Fast and accurate inference on gravitational waves from precessing compact binaries

Rory Smith,[1,*] Scott E. Field,[2,†] Kent Blackburn,[1] Carl-Johan Haster,[3] Michael Pürrer,[4] Vivien Raymond,[4] and Patricia Schmidt[1,5]

[1]*LIGO, California Institute of Technology, Pasadena, California 91125, USA*
[2]*Cornell Center for Astrophysics and Planetary Science, Cornell University, Ithaca, New York 14853, USA*
[3]*School of Physics and Astronomy, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom*
[4]*Albert-Einstein-Institut, Max-Planck-Institut für Gravitationsphysik, D-14476 Golm, Germany*
[5]*TAPIR, Walter Burke Institute for Theoretical Physics, California Institute of Technology, Pasadena, California 91125, USA*

(Received 29 April 2016; published 15 August 2016)

Inferring astrophysical information from gravitational waves emitted by compact binaries is one of the key science goals of gravitational-wave astronomy. In order to reach the full scientific potential of gravitational-wave experiments, we require techniques to mitigate the cost of Bayesian inference, especially as gravitational-wave signal models and analyses become increasingly sophisticated and detailed. Reduced-order models (ROMs) of gravitational waveforms can significantly reduce the computational cost of inference by removing redundant computations. In this paper, we construct the first reduced-order models of gravitational-wave signals that include the effects of spin precession, inspiral, merger, and ringdown in compact object binaries and that are valid for component masses describing binary neutron star, binary black hole, and mixed binary systems. This work utilizes the waveform model known as "IMRPhenomPv2." Our ROM enables the use of a fast *reduced-order quadrature* (ROQ) integration rule which allows us to approximate Bayesian probability density functions at a greatly reduced computational cost. We find that the ROQ rule can be used to speed-up inference by factors as high as 300 without introducing systematic bias. This corresponds to a reduction in computational time from around half a year to half a day for the longest duration and lowest mass signals. The ROM and ROQ rules are available with the main inference library of the LIGO Scientific Collaboration, `LALInference`.

DOI: 10.1103/PhysRevD.94.044031

I. INTRODUCTION

With the first gravitational-wave (GW) detection reported in February 2016, an exciting era of GW astronomy has begun [1]. The discovery of the GW source GW150914 with the advanced Laser Interferometer Gravitational-Wave Observatory (aLIGO) was shown to match the waveform predicted by general relativity for a pair of merging black holes (BBHs) [2]. Such compact binary coalescence (CBC) events, also including merging black hole neutron star (NSBH) or neutron star pairs (BNS), are expected to be the most abundant sources, with detection rates between a few and tens per year [3,4].

Detecting gravitational waves, and subsequently performing parameter estimation (PE) to infer the astrophysical parameters encoded in those waves, is a key goal of gravitational-wave astronomy. Spin-induced precession of the binaries is a generic feature of gravitational waves emitted from CBC events, and PE studies that neglect precession will ultimately suffer from (possibly large) systematic bias in the inferred parameter values [5,6]. However, including the effects of precession into template waveforms for PE carries a high computational cost associated with waveform generation and/or sufficiently sampling the astrophysical parameter space. The time it takes to complete an analysis scales (roughly) linearly with the waveform generation cost. There is, therefore, a need to incorporate precession effects in PE studies in a computationally efficient way. Unless abated, computational costs are likely to increase (i) as more detailed physical effects are added to waveform models, e.g. higher-order modes, and (ii) when in-band signals become longer as the detector's low-frequency sensitivity improves, making the detectors more sensitive to lower mass systems.

For parameter estimation studies, we are interested in computing the posterior probability density function (PDF),

$$p(\vec{\Lambda}|d) = \frac{\mathcal{P}(\vec{\Lambda})\mathcal{L}(d|\vec{\Lambda})}{e(d)}, \tag{1}$$

on the set of model parameters $\vec{\Lambda}$, where $\mathcal{P}(\vec{\Lambda})$ is the prior probability on the model parameters, $\mathcal{L}(d|\vec{\Lambda})$ is the

*rory.smith@caltech.edu
†sfield@astro.cornell.edu

likelihood of the data and $e(d)$ is known as the Bayesian "evidence" and describes the probability of the data given the model. The evidence is typically used for model selection and enters only as an overall scaling in parameter estimation.

Assuming the detector data d contains the GW signal $h(\vec{\Lambda}_{\texttt{true}})$ and noise n, the log-likelihood function can be computed as

$$\log \mathcal{L}(d|\vec{\Lambda}) = -\frac{1}{2}(d - h(\vec{\Lambda}), d - h(\vec{\Lambda})), \tag{2}$$

where $d = h(\vec{\Lambda}_{\texttt{true}}) + n$ and (a, b) is an *overlap* integral:

$$(d, h(\vec{\Lambda})) = 4\Re\Delta f \sum_{k=1}^{L} \frac{\tilde{d}^*(f_k)\tilde{h}(f_k; \vec{\Lambda})}{S_n(f_k)}. \tag{3}$$

Here $\tilde{d}(f_k)$ and $\tilde{h}(f_k; \vec{\Lambda})$ are the discrete Fourier transforms at frequencies $\{f_k\}_{k=1}^{L}$, and $S_n(f_k)$ is the detector's noise power spectral density (PSD). For a given observation time $T = 1/\Delta f$ and detection frequency window $(f_{\texttt{high}} - f_{\texttt{low}})$, there are $L \sim \texttt{int}([f_{\texttt{high}} - f_{\texttt{low}}]T)$ sampling points in (3). When L is large and $\vec{\Lambda}$ must be sampled extensively, there are three bottlenecks: (i) evaluation of the model at each f_k, (ii) numerically computing the sum in the likelihood (2), and (iii) repeated evaluation of the likelihood.

These bottlenecks compound to escalate the cost of a typical parameter estimation analysis, even for otherwise fast-to-compute waveform models. Consider that a typical analysis can require computing several tens of millions of templates [5], and in principle these templates cannot be computed in parallel. Hence, a single likelihood evaluation must be on the order of a millisecond for the PE analyses to be on the order of tens of hours. But this is often not the case. Evaluating the closed-form frequency-domain waveform model known as IMRPhenomPv2 [7]—as implemented in the LIGO Algorithm Library [8]—takes around half a second for a low mass systems starting from 20 Hz. These numbers imply PE run times on the order of six months.[1] Other commonly used waveform families incur similar or even higher computational costs. For example, the waveform family known as "SEOBNRv2_ROM" [10]—a "reduced-order model" of the aligned-spin waveform computed within the effective one body framework, and calibrated to numerical relativity—is only around a factor of 4 less expensive than IMRPhenomPv2. Conversely, the waveform family known as "SEOBNRv3" [11]—a precessing-spin waveform family computed within the effective one body framework, and calibrated to numerical relativity simulations—is around 170 times *more* expensive to compute than IMRPhenomPv2.

Reduced-order modeling (ROM) is a promising technique for mitigating the computational cost of gravitational-wave parameter estimation. A ROM approach seeks to find a computationally efficient representation of the waveform model. If a set of $N < L$ basis elements can be found which accurately spans the continuum template space, it is possible to replace the overlap (3) with a quadrature rule containing only N terms, reducing the overall cost by a factor of L/N. This cost reduction has been demonstrated in the context of gravitational waves from nonprecessing CBCs [12], but it was hitherto unclear that templates in the precessing case were also amenable to such *linear* dimensional reduction. Here, linear refers to an approximation that is expressed as a linear superposition of basis elements. Nonlinear dimensional reduction tools described in Refs. [10,13,14] are not directly applicable for *compressed* overlap integrals.

A variety of ROM-type techniques have recently appeared in the GW literature [10,12–17]. We shall use a combination of the *reduced-basis method* and the *empirical interpolation method*, whose favorable computational efficiency, ease-of-parallelization and numerical stability make them attractive candidates for tackling precessing waveform systems and other challenging models. The reduced-basis method constructs a basis set of N elements whose span reproduces the GW model within a specified accuracy. The empirical interpolation method then uses this model-specific basis to construct an N-point interpolant defined on the model space. Substituting the empirical interpolant representation into Eq. (2) yields the *reduced-order quadrature* (ROQ) rule [12,18,19], which ultimately provides the performance gain of L/N.

One of the caveats of the ROQ method is that, in order to realize the promised L/N speed-up, we must be able to directly evaluate the waveform model at special interpolation nodes in time or frequency. Typically, this means that the model is described by a closed-form expression. Nevertheless, for other models, such as those described by differential equations, direct evaluation may be accomplished using *surrogates* [10,14,15,20]. Although surrogate models have been constructed for nonspinning [15,20] and spin-aligned waveform models [10,14], it is not obvious that they can be (easily) constructed for precessing waveform models because surrogates rely on some form of high-dimensional fitting or interpolation. We return to this issue in the conclusion.

One of the main results of this paper is to apply the reduced-basis and empirical interpolation methods to gravitational waveform models from CBCs with precessing spins. That this is possible should not be taken for granted. First, there are significant computational costs associated

[1]This time was the average of 100 waveform evaluations and overlap computations. For each evaluation, we considered binary configurations with component masses of 1 $M_\odot$ and 4 $M_\odot$ and used random spin magnitudes and orientations on each iteration. The frequency resolution of the waveform was $\Delta f = 1/128$ Hz which assumes an in-band signal duration—rounded to the next-highest power of two—of 128s from 20 Hz, which is reasonable for such a binary configuration [9]. All timing experiments, including this one, are performed using an Intel Xeon CPU with a 2.70 GHz clock speed.

with long waveforms with multiple intrinsic parameters. To overcome this challenge, we have developed and used a code called GREEDYCPP that employs fast algorithms and possesses good scalability up to at least 32,000 cores [21,22]. Specially tailored parametric and frequency sampling strategies, discussed in Sec. III D, provide additional benefits. Second, although previous results show the existence of a compact basis for spin-aligned systems [14,23], one may be worried that the complex waveform morphologies characteristic of precessing CBCs could result in a substantial increase in the basis size. This work demonstrates that there is no such increase.

Assuming that waveform generation and likelihood computation comprises the full cost of a PE study, we find theoretical speed-up improvements between a factor of 4 (for short BBH signals) and 300 (for long BNS signals). The full range of speed-up factors, which assumes that the signal is in-band starting at 20 Hz, is shown in Fig. 10. Although we assume $f_{\texttt{low}} = 20$ Hz throughout this paper, we anticipate our speed-up factors would increase (decrease) as $f_{\texttt{low}}$ is lowered (raised) for a fixed value of the binary's masses (See Fig. 1 of Ref. [12]). If the entirety of the cost of parameter estimation is assumed to be the waveform and likelihood computations, we estimate a minimum run time of analyses from 6 hours (for analyses on BBH signals up to 4s in duration) to 12 hours (for BNS/NSBH signals up to 128s in duration). The speed-up factors imply that run times without the ROQ could be on the order of 1 day to around 6 months using similar computer hardware and codes. We also show that modeling errors in the ROQ do not introduce additional systematic bias into PE, as shown in Sec. V.

The paper is outlined as follows. In Sec. II, we summarize the basics of ROQs for precessing gravitational waveform families. In Sec. III, we describe our strategies for working with a high-dimensional waveform model space. In Sec. IV, we describe the results of running our basis-building pipeline and show the accuracy of the reduced-basis and empirical interpolant. Using the `LALInference` library, in Sec. V we compare the accuracy of using the ROQ in a PE analysis to the Full likelihood function, for a simulated signal injected into recolored Gaussian noise designed to mimic early aLIGO data [24]. We also describe the speed-up one could achieve by using the ROQ in PE analyses and we set a conservative performance benchmark for the run times of efficient PE codes. In the Appendix, we describe a novel use of the reduced-basis method as a diagnostic tool for waveform models and discuss its application to IMRPhenomPv2.

II. PRELIMINARIES

A. ROQ rules for precessing multimodal gravitational wave models

A gravitational-wave strain signal $h(t)$ detected by a ground-based interferometer has the form

$$h(t;\vec{\Lambda}) = F_+(\mathrm{ra},\mathrm{dec},\psi,r)h_+(t;\phi_c,t_c,\vec{\lambda}) + F_\times(\mathrm{ra},\mathrm{dec},\psi,r)h_\times(t;\phi_c,t_c,\vec{\lambda}), \quad (4)$$

where the antenna patterns $F_{(+,\times)}$ project the gravitational wave's +- and ×-polarization states, $h_{(+,\times)}$, into the detector's frame. The antenna patterns are functions of variables which specify the orientation of the detector with respect to the binary: the distance to the source (r) as well as the right ascension (ra), declination (dec) and polarization (ψ) angles. These four variables, along with the coalescence time (t_c) and its orbital phase at coalescence (ϕ_c), describe the signal's dependence on parameters that have a trivial effect on the waveform's amplitude and phase. We shall use $\vec{\lambda}$ to denote the signal's dependence on parameters that have a nontrivial effect on the waveform's amplitude and phase, such as its masses, spin magnitude and spin orientation.[2] The strain, and consequently the likelihood (2), depends on the full set of parameters $\vec{\Lambda} = \{\mathrm{ra},\mathrm{dec},\psi,r,t_c,\phi_c,\vec{\lambda}\}$.

When discussing waveform models, it is common practice to first introduce a complex gravitational wave strain,

$$h_+(t;\phi_c,t_c,\vec{\lambda}) - \mathrm{i}h_\times(t;\phi_c,t_c,\vec{\lambda}) = \sum_{\ell=2}^{\infty}\sum_{m=-\ell}^{\ell} h^{\ell m}(t;\phi_c,t_c,\vec{\lambda})_{-2}Y_{\ell m}, \quad (5)$$

which is subsequently decomposed into a basis of spin-weighted spherical harmonics. Most gravitational waveform models make predictions for the modes $h^{\ell m}(t;\vec{\lambda})$, from which a model of what a noise-free detector records, $h(t;\vec{\Lambda})$, is readily recovered.

The remainder of this subsection sketches the steps leading to the reduced-order quadrature rule. To build computationally efficient approximations to (2), we work directly with the Fourier transform of the strain,

$$\begin{aligned}\tilde{h}(f;\vec{\Lambda}) &= \int_{-\infty}^{\infty} h(t;\vec{\Lambda})\mathrm{e}^{2\pi\mathrm{i}ft}dt \\ &= F_+\tilde{h}_+(f;\phi_c,t_c,\vec{\lambda}) + F_\times\tilde{h}_\times(f;\phi_c,t_c,\vec{\lambda}) \\ &= \mathrm{e}^{-2\pi\mathrm{i}ft_c}[F_+\tilde{h}_+(f;\phi_c,0,\vec{\lambda}) + F_\times\tilde{h}_\times(f;\phi_c,0,\vec{\lambda})]\end{aligned} \quad (6)$$

[2]Note that we refrain from discussing "intrinsic" and "extrinsic" parameters because for precessing systems, extrinsic parameters like the binary's orbital inclination can produce nontrivial effects in the waveform's amplitude and phase and so they do not simply enter as scaling factors as in nonprecessing systems. Our ROQ rule is trained over the subset of parameters $\vec{\lambda}$ but applies to the full set $\vec{\Lambda}$ (cf. Sec. II B).

where the antenna pattern's arguments are omitted for brevity. The last equality follows from $h(t;t_c) = h(t-t_c;0)$, as a nonzero coalescence time t_c simply offsets the signal's time-of-arrival. Because $\tilde{h}_{(+,\times)}$ enters linearly into (d,h) and quadratically into (h,h), one of the goals of this paper is to build (temporarily focusing on the model's internal parameterization $\vec{\lambda}$) an approximation

$$\tilde{h}_{\rm A}(f_i;\vec{\lambda}) \approx \sum_{j=1}^{N_{\rm L}} B_j(f_i)\tilde{h}_{\rm A}(F_j;\vec{\lambda}), \quad \text{with} \quad {\rm A}\in\{+,\times\}, \tag{7a}$$

$$\begin{aligned}&\Re[\tilde{h}_{\rm A}(f_i;\vec{\lambda})\tilde{h}^*_{\rm B}(f_i;\vec{\lambda})]\\ &\quad\approx \sum_{k=1}^{N_{\rm Q}} C_k(f_i)\Re[\tilde{h}_{\rm A}(\mathcal{F}_k;\vec{\lambda})\tilde{h}^*_{\rm B}(\mathcal{F}_k;\vec{\lambda})],\\ &\quad\text{with}\quad {\rm A,B}\in\{+,\times\},\end{aligned} \tag{7b}$$

that accurately approximates both the polarization states and their products. Here the labels A and B take the values $(+,\times)$, $\{B_j\}_{j=1}^{N_{\rm L}}$ is the reduced basis (RB) for the polarizations and $\{C_k\}_{k=1}^{N_{\rm Q}}$ is the RB for the real part of all possible products of the polarizations. Notice that in Eq. (7a) $\tilde{h}_+$ and $\tilde{h}_\times$ share the *same* basis $\{B_j\}_{j=1}^{N_{\rm L}}$. Similarly the approximation to the products in Eq. (7b) $\tilde{h}_+\tilde{h}^*_+$, $\tilde{h}_\times\tilde{h}^*_\times$ and $\Re\tilde{h}_+\tilde{h}^*_\times$ also share a basis $\{C_k\}_{k=1}^{N_{\rm Q}}$. The values $\tilde{h}_{\rm A}(\vec{\lambda};F_j)$ are evaluations of the A-polarization states at the *empirical interpolation nodes* $\{F_j\}_{j=1}^{N_{\rm L}}$. The location of these nodes are uniquely selected to yield accurate interpolation with the set of basis vectors $\{B_j\}_{j=1}^{N}$. Similarly, polarization products $\tilde{h}_{\rm A}(\mathcal{F}_k;\vec{\lambda})\tilde{h}^*_{\rm B}(\mathcal{F}_k;\vec{\lambda})$ are evaluated at a set of empirical interpolation nodes $\{\mathcal{F}_k\}_{k=1}^{N_{\rm Q}}$, which are distinct from $\{F_j\}_{j=1}^{N_{\rm L}}$. The approximation (7) is known as an *empirical interpolant*, and its substitution into (3) yields a *reduced-order quadrature* (ROQ) rule. The empirical interpolant constitutes a ROM of the waveform family. Sec. II C describes the algorithms we use to build (7). As described in Sec. II B, with the exception of t_c the approximation (7) automatically applies to the model's full parameterization $\vec{\Lambda}$ despite being built for the subset of internal model parameters $\vec{\lambda}$. In many of the expressions which follow, we shall use $\vec{\Lambda}$ to denote the full parameter vector but with t_c explicitly separated off.

We break the likelihood into those pieces which we can approximate using (7)

$$\begin{aligned}2\log\mathcal{L} &= 2(d,h) - (h,h) - (d,d)\\ &= 2F_+(d,h_+) + 2F_\times(d,h_\times) - |F_+|^2(h_+,h_+) - |F_\times|^2(h_\times,h_\times) - 2F_+F_\times(h_+,h_\times) - (d,d)\\ &\approx 2F_+(d,h_+)_{\rm ROQ} + 2F_\times(d,h_\times)_{\rm ROQ} - |F_+|^2(h_+,h_+)_{\rm ROQ} - |F_\times|^2(h_\times,h_\times)_{\rm ROQ} - 2F_+F_\times(h_+,h_\times)_{\rm ROQ} - (d,d)\\ &= 2\log\mathcal{L}_{\rm ROQ}.\end{aligned} \tag{8}$$

The linear,

$$(d,h_{\rm A}(\vec{\lambda}))_{\rm ROQ} \approx \sum_{j=1}^{N_{\rm L}} \omega_j(t_c)\tilde{h}_{\rm A}(F_j;\vec{\lambda}), \tag{9a}$$

$$\omega_j(t_c) = 4\Re\Delta f\sum_{i=1}^{L}\frac{\tilde{d}^*(f_i)B_j(f_i)}{S_n(f_i)}{\rm e}^{-2\pi{\rm i}t_c f_i}, \tag{9b}$$

and quadratic,

$$(h_{\rm A}(\vec{\lambda}),h_{\rm B}(\vec{\lambda}))_{\rm ROQ} \approx \sum_{k=1}^{N_{\rm Q}} \psi_k\tilde{h}_{\rm A}(\mathcal{F}_k;\vec{\lambda})\tilde{h}^*_{\rm B}(\mathcal{F}_k;\vec{\lambda}), \tag{10a}$$

$$\psi_k = 4\Re\Delta f\sum_{i=1}^{L}\frac{C_k(f_i)}{S_n(f_i)}, \tag{10b}$$

ROQ rules are straightforward to derive: simply substitute the relevant approximations (7) into each of the five overlaps (3) appearing after the second equality in (8). Notice that the *data-dependent* weights ω_j are composed of full overlaps (3) between all the basis elements and the whitened data. While the weights ψ_k in the quadratic ROQ rule do not depend on the data stream $\tilde{d}(f)$, they do depend on the power spectral density $S_n(f)$ which, for the most realistic scenarios, is experimentally estimated. The next section describes our approach for the dependence of (9) on t_c. Generation of both flavors of weights comprises the ROQ *start-up cost*. Once the weights are known, computing the ROQ likelihood only requires $N_{\rm L}+N_{\rm Q}$ terms (hence, only the $N_{\rm L}+N_{\rm Q}$ waveform model evaluations), thereby reducing the cost of (3) by a factor of $L/(N_{\rm L}+N_{\rm Q})$.

Using the definition of the weights (9b) and (10b) and the reality of the basis set $\{C_k\}_{k=1}^{N_{\rm Q}}$, expression (8) can be written in a convenient form for numerical implementation as

$$2\log\mathcal{L}(d|\vec{\Lambda})_{\rm ROQ} + (d,d) = 2\Re\sum_{j=1}^{N_{\rm L}}\omega_j(t_c)\tilde{h}(F_j;\vec{\Lambda}) - \sum_{k=1}^{N_{\rm Q}}\psi_j\tilde{h}(\mathcal{F}_k;\vec{\Lambda})\tilde{h}^*(\mathcal{F}_k;\vec{\Lambda}). \tag{11}$$

Compared to the usual likelihood expression (2) using the typical overlap (3),

$$2\log\mathcal{L}(d|\vec{\Lambda}) + (d,d) = 2\Re\sum_{l=1}^{L}\frac{4\Delta f\tilde{d}^*(f_l)}{S_n(f_l)}\tilde{h}(f_l;\vec{\Lambda}) - \sum_{l=1}^{L}\frac{4\Delta f}{S_n(f_l)}\tilde{h}(f_l;\vec{\Lambda})\tilde{h}^*(f_l;\vec{\Lambda}), \tag{12}$$

shows the ROQ rule to be similar to the standard evaluation pattern, thereby allowing existing codes to easily implement these tools. The simplified expression (11) necessarily requires our basis to permit approximations of the form (7). In particular, had we instead built a separate basis for each polarization and product piece, we would have been forced to retain all five terms originally present in Eq. (8).

B. Trivial and nontrivial parameters

Certain parameters need not be included in the training of the ROM representation (7). In practice, this means we can explicitly set these "neglected" parameters to a fixed constant. In most cases, this is the correct thing to do. The distance to the source, for example, affects the strain as multiplication by an overall constant. Consequently, if $\tilde{h}(f;r=1,\ldots)$ can be accurately integrated with a ROQ rule, then so can $\tilde{h}(f;r\neq 1,\ldots)$. We simply evaluate Eq. (11) at the desired value of r. Sky position, orientation and orbital phase at coalescence affect the strain in a similar, frequency-independent manner.[3]

A notable exception is the signal's arrival time. Our approach for the dependence of (9) on t_c follows Ref. [19]: a unique set of ROQ weights is constructed for n_c equally spaced values of t_c sampling the interval $[t_{\texttt{trigger}}-W, t_{\texttt{trigger}}+W]$, where an estimate for the time window W centered around the coalescence time $t_{\texttt{trigger}}$ is given by the GW search pipeline. Instead of using nearest-neighbor interpolation, as was done in Ref. [19], we use spline interpolation to evaluate the weights at arbitrary values of t_c. Since the weights $\omega_j(t_c)$ are smooth functions of t_c they are well suited for higher-order interpolation. This means, as compared to nearest-neighbor interpolation, significantly higher accuracies and/or use significantly smaller values of n_c are achieved with a spline.

When data are recorded at multiple detectors, inference is carried out using a model whose parameterization is again given by Eq. (6). The ROQ works the same as before, so long as one takes into account the possible time-of-arrival offsets when computing $\omega_j(t_c)$. To handle this, we pad the time window estimates W by ± 26 ms, which is the duration required for a classical gravitational wave to travel from the Earth's geocenter to any conceivable earth-based GW detector. This allows the t_c-dependent ROQ weights to be applicable for all network detectors.

C. Numerical algorithms

The reduced-order quadrature rule is trained on a dense training set of waveforms using the algorithms of Refs. [12,18,19] which have been implemented in C++ and parallelized with message passing interface [21,22]. First, on this training set, we apply a greedy algorithm (see algorithm 1 of Ref. [19]) to construct a nearly optimal reduced basis for the waveform family [17,25]. The algorithm proceeds from a linear basis constructed from i waveforms already chosen. For each training set waveform, we compute the best possible approximation given as a linear combination of the basis elements. The approximate waveform with the largest error is added to the basis as its $i+1$ element. Next, given N basis elements we find the N uniquely determined empirical interpolation nodes with another, different greedy strategy [26,27]. Our implementation of the empirical interpolation method uses the modification suggested by Ref. [18] which reduces the overall cost from $\mathcal{O}(N^4)$ to $\mathcal{O}(N^3)$ (see algorithm 2 of Ref. [19]).

Out-of-training-set validation is carried out by computing the approximation error of randomly sampled waveforms. Typically we use $\approx 10^7$ random samples, which trivially parallelizes with OpenMP within each compute node. We record errors larger than 10^{-6}, adding these waveforms back into the original training set. On this enriched set, we reapply the greedy basis-building algorithm, thereby producing a more accurate basis. The ROQ building procedure, as just described, is largely automated [21].

D. Phenomenological model for precessing inspiral, merger and ringdown waveforms

Waveform models are available for a variety of binary configurations. The most general models include configurations in which the individual spin angular momenta S_i of the compact objects are allowed to be misaligned with the orbital angular momentum $\hat{L}$ of the binary. This spin misalignment is the source of more complex binary

[3]Conversely, the inclination angle which is normally considered "extrinsic" in nonprecessing models is "promoted" to an intrinsic parameter in precessing models because it is frequency dependent. As such, the extension of inclination to precessing systems is included in the parameter vector $\vec{\lambda}$ in Eq. (7).

dynamics which causes the orbital plane, as well as the individual spins, to precess [28–30]. Depending on the relative orientation between the source and the observer, mild to strong amplitude and phase modulations are observed in the GW signal (see, e.g., Ref. [31] for an illustration). Only recently have precessing waveform models describing an approximate inspiral-merger-ringdown (IMR) signal become available [7,32].

The waveform model used in this paper is a phenomenological waveform model known as IMRPhenomPv2 as implemented in the LIGO Algorithm Library (LAL) [8]. This model describes an approximate IMR signal of precessing binary black holes by appropriately rotating the waveforms of an aligned-spin system by means of Euler rotations into the modes exhibited by a precessing system [7]. Schematically, this "twisting up" procedure may be expressed as [31]

$$h^{\rm prec}_{\ell m} = \sum_m \mathbf{R}_{\ell m} h^{\rm aligned}_{\ell m}, \tag{13}$$

where $\mathbf{R}_{\ell m}$ denotes the operator which encoded the relevant Euler rotations. This requires three main ingredients: an accurate aligned-spin model, a description of the orbital precession dynamics and a prediction for the spin and mass of the resulting black hole remnant.

The underlying aligned-spin IMR waveform model is IMRPhenomD [33,34], an aligned-spin waveform model which provides only the $(2, |2|)$ modes of the GW signal. Its inspiral portion has been extensively calibrated to effective-one-body waveforms [35], and the merger part to numerical relativity (NR) waveforms for binary configurations with dimensionless spin magnitudes between -0.95 and 0.98 and mass ratios between 1 and 18.

To model the precession of the orbital plane, analytic post-Newtonian (PN) expressions through second post-Newtonian (PN) order in spin-orbit terms[4] are used [36]. The "twisting-up" procedure Eq. (13) results in a precessing waveform model which contains all $\ell = 2$ waveform modes. However, the absence of the $m = 0$ and $m = \pm 1$ modes in IMRPhenomD leads to approximate precessing modes. The spin and mass of the final black hole are obtained from fits to NR data [36]. We note that IMRPhenomPv2 has not been directly calibrated against precessing NR waveforms and does not include any tidal effects.

To compute the gravitational-wave polarizations h_+ and $h_\times$, it is convenient to adopt a time-independent Cartesian source frame attached to the binary. For aligned-spin binaries, a common choice is a coordinate frame such that $\hat{L} \equiv \hat{z}$. In the case of precession, however, $\hat{L}$ evolves with time, but the direction of the total angular momentum,

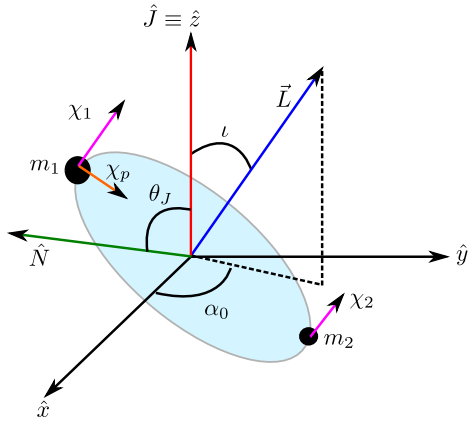

FIG. 1. The J-aligned source frame of a precessing binary. IMRPhenomPv2 uses a single precessing spin approximation to describe the inspiral, merger and ringdown and is described by the parameter vector $\vec{\lambda} = (m_1, m_2, \chi_1, \chi_2, \chi_p, \theta_J, \alpha_0)$, with $\hat{N}$ in the x-z plane. Here, χ_1 and χ_2 are the spin components that lie parallel to $\vec{L}$ on the heavier (χ_1) and lighter (χ_2) compact object; and the perpendicular spin parameter χ_p is the spin component that lies in the orbital plane and is associated with the heavier body m_1.

$\vec{J} = \vec{L} + \vec{S}_1 + \vec{S}_2$, stays approximately fixed during the binary's orbital evolution. A natural choice for the binary source frame therefore is a Cartesian coordinate system, where $\hat{J}$ at some reference gravitational-wave frequency $f_{\rm ref}$ defines the z-axis. This source frame is depicted in Fig. 1.

In the following, we denote the angle between the line-of-sight $\hat{N}$ and $\hat{J}$ by θ_J. The relative orientation between J and the GW detector significantly affects the morphology of a precessing signal. The parameter θ_J represents the natural generalization of the inclination of the orbital plane and is therefore an important parameter to be taken into account when building the ROM/ROQ.[5] Another relevant parameter is the azimuthal orientation of the orbital angular momentum L at the reference gravitational-wave frequency $f_{\rm ref}$ denoted by α_0. The evolution of $\alpha(t)$ encodes the precession of L around J and is thus often referred to as the "precession angle" [28]. Together, the Euler angles $\alpha(t)$ and $\iota(t)$ (defined in Fig. 1) "twist-up" the nonprecessing carrier model IMRPhenomD. The other model parameters are the component masses, m_1 and m_2 with $m_1 \geq m_2$, the dimensionless spin magnitudes projected onto the orbital angular momentum $\hat{L}$, χ_1 and χ_2, and one "effective" precessing spin parameter χ_p [37] defined as

$$\chi_p = \frac{\max\left(A_1 m_1^2 \chi_{1\perp}, A_2 m_2^2 \chi_{2\perp}\right)}{A_1 m_1^2}, \tag{14}$$

[4]The orbital angular momentum, however, uses a 2PN expression without any contribution from the spin terms.

[5]Alternatively, one could build an ROQ for the individual modes.

where $A_1 = 2 + 3m_2/2m_1$, $A_2 = 2 + 3m_1/2m_2$ and $\chi_{i\perp}$ are the magnitudes of the spin vectors perpendicular to $\hat{L}$, i.e., the spin projections into the orbital plane. The motivation for this choice of effective parameterization is the following: In general, a precessing binary can have up to four spin components orthogonal to L, which are *all* the source of precession. However, these can be combined efficiently into a single precessing spin parameter, χ_p, which when applied to the heavier body (m_1), captures the average precession exhibited by the system with all four in-plane spin components [37].

The relevant IMRPhenomPv2 parameters are given by $\vec{\lambda} = (m_1, m_2, \chi_1, \chi_2, \chi_p, \theta_J, \alpha_0)$. Other parameters that enter as an overall scaling, such as distance to the source or its position in the sky, are omitted in the waveform model itself as these can be included trivially. The model parameters in the $\hat{J}$-aligned source binary frame are shown in Fig. 1 which is adapted with permission from Ref. [37].

Various simplifying assumptions have been made in the current implementation[6] of the IMRPhenomPv2 waveform model. One is that $\hat{J}$ is kept constant, and that the angle between $\hat{L}$ and $\hat{J}$ is small. We, therefore, do not expect that IMRPhenomPv2 accurately models precessing cases where $J \sim 0$. Such cases, observed for highly antialigned spins with moderate mass ratios and only a small value of χ_p, are known as transitional precession as $\hat{J}$ undergoes a "flip" and completely changes its orientation [28]. We also do not expect waveforms of systems with higher mass ratios and large values of χ_p to be modeled accurately by IMRPhenomPv2 as the angle between L and J can be large for such cases. However, for some of these cases the model may still produce acceptable results, but detailed checks across the parameter space have not yet been performed.

III. STRATEGIES FOR BUILDING HIGH DIMENSIONAL GW ROMS

Previous work [10,12–16,23] on constructing reduced bases of waveform models have considered waveforms described by only a few *intrinsic* parameters or short signals. The IMRPhenomPv2 waveform family is described by seven parameters and the waveform morphologies are inherently more complex than in the nonprecessing case. The increase in the size of the parameter space, together with the greater variety of waveform morphologies, means that constructing a faithful training space is more difficult than in previous work.

Another concern has to do with the fact that we would like the ROQ to be useful for a very large range of astrophysically relevant parameters; from binary neutron stars with a total mass of around 2 $M_\odot$ to binary black holes with total masses of several tens of solar masses. The signals associated with these different ends of the mass spectrum have very different in-band durations.

In this section, we describe our strategy for dealing with these issues as they relate to populating a faithful training set. We also provide a short review of approaches used in previous work.

A. Mass and frequency partitions

We would like our ROQ to be valid for BNS, NSBH and BBH systems with as few basis elements as possible. In addition, we want to be able to exploit the lowest sensitive frequency of the detectors. To ensure these conditions are met, we find it useful to partition the mass space into (overlapping) regions in chirp mass. These overlapping regions are defined by

$$\mathcal{M}(T = 2^{n+1}s) \leq \mathcal{M} \leq 1.2\mathcal{M}(T = 2^n s), \quad (15)$$

where T is the waveform duration [8] from 20 Hz, $\mathcal{M} = (m_1 m_2)^{3/5}/(m_1 + m_2)^{1/5}$ is the chirp mass, which specifies the waveform duration to leading order, and $q = m_1/m_2 \geq 1$ is the mass ratio chosen between 1 and 9. To interpret $\mathcal{M}$ as a function of time (and vice versa), we build an interpolant of $\mathcal{M}(T)$ using the LAL function `SimIMRSEOBNRv2ChirpTimeSingleSpin`. We compute the signal duration for a given chirp mass, fixing the spins to be maximally prograde and the mass ratio to be 9, which produces the longest inspirals [38]. We consider the following powers of 2: $n = 2, 3, \ldots, 6$, corresponding to regions in $\mathcal{M}$-space describing signals with durations; 1.5 s $\leq T \leq$ 4 s; 3 s $\leq T \leq$ 8 s; 6 s $\leq T \leq$ 16 s; 12 s $\leq T \leq$ 32 s; 23.8 s $\leq T \leq$ 64 s; 47.5 s $\leq T \leq$ 128 s. The union of the overlapping regions in chirp mass capture binary systems with signal durations between slightly less than 2 s up to 128 s starting from 20 Hz. The upper frequencies for the cases in Table I correspond to the maximum-over-configuration ringdown frequency, rounded to the next-highest power of two.

Our particular choices have been guided by the expectation that, typically, a stochastic sampler will stay confined to a given partition or two. Future improvements to the ROQ method presented here, and more generally ROM building, may find different partition strategies to work better. Notice that the finer we make our mass partition, the fewer basis elements will be needed in each partition and, hence, the greater the ROQ compression. Finer mass partitions also reduce the *offline* cost associated with building the basis in a given partition. On the other hand, if we add up all the basis elements from all the partitions, we should expect to find this total to be larger than a corresponding basis resulting from one large partition of equivalent extent. Both small [39] and large [17] partitions have been considered in other contexts.

[6]The results of this paper use the phenomPv2 model as implemented in the LAL with a git hash of a50aca13b97412999-fad03a073a6a5b319fd5bc4.

TABLE I. Regions in parameter- and mass-frequency space in which we build a distinct ROM. Each case corresponds to an overlapping region in chirp-mass ($\mathcal{M}$) space. In all cases, the bases/interpolants are valid in the mass-ratio interval $1 \leq q \leq 9$ which is within IMRPhenomPv2's calibration-range [34]. This range in mass ratio allows us to describe BNS systems, NSBH systems and BBH systems. Additionally, we impose the constraint on the component masses $m_1 \geq m_2 \geq 1\ M_\odot$. For each case, we limit the magnitudes of the spin-related parameters (χ_1, χ_2, χ_p) to lie within the range $(-0.9, -0.9, 0) \leq (\chi_1, \chi_2, \chi_p) \leq (0.9, 0.9, 0.9)$ and we use the full range for the spin angles: $(0, 0) \leq (\theta_J, \alpha_0) \leq (\pi, 2\pi)$. Cases A′–F′ show how the mass and frequency ranges of cases A–F can be scaled (See Sec. III F) to 10 Hz without any additional computational effort.

		f (Hz)		Waveform		$\mathcal{M}$ ($M_\odot$)		Basis size		
Case	Build strategy	Min	Max	duration T	Δf (Hz)	Min	Max	Linear	Quadratic	Speed-up
A	Enriched greedy	20	1024	1.5 s $\leq T \leq$ 4 s	1/4	12.3	23	300	197	8
B	Enriched greedy	20	1024	3 s $\leq T \leq$ 8 s	1/8	7.9	14.8	388	278	12
C	Enriched greedy	20	2048	6 s $\leq T \leq$ 16 s	1/16	5.2	9.5	360	233	54
D	Enriched greedy	20	2048	12 s $\leq T \leq$ 32 s	1/32	3.4	6.2	524	254	83
E	Enriched greedy	20	2048	23.8 s $\leq T \leq$ 64 s	1/64	2.2	4.2	749	270	127
F	Enriched greedy	20	4096	47.5 s $\leq T \leq$ 128 s	1/128	1.4	2.6	1253	487	300
A′	A scaled	10	512	3 s $\leq T \leq$ 8 s	1/8	24.6	46	300	197	8
B′	B scaled	10	512	6 s $\leq T \leq$ 16 s	1/16	15.8	29.6	288	278	12
C′	C scaled	10	1024	12 s $\leq T \leq$ 32 s	1/32	10.4	19	360	233	54
D′	D scaled	10	1024	23.8 s $\leq T \leq$ 64 s	1/64	6.8	12.4	524	254	83
E′	E scaled	10	1024	47.5 s $\leq T \leq$ 128 s	1/128	4.4	8.4	749	270	127
F′	F scaled	10	2048	95 s $\leq T \leq$ 256 s	1/256	2.8	5.2	1253	487	300

B. Ranges in mass ratio and spin

Unlike our treatment of the chirp mass, we do not use any special partitioning strategy for the six remaining parameters. Table I and its caption summarizes the default parameter intervals used for the mass ratio (q) and the spin-related parameters $(\chi_1, \chi_2, \chi_p, \theta_J, \alpha_0)$.

Since we are working with internal IMRPhenomPv2 parameters, we have to impose constraints on some of the model's spin-related parameters [36]. These constraints eliminate unphysical systems with spins above the Kerr limit. The original physical BH binary can have in-plane spin components $\chi_{i,p}$ on either BH $i = 1, 2$. The spins must satisfy the Kerr limit on each BH: $\chi_{i,p}^2 + \chi_i^2 \leq 1$. Since the model's effective precessing spin satisfies $\chi_p \leq \max[\chi_{1,p}, W(q)\chi_{2,p}]$ with $W(q) = \frac{3q+4}{4q^2+3q}$, in practice simply excluding $\chi_p^2 + \chi_1^2 \geq 1$ is good enough.

We have had to place one additional restriction on the spins. Specifically, we exclude the region where and $\chi_1 \leq 0.4$–7η. This constraint arises because the model exhibits nonsmooth, rapidly changing behavior with parametric variation thereby precluding the existence of an accurate, sparse basis. We describe this problematic region in the Appendix.

C. Deterministic and random sampling of the parameter space

Previous work [10,13–16,23] has shown that a good strategy for sampling in the mass space is to sample uniformly in $\mathcal{M}^{3/5}$ as this is the leading-order mass term that enters into the waveform phasing. It has also been observed that the basis elements are preferentially selected from the boundary of the parameter space, suggesting an efficient training set would overpopulate these regions. Additionally, the authors of Ref. [13] considered a random greedy sampling strategy for precessing waveforms, parametrized by phase in a coprecessing frame. In this framework, a new training set is randomly generated at each iteration of the greedy algorithm thereby allowing for an effectively greater number of training waveforms. This strategy was motivated by the cost of storing the training set in memory which we overcome by using a parallelized code.

For cases A–C in Table I, we find that using just eight sample points on a uniform grid in $\mathcal{M}^{3/5}$ and η and a uniform grid of eight points in each of the remaining five spin-parameters yields a reasonably accurate basis. For the more challenging cases D–F in Table I, we increase our set to 64 sample points on a uniform grid in $\mathcal{M}^{3/5}$ and η while using the same sample strategy for the remaining five parameters.

In the validation step, we evaluate the model at randomly chosen parameter values. Parameter values at which the approximation error is greater than 10^{-6} are flagged. We combine these high-error points to the ones previously selected by the greedy algorithm; their union constitutes a new training set. Running the greedy algorithm on this new set produces an enriched basis with an improved error as judged by yet another series of validations. The validation → enrichment → validation → … iterations continue until the worst error is below 10^{-6}. This is somewhat similar in spirit to the sampling strategy of [13] described above.

Due to the fact that the validation step is embarrassingly parallel over the random samples (as opposed to the greedy algorithm, which requires a modest amount of communication), we can easily handle a large number of random

points. We typically consider roughly 10 to 15 million points per validation study.

D. Frequency resolution of the training set

To capture the main waveform features the training set must be faithfully sampled in both parametric and physical dimensions. This must be balanced against the size of the training spaces in physical memory. For example, storing a training set with $(64^2) \times (8^5)$ waveforms with a bandwidth of ~4096 Hz and a frequency resolution of $\Delta f = 1/64$ Hz would require around 500 terabytes of memory. Our training set waveforms use an adaptive frequency sampling strategy, $\Delta f(f)$, which significantly reduces the greedy algorithm's memory footprint to around 64 GB. We only apply this adaptive sampling to cases D–F in Table I as the other cases' training sets fit comfortably into memory.

Our choice of frequency resolution $\Delta f(f)$ comes from determining the longest signal duration for a given mass and frequency band. These are found empirically, first by finding the duration of the lightest binary system in a set of frequency bands for each of the cases in Table I, and then rounding this up to the next-highest power of two. The frequency resolution is taken to be the inverse of this duration. By selecting $\Delta f(f)$ in this way, we ensure that the waveforms are sampled above the Nyquist rate in each band. This can be applied across multiple bands (20–64 Hz, 64–128 Hz, etc.). This is a similar strategy to "multi-banding" which has been useful in other contexts e.g., the gravitational-wave search pipeline of [40]; see Table 3 of Ref. [40].

We stress that the adaptive frequency sampling described above is used for training set waveforms only. Once the greedy points are known, to collocate with the data on a set of equally spaced frequencies corresponding to the global Nyquist rate, we up-sample by direct evaluation of the waveform model. This does not cause a memory bottleneck, however, because the basis is significantly smaller than the training set. Refs. [12,18,23] used a similar strategy whereby the frequency interval was split with a domain decomposition following the local Nyquist frequency and employing Gaussian quadratures in each subdomain. Additional validation is needed to check that up-sampling does not introduce an unacceptably large error, which we demonstrate in Sec. IV C. Refs. [12,18] provide further discussions of subtleties related to up-sampled basis.

E. The basis building pipeline

The ROQ rule derived in Sec. II A requires a basis set for both the plus- and cross-polarizations, $\{B_j\}_{j=1}^{N_L}$ in Eq. (7a), and another, different basis set for the three product combinations of these polarizations, $\{C_k\}_{k=1}^{N_Q}$ in Eq. (7b). We build these linear and quadratic parts of the ROQ hierarchically in steps.

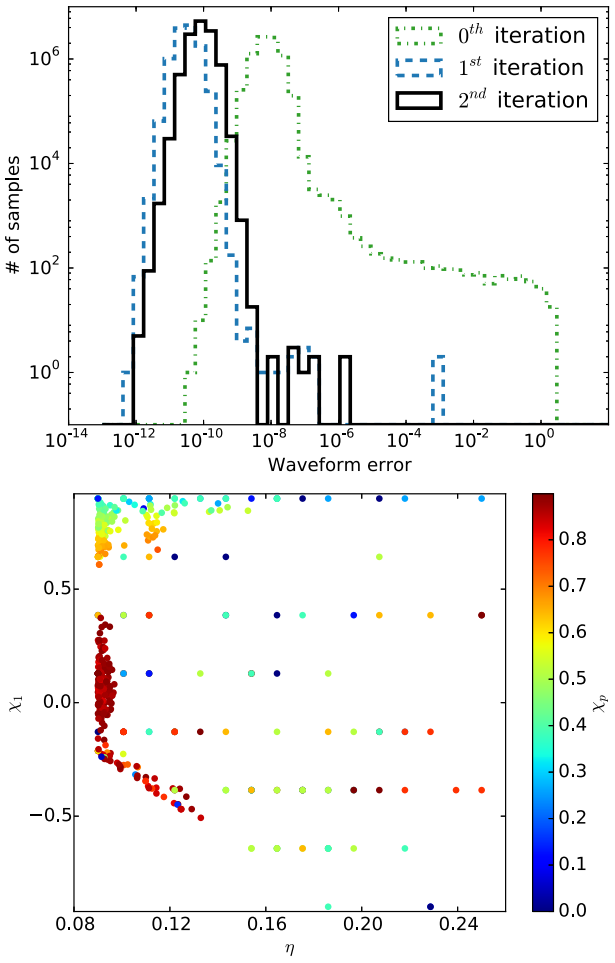

FIG. 2. An example of the basis enrichment strategy applied to case D from Table I (the plus- and cross-polarizations only). *Top*: A sequence of histograms showing the distribution of the reduced-basis approximation error for ≈10 million out-of-sample model evaluations. Notice a continual lowering of the *maximum* approximation error with each iteration. *Bottom*: The final distribution of greedy points in a three-dimensional subspace. This set is a mixture of the initial structured grid used in the zeroth iteration and random points identified through the enrichment process.

We start by building a basis for the linear part of the ROQ. Empirically, we have found that an accurate basis trained exclusively for the plus-polarization continues to approximate the cross-polarization with good accuracy, and vice versa. Consequently, we populate a training set for the $\tilde{h}_+$ mode of the strain only. A greedy algorithm identifies a "zeroth iteration" basis $\{B_j^0\}_{j=1}^{N_L^0}$. We then perform a validation of this basis against both polarizations. Waveform errors greater than $\epsilon = 10^{-6}$ are used in a basis enrichment step described above. We iterate until an ϵ-accurate basis is achieved (often one or two iterations suffice). Figure 2 displays a sequence of error histograms (top panel) and the final distribution of greedy points (bottom panel) in a three-dimensional subspace.

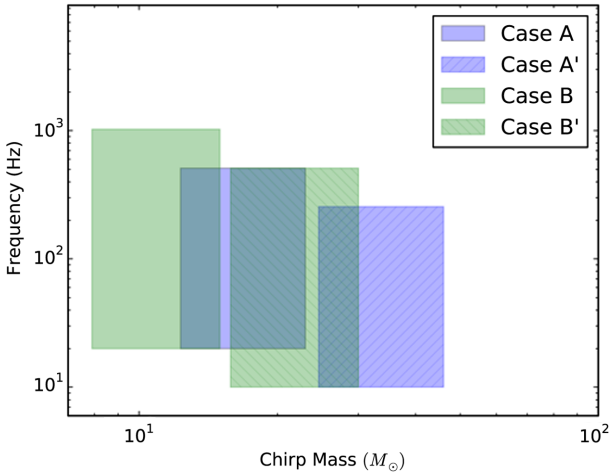

FIG. 3. Illustration of how to translate basis results to new regions of mass and frequency. By using the scaling described in the text, the basis for case A (blue unhatched region) maps on to case A′ (blue hatched region). Similarly, the unhatched and hatched green regions, respectively, correspond to cases B and B′ in Table I. These primed regions form a starting point for building a 10 Hz basis without extra computational effort.

Next, a basis for the quadratic part of the ROQ is built using previously computed information. To motivate our approach, notice that a tensor product of the linear basis is sufficient to describe $\tilde{h}_+^*\tilde{h}_+$, $\tilde{h}_\times^*\tilde{h}_\times$ and $\Re(\tilde{h}_+^*\tilde{h}_\times)$. Consequently, we take the greedy points which define the linear basis and form an ansatz training set consisting of $(\tilde{h}_+^* + \tilde{h}_\times^*)(\tilde{h}_+ + \tilde{h}_\times)$. The quadratic basis $\{C_i(f)\}_{i=1}^{N_Q}$ is built following the same iterative enrichment procedure used for the linear basis. A more direct (but more costly) two-step approach to treating these product terms is given in [18].

F. Translating basis results to new regions of mass and frequency

By exploiting a mass-frequency mapping allowed by the Einstein equation in vacuum we can extend the basis' region of validity over an enlarged mass and frequency range which would otherwise require extra computational effort to build. Recall that a waveform described by a particular chirp mass $\mathcal{M}$ and low and high frequencies $f_{\rm low}$ and $f_{\rm high}$ can be transformed into a waveform described by a chirp mass of $\mathcal{M}' = n\mathcal{M}$ and low and high frequencies of $f'_{\rm low} = f_{\rm low}/n$ and $f'_{\rm high} = f_{\rm high}/n$. As an example, consider the basis for case A (Table I)—which covers masses $12.3 \leq \mathcal{M}/M_\odot \leq 23$ and frequencies $20 \leq f/\mathrm{Hz} \leq 512$ at a resolution of $\Delta f = 1/4$ Hz—which can be mapped onto case A′ with masses $24.6 \leq \mathcal{M}/M_\odot \leq 46$, frequencies $10 \leq f/\mathrm{Hz} \leq 256$ with a frequency resolution of $\Delta f = 1/8$ Hz by setting $n = 2$. This procedure can be repeated to access higher masses and lower frequencies, or lower masses and higher frequencies. Table I summarizes one possible extension of a ROM/ROQ from $f_{\rm low} = 20$ Hz to $f_{\rm low} = 10$ Hz. Figure 3 depicts the appearance of gaps in the translated ROM, the filling of which would require additional numerical work, although significantly less than had the 10 Hz-basis been built from scratch. This technique, which necessarily requires our basis have been built *without* reference to any particular noise curve, has also been used in other ROMs [10,14,15,20].

IV. BUILDING AND VALIDATING THE EMPIRICAL INTERPOLANT

In this section, we numerically compute the requisite empirical interpolation representation (7) which, once the detector's data d are known (cf. Sec. V), will enable accelerated likelihood evaluations from Eq. (11). Since the ROQ's error, $|\log\mathcal{L}(d|\vec{\Lambda}) - \log\mathcal{L}(d|\vec{\Lambda})_{\rm ROQ}|$, is controlled by the empirical interpolant's error [18], we are especially interested in quantifying the latter error for any possible gravitational wave model evaluation. Given an approximation $\hat{a} \approx a$, which could stand for either the linear or quadratic parts, we report the error as the square of the unweighted ($S_n = 1$) norm of $\hat{a} - a$, which is related to the unweighted overlap, $(\hat{a}, a)$, by

$$\|\hat{a} - a\|^2 = (\hat{a} - a, \hat{a} - a),$$

where $\hat{a}$ and a are normalized. It is this "white-noise" error which directly controls the ROQ's log-likelihood approximation error. The next section describes parameter estimation studies for which, clearly, $S_n \neq 1$.

A. Linear parts

Our first task is to build the basis, $\{B_j\}_{j=1}^{N_L}$, and ROQ nodes, $\{F_j\}_{j=1}^{N_L}$. These pieces are required to form the part of the ROQ rule (9) which is linear in $\tilde{h}$.

To find the basis, we apply the greedy algorithm to training sets defined on each case A–F from Table I. As discussed in the previous section, our training set is iteratively enriched with random sampling. Figure 4 reports the greedy algorithm's error profile when applied to the final (and hence most dense) training set iteration. Figure 4 shows a similar behavior in all cases, namely, an initially slow fall-off in the representation error followed by an exponential decrease. This by-now common feature has been seen across different waveform models using different dimensional reduction algorithms [10,12–16,20,23,39]. Figure 4 also shows that the number of reduced-basis waveforms needed to approximate intervals describing successively smaller chirp mass values increases. A notable exception, however, is case B which for some error thresholds is actually larger than case C. One possible explanation is that the iteratively enriched basis is suboptimal as compared to a hypothetical basis built from

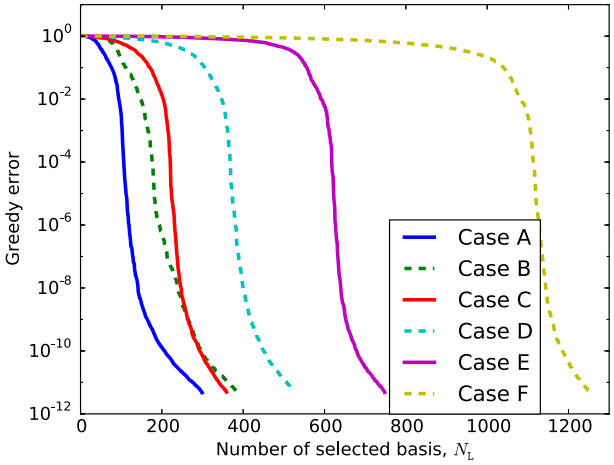

FIG. 4. Greedy error, defined as the maximum approximation error over the training set using the first N_L basis, versus basis number N_L. The error profile is fairly similar across all cases. Since the ROQ speed-up is (almost) proportional to N_L, further speed-up can be achieved at the expense of accuracy. Throughout the paper we select the first N_L basis satisfying a conservative 5×10^{-12} greedy error tolerance.

an arbitrarily dense training set. Nevertheless, even these sub-optimal basis provide excellent performance gains while being less demanding to compute. Table I summarizes the resulting linear bases corresponding to a greedy error of 5×10^{-12}.

To find the ROQ nodes, we apply the empirical interpolation method for each case defined in Table. I. As input to the algorithm we provide the reduced-basis vectors and the corresponding set of frequency points. Figure 5 depicts the distribution of selected ROQ nodes for the two most extreme cases A (top) and F (bottom). Notice the that EI method preferentially selects points at lower frequencies, which matches our expectation that the information carried by these waves is encoded in the cycles which "pile up" at lower frequencies.

Figure 6 reports the out-of-sample validation study, which uses ≈15 million random waveform evaluations not in the original training set. The errors $\epsilon_\times = (\hat{h}_\times - h_\times, \hat{h}_\times - h_\times)$ and $\epsilon_+ = (\hat{h}_+ - h_+, \hat{h}_+ - h_+)$ are found to be small in all cases. Thanks to the frequency-independence of the antenna patterns, one can directly relate these errors to the error in the linear part of the ROQ rule (9)

$$|(d,h) - (d,h)_{\rm ROQ}| \leq \mathcal{C}_1 |F_+| \epsilon_+ + \mathcal{C}_2 |F_\times| \epsilon_\times,$$

without any extra numerical work. Importantly, this avoids the computation of errors over an enlarged parameter space including ra, dec and ψ. The constants $\mathcal{C}_1$ and $\mathcal{C}_2$ are computable. Finally, Fig. 6 demonstrates that we incur a penalty factor of ≈100 when approximating by an empirical interpolant as opposed to orthogonal projection onto the

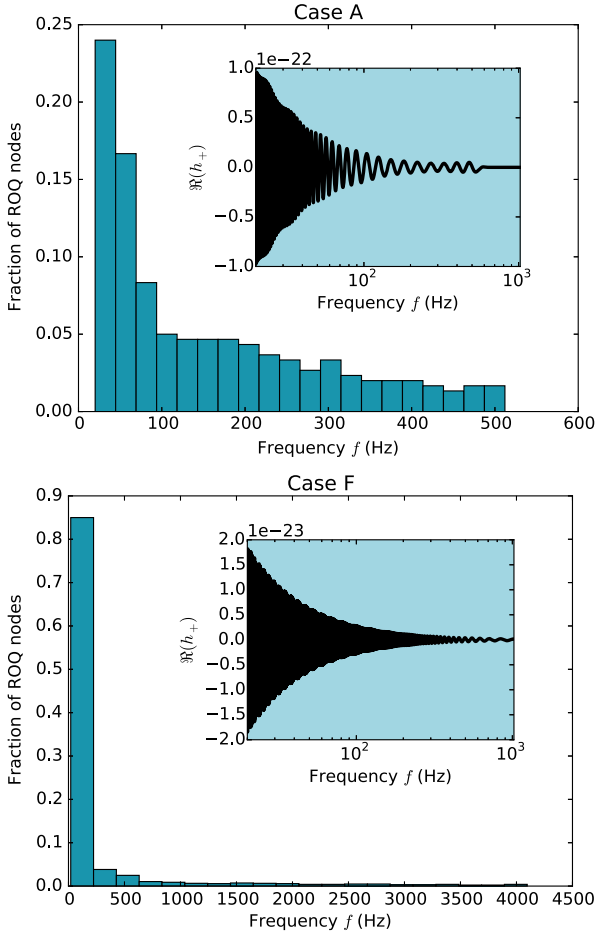

FIG. 5. Histogram of selected ROQ nodes and a representative waveform for case A (top) and case F (bottom). Evidently the selected frequency points cluster at small values. This is intuitively expected because lower frequency intervals contain a greater number of waveform cycles, a feature which is automatically detected by the empirical interpolation method. Histograms of those cases not shown are qualitatively similar, being a mixture of these two boundary cases.

basis which is guaranteed to yield the smallest possible error. We do not know ahead of time what this penalty factor might be; this further motivates our choice of working to small 5×10^{-12} accuracies in the basis building step.

B. Quadratic parts

Our next task is to build the basis, $\{C_j\}_{j=1}^{N_Q}$, and ROQ nodes, $\{\mathcal{F}_j\}_{j=1}^{N_Q}$. These pieces are required to form the part of the ROQ rule (10) which is quadratic in $\tilde{h}$. The steps are essentially the same as in the linear case just described. Table I summarizes the resulting quadratic bases corresponding to a greedy error of 5×10^{-12}.

We now skip directly to the approximation errors, quantified by yet another out-of-sample validation study.

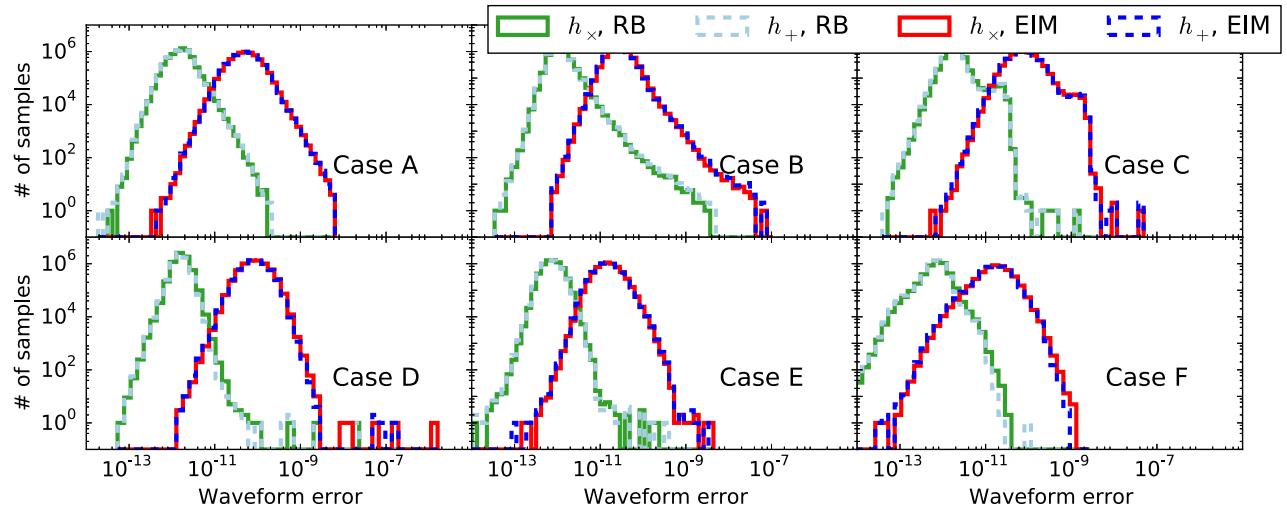

FIG. 6. Projection (RB) and empirical interpolation (EIM) errors (generically x-axis labeled as "Waveform error") for ≈15 million randomly drawn waveforms. Each subfigure reports on the errors for an approximation defined by the six cases listed in Table I. The validations are performed using the same adaptive frequency sampling strategy as was used to find the basis (cf. Sec. III D).

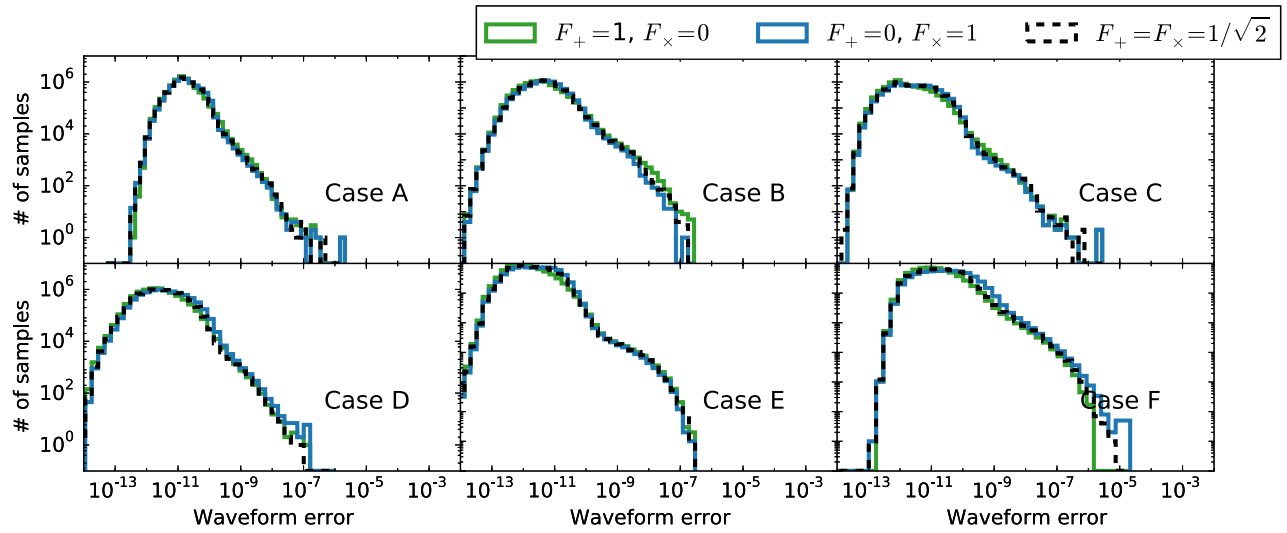

FIG. 7. Empirical interpolation errors (generically x-axis labeled as "Waveform error") for ≈15 million randomly drawn waveforms. Each subfigure reports on the errors for an approximation defined by the six cases listed in Table I. The validations are performed using an adaptive frequency sampling strategy and for three representative antenna pattern configurations.

As for the linear case, we would again like to relate the ROQ error to the errors due to approximation of each quadratic polarization parts. Due to the differences in sizes of each quadratic piece, computing relative errors are uninformative in this case.[7] We instead compute the error from the approximation of $(F_+\tilde{h}_+ + F_\times\tilde{h}_\times)(F_+\tilde{h}_+^* + F_\times\tilde{h}_\times^*)$ by its empirical interpolant for three representative cases. The results are shown in Fig. 7.

[7]Since $\|\Re(\tilde{h}_+^*\tilde{h}_\times)\| \ll \|\tilde{h}_+^*\tilde{h}_+\|$ in the nonprecessing limit, the relative approximation error of $\Re(\tilde{h}_+^*\tilde{h}_\times)$ may be large but insignificant insofar as likelihood accuracy is concerned.

C. Upsampling

As discussed in Sec. III D, in order to reduce the greedy algorithm's memory footprint to manageable sizes we use an adaptive frequency sampling strategy. Yet to compute the ROQ weights (9a) and (10a), the basis must be known at the same frequency values recorded by the detector. To collocate with the data on a set of equally spaced frequencies corresponding to the global Nyquist rate, we up-sample by direct evaluation of the waveform model at the greedy points and reorthogonalize the basis. Figure 8 reports the additional error due to up-sampling. That the errors remain similarly small is evidence that our training set waveforms are well resolved by the adaptive frequency grid.

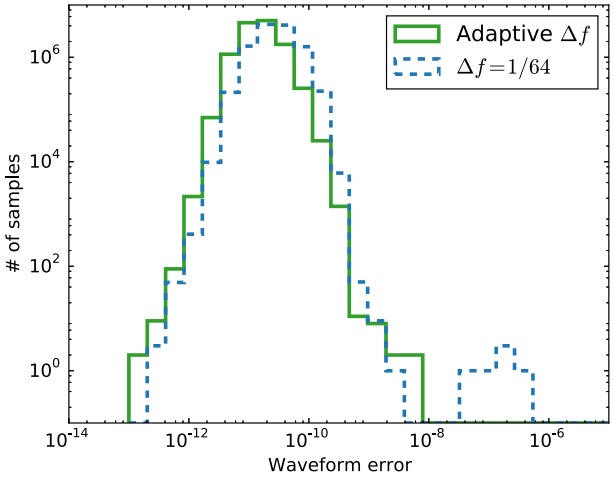

FIG. 8. Empirical interpolant approximation errors (the plus- and cross-polarizations only) when using an adaptive (solid green line) and uniform (dashed blue line) frequency sampling. The adaptive sampling is used during the ROQ building procedure. To compute log-likelihoods with our ROQ rule we upsample to a uniform frequency grid, and so this error (which constitutes the last in a series of approximations of the underlying model) is the most relevant for ROQ-accelerated inference studies. Results are shown for case E only; other cases are qualitatively similar. Maximum upsampled EIM errors of 6×10^{-9}, 1×10^{-7}, 1×10^{-5}, 7×10^{-8}, 4×10^{-7} and 1×10^{-9} were computed for cases A–F, respectively.

V. PARAMETER ESTIMATION

A. Accuracy comparisons

To determine how the empirical interpolation errors (as summarized in Figs. 6, 7, and 8) affect parameter estimation, we present a comparison between the recovered posterior PDFs using both the Full and the ROQ likelihood functions evaluated with `LALInferenceNest` [5], which is one of the stochastic samplers available with the `LALInference` library [8]. A simulated binary black hole signal represented by IMRPhenomPv2 and drawn from the parameter space defined by case A in Table I was injected coherently in the two LIGO detectors. To represent the nonstationarity of the detector noise the injection was made into real data from the sixth LIGO science run [41], recoloured to reflect the expected early aLIGO sensitivity (cf. Ref. [24,42] which used the same data for studying simulated binary neutron star detections).

Under the assumption that the ROQ is an approximation of the Full likelihood, the two methods are required to be statistically indistinguishable in order for the ROQ to qualify as a valid substitute to the Full likelihood function for parameter estimation. As is shown in Fig. 9, the Full and the ROQ methods recover posterior PDFs that are almost visually identical. We quantify the difference between the two sets of posterior PDFs by computing the KL divergence [43],

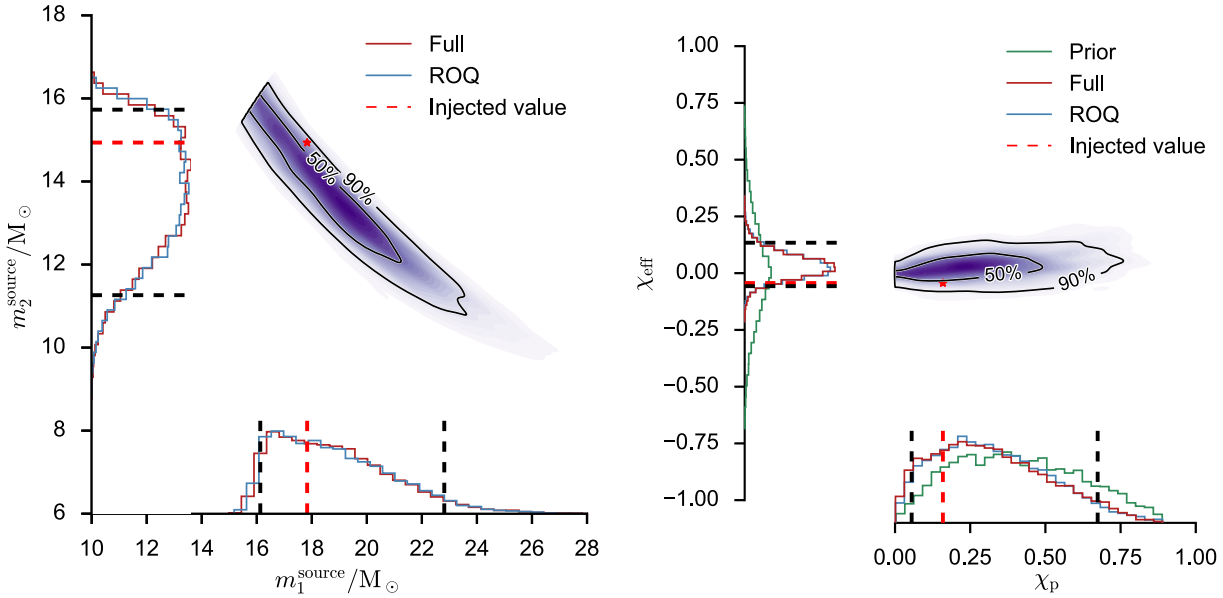

FIG. 9. Comparing the recovered posterior PDFs for the black hole masses measured in their source frame (left) and the two dominant spin parameters (right), c.f. [44]. The parameter $\chi_{\rm eff} = (m_1\chi_1 + m_2\chi_2)/(m_1 + m_2)$ is known as the "effective" spin and is a mass-weighted combination of the two spin components parallel to the orbital angular momentum. The dashed black lines mark the 90% credible interval for both the ROQ and Full likelihoods, which are the same within the statistical sampling uncertainty of ~1%. The posteriors do not peak exactly at the injected values due to the presence of detector noise.

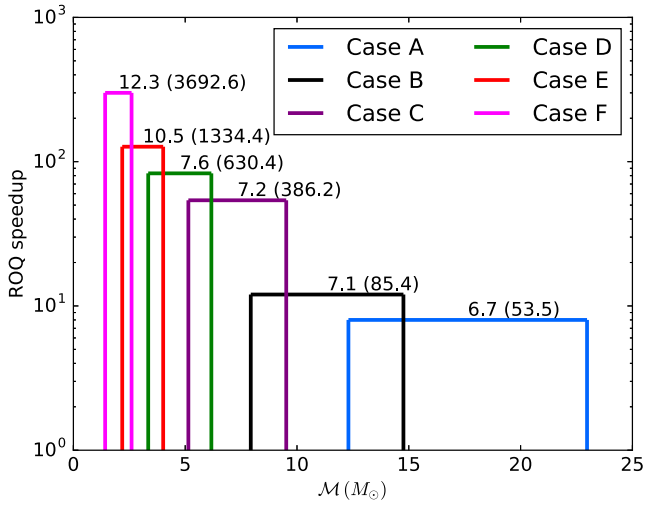

FIG. 10. Theoretical parameter estimation speed-up (using the ROQ) for cases A–F in Table I. The speed-up is calculated from the ratio $L/(N_{\rm L}+N_{\rm Q})$, where $L=(f_{\rm max}-f_{\rm min})/\Delta f$ is the number of quadrature points in the Full likelihood, $N_{\rm L}$ is the size of the linear basis and $N_{\rm Q}$ is the size of the quadratic basis. The sum $N_{\rm L}+N_{\rm Q}$ is the number of points in the ROQ likelihood (11). The plot is annotated with the time (in hours) to compute 2×10^7 ROQ (Full) likelihood evaluations, roughly the number of evaluations required for a typical PE analysis [5]. Our tests were performed using an Intel Xeon CPU with a 2.70 GHz clock speed.

$$D_{\rm KL}(P|Q)=\sum_i P_i\log\left(\frac{P_i}{Q_i}\right), \tag{16}$$

for all of the one-dimensional PDFs produced by the Full and ROQ analyses, including but not limited to the parameters shown in Fig. 9. The KL-divergence quantifies the relative entropy, in units of bits, between the probability distributions P and Q, or equivalently the amount of information lost when using Q as an approximation to P. For $(P,Q)=(\text{Full},\text{ROQ})$ the minimum, median and maximum $D_{\rm KL}$ are (0.0020, 0.0057, 0.0141) bits, respectively. This can be compared to the set of $D_{\rm KL}$ for $(P,Q)=(\text{Full},\text{prior})$ of $(0.016, 0.33, \infty)$ bits, which reflects the information gain contained in the likelihood on its own.

A comprehensive study of the parameter estimation capabilities using ROQs will be presented in [45].

B. Performance benchmarks

Having established the equivalence of the results for the Full and ROQ likelihoods, we now consider the performance gains afforded by the ROQ rule. In Fig. 10, we show the expected likelihood speed-up ratio L/N. Here L is the number of operations in the non-ROQ likelihood and $N=N_{\rm L}+N_{\rm Q}$ is the number of operations in the ROQ likelihood (11). The speed-up is seen to be as large as ≈300 for low mass systems. Assuming the entirety of the PE cost is in the form of waveform or likelihood evaluations, which scale linearly with L (Full) or N (ROQ), the ratio L/N provides the theoretical performance improvement for any hypothetical PE study.

We estimate the run time of parameter estimation studies by (i) computing the waveform at the empirical interpolation nodes for the linear and quadratic pieces of the ROQ and (ii) subsequently computing 2×10^7 evaluations of the ROQ-likelihood (11) for random-valued integration weights, which is a reasonable number of MCMC samples needed to produce a few thousand statistically independent samples using the `LALInference` code [5]. These timing results are also summarized in Fig. 10. We find that by using the ROQ, and assuming that the bulk of the cost of parameter estimation is in computing waveforms and overlap integrals, then the run time of PE codes should be between around six hours (for analyses that restrict themselves to chirp mass bins as in case A of Table I) to around twelve hours (for analyses that restrict themselves to chirp mass bins as in case F of Table I). Our tests were performed using a single core on an Intel Xeon CPU with a 2.70 GHz clock speed. The test used a stand-alone python script calling the `LALSimulation` library through its SWIG interface.

These timing experiments obviously depend strongly on the effort of (hardware-specific) optimization or parallelization schemes, such as offloading work to MIC processors [46–48], which we have not explored. Nevertheless, the quoted speed-up numbers are *independent* of these details.

Finally, we note that there is a once-per-analysis "start-up" cost of computing the set of ROQ weights (9b) and (10b). This cost, which amounts to $\mathcal{O}(10^4)$ overlaps (3) and parallelizes trivially, is negligible compared to a full inference simulation. As a representative example, we computed 10,000 sets of ROQ weights for a typical time window of 0.2s centered on the trigger-time, each associated with a unique value of the coalescence-time t_c within this window. Computing weights for 10,000 values of t_c corresponds to sampling the t_c at a constant rate $\Delta t_c=0.2/10^5=2\times10^{-6}$, which is around a thousand times smaller than the typical measurement uncertainty in t_c [12]. We find that the time to compute the ROQ weights is on the order of a few minutes for all cases in Table I, which is much smaller than both the estimated ROQ and Full inference run times.

VI. CONCLUSION

We have presented a method for building reduced-order models and quadrature rules of precessing, inspiral-merger-ringdown gravitational waveforms designed specifically to improve the efficiency of astrophysical inference. Our method, which is generic, was applied to the waveform family known as IMRPhenomPv2. We find that by using an IMRPhenomPv2-specific reduced-order quadrature rule, parameter estimation studies can be sped up by factors of 4 (for binary black holes) to 300 (for binary neutron

stars) in analyses starting from a low-frequency cutoff of 20 Hz; see Fig. 10. Crucially, this performance-boosting technique does not sacrifice the accuracy of parameter estimates as shown in Fig. 9 and discussed in Sec. V. We stress that nearly-indistinguishable PE results are a consequence of the high-accuracy ROM built in Sec. IV. Below we discuss extensions to the work presented here.

Larger parameter regions: The method presented here is generic and capable of handling large parameter domains. Recently, the nonprecessing IMRPhenomD model [33,34] underlying IMRPhenomPv2 has been calibrated up to mass ratios of $q = 18$ and aligned spins of ∼0.85 (0.98 at equal-mass). We hope to explore the application of our methods to these extremal values of the model, which might require more sophisticated parameter sampling and domain decomposition strategies.

Other waveform families: Some waveform families are described by costly differential equations. These could be effective-one-body models [32,35,49,50], PN models [51,52] or the Einstein equations. While in principle our techniques can be applied to these models to construct the reduced basis and empirical interpolation nodes, it is not clear how to directly evaluate the waveform model *at* the empirical interpolation nodes so that the ROQ can actually be used. As long as the ROM depends linearly on its basis, the surrogate modeling tools of Refs. [10,14,53–55] may be applicable. Common to these techniques is the construction of a closed-form expression capturing the parametric behavior of well-chosen waveform data, such as the amplitude and phase values at specially selected times or frequencies. Consequently, the cost of evaluating a surrogate model will necessarily grow with parametric dimensionality. The efficiencies of these models for precessing systems remains an open question (none have been built to date). Currently, then, closed-form phenomenological waveform families offer the best trade off for achieving rapid and accurate parameter estimation with an ROQ. We believe ROQs to be especially useful for long waveforms dominated by many inspiral cycles, where approximate methods are expected to be accurate and ROQ speed-ups are at their largest.

ACKNOWLEDGMENTS

We thank Harbir Antil, Jonathan Blackman, Thomas Dent, Chad Galley, Mark Hannam, Tom Loredo, Saul Teukolsky, Manuel Tiglio and Alan Weinstein for many useful discussions and encouragement throughout this project and Jonathan Blackman, Mike Boyle, Sascha Husa and Alejandro Bohé for help towards explaining features described in the Appendix. We would also like to thank our LIGO Presentation and Publication reviewer for their clear and detailed feedback on this manuscript. LIGO was constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the National Science Foundation (NSF) and operates under Cooperative Agreement No. PHY-0757058. M. P. was supported by STFC Grant No. ST/I001085/1 and the Max Planck Gesellschaft. S. F. was supported in part by NSF Grants No. PHY-1306125 and No. AST-1333129 to Cornell University and the Sherman Fairchild Foundation. P. S. was supported by the Sherman Fairchild Foundation and NSF Grants No. PHY-1404569 and No. PHY-1151197 at Caltech. Some of the computations were carried out using the high performance computing resources provided by Louisiana State University (http://www.hpc.lsu.edu), the Extreme Science and Engineering Discovery Environment (XSEDE) [56], and the Zwicky cluster at Caltech, which is supported by the Sherman Fairchild Foundation and by NSF Award No. PHY-0960291. We are grateful for computational resources provided by Cardiff University and funded by an STFC grant supporting UK Involvement in the Operation of Advanced LIGO. This paper carries LIGO Document No. P1600096.

APPENDIX: GREEDY FEATURE DETECTOR

Here we describe a novel use for the greedy algorithm which we believe might help waveform developers identify abrupt changes in behavior or discontinuities in waveform models.

One of the key criteria for the reduced-basis method to deliver a basis that exhibits exponentially fast error convergence is that the model space varies smoothly with

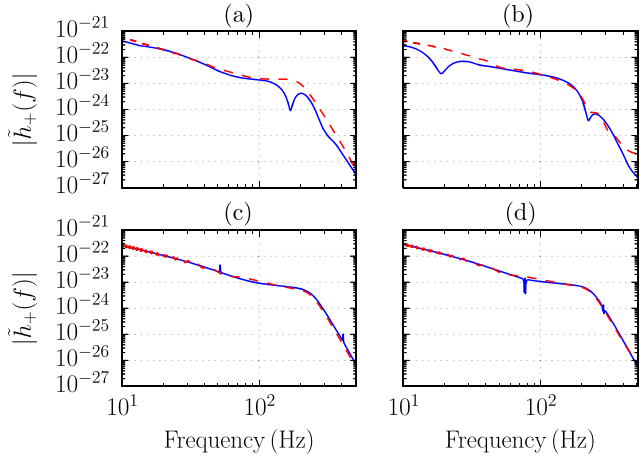

FIG. 11. Amplitudes of $\tilde{h}_+(f)$ for selected points in the $\chi_p \approx 0$ cluster shown in Fig. 12. The parameter values for configurations (a)–(d) are given in Table II. Abrupt, sharp features are clearly visible in the IMRPhenomPv2 amplitudes (blue solid lines), but are absent in the SEOBNRv3 amplitudes (red dashed lines). These features are difficult to capture with the reduced-basis method without sacrificing the sparsity and/or accuracy of the basis.

TABLE II. IMRPhenomPv2 parameters for the configurations shown in Fig. 11.

Case	$M_{\rm tot}$ $[M_\odot]$	η	χ_1	χ_2	$\chi_{\rm p}$	$\theta_{\rm J}$	α_0
(a)	65.054	0.15	−0.773	0.054	−0.161	−0.44	−0.039
(b)	62.748	0.144	−0.772	−0.153	−0.134	1.084	2.773
(c)	53.375	0.148	−0.78	0.113	−0.0	1.594	2.338
(d)	55.583	0.171	−0.874	−0.636	0.001	1.58	1.169

respect to parameter variations. When this criterion is not met, and the model space exhibits abrupt or discontinuous behavior, we typically find that the greedy algorithm selects basis elements from regions in parameter space where the nonsmoothness occurs.

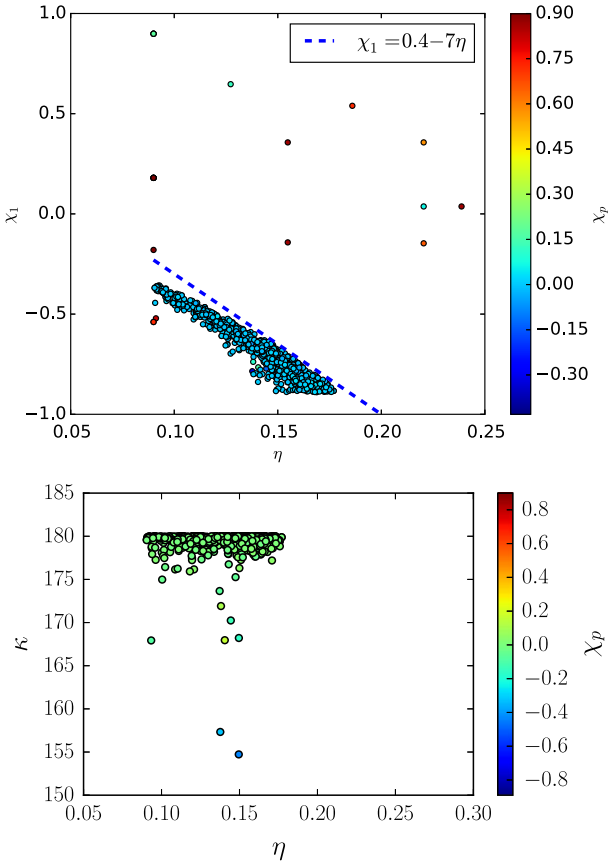

FIG. 12. *Top*: When applied to the full seven-dimensional parameter space, the greedy algorithm identifies a "feature cluster" where the model exhibits fast changing (potentially nonsmooth) behavior. The cluster directly below the dashed blue line arises for values $\chi_p \approx 0$ large antialigned spin χ_1 for unequal mass ratios. Figure 11 shows a few waveforms from this region. *Bottom*: Values of κ, the angle between the orbital angular momentum L and the total spin $S = S_1 + S_2$, as a function of the symmetric mass ratio η for the same cyan ($\chi_p \approx 0$) cluster as shown in the top panel. We observe a clear clustering between 175° and 180°.

We can use this to our advantage by simply inspecting the location of points selected by the greedy algorithm and monitoring for high density clusters. This technique was previously used to find a problem in SEOBNRv1 [35] (see Fig. 15 in Ref. [10]).

Below we show an example of the greedy feature detector for case A in Table I. Figure 12 (top) shows a cluster that was identified in the enrichment step of our basis building pipeline. The cluster (cyan circles) corresponds to a subspace that we approximate as $\chi_1 < 0.4 - 7\eta$. For reasons previously discussed, such clusters are problematic for building ROQs. By removing this cluster from the parameter space in all the cases in Table I, we are able to maintain a sparse and accurate basis and empirical interpolant.

The lower panel in Fig. 12 plots the value of κ, which denotes the angle between L and the total spin S at the reference frequency $f_{\rm ref}$, from the $\chi_{\rm p} \sim 0$ cluster. We find that the majority of waveforms from this cluster satisfies $175° \leq \kappa \leq 180°$, which is consistent with the condition for the occurrence of transitional precession [28] (which, in this case, may or may not be of a physical origin). It was shown in [28] that a requirement for the system to undergo transitional precession is $\kappa \geq 164°$. Transitional precession is more likely to occur in binary systems with high mass ratios and initial conditions where the magnitudes of $\vec{L}$ and $\vec{S}$ are similar and point in nearly opposite directions. Such cases are not correctly described by the IMRPhenomPv2 waveform model, and (unphysical) sharp features in this region of the parameter space are identified by the greedy algorithm as shown in the bottom panel of Fig. 12.

As discussed in Sec. II D, the waveform model under consideration, IMRPhenomPv2, does not faithfully model these cases and therefore the occurrence of sharp features in this region of the parameter space may be possible. To illustrate this, Fig. 11 explicitly shows examples (see Table II) of the abrupt features in the IMRPhenomPv2 amplitudes. For comparison, we also plot SEOBNRv3 amplitudes[8] which behave smoothly for those cases.

[8]The mapping from the general spin information used by SEOBNRv3 to IMRPhenomPv2's internal parameters is surjective. To find parameters for SEOBNRv3 this mapping was inverted with the following choice for the spin components in a frame aligned with $\hat{L}_N$ at $f_{\rm ref} = 20$ Hz: $S_{1x} = \cos(\alpha_0)\chi_{\rm p}$, $S_{1y} = \sin(\alpha_0)\chi_{\rm p}$, $S_{1z} = \chi_1$, $S_{2x} = S_{2y} = 0$ and $S_{2z} = \chi_2$. Explicitly, the mapping is given by $(\vec{S}_1, \vec{S}_2, \hat{L}_N, f_{\rm ref}, m_1, m_2) \twoheadrightarrow (\chi_1, \chi_2, \chi_{\rm p}, \theta_{\rm J}, \alpha_0, f_{\rm ref}, m_1, m_2)$, where $\hat{L}_N = (\sin(\iota), 0, \cos(\iota))$ (in a frame aligned with the view direction), ι is the angle between $\hat{L}_N$ and the line of sight and $\theta_{\rm J}$ is the angle between $\vec{J}$ and the line of sight.

[1] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. Lett. **116**, 061102 (2016).

[2] B. P. Abbott *et al.* (Virgo, LIGO Scientific), Phys. Rev. Lett. **116**, 221101 (2016).

[3] J. Abadie, B. P. Abbott, R. Abbott, M. Abernathy, T. Accadia, F. Acernese, C. Adams, R. Adhikari, P. Ajith, B. Allen *et al.*, Classical Quantum Gravity **27**, 173001 (2010).

[4] B. P. Abbott *et al.* (Virgo, LIGO Scientific), arXiv: 1602.03842.

[5] J. Veitch, V. Raymond, B. Farr, W. Farr, P. Graff, S. Vitale, B. Aylott, K. Blackburn, N. Christensen, M. Coughlin *et al.*, Phys. Rev. D **91**, 042003 (2015).

[6] J. Aasi, J. Abadie, B. P. Abbott, R. Abbott, T. D. Abbott, M. Abernathy, T. Accadia, F. Acernese, C. Adams, T. Adams *et al.* (LIGO-Virgo Scientific Collaboration), Phys. Rev. D **88**, 062001 (2013).

[7] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Phys. Rev. Lett. **113**, 151101 (2014).

[8] Ligo/lsc algorithms library, https://www.lsc-group.phys.uwm.edu/daswg/projects/lalsuite.html.

[9] B. S. Sathyaprakash and B. F. Schutz, Living Rev. Relativ. **12**, 18 (2009).

[10] M. Pürrer, Classical Quantum Gravity **31**, 195010 (2014).

[11] C. Devine, Z. B. Etienne, and S. T. McWilliams, Classical Quantum Gravity **33**, 125025 (2016).

[12] P. Canizares, S. E. Field, J. Gair, V. Raymond, R. Smith, and M. Tiglio, Phys. Rev. Lett. **114**, 071104 (2015).

[13] J. Blackman, B. Szilagyi, C. R. Galley, and M. Tiglio, Phys. Rev. Lett. **113**, 021101 (2014).

[14] M. Pürrer, Phys. Rev. D **93**, 064041 (2016).

[15] S. E. Field, C. R. Galley, J. S. Hesthaven, J. Kaye, and M. Tiglio, Phys. Rev. X **4**, 031006 (2014).

[16] S. Caudill, S. E. Field, C. R. Galley, F. Herrmann, and M. Tiglio, Classical Quantum Gravity **29**, 095016 (2012).

[17] S. E. Field, C. R. Galley, F. Herrmann, J. S. Hesthaven, E. Ochsner, and M. Tiglio, Phys. Rev. Lett. **106**, 221102 (2011).

[18] H. Antil, S. Field, F. Herrmann, R. Nochetto, and M. Tiglio, J. Sci. Comput. **57**, 604 (2013).

[19] P. Canizares, S. E. Field, J. R. Gair, and M. Tiglio, Phys. Rev. D **87**, 124005 (2013).

[20] J. Blackman, S. E. Field, C. R. Galley, B. Szilágyi, M. A. Scheel, M. Tiglio, and D. A. Hemberger, Phys. Rev. Lett. **115**, 121102 (2015).

[21] GREEDYCPP, https://bitbucket.org/sfield83/greedycpp/.

[22] H. Antil, D. Chen, and S. Field (to be published).

[23] S. E. Field, C. R. Galley, and E. Ochsner, Phys. Rev. D **86**, 084046 (2012).

[24] C. P. L. Berry *et al.*, Astrophys. J. **804**, 114 (2015).

[25] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk, SIAM J. Math. Anal. **43**, 1457 (2011).

[26] Y. Maday, N. C. Nguyen, A. T. Patera, and S. H. Pau, Commun. Pure Appl. Anal. **8**, 383 (2009).

[27] S. Chaturantabut and D. C. Sorensen, SIAM J. Sci. Comput. **32**, 2737 (2010).

[28] T. A. Apostolatos, C. Cutler, G. J. Sussman, and K. S. Thorne, Phys. Rev. D **49**, 6274 (1994).

[29] L. E. Kidder, Phys. Rev. D **52**, 821 (1995).

[30] S. Ossokine, M. Boyle, L. E. Kidder, H. P. Pfeiffer, M. A. Scheel, and B. Szilágyi, Phys. Rev. D **92**, 104028 (2015).

[31] P. Schmidt, M. Hannam, and S. Husa, Phys. Rev. D **86**, 104063 (2012).

[32] Y. Pan, A. Buonanno, A. Taracchini, L. E. Kidder, A. H. Mroué, H. P. Pfeiffer, M. A. Scheel, and B. Szilágyi, Phys. Rev. D **89**, 084006 (2014).

[33] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. J. Forteza, and A. Bohé, Phys. Rev. D **93**, 044006 (2016),.

[34] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. J. Forteza, and A. Bohé, Phys. Rev. D **93**, 044007 (2016).

[35] A. Taracchini, Y. Pan, A. Buonanno, E. Barausse, M. Boyle, T. Chu, G. Lovelace, H. P. Pfeiffer, and M. A. Scheel, Phys. Rev. D **86**, 024011 (2012).

[36] A. Bohé *et al.* (to be published).

[37] P. Schmidt, F. Ohme, and M. Hannam, Phys. Rev. D **91**, 024043 (2015).

[38] C. Misner, K. Thorne, and J. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).

[39] K. Cannon, A. Chapman, C. Hanna, D. Keppel, A. C. Searle, and A. J. Weinstein, Phys. Rev. D **82**, 044025 (2010).

[40] K. Cannon, R. Cariou, A. Chapman, M. Crispin-Ortuzar, N. Fotopoulos, M. Frei, C. Hanna, E. Kara, D. Keppel, L. Liao *et al.*, Astrophys. J. **748**, 136 (2012).

[41] J. Aasi, J. Abadie, B. P. Abbott, R. Abbott, T. Abbott, M. R. Abernathy, T. Accadia, F. Acernese, C. Adams, T. Adams *et al.*, Classical Quantum Gravity **32**, 115012 (2015).

[42] J. Aasi, J. Abadie, B. P. Abbott, R. Abbott, T. Abbott, M. R. Abernathy, T. Accadia, F. Acernese, C. Adams, T. Adams *et al.*, Tech. Rep. LIGO Document No. T1100338-v13, 2012.

[43] S. Kullback and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1951).

[44] B. P. Abbott *et al.* (Virgo, LIGO Scientific), Phys. Rev. Lett. **116**, 241102 (2016).

[45] C. Haster, R. Smith, K. Blackburn, S. Field, V. Raymond, M. Pürrer, and P. Schmidt (to be published).

[46] Intel many integrated core architecture (intel mic architecture), https://software.intel.com/en-us/forums/intel-many-integrated-core, accessed: 2016-03-07.

[47] Intel c++ compilers, https://software.intel.com/en-us/c-compilers, accessed: 2016-03-07.

[48] Intel math kernel library (intel mkl), https://software.intel.com/en-us/intel-mkl, accessed: 2016-03-07.

[49] A. Buonanno and T. Damour, Phys. Rev. D **59**, 084006 (1999).

[50] T. Damour, A. Nagar, E. N. Dorband, D. Pollney, and L. Rezzolla, Phys. Rev. D **77**, 084017 (2008).

[51] L. Blanchet, Living Rev. Relativ. **9**, 4 (2006).

[52] A. Buonanno, B. R. Iyer, E. Ochsner, Y. Pan, and B. S. Sathyaprakash, Phys. Rev. D **80**, 084043 (2009).
[53] R. J. E. Smith, K. Cannon, C. Hanna, D. Keppel, and I. Mandel, Phys. Rev. D **87**, 122002 (2013).
[54] K. Cannon, J. D. Emberson, C. Hanna, D. Keppel, and H. P. Pfeiffer, Phys. Rev. D **87**, 044008 (2013).
[55] K. Cannon, C. Hanna, and D. Keppel, Phys. Rev. D **85**, 081504 (2012).
[56] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson *et al.*, Comput. Sci. Eng. **16**, 62 (2014).