDFs peak around the same mass point. We still see a widening of the curves with increasing $f_{\rm min}$; this corresponds to a wider and shorter PDF. We calculate D_{KL} and CI for this case as well, see Table XI. These values shows that $f_{\rm min} = 10, 20$Hz are relatively similar while the higher frequencies are significantly different.

V. RECONSTRUCTING PROPERTIES OF SYNTHETIC DATA I: ZERO, ALIGNED, AND PRECESSING SPIN

This section is dedicated to end-to-end demonstrations of this parameter estimation technique. Unless otherwise specified, we adopt a total binary mass of $M = 70M_\odot$ and use the fiducial early-O1 PSD [46] to qualitatively reproduce the characteristic features of data analysis for GW150914. Without loss of generality and consistent with common practice, we adopt a "zero noise" realization (i.e., the data used for each instrument is equal to its expected response to our synthetic source). Table I is a list of simulations we have used as sources in our end-to-end runs; these include zero, aligned, and precessing systems all at different inclinations. Here we start with a end-to-end demonstration with zero spin from SXS.

A. Zero Spin: A fiducial example demonstrating the method's validity

We first illustrate the simplest possible and most-well-studied scenario: a compact binary with zero spin and equal mass, as represented here by SXS-1. To enable comparison with other cases where higher-order modes will be more significant, we adopt inclinations $\imath = 0, 0.5, 0.785, 1.0, 1.5, 2.35$. For the purposes of illustration, we present our end-to-end plots using an inclination $\imath = 0$.

The left panel of Figure 11 shows $\chi_{\rm eff}$ vs 1/q; the points represent the maximum log likelihood $\ln \mathcal{L}_{\rm marg}$ of all the different *ILE* runs across parameter space. The green contour is the 90% CI derived using the quadratic fit to $\ln \mathcal{L}_{\rm marg}$ for nonprecessing systems only. The colored points represent points that fall in $\ln \mathcal{L}_{\rm marg} < 127$ region with the red points representing higher $\ln \mathcal{L}_{\rm marg}$ and violet represent lower $\ln \mathcal{L}_{\rm marg}$. The gray points represent points that fall between $\ln \mathcal{L}_{\rm marg} = 130$ and $\ln \mathcal{L}_{\rm marg} = 127$. The black points represent points that fall in $\ln \mathcal{L}_{\rm marg} > 130$. These intervals were determined using the inverse χ^2 distribution [see Eq. (10)] adopting $d = 4$ (two masses with aligned spin) for the black points and $d = 8$ (two masses with precessing spins). This CI is consistent with the point distribution $\ln \mathcal{L}_{\rm marg} > 130$ (i.e. black points), which represents the points closest to the maximum. The right panel of Figure 11 shows the $\chi_{\rm eff}$ vs M with the same green contour and black point distribution. As with the left panel, the green contour is consistent with the black point distribution. Both plots recover the true parameters (indicated by the big red dot) with regards to the confidence interval and the black point distributions.

The left panel of Figure 12 shows the χ_{1z} vs χ_{2z} where $\chi_{1z,2z}$ is the z component of the dimensionless spin [see Eq. (1)]. All the colors here represent the same as in Figure 11. We again see that the green contour is consistent with the black point distribution. The right panel of Figure 12 shows the 1D posteriors for $1/q$ for six different inclinations. These produce distributions we expect to see; all the curves from the different inclinations lie on top of each other. This implies that higher order modes for this particular case are not expected to provide any extra information. By construction, this source needs no higher order modes to completely recover the parameters. Since all inclinations have the same distribution shape, the results here are independent of inclination at a fixed SNR.

B. Nonprecessing binaries: unequal mass ratios and aligned spin

In the previous zero spin case, the higher order modes had a minimal impact. Now we introduce an aligned spin GW150914-like simulation as the source, SXS-0233. For our total mass of $M = 70M_\odot$, we expect that the impact of higher order modes border on being significant. Because of this, we did 2 end-to-end runs with SXS-0233: one with $l \leq 2$ and the other with $l \leq 3$. The panels in Figure 13 are the same type of plots as in the previous case; however, we have also included a contour representing the 90% CI for $l \leq 3$ (green dashed line). In the left panel of Figure 13, the posterior corresponding to $l \leq 3$ better constrains the mass ratio than that of the posterior corresponding to $l \leq 2$. In this case, including higher order modes provides more information about the mass ratio, allowing us to constrain it more tightly. The right panel of Figure 13 is the same type of plot as the bottom panel of Figure 11; however, this includes the results from the $l \leq 3$ runs. Since the $\ln \mathcal{L}_{\rm marg}$ was higher, the number of black and gray points slightly decreased. It is clear from these two plots that higher order modes are significant and need to be included for this source to get the best possible constrains on the parameters. The right panel in Figure 13 shows the $\chi_{\rm eff}$ vs M; these show little difference between the $l \leq 2$ and the $l \leq 3$ contours. The contours agree very well with each as well as the black points' distribution in both panels of Figure 13. We recover the true parameters in both plots and with $l \leq 2$ and $l \leq 3$; however, we can better constrain q with higher order modes.

As with the zero spin case, we plot $\ln \mathcal{L}_{\rm marg}$ as a function of χ_{1z} and χ_{2z} in the left panel in Figure 14. Here again the dashed and solid green contour represents the confidence interval for $l \leq 2$ and $l \leq 3$ respectively and are largely consistent with each other. The right panel of Figure 14 shows the 1D distributions for 1/q for different inclination values. The difference in the curves here could be explained by higher order modes; however, more needs to be done to corroborate this hypothesis.

In this particular case, higher order modes have a relatively modest impact on the posterior. The minimal impact is by design: moving away from zero spin and equal mass within the posterior of GW150914, we have explicitly selected a point in parameter space where higher-order modes have just become marginally significant. Even remaining within the posterior of GW150914, as we move towards more extreme antisymmetric spins and mass ratios, higher-order modes can play an increasingly significant role. We will address this issue further in subsequent work.

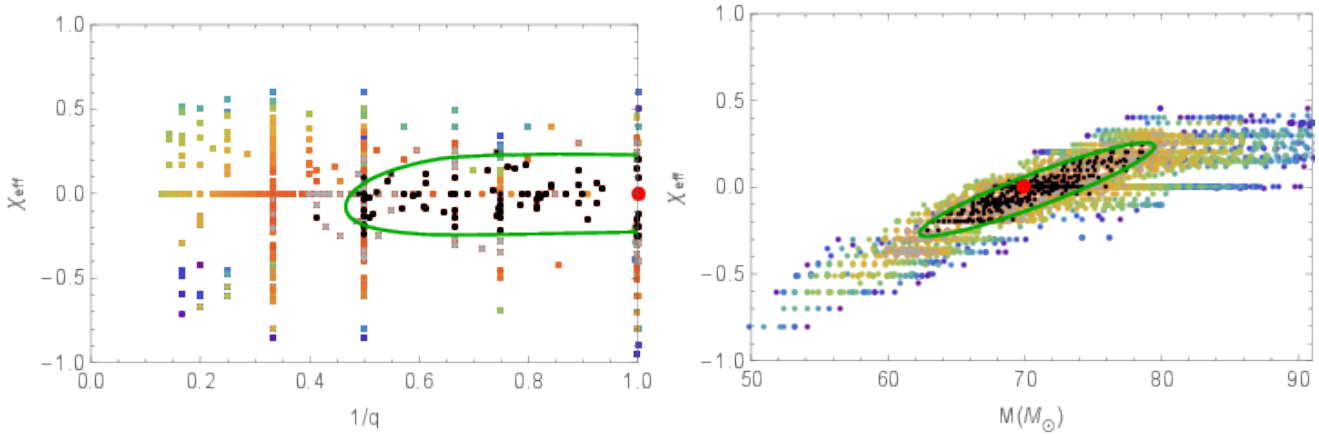

FIG. 11: **Parameter recovery for zero spin equal mass binary I**: Each point represents a NR simulation and a particular total mass compared against a SXS-1 source. The left panel shows $\chi_{\rm eff}$ vs 1/q with q=m_1/m_2 and $\chi_{\rm eff}$ defined in Eq. (2), and the right panel shows $\chi_{\rm eff}$ vs M. The gray points represent points that fall between $\ln \mathcal{L}_{\rm marg} = 130$ and $\ln \mathcal{L}_{\rm marg} = 127$. The black points represent points that fall in $\ln \mathcal{L}_{\rm marg} > 130$, i.e. templates that best match the source. The peak value with this run was $\ln \mathcal{L}_{\rm marg} = 134$. These intervals were determined using the inverse χ^2 distribution (see Eq. 10). The rest of the colors represent all the points $\ln \mathcal{L}_{\rm marg} < 127$ with the red represent the highest in the region. The green contour is the 90% CI derived using the quadratic fit to $\ln \mathcal{L}_{\rm marg}$ for nonprecessing systems only. The big red dot represents the true parameters of the source. We are able to recover the 2D posterior distribution that is consistent with the distributions with $\ln \mathcal{L}_{\rm marg} > 130$ (black points).

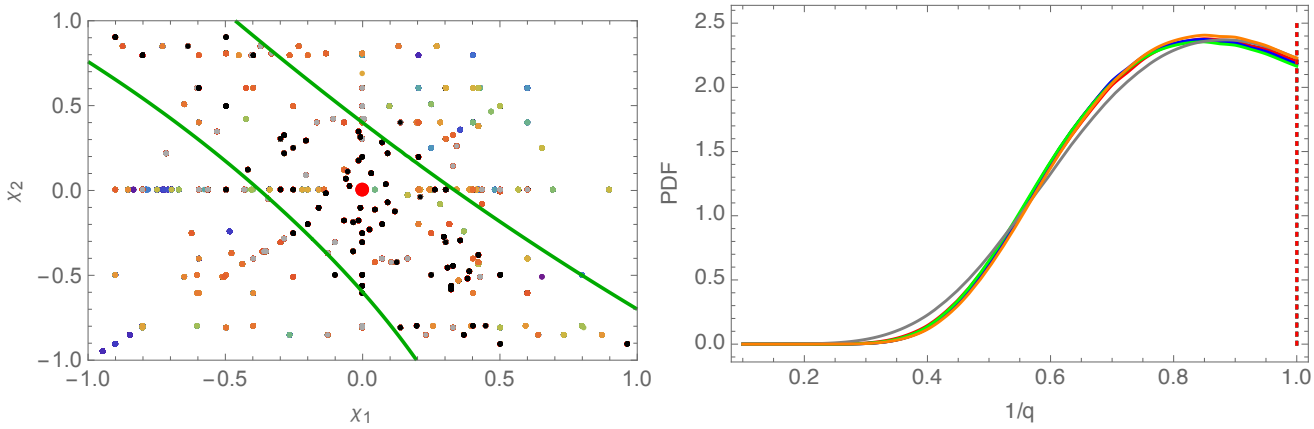

FIG. 12: **Parameter recovery for zero spin equal mass binary II**: The left panel shows the $\ln \mathcal{L}_{\rm marg}$ as a function of χ_{1z} and χ_{2z}. The rainbow, gray, and black points represent the same intervals as in Figure 11. The green contour also represents the same CI as Figure 11. The right panel shows the 1D posterior distribution for $1/q$. This 1D posterior was derived from the quadratic fit of to $\ln \mathcal{L}_{\rm marg}$ for nonprecessing systems only. Here we show results for six inclinations: $\iota = 0.0$ (black), $\iota = 0.5$ (red), $\iota = 0.785$ (blue), $\iota = 1.0$ (green), $\iota = 1.5$ (gray), $\iota = 2.35$ (orange). We see that the results from all the inclinations are the same, i.e. no more information can be obtained with higher order modes.

C. Precessing binaries: unequal mass ratios and precessing spin, but short duration

Since all the fits in this study have only used the non-precessing binaries, one might come to the conclusion that this limits us to analyzing only zero spin and aligned source. We can potentially recover parameters of precessing sources if the duration of these sources are short enough; this translates to only a few cycles and therefore little to no precession before merger, see before Eq. (9) in [46]. Figure 15 are the same type of plots as in Figure 11. Here the gray points represent points that fall between $\ln \mathcal{L}_{\rm marg} = 165$ and $\ln \mathcal{L}_{\rm marg} = 163$, and the black points represent points that fall in $\ln \mathcal{L}_{\rm marg} > 165$. The colored points represent the points that fall in the region $\ln \mathcal{L}_{\rm marg} < 163$ with the red points represent the higher $\ln \mathcal{L}_{\rm marg}$ values. As with the previous cases, these intervals were determined using the inverse χ^2 distribution [see Eq. (10)] adopting $d = 4$ (two masses with aligned spin) for the black points and $d = 8$ (two masses with precessing spins) for the gray points. As we expected, the short duration of this source allows us to recover the parameters with a fit that only uses the nonprecessing cases as shown in the left panel of Figure 17. Here we plot the $\ln \mathcal{L}_{\rm marg}(M)$ of a single null run of *ILE* comparing SXS-0234v2 with itself

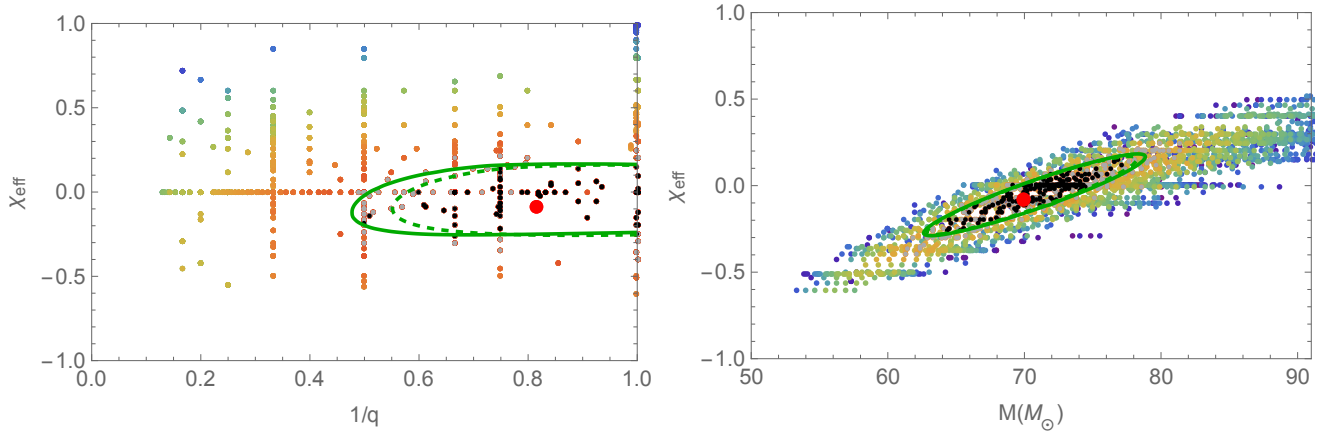

FIG. 13: **Parameter recovery for an aligned, GW150914-like unequal mass binary I**: Each point represents a NR simulation and a particular total mass compared against a SXS-0233 source. The left panel shows $\chi_{\rm eff}$ vs 1/q with q=m_1/m_2, and the right panel shows $\chi_{\rm eff}$ vs M with $\chi_{\rm eff}$ defined in Eq. (2). The gray points represent points that fall between $\ln \mathcal{L}_{\rm marg} = 167$ and $\ln \mathcal{L}_{\rm marg} = 165$. The black points represent points that fall in $\ln \mathcal{L}_{\rm marg} > 167$, i.e. templates that best match the source. The rest of the colors represent all the points $\ln \mathcal{L}_{\rm marg} < 165$ with the red represent the highest in the region. The green contours are the 90% CI derived using the quadratic fit to $\ln \mathcal{L}_{\rm marg}$ for nonprecessing systems only. The dash line is the CI for $l \leq 3$, and the solid line is the CI for $l \leq 2$. The big red dot represents the true parameters of the source. We are able to better constrain the posterior by using higher modes for this system.

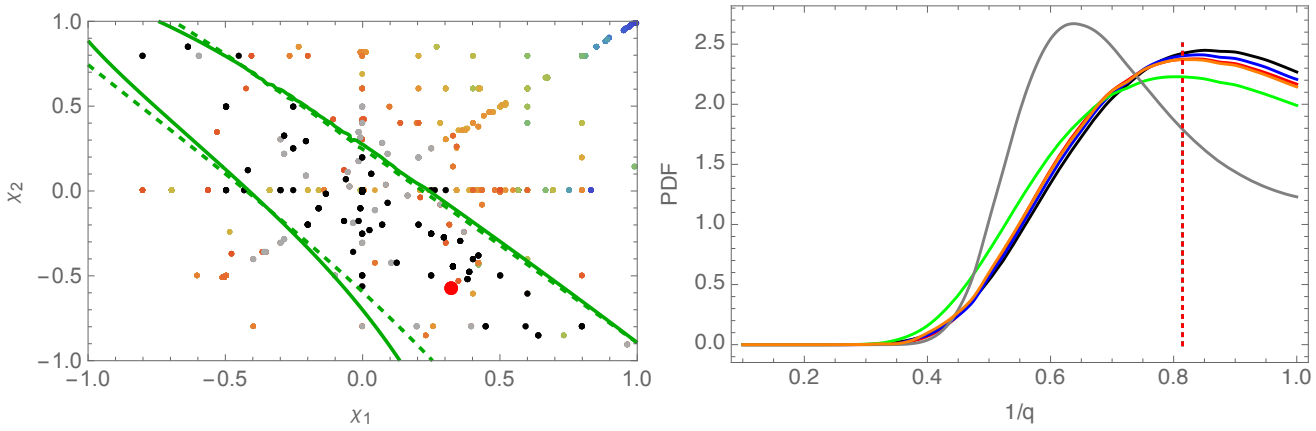

FIG. 14: **Parameter recovery for an aligned, GW150914-like unequal binary II**: The left panel shows the $\ln \mathcal{L}_{\rm marg}$ as a function of χ_{1z} and χ_{2z}. The colored, gray, and black points represent the same intervals as in Figure 13. The green contours also represents the same CI as Figure 11. The big red dot represents the true parameters of the source. The right panel shows the 1D posterior distribution for 1/q. This 1D posterior was derived from the quadratic fit of to $\ln \mathcal{L}_{\rm marg}$ for nonprecessing systems only. Here we show results for six inclinations all represented by the same colors as the zero spin case, see Figure 12. In this case, we see significant differences between the curves implying that higher order modes could important for accurate analysis of this source.

(black) and the whole end-to-end $\ln \mathcal{L}_{\rm marg}(M)$ using SXS-0234v2 as the source. By construction, the $\ln \mathcal{L}_{\rm marg}$ from the null run of SXS-0234v2 is the highest $\ln \mathcal{L}_{\rm marg}(M)$ possible. If the maximum $\ln \mathcal{L}_{\rm marg}$ from the whole end-to-end run is close ($\Delta \ln L \leq 1$), we can recover the parameters of the simulations without fitting with the precessing systems. In this case, the $\Delta \ln L = 0.97$. We can therefore accurately recover the parameters of this precessing system as evident by Figure 15.[3]

We again show $\ln \mathcal{L}_{\rm marg}$ as a function of χ_{1z} and χ_{2z} in the left panel of Figure 16 with all the colors and contours representing the as in Figure 12. The green contour are consistent with the black point distribution. We again plot the 1D dis-

[3] When interpreting the above statement, however, it is important to note our analysis by construction uses only information $f > 30$Hz. If we had access to a wider range of long simulations, we could have access to information from precession cycles between $10 - 30$Hz, even for sources of this kind and in this data. More work is needed to assess the prospects for recovery for longer, more generic sources.

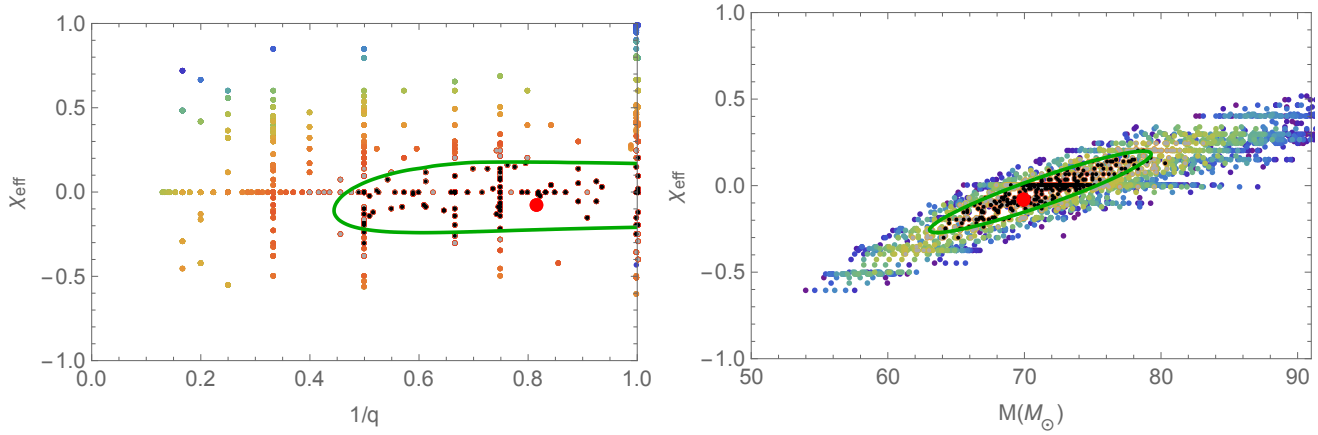

FIG. 15: **Parameter recovery for an precessing, short, unequal mass binary I**: Each point represents a NR simulation and a particular total mass compared against a SXS-0234v2 source with $l \leq 2$ modes. The left panel shows the $\chi_{\rm eff}$ vs 1/q with q=m_1/m_2 and $\chi_{\rm eff}$ defined in Eq. (2), and the right panel shows the $\chi_{\rm eff}$ vs M. The gray points represent points that fall between $\ln \mathcal{L}_{\rm marg} = 165$ and $\ln \mathcal{L}_{\rm marg} = 163$. The black points represent points that fall in $\ln \mathcal{L}_{\rm marg} > 165$, i.e. templates that best match the source. The rest of the colors represent all the points $\ln \mathcal{L}_{\rm marg} < 163$ with the red represent the highest in the region. The green contour is the 90% CI derived using the quadratic fit to $\ln \mathcal{L}_{\rm marg}$ for nonprecessing systems only. The big red dot represents the true parameters of the source. We are able to recover the 2D posterior distribution that is consistent with the distributions with $\ln \mathcal{L}_{\rm marg} > 165$ (black points).

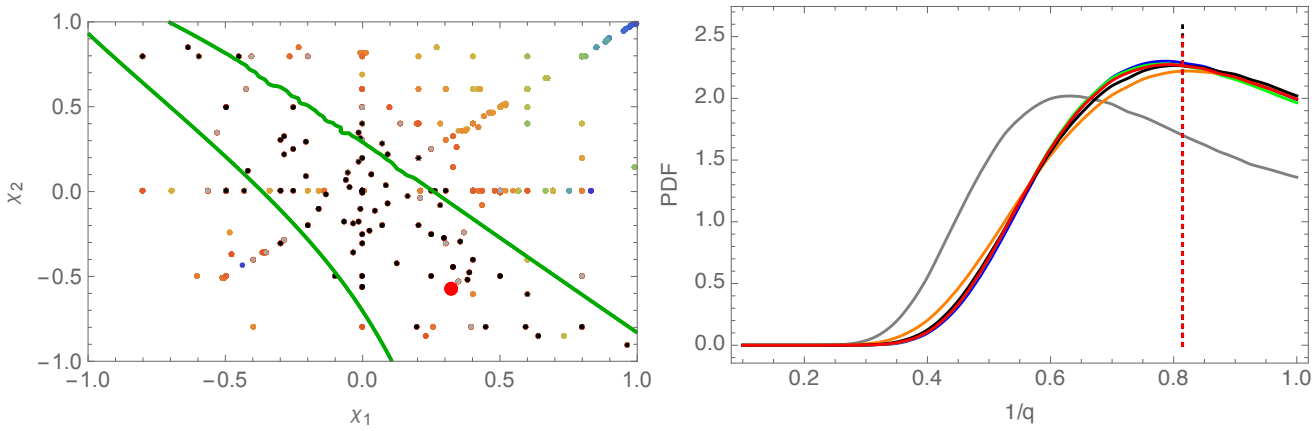

FIG. 16: **Parameter recovery for an precessing, short, unequal mass binary II**: The left panel shows the $\ln \mathcal{L}_{\rm marg}$ as a function of χ_{1z} and χ_{2z}. The gray, black, and other color points represent the same intervals as in Figure 15. The green contour represents the same contour as in Figure 15. The big red dot represents the true parameters of the source. The green contour is consistent with the black point distribution. The right panel shows the 1D posterior distribution for 1/q. This 1D posterior was derived from the quadratic fit of to $\ln \mathcal{L}_{\rm marg}$ for nonprecessing systems only. Here we show results for the same 6 inclinations all represented by the same colors as the zero spin case, see Figure 12. In this case, we see significant differences between the curves implying that higher order modes could important for accurate analysis of this source. He also see a large discrepancies between the $\iota = 1.5$ distribution and the other inclinations. See Figure 17 and Figure 18 for further analyses.

tribution for 1/q for different inclinations in the right panel of Figure 16 with all the colors corresponding to the same inclinations as in the right panel of Figure 12. Here we see relative consistency between the different inclinations, with a consistent trend towards extracting marginally more information as the inclination increases. We have an outlier for $\iota = 1.5$: a nearly edge-on line of sight. For such a line of sight, keeping in mind we tune the source distance to fix the network SNR, precession-induced modulations are amplified; this outlier *could* and probably does represent the impact of precession. To investigate this further, we again plot $\ln \mathcal{L}_{\rm marg}(M)$ of a single null run of *ILE* comparing SXS-0234v2 with itself (black) and the whole end-to-end $\ln \mathcal{L}_{\rm marg}(M)$ using SXS-0234v2 with $\iota = 1.5$ as the source, see the right panel of Figure 17. By construction, the $\ln \mathcal{L}_{\rm marg}$ from the null run of SXS-0234v2 is the highest $\ln \mathcal{L}_{\rm marg}(M)$ possible. Here we find a bigger difference between $\ln \mathcal{L}_{\rm marg}$ of the null run and $\ln \mathcal{L}_{\rm marg}$ of the entire end-to-end run: $\Delta \ln L \sim 1.8$. We then take all the individual runs from the end-to-end runs that compared 0234v2 to itself and plot $\ln \mathcal{L}_{\rm marg}(M)$ for each inclination. As evident in Figure 18, the $\iota = 1.5$ curve lies well below the rest of the inclinations. More investigations are needed to be done to figure out this discrepancy; however, this could imply SXS-0234v2 has many modes that are rele-

vant, reflecting precession-induced modulation most apparent perpendicular to $\bar{J}$ the total angular momentum vector. In future work, where we attempt to recover all spin degrees of freedom for precessing sources, we will focus in particular on edge-on lines of sight like this.

VI. CONCLUSIONS

We have presented and assessed a method to directly interpret real gravitational wave data by comparison to numerical solutions of Einstein's equations. This method can employ existing harmonics and physics that has been or can be modeled. While any other method can do so as well if suitable models have been developed and calibrated, this method skips the step of translating NR results into model improvements, circumventing the effort and potential biases introduced in doing so.

We also provided a detailed systematic study of the potential errors introduced in our method. We first used the overlap or mismatch to assess the difference between different simulations along fiducial lines of sight. As noted in Eq. (20), we expect that $\ln \mathcal{L}$ is approximately proportional to the mismatch by an overall constant. We demonstrate this relationship explicitly, using NR sources and synthetic data. Once we obtained $\ln \mathcal{L}_{\rm marg}$, we fitted with a simple quadratic and derived a PDF using Eq. (22) with its corresponding 90% CI. Using the PDFs, we can graphically see any errors that would have been propagated through. To quantify this change, we calculated a KL Divergence between two PDFs [see Eq. (23)]. By using these diagnostics, we addressed and quantified systematic errors that could affect our parameter estimation results.

Our validation studies systematically assessed the impact of (a) Monte Carlo error, (b) waveform extraction error, (c) simulation resolution, and (d) low frequency cutoff/signal duration via our diagnostics.

- (a) Based on our results from our examples, we were confident that the error from our Monte Carlo integration would be small. To quantify the results that seem apparent by eye, we applied our diagnostics (omitting the mismatch) and found the D_{KL} between the PDFs (i.e. D_{KL}(v1,v1), D_{KL}(v1,v2), D_{KL}(v1,v3)) to be all $D_{KL} \sim 10^{-5}$.

- (b) In a similar fashion, we applied our diagnostics to GW150914-like simulations from the SXS and RIT NR groups. We validated the utility of the perturbative extraction technique but noted some differences between the strain provided by SXS and perturbative extraction applied to their ψ_4 data. Based on excellent agreement between RIT (with perturbative extraction) and SXS provided strain, we expect the discrepancies relate to improper assumptions regarding SXS coordinates. More needs to be done to discover the origin of this disparity. From our match study, we determined that the impact of the error due to waveform extraction is insignificant at a large enough extraction radius. This was validated via the D_{KL} between three PDFs with the highest possible extraction radii, which were all around $10^{-2} - 10^{-3}$.

- (c) When using our mismatch study to assess the impact of resolution error, it was determined that the mismatch for all the different resolution was $\mathcal{M} \sim 10^{-5}$. This seemingly small difference in the waveform was then reaffirmed by the corresponding $D_{KL} \sim 10^{-4} - 10^{-5}$. From our diagnostics, it was clear that the error introduced by numerical resolution was negligible.

- (d) We finally used our diagnostics to the assess impact of low frequency cutoffs and signal duration. For both NR and analytic models, the available frequency content provided can significantly affect our results. After deriving our PDFs and calculating the D_{KL}, we found the lower $f_{\rm min}(10, 20{\rm Hz})$ were very similar with a narrow PDF and a high peak while the higher $f_{\rm min}(30, 40{\rm Hz})$ produced a wider PDF with a lower peak. We stress the importance of the hybridization of the NR waveforms to allow for a low $f_{\rm min}$ to standardization NR waveforms while providing the longest waveform possible.

We also provided three end-to-end examples with three different types of sources. First, we used a simple example – zero spin equal mass, where no significant higher order modes complicate our interpretation – to show our method works. Second, we examined an aligned, GW150914-like, unequal mass source. Though the leading-order quadruple radiation from such a source is nearly degenerate with an equal mass, zero spin system, this binary has asymmetries which produce higher order modes. We used our method with the $l \leq 2$ as well as the $l \leq 3$ modes and found we could better constrain q using higher modes. We also found significant differences between the 1D probability distributions for 1/q; this implied that higher modes were significant. Third, we used our method on a precessing but short unequal mass source. Due to its short duration of the observationally accessible signal, this comparable-mass binary has little to no time to precess in band. This allows us to recover the parameters of the binary even though we construct a fit based on the nonprecessing binaries. Even though the recovery of parameters was possible, the edge-on case for our 1D distributions were significantly different than the rest. For this line of sight, precession-induced modulations are most significant; the simplifying approximation that allowed success for the other lines of sight break down. Even though we suspect this is also due to higher order modes, more needs to be done to validate this claim. In the future, we will extend this strategy to recover parameters of generic precessing sources.

The method presented here relies on interpolation between existing simulations of quasi-circular black hole binary mergers. For nonprecessing binaries, this three-dimensional space has been reasonably well-explored. For generic quasi-circular mergers, however, substantially more simulations may be required to fill the seven-dimensional parameter space sufficiently for this method. Fortunately, targeted followup numerical simulations of heavy binary black holes are always

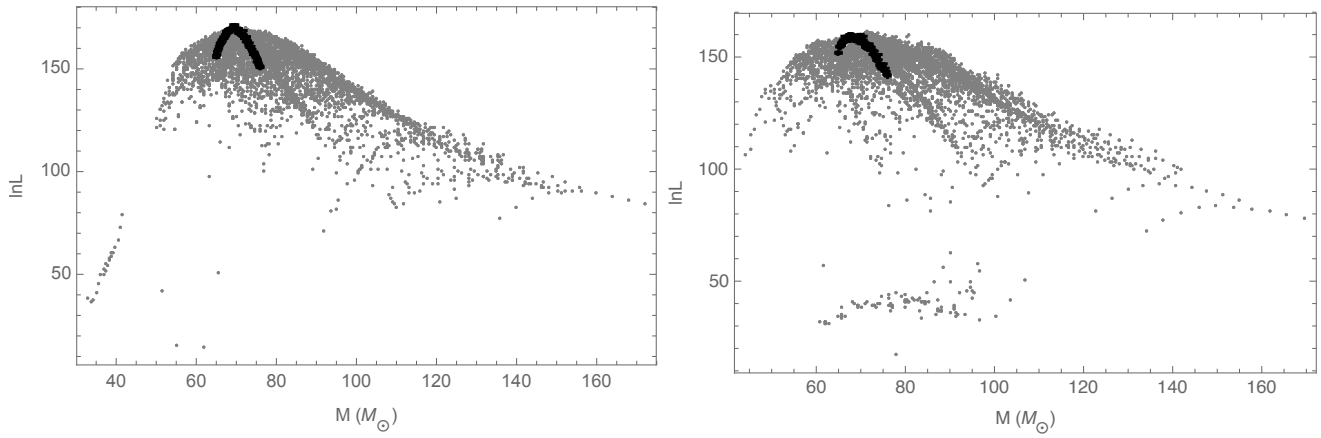

FIG. 17: **Proof of parameter recovery for an precessing, short, unequal mass binary**: Here is $\ln \mathcal{L}_{\rm marg}(M)$ of a single *ILE* null run comparing SXS-0234v2 with itself (black) and the $\ln \mathcal{L}_{\rm marg}(M)$ for the full end-to-end run with SXS-0234v2 as its source (gray). The left panel represent runs with a source with $\imath = 0.0$, and the right panel represent runs with a source with $\imath = 1.5$. The gray points only include the nonprecessing templates. If we take the difference between the $\ln \mathcal{L}_{\rm marg}$ from the whole end-to-end run and the $\ln \mathcal{L}_{\rm marg}$ from the null run, we get a $\Delta \ln L \sim 0.97$ for $\imath = 0.0$ and $\Delta \ln L \sim 1.8$ for $\imath = 1.5$. Even if we were to include the best template in our end-to-end runs (which is itself), we only get a slight increase in the $\ln \mathcal{L}_{\rm marg}$ for the face-on inclination. However, the edge-on case change seems significant; see Figure 18 for an investigation focusing on the peak values.

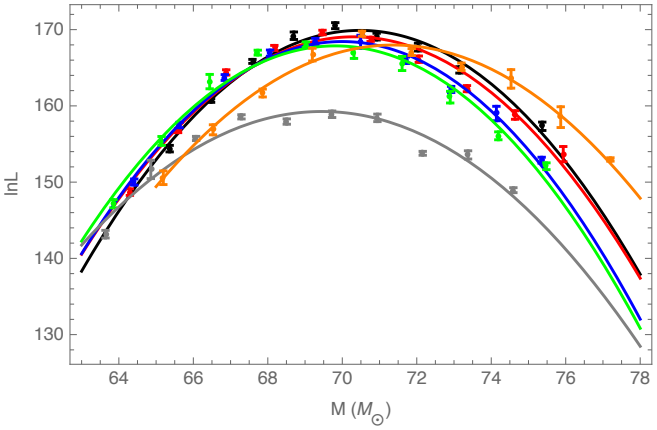

FIG. 18: **Discrepancy in** $\ln \mathcal{L}_{\rm marg}(M)$ **for** $\imath = 1.5$: This is a plot of multiple $\ln \mathcal{L}_{\rm marg}(M)$ comparing SXS-0234v2 with itself at different inclinations. Here $\imath = 0.0$ is black, $\imath = 0.5$ is red, $\imath = 0.785$ is blue, $\imath = 1.0$ is green, $\imath = 1.5$ is gray, and $\imath = 2.35$ is orange. The edge-on case is clearly different than the rest of inclinations; more needs to be done to discover the origin of this discrepancy; however, this could be due to many significant higher modes.

possible. These simulations will be incredibly valuable to validate any inferences about binary black hole mergers, from this or any other method. For this method in particular, followup simulations can be used to directly assess our estimates, and revise them. We will outline followup strategies and iterative fitting procedures in subsequent work.

Acknowledgments

The RIT authors gratefully acknowledge the NSF for financial support from Grants: No. PHY-1505629, No. AST-1664362 No. PHY-1607520, No. ACI-1550436, No. AST-1516150, and No. ACI-1516125. Computational resources were provided by XSEDE allocation TG-PHY060027N, and by NewHorizons and BlueSky Clusters at Rochester Institute of Technology, which were supported by NSF grant No. PHY-0722703, DMS-0820923, AST-1028087, and PHY-1229173. This research was also part of the Blue Waters sustained-petascale computing NSF projects ACI-0832606, ACI-1238993, and OCI-1515969, OCI-0725070.

The SXS collaboration authors gratefully acknowledge the NSF for financial support from Grants: No. PHY-1307489, No. PHY-1606522, PHY-1606654, and AST- 1333129. They also gratefully acknowledge support for this research at CITA from NSERC of Canada, the Ontario Early Researcher Awards Program, the Canada Research Chairs Program, and the Canadian Institute for Advanced Research. Calculations were done on the ORCA computer cluster, supported by NSF grant PHY-1429873, the Research Corporation for Science Advancement, CSU Fullerton, the GPC supercomputer at the SciNet HPC Consortium [47]; SciNet is funded by: the Canada Foundation for Innovation (CFI) under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund (ORF) – Research Excellence; and the University of Toronto. Further calculations were performed on the Briarée cluster at Sherbrooke University, managed by Calcul Québec and Compute Canada and with operation funded by the Canada Foundation for Innovation (CFI), Ministére de l'Économie, de l'Innovation et des Exportations du Quebec (MEIE), RMGA and the Fonds de recherche du Québec - Nature et Technologies (FRQ-NT).

The GT authors gratefully acknowledge the NSF for financial support from Grants: No. ACI-1550461 and No. PHY-1505824. Computational resources were provided by XSEDE and the Georgia Tech Cygnus Cluster.

Finally, the authors are grateful for computational resources used for the parameter estimation runs provided by the Leonard E Parker Center for Gravitation, Cosmology and Astrophysics at the University of Wisconsin-Milwaukee; the Albert Einstein Institute at Hanover, Germany; and the California Institute of Technology at Pasadena, California.

[1] The LIGO Scientific Collaboration and the Virgo Collaboration. Direct Observation of Gravitational Waves from a Binary Black Hole Merger. *Phys. Rev. Lett*, 16:061102–+, February 2016.

[2] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, and et al. Binary Black Hole Mergers in the First Advanced LIGO Observing Run. *Physical Review X*, 6(4):041015, October 2016.

[3] A. Taracchini, A. Buonanno, Y. Pan, T. Hinderer, M. Boyle, D. A. Hemberger, L. E. Kidder, G. Lovelace, A. H. Mroué, H. P. Pfeiffer, M. A. Scheel, B. Szilágyi, N. W. Taylor, and A. Zenginoglu. Effective-one-body model for black-hole binaries with generic mass ratios and spins. *Phys. Rev. D*, 89(6):061502, March 2014.

[4] M. Pürrer. Frequency-domain reduced order models for gravitational waves from aligned-spin compact binaries. *Classical and Quantum Gravity*, 31(19):195010, October 2014.

[5] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer. Simple Model of Complete Precessing Black-Hole-Binary Gravitational Waveforms. *Physical Review Letters*, 113(15):151101, October 2014.

[6] The LIGO Scientific Collaboration, the Virgo Collaboration, B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, and et al. Directly comparing GW150914 with numerical solutions of Einstein's equations for binary black hole coalescence. *ArXiv e-prints*, June 2016.

[7] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, and et al. Effects of waveform model systematics on the interpretation of GW150914. *Classical and Quantum Gravity*, 34(10):104002, May 2017.

[8] T. Damour. Coalescence of two spinning black holes: An effective one-body approach. *Phys. Rev. D*, 64(12):124013, December 2001.

[9] É. Racine. Analysis of spin precession in binary black hole systems including quadrupole-monopole interaction. *Phys. Rev. D*, 78(4):044021, August 2008.

[10] P. Ajith, M. Hannam, S. Husa, Y. Chen, B. Brügmann, N. Dorband, D. Müller, F. Ohme, D. Pollney, C. Reisswig, L. Santamaría, and J. Seiler. Inspiral-Merger-Ringdown Waveforms for Black-Hole Binaries with Nonprecessing Spins. *Physical Review Letters*, 106(24):241101, June 2011.

[11] C. Pankow, P. Brady, E. Ochsner, and R. O'Shaughnessy. Novel scheme for rapid parallel parameter estimation of gravitational waves from compact binary coalescences. *Phys. Rev. D*, 92(2):023002, July 2015.

[12] D. Shoemaker, B. Vaishnav, I. Hinder, and F. Herrmann. Numerical relativity meets data analysis: spinning binary black hole case. *Classical and Quantum Gravity*, 25(11):114047, June 2008.

[13] E. Berti, V. Cardoso, J. A. Gonzalez, U. Sperhake, M. Hannam, S. Husa, and B. Brügmann. Inspiral, merger, and ringdown of unequal mass black hole binaries: A multipolar analysis. *Phys. Rev. D*, 76(6):064034, September 2007.

[14] J. D. Schnittman, A. Buonanno, J. R. van Meter, J. G. Baker, W. D. Boggs, J. Centrella, B. J. Kelly, and S. T. McWilliams. Anatomy of the binary black hole recoil: A multipolar analysis. *Phys. Rev. D*, 77(4):044031, February 2008.

[15] R. O'Shaughnessy, B. Vaishnav, J. Healy, and D. Shoemaker. Intrinsic selection biases of ground-based gravitational wave searches for high-mass black hole-black hole mergers. *Phys. Rev. D*, 82(10):104006, November 2010.

[16] L. E. Kidder. Coalescing binary systems of compact objects to (post)$^{5/2}$-Newtonian order. V. Spin effects. *Phys. Rev. D*, 52:821–847, July 1995.

[17] L. Blanchet. Gravitational Radiation from Post-Newtonian Sources and Inspiralling Compact Binaries. *Living Reviews in Relativity*, 17:2, February 2014.

[18] A. H. Mroué, M. A. Scheel, B. Szilágyi, H. P. Pfeiffer, M. Boyle, D. A. Hemberger, L. E. Kidder, G. Lovelace, S. Ossokine, N. W. Taylor, A. Zenginoğlu, L. T. Buchman, T. Chu, E. Foley, M. Giesler, R. Owen, and S. A. Teukolsky. Catalog of 174 Binary Black Hole Simulations for Gravitational Wave Astronomy. *Physical Review Letters*, 111(24):241104, December 2013.

[19] J. Healy, C. O. Lousto, Y. Zlochower, and M. Campanelli. The RIT binary black hole simulations catalog. *ArXiv e-prints*, March 2017.

[20] K. Jani, J. Healy, J. A. Clark, L. London, P. Laguna, and D. Shoemaker. Georgia tech catalog of gravitational waveforms. *Classical and Quantum Gravity*, 33(20):204001, October 2016.

[21] N. T. Bishop and L. Rezzolla. Extraction of Gravitational Waves in Numerical Relativity. *ArXiv e-prints*, June 2016.

[22] H. Nakano, J. Healy, C. O. Lousto, and Y. Zlochower. Perturbative extraction of gravitational waveforms generated with numerical relativity. *Phys. Rev. D*, 91(10):104022, May 2015.

[23] C. Reisswig and D. Pollney. Notes on the integration of numerical relativity waveforms. *Classical and Quantum Gravity*, 28(19):195015, October 2011.

[24] P. Kumar, I. MacDonald, D. A. Brown, H. P. Pfeiffer, K. Cannon, M. Boyle, L. E. Kidder, A. H. Mroué, M. A. Scheel, B. Szilágyi, and A. Zenginoğlu. Template banks for binary black hole searches with numerical relativity waveforms. *Phys. Rev. D*, 89(4):042002, February 2014.

[25] C. Cutler and M. Vallisneri. LISA detections of massive black hole inspirals: Parameter extraction errors due to inaccurate template waveforms. *Phys. Rev. D*, 76(10):104018, November 2007.

[26] H.S. Cho, E. Ochsner, R. O'Shaughnessy, C. Kim, and C.H. Lee. Gravitational waves from BH-NS binaries: Phenomenological Fisher matricies and parameter estimation using higher harmonics. *Phys. Rev. D*, 87:02400–+, January 2013.

[27] R. O'Shaughnessy, B. Farr, E. Ochsner, H.-S. Cho, C. Kim, and C.-H. Lee. Parameter estimation of gravitational waves from nonprecessing black hole-neutron star inspirals with higher harmonics: Comparing Markov-chain Monte Carlo posteriors to

an effective Fisher matrix. *Phys. Rev. D*, 89(6):064048, March 2014.

[28] C. Cutler and É. E. Flanagan. Gravitational waves from merging compact binaries: How accurately can one extract the binary's parameters from the inspiral waveform\? *Phys. Rev. D*, 49:2658–2697, March 1994.

[29] L. Lindblom, B. J. Owen, and D. A. Brown. Model waveform accuracy standards for gravitational wave data analysis. *Phys. Rev. D*, 78(12):124020, December 2008.

[30] C. Markakis, J. S. Read, M. Shibata, K. Uryū, J. D. E. Creighton, J. L. Friedman, and B. D. Lackey. Neutron star equation of state via gravitational wave observations. In *Journal of Physics Conference Series*, volume 189 of *Journal of Physics Conference Series*, page 012024, October 2009.

[31] J. S. Read, C. Markakis, M. Shibata, K. Uryū, J. D. E. Creighton, and J. L. Friedman. Measuring the neutron star equation of state with gravitational wave observations. *Phys. Rev. D*, 79(12):124033, June 2009.

[32] Z. Ivezic, A. J. Connolly, and VanderPlas, J. T. and Gray, A. *Statistics, Data Mining, and Machine Learning in Astronomy*. Princeton Press, Cambridge, UK, 2014.

[33] G. Lovelace, C. Lousto, J. Healy, M. Scheel, A. Garcia, R. O'Shaughnessy, M. Boyle, M. Campanelli, D. Hemberger, L.E. Kidder, H. Pfeiffer, B. Szilagyi, S. Teukolsky, and Y. Zlochower. Modeling the source of GW150914 with targeted numerical-relativity simulations. *Submitted to CQG; available as arxiv:1607.05377*, jun 2016.

[34] P. Kumar, K. Barkett, S. Bhagwat, N. Afshari, D. A. Brown, G. Lovelace, M. A. Scheel, and B. Szilágyi. Accuracy and precision of gravitational-wave models of inspiraling neutron star-black hole binaries with spin: Comparison with matter-free numerical relativity in the low-frequency regime. *Phys. Rev. D*, 92(10):102001, November 2015.

[35] V. Varma, P. Ajith, S. Husa, J. C. Bustillo, M. Hannam, and M. Pürrer. Gravitational-wave observations of binary black holes: Effect of nonquadrupole modes. *Phys. Rev. D*, 90(12):124004, December 2014.

[36] V. Varma and A. Parameswaran. Effects of non-quadrupole modes in the detection and parameter estimation of black hole binaries with nonprecessing spins. *In preparation (LIGO P1600332)*, 2016.

[37] J. Calderón Bustillo, S. Husa, A. M. Sintes, and M. Pürrer. Impact of gravitational radiation higher order modes on single aligned-spin gravitational wave searches for binary black holes. *Phys. Rev. D*, 93(8):084019, April 2016.

[38] A. Taracchini, Y. Pan, A. Buonanno, E. Barausse, M. Boyle, T. Chu, G. Lovelace, H. P. Pfeiffer, and M. A. Scheel. Prototype effective-one-body model for nonprecessing spinning inspiral-merger-ringdown waveforms. *Phys. Rev. D*, 86(2):024011, July 2012.

[39] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, and et al. Properties of the Binary Black Hole Merger GW150914. *Physical Review Letters*, 116(24):241102, June 2016.

[40] M. Hannam, S. Husa, J. A. González, U. Sperhake, and B. Brügmann. Where post-Newtonian and numerical-relativity waveforms meet. *Phys. Rev. D*, 77(4):044020, February 2008.

[41] M. Boyle. *Accurate gravitational waveforms from binary black-hole systems*. PhD thesis, California Institute of Technology, 2008.

[42] P. Ajith, S. Babak, Y. Chen, M. Hewitson, B. Krishnan, A. M. Sintes, J. T. Whelan, B. Brügmann, P. Diener, N. Dorband, J. Gonzalez, M. Hannam, S. Husa, D. Pollney, L. Rezzolla, L. Santamaría, U. Sperhake, and J. Thornburg. Template bank for gravitational waveforms from coalescing binary black holes: Nonspinning binaries. *Phys. Rev. D*, 77(10):104017, May 2008.

[43] I. MacDonald, A. H. Mroué, H. P. Pfeiffer, M. Boyle, L. E. Kidder, M. A. Scheel, B. Szilágyi, and N. W. Taylor. Suitability of hybrid gravitational waveforms for unequal-mass binaries. *Phys. Rev. D*, 87(2):024009, January 2013.

[44] N. T. Bishop and L. Rezzolla. Extraction of Gravitational Waves in Numerical Relativity. *ArXiv e-prints*, June 2016.

[45] J. Calderón Bustillo, A. Bohé, S. Husa, A. M. Sintes, M. Hannam, and M. Pürrer. Comparison of subdominant gravitational wave harmonics between post-Newtonian and numerical relativity calculations and construction of multi-mode hybrids. *ArXiv e-prints*, January 2015.

[46] The LIGO Scientific Collaboration and the Virgo Collaboration. GW150914: A merging binary black hole at redshift 0.1. *Available at https://dcc.ligo.org/LIGO-P1500218*, February 2016.

[47] Chris Loken, Daniel Gruner, Leslie Groer, Richard Peltier, Neil Bunn, Michael Craig, Teresa Henriques, Jillian Dempsey, Ching-Hsing Yu, Joseph Chen, L Jonathan Dursi, Jason Chong, Scott Northrup, Jaime Pinto, Neil Knecht, and Ramses Van Zon. SciNet: Lessons Learned from Building a Power-efficient Top-20 System and Data Centre. *J. Phys.: Conf. Ser.*, 256:012026, 2010.

Appendix A: Exploring the parameter space

In this appendix, we provide additional examples of our method using numerical relativity simulations in different regions of parameter space. We demonstrate our method works reliably for extreme black hole spins (Figure 20) as well as in regions where few simulations with comparable parameters are available (Figures 21 and 22). For the parameters of each source, see the following source labels (in order as they appear) in Table I: RIT-5, SXS-high-antispin, SXS-$\chi_{\rm eff}$0.4, and RIT-2.

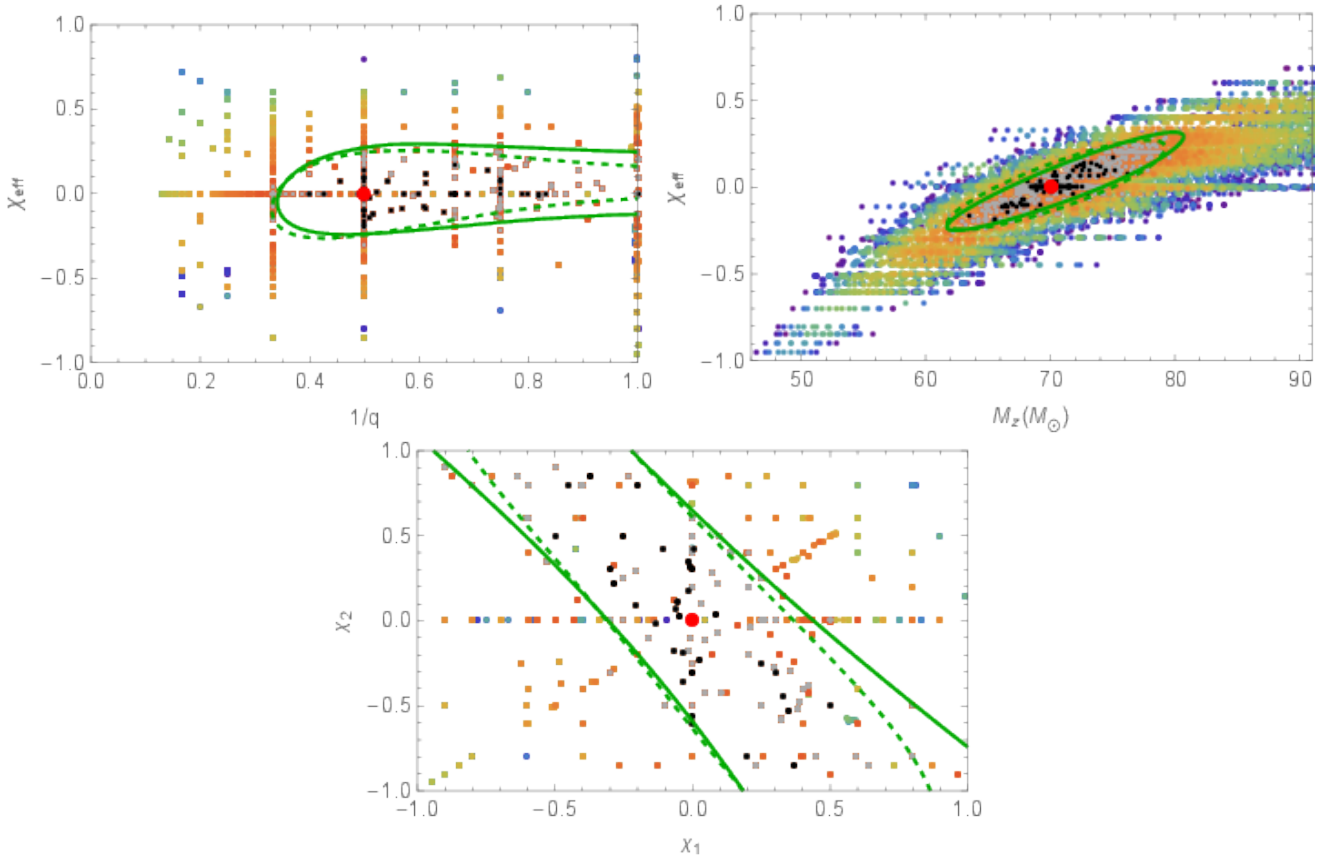

FIG. 19: **Parameter recovery for a zero spin,** $q = 2$ **binary**: Each point represents a NR simulation and a particular total mass compared against a RIT-5 source. The top left panel shows the $\chi_{\rm eff}$ vs 1/q with q=m_1/m_2 and $\chi_{\rm eff}$ defined in Eq. (2), the top right panel shows the $\chi_{\rm eff}$ vs M, and the bottom panel shows the χ_1 vs χ_2. The gray points represent points that fall between $\ln \mathcal{L}_{\rm marg} = 166$ and $\ln \mathcal{L}_{\rm marg} = 164$. The black points represent points that fall in $\ln \mathcal{L}_{\rm marg} > 167$, i.e. templates that best match the source. The rest of the colors represent all the points $\ln \mathcal{L}_{\rm marg} < 164$ with the red represent the highest in the region. The green contour is the 90% CI derived using the quadratic fit to $\ln \mathcal{L}_{\rm marg}$ for nonprecessing systems only. The dash line is the CI for $l \leq 3$, and the solid line is the CI for $l \leq 2$. The big red dot represents the true parameters of the source. We are able to recover the 2D posterior distribution that is consistent with the distributions with $\ln \mathcal{L}_{\rm marg} > 167$ (black points).

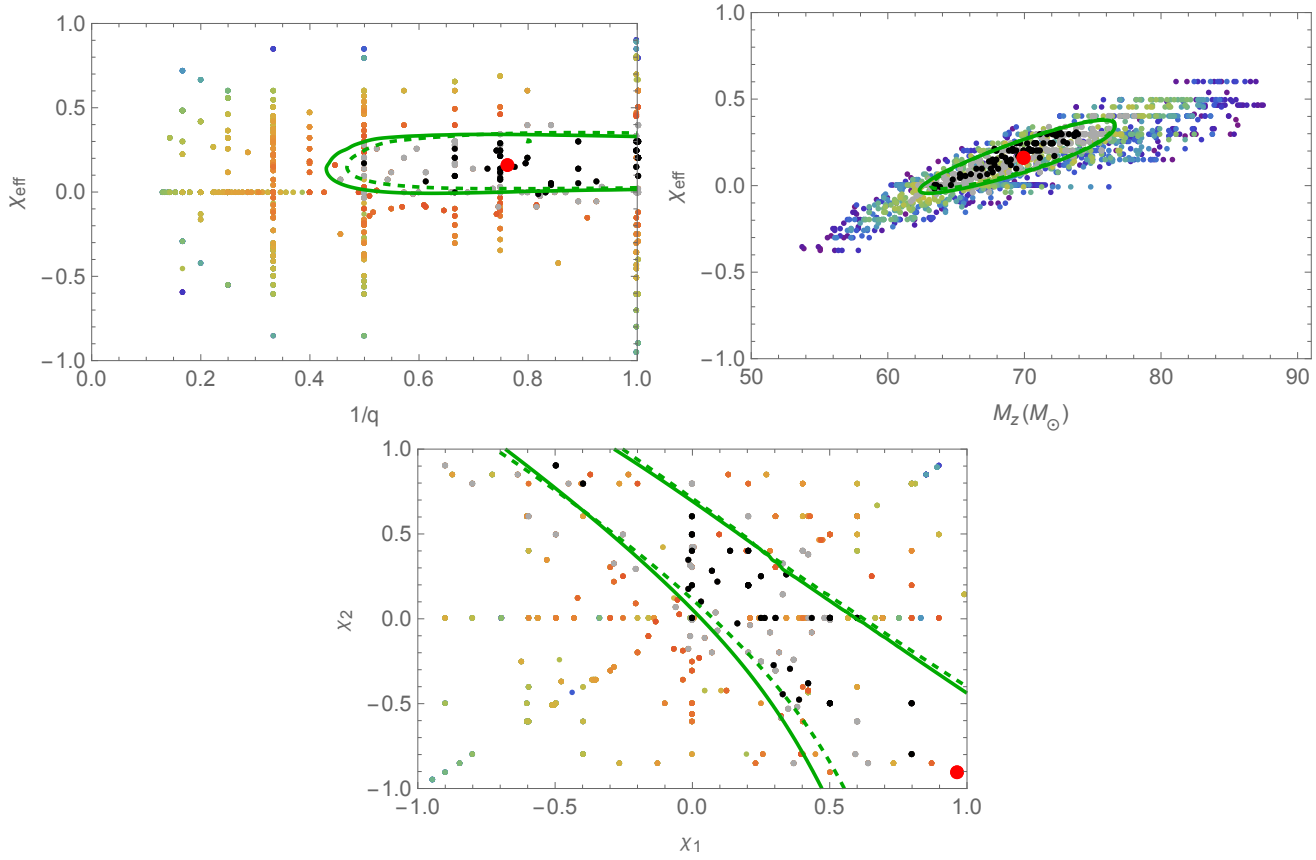

FIG. 20: **Parameter recovery for a high, anti-aligned spin** $q = 1.31$ **binary**: Each point represents a NR simulation and a particular total mass compared against a SXS-high-antispin source. The top left panel shows the $\chi_{\rm eff}$ vs 1/q with q=m_1/m_2 and $\chi_{\rm eff}$ defined in Eq. (2), the top right panel shows the $\chi_{\rm eff}$ vs M, and the bottom panel shows the χ_1 vs χ_2. The gray points represent points that fall between $\ln \mathcal{L}_{\rm marg} = 167$ and $\ln \mathcal{L}_{\rm marg} = 164$. The black points represent points that fall in $\ln \mathcal{L}_{\rm marg} > 167$, i.e. templates that best match the source. The rest of the colors represent all the points $\ln \mathcal{L}_{\rm marg} < 164$ with the red represent the highest in the region. The green contour is the 90% CI derived using the quadratic fit to $\ln \mathcal{L}_{\rm marg}$ for nonprecessing systems only. The dash line is the CI for $l \leq 3$, and the solid line is the CI for $l \leq 2$. The big red dot represents the true parameters of the source. We are able to recover the 2D posterior distribution that is consistent with the distributions with $\ln \mathcal{L}_{\rm marg} > 167$ (black points).

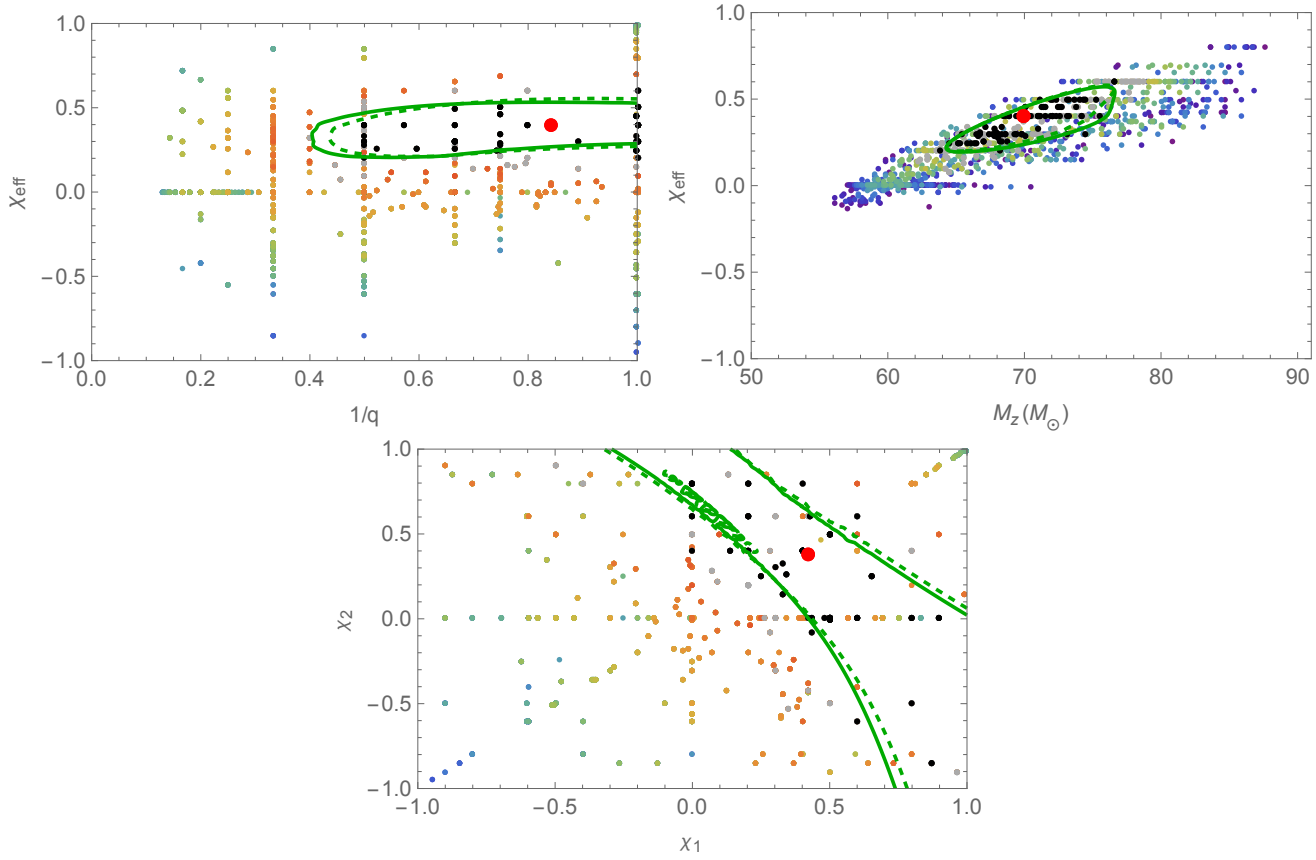

FIG. 21: **Parameter recovery for a $\chi_{\rm eff} = 0.4$ spin $q = 1.19$ binary**: Each point represents a NR simulation and a particular total mass compared against a SXS-$\chi_{\rm eff}$0.4 source. The top left panel shows the $\chi_{\rm eff}$ vs 1/q with q=m_1/m_2 and $\chi_{\rm eff}$ defined in Eq. (2), the top right panel shows the $\chi_{\rm eff}$ vs M, and the bottom panel shows the χ_1 vs χ_2. The gray points represent points that fall between $\ln \mathcal{L}_{\rm marg} = 167$ and $\ln \mathcal{L}_{\rm marg} = 164$. The black points represent points that fall in $\ln \mathcal{L}_{\rm marg} > 167$, i.e. templates that best match the source. The rest of the colors represent all the points $\ln \mathcal{L}_{\rm marg} < 164$ with the red represent the highest in the region. The green contour is the 90% CI derived using the quadratic fit to $\ln \mathcal{L}_{\rm marg}$ for nonprecessing systems only. The dash line is the CI for $l \leq 3$, and the solid line is the CI for $l \leq 2$. The big red dot represents the true parameters of the source. We are able to recover the 2D posterior distribution that is consistent with the distributions with $\ln \mathcal{L}_{\rm marg} > 167$ (black points).

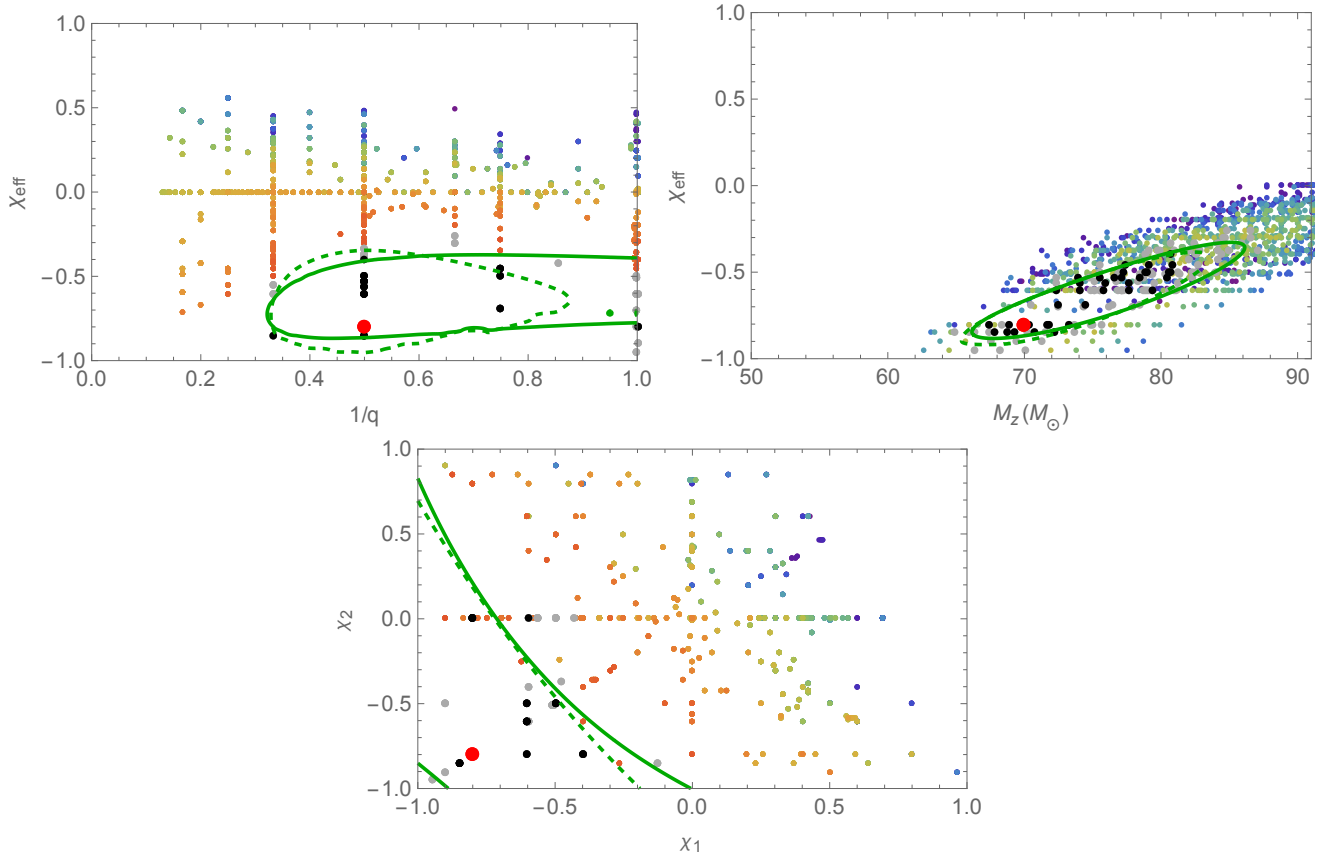

FIG. 22: **Parameter recovery for a $\chi_1 = \chi_2 = -0.8$ spin $q = 2.0$ binary**: Each point represents a NR simulation and a particular total mass compared against a RIT-2 source. The top left panel shows the $\chi_{\rm eff}$ vs 1/q with q=m_1/m_2 and $\chi_{\rm eff}$ defined in Eq. (2), the top right panel shows the $\chi_{\rm eff}$ vs M, and the bottom panel shows the χ_1 vs χ_2. The gray points represent points that fall between $\ln \mathcal{L}_{\rm marg} = 165$ and $\ln \mathcal{L}_{\rm marg} = 162$. The black points represent points that fall in $\ln \mathcal{L}_{\rm marg} > 165$, i.e. templates that best match the source. The rest of the colors represent all the points $\ln \mathcal{L}_{\rm marg} < 162$ with the red represent the highest in the region. The green contour is the 90% CI derived using the quadratic fit to $\ln \mathcal{L}_{\rm marg}$ for nonprecessing systems only. The dash line is the CI for $l \leq 3$, and the solid line is the CI for $l \leq 2$. The big red dot represents the true parameters of the source. We are able to recover the 2D posterior distribution that is consistent with the distributions with $\ln \mathcal{L}_{\rm marg} > 165$ (black points).