Surrogate model of hybridized numerical relativity binary black hole waveforms

Vijay Varma,1 Scott E. Field,2 Mark A. Scheel,1 Jonathan Blackman,1 Lawrence E. Kidder,3 and Harald P. Pfeiffer4,5

1Theoretical Astrophysics, California Institute of Technology, Pasadena, CA 91125, USA
2Department of Mathematics, Center for Scientific Computing and Visualization Research, University of Massachusetts, Dartmouth, MA 02747, USA
3Center for Radiophysics and Space Research, Cornell University, Ithaca, New York 14853, USA
4Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Am Mühlenberg 1, Potsdam 14476, Germany
5Canadian Institute for Theoretical Astrophysics, 60 St. George Street, University of Toronto, Toronto, ON M5S 3H8, Canada

(Dated: December 20, 2018)

Numerical relativity (NR) simulations provide the most accurate binary black hole gravitational waveforms, but are prohibitively expensive for applications such as parameter estimation. Surrogate models of NR waveforms have been shown to be both fast and accurate. However, NR-based surrogate models are limited by the training waveforms’ length, which is typically about 20 orbits before merger. We remedy this by hybridizing the NR waveforms using both post-Newtonian and effective one body waveforms for the early inspiral. We present NRHybSur3dq8, a surrogate model for hybridized nonprecessing numerical relativity waveforms, that is valid for the entire LIGO band (starting at 20 Hz) for stellar mass binaries with total masses as low as 2.25 $M_\odot$. We include the $\ell \leq 4$ and (5,5) spin-weighted spherical harmonic modes but not the (4,1) or (4,0) modes. This model has been trained against hybridized waveforms based on 104 NR waveforms with mass ratios $q \leq 8$, and $|\chi_{1z}|, |\chi_{2z}| \leq 0.8$, where $\chi_{1z}$ ($\chi_{2z}$) is the spin of the heavier (lighter) BH in the direction of orbital angular momentum. The surrogate reproduces the hybrid waveforms accurately, with mismatches $\lesssim 3 \times 10^{-4}$ over the mass range 2.25$M_\odot \leq M \leq 300M_\odot$. At high masses ($M \gtrsim 40M_\odot$), where the merger-ringdown is more prominent, we show roughly two orders of magnitude improvement over existing waveform models. We also show that the surrogate works well even when extrapolated outside its training parameter space range, including at spins as large as 0.998. Finally, we show that this model accurately reproduces the spheroidal-spherical mode mixing present in the NR ringdown signal.

I. INTRODUCTION

The era of gravitational wave (GW) astronomy has been emphatically unveiled with the recent detections [1–7] by LIGO [8] and Virgo [9]. The detection of gravitational wave signals from compact binary sources is expected to become a routine occurrence as the advanced detectors reach their design sensitivity [10, 11]. The possible science output from these events crucially depends on the availability of an accurate waveform model to compare against observed signals.

Numerical relativity (NR) is the only ab initio approach that accurately produces waveforms from the merger of a binary black hole (BBH) system. However, because NR simulations are computationally expensive, it is impractical to use them directly for applications such as parameter estimation, which can require upwards of $10^9$ waveform evaluations. Therefore, the GW community has developed several approximate waveform models [12–21], some of which are fast to evaluate. These models make certain physically-motivated assumptions about the underlying phenomenology of the waveforms, and they fit for any remaining free parameters using NR simulations.

Surrogate modeling [22, 23] is an alternative approach that doesn’t assume an underlying phenomenology and has been applied to a diverse range of problems [22–31]. NR Surrogate models follow a data-driven approach, directly using the NR waveforms to implicitly reconstruct the underlying phenomenology. Three NR surrogate models have been built so far [26–28], including a 7-dimensional (mass ratio $q$ and two spin vectors) model for generically precessing systems in quasi-circular orbit [28]. Through cross-validation studies, these models were shown to be nearly as accurate as the NR waveforms they were trained against.

Despite the success of the surrogate modeling approach, existing surrogate models have two important limitations: (1) Because they are based solely on NR simulations, which typically are only able to cover the last $\sim 20$ orbits of a BBH inspiral, they are not long enough to span the full LIGO band for stellar mass binaries. (2) Apart from the first non-spinning model [26], these models have been restricted to mass ratios $q \leq 2$ [1]. There are two reasons for this: (i) The 7d parameter space is vast, requiring at least a few thousand simulations to sufficiently cover it. (ii) Because of the smaller length scale introduced by the lighter black hole, NR simulations become increasingly more expensive with mass ratio.

In this work we address these limitations in the context of nonprecessing BBH systems. First, to include the early inspiral we “hybridize” the NR waveforms: each full waveform consists of a post-Newtonian (PN) and effective one body (EOB) waveform at early times that is

$1$ We use the convention $q = m_1/m_2$, where $m_1$ and $m_2$ are the masses of the component black holes, with $m_1 \geq m_2$. 
smoothly attached to an NR waveform at late times. Second, since we restrict ourselves to the 3-dimensional space of nonprecessing BBHs, fewer simulations are necessary compared to the 7-dimensional case, and therefore we can direct computational resources to simulations with higher mass ratios. The resulting model, NRHybSur3dq8, is the first NR-based surrogate model to span the entire LIGO frequency band for stellar mass binaries; assuming a detector low-frequency cut-off of 20 Hz, this model is valid for total masses as low as \(2.25M_\odot\). This model is based on 104 NR waveforms in the parameter range \(q \leq 8\), and \(|\chi_{1z}|, |\chi_{2z}| \leq 0.8\), where \(\chi_{1z}, (\chi_{2z})\) is the dimensionless spin of the heavier (lighter) black hole (BH).

The plus \((h_+)\) and cross \((h_\times)\) polarizations of GWs can be conveniently represented by a single complex timedomain waveform \(h(t)\) that can be decomposed into a sum of spin-weighted spherical harmonic modes \(h_{\ell m}\) [32, 33], so that the waveform along any direction \((\iota, \varphi_0)\) in the binary’s source frame is given by

\[
h(t, \iota, \varphi_0) = \sum_{\ell=-\infty}^{\infty} \sum_{m=-\ell}^{\ell} h_{\ell m}(t) - 2Y_{\ell m}(\iota, \varphi_0),
\]

where \(-2Y_{\ell m}\) are the spin \(-2\) weighted spherical harmonics, \(\iota\) is the inclination angle between the orbital angular momentum of the binary and line-of-sight to the detector, and \(\varphi_0\) is the initial binary phase. \(\varphi_0\) can also be thought of as the azimuthal angle between the \(x\)-axis of the source frame and the line-of-sight to the detector. We define the source frame as follows: The \(z\)-axis is along the orbital angular momentum direction, which is constant for nonprecessing BBH. The \(x\)-axis is along the line of separation from the lighter BH to the heavier BH at some reference time/frequency. The \(y\)-axis completes the triad.

The \(\ell = |m| = 2\) terms typically dominate the sum in Eq. (1), and are referred to as the quadrupole modes. Studies [34-41] have shown that the nonquadrupole modes, while being subdominant, can play a nonnegligible role in detection and parameter estimation of GW sources, particularly for large signal to noise ratio (SNR), large total mass, large mass ratio, or large inclination angle \(\iota\). For the first event, GW150914 [1], the systematic errors due to the quadrupole-mode-only approximation are generally smaller than the statistical errors [42, 43], although higher modes may lead to modest changes in some of the extrinsic parameter values [44]. However, as the detectors approach the design sensitivity [10], one should prepare for high-SNR sources (particularly at larger mass ratios than those seen so far), where the quadrupole-mode-only approximation breaks down. In addition, nonquadrupole modes can help break the degeneracy between the binary inclination and distance, which is present for quadrupole-mode-only models (see e.g. [14, 45, 46]).

In this work, we model the following spin-weighted spherical harmonic modes: \(\ell \leq 4\) and \((5,5)\), but not the \((4,1)\) or \((4,0)\) modes \(^2\). Several inspiral-merger-ringdown waveforms [14, 15, 20, 21] that include nonquadrupole modes have been developed in recent years; however, compared to those models we show an improved accuracy and we include more modes.

The rest of the paper is organized as follows. In Sec. II we choose the parameters at which to perform NR simulations, which will be used for training the surrogate model. Sec. III describes the NR simulations. Sec. IV describes our procedure to compute the waveform for the early inspiral using PN and EOB waveforms. Sec. V describes our hybridization procedure to attach the early inspiral waveform to the NR waveforms. Sec. VI describes the construction of the surrogate model. In Sec. VII, we test the surrogate model by comparing against NR and hybrid waveforms. We end with some concluding remarks in Sec. VIII. We make our model available publicly through the easy-to-use Python package gwsurrogate [47].

II. TRAINING SET GENERATION

A. Greedy parameters from PN surrogate model

We do not know a priori the distribution or number of NR simulations required to build an accurate surrogate model. Furthermore, we hope to select a representative distribution that will allow for an accurate surrogate to be built with as few NR simulations as possible. Therefore, we estimate this distribution by first building a surrogate model for PN waveforms: we find that parameters suitable for building an accurate PN surrogate are also suitable for building an NR or a hybrid NR-PN surrogate.

We use the same methods to build the PN surrogate as we use for the hybrid surrogate (cf. Sec. VI). We use the PN waveforms described in Sec. IV A; however, for simplicity we only model the \((2,2)\) mode. In addition, we restrict the length of the PN waveforms to be 5000\(M\), terminating at the innermost-stable-circular-orbit’s orbital frequency, \(\omega_{\text{orb}} = 6^{-3/2} \, \text{rad} / M\), where \(M\) is the total mass of the binary.

We determine the desired training data set of parameters as follows. We begin with just the corner cases of the parameter space; for the 3d case considered here, that consists of 8 points at \((q, \chi_{1z}, \chi_{2z}) = (1 \text{ or } 8, \pm 0.8, \pm 0.8)\). We build up the desired set of parameters iteratively, in a greedy manner: At each iteration we build a PN surrogate using the current training data set and test the model against a much larger (~10 times) validation data set. The validation data set is generated by randomly resampling the parameter space at each iteration. Since the boundary cases are expected to be more important, for 30% of the points in the validation set we

\(^2\) Because of the symmetries of nonprecessing BBHs (see Eq. (23)), the \(m < 0\) modes contain the same information as the \(m > 0\) modes, and do not need to be modeled separately.
sample only from the boundary of the parameter space, which corresponds to the faces of a cube in the 3d case. We select the parameter in the validation set that has the largest error (cf. Eq. (2)), and add this to our training set (hence the name greedy parameters). We repeat until the validation error reaches a certain threshold.

In order to estimate the difference between two complexified waveforms, \( h_1 \) and \( h_2 \), we use the time-domain mismatch,

\[
\mathcal{M}_M = 1 - \frac{\langle h_1, h_2 \rangle}{\sqrt{\langle h_1, h_1 \rangle \langle h_2, h_2 \rangle}},
\]

\[
\langle h_1, h_2 \rangle = \left| \int_{t_{\text{min}}}^{t_{\text{max}}} h_1(t) h_2^*(t) \, dt \right|,
\]

where * indicates a complex conjugation, and | | indicates the absolute value. Note that in this section, we do not perform an optimization over time and phase shifts. In addition, we assume a flat noise curve.

**FIG. 1.** Largest mismatch of the surrogate (over the entire validation set) as a function of number of greedy parameters used to train the PN surrogate. The PN surrogate is seen to converge to the validation waveforms as the size of the training data set increases.

Figure 1 shows how the maximum validation error decreases as we add greedy parameters to our training data set. For our case, we stop at 100 greedy parameters (at which point the mismatch is \(< 10^{-6}\)) and use those parameters to perform the NR simulations. Note that we don’t expect 100 NR simulations to produce an NR surrogate with comparable accuracy, \( \mathcal{M}_M < 10^{-6} \), for two reasons. First, unlike the PN waveforms used here, the NR simulations also include the merger-ringdown part, which we expect to be more difficult to model. Second, the NR numerical truncation error is typically higher than \( 10^{-6} \) in mismatch, therefore the numerical noise will limit the accuracy.

### III. NR SIMULATIONS

The NR simulations for this model are performed using the Spectral Einstein Code (SpEC) [48–53] developed by the SXS [54] collaboration. Of the 100 cases determined in Sec. II, only 91 simulations were successfully completed. These simulations have been assigned the identifiers SXS:BBH:1419 - SXS:BBH:1509, and will be made publicly available in the upcoming update of the SXS public catalog [55]. For cases with equal mass, but unequal spins, we can exchange the two BHs to get an extra data point. There are 13 such cases, leading to a total of 104 NR waveforms. These are shown as circular markers in Fig. 2.

**FIG. 2.** The parameter space covered by the 104 NR waveforms (circle markers) used in the construction of the surrogate model in Sec. VI. We also show the 9 long NR waveforms (square markers) used to test hybridization in Sec. VII B, and the 8 NR waveforms (triangle markers) used to test extrapolation in Sec. VII C. The axes show the mass ratio and the spin on the heavier BH, while the colors indicate the spin on the lighter BH. The black rectangle indicates the bounds of the training region: \( 1 \leq q \leq 8, \ -0.8 \leq \chi_{1z}, \chi_{2z} \leq 0.8 \).

The start time of these simulations varies between 4270M and 5227M before the peak of the waveform amplitude (defined in Eq. (38)), where \( M = m_1 + m_2 \) is the total Christodoulou mass measured after the initial burst of junk radiation. The algorithm for choosing a fiducial time at which junk radiation ends is discussed in Ref. [55]. The initial orbital parameters are chosen through an iterative procedure [56] such that the orbits are quasicircular; the largest eccentricity for these simulations is \( 7.5 \times 10^{-4} \), while the median value is \( 4.2 \times 10^{-4} \). The waveforms

\(^3\) The main reason for failure is large constraint violation as the binary approaches merger. We believe a better gauge condition may be needed for some of these simulations.
are extracted at several extraction surfaces at varying finite radii form the origin and then extrapolated to future null infinity [57]. Finally, the extrapolated waveforms are corrected to account for the initial drift of the center of mass [58, 59]. The time steps during the simulations are chosen nonuniformly using an adaptive time-stepper [55]. We interpolate these data to a uniform time step of $0.1M$; this is dense enough to capture all frequencies of interest, including near merger.

IV. INSPIRAL WAVEFORMS

While NR provides accurate waveforms, computational constraints limit NR to only the late inspiral, merger, and ringdown phases. Fortunately, PN/EOB waveforms are expected to be accurate in the early inspiral. Hence we can “stitch” together an early inspiral waveform and an NR waveform, to get a hybrid waveform [39, 60–67] that spans the entire frequency range relevant for ground-based detectors. In this section, we describe the waveforms we use for the early inspiral, leaving the hybridization procedure for the next section.

A. PN waveforms

We first generate PN waveforms as implemented in the GWFrames package [68]. For the orbital phase we include nonspinning terms up to 4 PN order [69–73] and spin terms up to 2.5 PN order [74–76]. We use the TaylorT4 [77] approximant to generate the PN phase; however, as described below, we replace this phase with an EOB-derived phase. For the amplitudes, we include terms up to 3.5 PN order [78–80].

The spherical harmonic modes of the PN waveform can be written (after rescaling to unit total mass and unit distance) as [70, 78],

$$h^\text{PN}_{lm} = 2 \eta (v^\text{PN})^2 \frac{16\pi}{3} H_{lm}^\text{PN} e^{-i m \phi^\text{PN}_{\text{orb}}},$$

where $\eta = q/(1+q)^2$ is the symmetric mass ratio, $v^\text{PN}$ is the characteristic speed that sets the perturbation scale in PN, $\phi^\text{PN}_{\text{orb}}$ is the (real) orbital phase, and $H_{lm}^\text{PN}$ are the complex amplitudes of different modes. Note that we ignore the tail distortions [81, 82] to the orbital phase as these are 4 PN corrections (see e.g. [83]).

The complex strain $h^\text{PN}_{lm}$ is obtained as a time series from GWFrames. We can absorb the complex part of the amplitudes into the phases and rewrite the strain as

$$h^\text{PN}_{lm} = A^\text{PN}_{lm} e^{-i \phi^\text{PN}_{lm}},$$

$$\phi^\text{PN}_{lm} = m \phi^\text{PN}_{\text{orb}} + \xi^\text{PN}_{lm},$$

$$\phi^\text{PN}_{\text{orb}} = \frac{\phi^\text{EOB}_{22}}{2},$$

where $A^\text{PN}_{lm}$ and $\phi^\text{PN}_{lm}$ are the real amplitude and phase of a given mode, and $\xi^\text{PN}_{lm}$ is an offset that captures the complex part of $H^\text{PN}_{lm}$. Note that Eqs. (6) and (7) together imply $\xi^\text{PN}_{22} = 0$; $H^\text{PN}_{22}$ contains complex terms starting at 2.5PN, but these appear as 5PN corrections in the phase (see e.g. [83]), which we can safely ignore.

At this stage, $A^\text{PN}_{lm}$, $\phi^\text{PN}_{lm}$, and $\xi^\text{PN}_{lm}$ are functions of time. But they can be recast as functions of the characteristic speed by first computing

$$v^\text{PN}(t) = \left(\frac{d\phi^\text{PN}_{\text{orb}}}{dt}\right)^{1/3},$$

where the derivative is performed numerically, and then inverting Eq. (8) to obtain $t(v^\text{PN})$. Then we define

$$A^\text{PN}_{lm}(v^\text{PN}) = |h^\text{PN}_{lm}(t(v^\text{PN}))|,$$

$$\xi^\text{PN}_{lm}(v^\text{PN}) = \phi^\text{PN}_{lm}(t(v^\text{PN})) - m \phi^\text{PN}_{\text{orb}}(t(v^\text{PN})).$$

Note that the PN waveform is generated in the source frame defined such that the reference time is the initial time. This also ensures that the heavier BH is on the positive $x$–axis at the initial time, and the initial orbital phase is zero.

To summarize: From the GWFrames package, we obtain the complex time series $h^\text{PN}_{lm}$ (Eq. (5)). We compute the orbital phase (Eq. (7)), the real amplitudes (Eq. (9)), and the phase offsets (Eq. (10)). These three quantities are obtained as a time series but can be represented as functions of the characteristic speed using Eq. (8).

B. EOB correction

As was done in previous works [34, 39], we find that the accuracy of the inspiral waveform can be improved by replacing the PN phase with the phase derived from an NR-calibrated EOB model. For this work we use SEOBNRv4 [17].

SEOBNRv4 is a time domain model that includes only the $(2,2)$ mode, which we can decompose as follows:

$$h^\text{EOB}_{22} = A^\text{EOB}_{22} e^{-i \phi^\text{EOB}_{22}},$$

where $A^\text{EOB}_{22}$ and $\phi^\text{EOB}_{22}$ are the real amplitude and phase of the $(2,2)$ mode. These are functions of time, but following the same procedure as earlier, they can be recast in terms of the characteristic speed:

$$\phi^\text{EOB}_{\text{orb}}(t) = \frac{\phi^\text{EOB}_{22}(t)}{2},$$

$$v^\text{EOB}(t) = \left(\frac{d\phi^\text{EOB}_{\text{orb}}}{dt}\right)^{1/3},$$

where the derivative is performed numerically, and we invert Eq. (13) to obtain $t(v^\text{EOB})$. We replace $v^\text{PN} \rightarrow v^\text{EOB}$...
see that the PN waveform becomes inaccurate at late times, as expected. The EOB-corrected waveform, on the other hand, remains faithful to the NR waveform until much later times.
A. Choice of cost function

We use the following cost function when comparing two waveforms, \( h \) and \( \tilde{h} \), in the matching region:

\[
\mathcal{E}[h, \tilde{h}] = \frac{1}{2} \sum_{\ell,m} \int_{t_1}^{t_2} \left| h_{\ell m}(t) - \tilde{h}_{\ell m}(t) \right|^2 dt,
\]

where \( t_1 \) and \( t_2 \) denote the start and end of the matching region, to be defined in Sec. V C, and the sum does not include \( m = 0 \) modes for reasons described in Sec. V B. This cost function was introduced in Ref. [27] and is shown to be related to the weighted average of the mismatch over the sky.

We minimize the cost function by varying the time and frame shifts between the NR and inspiral waveforms

\[
\min_{t_0, \varphi_0, \psi} \mathcal{E}[\tilde{h}^\text{NR}(t; \varphi_0, \psi), h^\text{ins}(t; t_0)],
\]

where

\[
\tilde{h}_{\ell m}(t; \varphi_0, \psi) = h_{\ell m}^\text{NR}(t) e^{i \varphi_0} e^{2i \psi},
\]

\[
h_{\ell m}^\text{ins}(t; t_0) = \tilde{h}_{\ell m}(t - t_0).
\]

We perform the time shifts on the inspiral waveform so that the matching region always corresponds to the same segment of the NR waveform. The frame shifts are performed on the NR waveform so as to preserve the initial frame alignment of the inspiral waveform (cf. Sec. IV A). This alignment gets inherited by the hybrid waveform, and is important in the surrogate construction.

B. \( m = 0 \) modes

We find that the \( m = 0 \) modes of the inspiral waveforms do not agree very well with the NR waveforms. There are several possible reasons for this [86]: (1) The NR waveform does not have the correct “memory” contribution since this depends on the entire history of the system starting at \( t = -\infty \), while the NR simulation covers only the last few orbits. (2) The extrapolation to future null infinity does not work as well for these modes [55]. This could be improved in the future with Cauchy Characteristic Extraction [87–90]. (3) The amplitude of these modes is very small except very close to merger; therefore, the early part of the NR waveform where we compare with the inspiral waveforms is contaminated by numerical noise.

Therefore, when constructing the hybrid waveforms, we set the entire inspiral waveform to zero for these modes,

\[
h_{\ell m=0}^{\text{ins}} = 0.
\]

When computing the cost function (Eq. (18)), we ignore the \( m = 0 \) modes.

This means that our hybrid waveforms for these modes are equivalent to the NR waveforms. In addition, the main contribution for these modes comes from the region close to merger, which does not correspond to a memory signal, but instead is due to axisymmetric excitations near merger (cf. bottom panel of Fig. 5).

C. Choice of matching region

There are several considerations to take into account when choosing a matching region \([t_1, t_2]\) for the cost function (Eq. (18)): (1) The NR and inspiral waveforms should agree with each other reasonably in this region; at early times the NR waveform is contaminated by junk radiation while at late times the inspiral waveform deviates from NR (cf. Figs. 3 and 5). (2) The matching region should be wide enough that the cost function is meaningful.

Our matching region starts at \( 1000M \) after the start of the NR waveform; we find that this is necessary to avoid noise due to junk radiation in some of the higher order modes. The length of the NR waveforms from the start of the matching region to the peak of the waveform amplitude varies between \( 3270M \) and \( 4227M \). The width of the matching region is then chosen to be equal to the time taken for 3 orbits of the binary. We use the phase of the \((2,2)\) mode of the NR waveform to determine this. This choice ensures the width of the matching region scales appropriately with the NR starting frequency, so that we get wider matching regions when the NR waveform starts early in the inspiral.

D. Allowed ranges for frame and time shifts

The allowed range for \( \varphi_0 \) is \([0, 2\pi]\). For nonprecessing binaries the allowed values for \( \psi \) can be restricted by taking into account the symmetries of the system. We will show that this restriction is a consequence of the well-known relationship

\[
\dot{h}_{\ell, -m} = (-1)^\ell \dot{h}_{\ell, m},
\]

between the \( m < 0 \) modes and the \( m > 0 \) modes for nonprecessing binaries. We compute the shifted waveform

\[
\dot{h}_{\ell, -m}(t) e^{-im\varphi_0} e^{2i\psi} = \dot{h}_{\ell, -m}(t; \varphi_0, \psi)
\]

\[
= (-1)^\ell (\dot{h}_{\ell, m}(t; \varphi_0, \psi))^* = (-1)^\ell e^{-2i\psi} e^{-im\varphi_0} \dot{h}_{\ell, m}(t)
\]

\[
= e^{-2i\psi} e^{-im\varphi_0} \dot{h}_{\ell, -m}(t)
\]

\[
\Rightarrow e^{2i\psi} = e^{-2i\psi}.
\]

Eq. (24) implies that the only allowed values for \( \psi \) are 0 and \( \pi/2 \). 4 If the inspiral waveform and the NR waveform have the same sign convention, then \( \psi = 0 \). Unfortunately, not all NR catalogs and PN-waveform codes

4 \( \psi = \pi \) is also allowed, but it is degenerate with \( \psi = 0 \).
use the same sign convention, so we allow the possibility of $\psi = \pi/2$ to account for this.

To set the allowed range for $t_0$, we begin by computing the orbital frequency of the inspiral waveform, $\omega^\text{ins}$, as half the frequency of the (2,2) mode. Similarly, we compute the orbital frequency of the NR waveform, $\omega^\text{NR}$. We first time-align the NR and inspiral waveforms such that their frequencies match at the start of the matching region. This gives us a good starting point to vary the time shift.

We also define,

$$\omega^\text{ins}_{\text{mid}} = \omega^\text{NR}(t = t_1),$$

$$\omega^\text{ins}_{\text{low}} = 0.995 \times \omega^\text{ins}_{\text{mid}},$$

$$\omega^\text{ins}_{\text{hi}} = 1.005 \times \omega^\text{ins}_{\text{mid}},$$

where $\omega^\text{NR}(t = t_1)$ is the NR frequency at the start of the matching region. The allowed range for time shifts $t_0$ is restricted to lie in the interval $[\omega^\text{ins}_{\text{low}}, \omega^\text{ins}_{\text{mid}}, \omega^\text{ins}_{\text{hi}}]$, where $\omega^\text{ins}_{\text{low}}$, $\omega^\text{ins}_{\text{mid}}$ and $\omega^\text{ins}_{\text{hi}}$ are the times at which $\omega^\text{ins}(t)$ is equal to $\omega^\text{ins}_{\text{low}}$, $\omega^\text{ins}_{\text{mid}}$ and $\omega^\text{ins}_{\text{hi}}$ respectively. In other words, the allowed range for $t_0$ is a region near $t_0 = 0$. $t_0 = 0$ is the case when the frequencies of the inspiral and the NR waveforms match at $t_1$, the start of the matching region. The lower (upper) limit for $t_0$ is chosen such that the inspiral waveform has a frequency equal to 0.995 (1.005) times the NR frequency at $t_1$.

The factors in Eqs. (26) and (27) are chosen such that the time shift that minimizes the cost function is always well within the range of allowed time shifts. Hence, choosing a wider range (i.e. values of these factors farther from unity) does not improve the hybridization procedure. Note also that, like the width of the matching region in Sec. V C, setting the range of time shifts based on the orbital frequency ensures that it scales appropriately with the start frequency of the NR waveform.

The minimization in Eq. (19) is performed as follows. We vary the time shift $t_0$ over 500 uniformly spaced values in the above mentioned time range $^5$. For each of these time shifts $t_0$, we try both allowed values of $\psi \in \{0, \pi/2\}$. For each $t_0$ and $\psi$, we minimize the cost function over $\varphi_0$ using the Nelder-Mead downhill simplex minimization algorithm as implemented in Scipy [84]. To avoid local minima in the $\varphi_0$ minimization, we perform 10 searches with different initial guesses, which are sampled from a uniform random distribution in the range $[0, 2\pi]$.

E. Stitching NR and inspiral waveforms.

Having obtained the right frame and time shifts between the NR and inspiral waveforms, the final step is to smoothly stitch the inspiral waveform to the shifted NR waveform. The stitching is done using a smooth blending function:

$$\tau(t) = \begin{cases} 
0, & \text{if } t < t_1 \\
\sin^2 \left( \frac{\pi}{2} \frac{t - t_1}{t_2 - t_1} \right), & \text{if } t_1 \leq t \leq t_2 \\
1, & \text{if } t > t_2,
\end{cases}$$

where $t_1$ and $t_2$ take on the same values as those appearing in Eq. (18). Different blending functions have been proposed in the literature [60, 63, 65, 91]. Our choice is equivalent to the blending function defined in Ref. [63]. We find that our results are not sensitive to the choice of blending function.

In what follows, for brevity, we drop the hybridization parameters $\varphi_0$, $\psi$, $t_0$ with the understanding that the models are stitched together after transforming into hybridization frame,

$$h^\text{ins}_{\ell m}(t) \equiv h^\text{ins}_{\ell m}(t; t_0),$$

$$h^\text{NR}_{\ell m}(t) \equiv h^\text{NR}_{\ell m}(t; \varphi_0, \psi).$$

Given the shifted waveforms and the blending function, there are still several ways in which one can stitch the waveforms together.

1. Inertial frame stitching

One could work with the complex waveform strain and define:

$$h^\text{Hyb}_{\ell m} = (1 - \tau(t)) h^\text{ins}_{\ell m}(t) + \tau(t) h^\text{NR}_{\ell m}(t).$$

With this choice, by construction, the complex strain transitions smoothly from the inspiral part to the NR part over the matching region. However, the transition is more complicated for the frequency, since it involves time derivatives of the complex argument of the strain; the time derivatives of the blending function do not behave like a smooth blending function. This is demonstrated in the left panel of Fig. 4: the inspiral and NR frequencies agree well in the matching region but the frequency of the hybrid wave deviates from this.

2. Amplitude-Frequency stitching

To avoid the undesirable artifacts described above, we choose to perform the inspiral-NR stitching using the amplitude and frequency rather than the inertial frame strain.

We begin by decomposing the NR and inspiral waveforms into their respective amplitude and phase:

$$h^\text{NR}_{\ell m}(t) = A^\text{NR}_{\ell m} e^{-i \phi^\text{NR}_{\ell m}}, \quad h^\text{ins}_{\ell m}(t) = A^\text{ins}_{\ell m} e^{-i \phi^\text{ins}_{\ell m}}.$$
The frequency of each mode
\[ \omega_{\ell m}^{\text{NR}} = \frac{d\phi_{\ell m}^{\text{NR}}}{dt}, \quad \omega_{\ell m}^{\text{ins}} = \frac{d\phi_{\ell m}^{\text{ins}}}{dt}, \] is then numerically computed from 4th-order finite difference approximations to the time derivative. Finally, we stitch the amplitude and frequency of each mode to get their hybrid versions:
\[ A_{\ell m}^{\text{Hyb}} = (1 - \tau(t)) A_{\ell m}^{\text{ins}}(t) + \tau(t) A_{\ell m}^{\text{NR}}, \]
\[ \omega_{\ell m}^{\text{Hyb}} = (1 - \tau(t)) \omega_{\ell m}^{\text{ins}}(t) + \tau(t) \omega_{\ell m}^{\text{NR}}. \]

To get the inertial frame strain we first need to integrate the frequency to get the phase. However, we already know the phase in the region before (only inspiral) and after (only NR) the matching region. So, we integrate the hybrid frequency
\[ \phi_{\ell m}^{\text{Hyb-match-region}} = \int_{t_1}^{t_2} \omega_{\ell m}^{\text{Hyb}} dt, \]
and after (only NR) the matching region. So, we integrate the hybrid frequency
\[ \phi_{\ell m}^{\text{Hyb}} = \begin{cases} \phi_{\ell m}^{\text{ins}} + \delta_{\ell m}^1, & \text{if } t < t_1 \\ \phi_{\ell m}^{\text{Hyb-match-region}} + \delta_{\ell m}^2, & \text{if } t_1 \leq t \leq t_2 \\ \phi_{\ell m}^{\text{NR}}, & \text{if } t > t_2, \end{cases} \]
where \( \delta_{\ell m}^1 \) and \( \delta_{\ell m}^2 \) are chosen such that \( \phi_{\ell m}^{\text{Hyb}} \) is continuous at \( t_1 \) and \( t_2 \).

Since, by construction, the frequency transitions smoothly from the inspiral-waveform to NR data, we can eliminate the artifact seen in the bottom left panel of Fig 4 (dashed line), as demonstrated in the right panel of Fig. 4.

We note that since the \( m = 0 \) modes are purely real/imaginary and nonoscillatory for nonprecessing systems, they do not have a frequency associated with them, therefore we use the inertial frame stitching of Sec. V E 1 for these modes. For these modes the waveform goes from zero to the NR value over the matching region.

In the hybridized waveform we include the \( \ell \leq 4 \) and (5, 5) modes, but not the (4, 1) or (4, 0) modes. For the (4, 1) and (4, 0) modes we find that the inspiral and NR waveforms do not agree very well. An example of the final NR, inspiral and hybrid waveforms is shown in Fig. 5.

VI. BUILDING THE SURROGATE MODEL

Starting from the 104 NR waveforms described in Sec. II and Sec. III, we construct hybrid waveforms as described in Sec. V. In this section we describe our method to construct a surrogate model for these hybrid waveforms.

A. Processing the training data

Before building a surrogate model, we process the hybrid waveforms as follows.
\[ q = 8.0 \, \chi_{1z} = -0.8 \, \chi_{2z} = 0.8 \]

**FIG. 5.** An example hybrid waveform used in this work. We show the \( \ell = 2 \) modes of the inspiral, NR and hybrid waveforms. The binary parameters are shown on the top of the plot. The vertical red dashed lines indicate the matching region. Note that this plot shows the inspiral and NR waveforms after the time and frame shifts are done.

1. **Time shift**

   We shift the time arrays of the hybrid waveforms such that the peak of the total amplitude
   \[
   A_{tot} = \sqrt{\sum_{l,m} |h_{lm}|^2},
   \]  
   occurs at \( t = 0 \) for each waveform.

2. **Frequency and mass ranges of validity**

   The length of a hybrid waveform is set by choosing a starting orbital frequency for the inspiral waveform; we use \( \omega_0 = 2 \times 10^{-4} \) rad/M for all waveforms. However, for the same starting frequency, the length in time of the waveform is different for different mass ratios and spins. Since we want to construct a time-domain surrogate model, we require a common time array for all hybrid waveforms. The initial time for the surrogate is determined by the shortest hybrid waveform in the training data set; this waveform begins at a time \( \sim 5.4 \times 10^8 M \) before the peak. We truncate all hybrid waveforms to this initial time value.

   The largest starting orbital frequency among the truncated hybrid waveforms is \( \omega_0 = 2.9 \times 10^{-4} \) rad/M, which sets the low frequency limit of validity of the surrogate. For LIGO, assuming a starting GW frequency of 20 Hz, the \( (2, 2) \) mode of the surrogate is valid for total masses \( M \geq 0.9 M_\odot \). The highest spin-weighted spherical harmonic mode we include in the surrogate model is \( (5, 5) \), for which the frequency is \( 5/2 \) times that of the \( (2, 2) \) mode. Therefore, all modes of the surrogate are valid for \( M \geq 2.25 M_\odot \). This coverage of total mass is sufficient to model all stellar mass binaries of interest for ground based detectors; for an equal mass binary neutron star system, the total mass is \( M \sim 2.7 M_\odot \).

3. **Downsampling and common time samples**

   Because the hybrid waveforms are so long, it is not practical to sample the entire waveform with the same step size we use for the NR waveforms (0.1M). Fortunately, the early low-frequency portion of each waveform requires sparser sampling than the later high-frequency portion. We therefore down-sample the time arrays of the truncated hybrid waveforms to a common set of time samples. We choose these samples so that there are 5 points per orbit for the above-mentioned shortest hybrid waveform in the training data set, except for \( t \geq -1000 M \) we choose uniform time samples separated by 0.1M. This ensures that we have a denser sampling rate at late times when the frequency is higher. We retain times up to 135M, which is sufficient to capture the
Before down-sampling, we first transform the waveform into the co-orbital frame, defined as:

\[
\begin{align*}
\hat{h}_m^C &= h_m \ e^{im\phi_{\text{orb}}}, \\
\hat{h}_{22} &= A_{22} \ e^{-i\phi_{22}}, \\
\phi_{\text{orb}} &= \frac{\phi_{22}}{2},
\end{align*}
\]  

where \( h_m \) is the inertial frame hybrid waveform, \( \phi_{\text{orb}} \) is the orbital phase, and \( \phi_{22} \) is the phase of the (2,2) mode. The co-orbital frame can be thought of as roughly co-rotating with the binary, since we perform a time-dependent rotation given by the instantaneous orbital phase. Therefore the waveform is a slowly varying function of time in this frame, increasing the accuracy of interpolation to the chosen common time samples. For the (2,2) mode we save the downsampled amplitude \( A_{22} \) and phase \( \phi_{22} \), while for all other modes we save \( \hat{h}_m^C \). We find that this down-sampling results in interpolation errors \( \epsilon \lesssim 10^{-10} \) (defined in Eq. (18)) for all hybrid waveforms.

4. Phase alignment

After down-sampling to the common temporal grid of the surrogate, we rotate the waveforms about the z-axis such that the orbital phase \( \phi_{\text{orb}} \) is zero at \( t = -1000M \). Note that this by itself would fix the physical rotation up to a shift of \( \pi \). When generating the inspiral waveforms for hybridization, we align the system such that the heavier BH is on the positive \( x \)-axis at the initial frequency; this fixes the \( \pi \) ambiguity. Therefore, after this phase rotation, the heavier BH is on the positive \( x \)-axis at \( t = -1000M \) for all waveforms\(^6\).

B. Decomposing the data

It is much easier to build a model for slowly varying functions of time. Therefore, rather than work with the inertial frame strain \( h_m \), which is oscillatory, we work with simpler “waveform data pieces”\(^7\), as explained below. We build a separate surrogate for each waveform data piece. When evaluating the full surrogate model, we first evaluate the surrogate of each data piece and then recombine the data pieces to get the inertial frame strain.

A common choice in literature when working with non-precessing waveforms has been to decompose the complex strain into an amplitude and phase, each of which is a slowly varying function of time:

\[
h_{\ell m} = A_{\ell m} e^{-i\phi_{\ell m}}.
\]  

However, when \( q = 1 \) and \( \chi_{12} = \chi_{22} \), the amplitude of odd-\( m \) modes becomes zero due to symmetry. This means that the phase becomes meaningless, so one has to treat such cases separately. For example, Ref\[26\] used specialized basis functions for the odd-\( m \) modes that captured the divergent behavior of the phase in the equal-mass limit.

To avoid this issue, instead of using the amplitude and phase we use the real and imaginary parts of the co-orbital frame strain \( \hat{h}_m^C \), defined in Eq. (39), for all non-quadrupole modes. The co-orbital frame strain is always meaningful: in the special, symmetric case mentioned above, the co-orbital frame strain for the odd-\( m \) modes just goes to zero, rather than diverge. For the (2,2) mode we use the amplitude\(^7\) \( A_{22} \) and phase \( \phi_{22} \).

As mentioned above, our hybrid waveforms are very long, typically containing \( \sim 3 \times 10^4 \) orbits. This presents new challenges that are not present for pure-NR surrogates. For instance, \( \phi_{22} \) sweeps over \( \sim 4 \times 10^5 \) radians for a typical hybrid waveform. We find that the accuracy of the surrogate model at early times improves if we first subtract a PN-derived approximation to the phase, model the phase difference rather than \( \phi_{22} \), and then add back the PN contribution when evaluating the surrogate model. In particular, we use the leading order TaylorT3 approximant\[92\]. For this approximant, the phase is given as an analytic, closed-form, function of time. Therefore, even though TaylorT3 is known to be less accurate than some other approximants\[93\], its speed makes it ideal for our purpose as we only need it to capture the general trend. At leading order, the TaylorT3 phase is given by:

\[
\phi_{22}^{T3} = \phi_{\text{ref}}^{T3} - \frac{2}{\eta} \frac{\theta}{\theta^5},
\]

where \( \phi_{\text{ref}}^{T3} \) is an arbitrary integration constant, \( \theta = [\eta \ (t_{\text{ref}} - t)/(5M)]^{-1/8}, \) \( t_{\text{ref}} \) is an arbitrary time offset, and \( \eta \) is the symmetric mass ratio. Note that \( \phi_{22}^{T3} \) diverges at \( t = t_{\text{ref}} \). We choose \( t_{\text{ref}} = 1000M, \) long after the end of the waveform (recall that the peak is at \( t = 0 \)), to ensure that we are always far away from this divergence. We choose \( \phi_{\text{ref}}^{T3} \) such that \( \phi_{22}^{T3} = 0 \) at \( t = -1000M \); this is the same time at which we align the hybrid phase in Sec. VI A 4.

Instead of modeling \( \phi_{22} \), we model the residual

\[
\phi_{\text{res}}^{T3} = \phi_{22} - \phi_{22}^{T3},
\]

after removing the leading-order contribution \( \phi_{22}^{T3} \). By construction, \( \phi_{\text{res}}^{T3} \) goes to zero at \( t = -1000M \). We find

\[\text{Note that for the (2,2) mode } A_{22} = 6 \phi_{22}^2.\]

\[\text{Here the BH positions at } t = -1000M \text{ are defined from the waveform at future null infinity, using a phase rotation relative to the early inspiral where the BH positions are well-defined in PN theory; these positions do not necessarily correspond to the (gauge-dependent) coordinate BH positions in the NR simulation.}\]
that after removing the leading order TaylorT3 phase, the scale of $\phi_{22}^{T3}$ for a typical hybrid is $\sim 10^3$ radians, as compared to $\sim 4 \times 10^5$ radians for $\phi_{22}$. In essence, this captures almost all of the phase evolution in the early inspiral, simplifying the problem of modeling the phase to the same as modeling the phase of late-inspiral NR waveforms. We stress that the exact form of $\phi_{22}^{T3}$ (or its physical meaning) is not important, as long as it captures the general trend, since we add the exact same $\phi_{22}^{T3}$ to our model of $\phi_{22}^{\text{res}}$ when evaluating the surrogate. In fact, we find that adding higher order PN terms in Eq. (43) does not improve the accuracy of the surrogate.

To summarize, we decompose the hybrid waveforms into the following waveform data pieces, each of which is a smooth, slowly varying function of time: $(A_{22}, \phi_{22}^{\text{res}})$ for the $(2,2)$ mode, and the real and imaginary parts of $\kappa_{\ell m}^C$ for all other modes$^8$.

### C. Building the surrogate

Once we have the waveform data pieces, we build a surrogate model for each data piece using the procedure outlined in Refs. [22, 27], which we only briefly describe here. Note that the steps below are applied independently for each waveform data piece.

#### 1. Greedy basis

We first construct a greedy reduced-basis [94] such that the projection errors (cf. Eq. (5) of Ref. [27]) for the entire data set onto this basis are below a given tolerance. For the basis tolerances we use $10^{-2}$ for the $\phi_{22}^{\text{res}}$ data piece, $2 \times 10^{-5}$ for $A_{22}$, and $8 \times 10^{-6}$ for all other data pieces. These are chosen through visual inspection of the basis functions to ensure they are not noisy, and based on the expected truncation error of the NR waveforms. For instance, we expect the error in phase to be about $10^{-2}$ radians.

The greedy procedure is initialized with a single basis function as described in Ref. [27]. Then at each step in the greedy procedure, the waveform with the highest projection error onto the current basis is added to the basis. Previous work has shown that the resulting greedy reduced-basis is robust to different choices of initialization [95]. When computing the basis projection errors, we only include data up to $50M$ after the peak. We find that this helps avoid noisy basis functions. This is particularly important for the phase data piece as this becomes meaningless at late times, when the waveform amplitude becomes very small.

$^8$ For $m = 0$ modes of nonprecessing systems, $\kappa_{\ell m}^C$ is purely real (imaginary) for even (odd) $\ell$, so we ignore the imaginary (real) part for $m = 0$ modes.

#### 2. Empirical interpolation

Next, using a different greedy procedure, we construct an empirical interpolant [96–98] in time. This picks out the most representative time nodes, where the number of time nodes is the same as the number of greedy basis functions. We require that the start of the waveform always be included as a time node for all data pieces. This is a useful modeling choice because the magnitude of the waveform data pieces in the very early inspiral can be smaller than the basis tolerances mentioned above. By requiring the first index to be an empirical time node, we enforce an anchor point that ensures the waveform data piece has the right magnitude at the start of the waveform. Furthermore, we do not allow any empirical time nodes at times $> 50M$, since we expect this part to be dominated by noise (especially for the phase data piece).

#### 3. Parametric fits

Finally, for each time node, we construct a fit across the parameter space. The fits are done using the Gaussian process regression fitting method described in the supplemental material of Ref. [99]. Following Ref. [99], we parameterize our fits using $\log(q), \tilde{\chi}$, and $\chi_a$. Here $\tilde{\chi}$ is the spin parameter entering the GW phase at leading order [16, 100–102] in the PN expansion,

$$\chi_{\text{eff}} = \frac{q \chi_{1z} + \chi_{2z}}{1 + q},$$

$$\tilde{\chi} = \frac{\chi_{\text{eff}} - 38\eta(\chi_{1z} + \chi_{2z})/113}{1 - 76\eta/113},$$

and $\chi_a$ is the “anti-symmetric spin”,

$$\chi_a = \frac{1}{2}(\chi_{1z} - \chi_{2z}).$$

The fit accuracy, and as a result the accuracy of the surrogate model, improves noticeably when using $\log(q)$, compared to $q$ or $\eta$.

### D. Evaluating the surrogate

When evaluating the surrogate waveform, we first evaluate each surrogate waveform data piece. Next, we compute the phase of the $(2,2)$ mode,

$$\phi_{22} = \phi_{22}^{\text{res.S}} + \phi_{22}^{T3},$$

where $\phi_{22}^{\text{res.S}} \approx \phi_{22}^{\text{res}}$ is the surrogate model for $\phi_{22}^{\text{res}}$ and $\phi_{22}^{T3}$ is given in Eq. (43). If the waveform is required at a uniform sampling rate, we interpolate each waveform data piece from the sparse time samples used to construct the model to the required time samples, using a cubic-spline interpolation scheme. Finally, we use Eqs. (39), (40), and (41) to reconstruct the surrogate prediction for the inertial frame strain.
mismatch values over points on the sky as well as different hybrid waveforms.

In order to estimate the difference between two waveforms, \( h_1 \) and \( h_2 \), we use the mismatch, defined in Eq. (2), but in this section instead of Eq. (3) we use the frequency-domain inner-product

\[
\langle h_1, h_2 \rangle = 4 \text{Re} \int_{f_{\text{min}}}^{f_{\text{max}}} \frac{\tilde{h}_1(f) \tilde{h}_2^*(f)}{S_n(f)} df,
\]

where \( \tilde{h}(f) \) indicates the Fourier transform of the complex strain \( h(t) \), \( ^* \) indicates a complex conjugation, Re indicates the real part, and \( S_n(f) \) is the one-sided power spectral density of a GW detector. We taper the time domain waveform using a Planck window [103], and then zero-pad to the nearest power of two. We further zero-pad the waveform to increase the length by a factor of eight before performing the Fourier transform. The tapering at the start of the waveform is done over 1.5 cycles of the (2, 2) mode. The tapering at the end is done over the last 20\( M \). Note that our model contains times up to 135\( M \) after the peak of the waveform amplitude, and the signal has essentially died down by the last 20\( M \).

We compute mismatches following the procedure described in Appendix D of Ref. [27]: the mismatches are optimized over shifts in time, polarization angle, and initial orbital phase. Both plus and cross polarizations are treated on an equal footing by using a two-detector setup where one detector sees only the plus and the other only the cross polarization. We compute the mismatches at 37 points uniformly distributed on the sky in the source frame, and we use all available modes of a given waveform model.

When computing flat noise mismatches (\( S_n = 1 \)), we take \( f_{\text{min}} \) to be the frequency of the (2, 2) mode at the end of the initial tapering window, and \( f_{\text{max}} = 5f_{22}^{\text{peak}} \), where \( f_{22}^{\text{peak}} \) is the frequency of the (2, 2) mode at its peak. This choice of \( f_{\text{max}} \) ensures that we capture the peak frequencies of all modes considered in this work, including the (5, 5) mode, whose frequency has the highest multiple of the (2, 2) mode frequency of all the modes we model. We also compute mismatches with the Advanced-LIGO design sensitivity Zero-Detuned-HighP noise curve [104]. For this we use \( f_{\text{min}} = 20\text{Hz} \) and \( f_{\text{max}} = 2000\text{Hz} \).

A. Surrogate errors

We evaluate the accuracy of our new surrogate model, NRHybSur3dq8, by computing mismatches against hybrid waveforms. For this, we compute “out-of-sample” errors as follows. We first randomly divide the 104 training waveforms into groups of \( \sim5 \) waveforms each. For each group, we build a trial surrogate using the remaining \( \sim99 \) training waveforms and test against these five validation ones. We also compute the mismatch between an existing higher-mode waveform model, SEOBNRv4HM [15], and the hybrid waveforms.

Figure 6 summarizes mismatches of both NRHybSur3dq8 and SEOBNRv4HM versus the hybrid waveforms. We use all available modes for each waveform model. In the left panel we show mismatches computed...
FIG. 7. The plus polarization of the waveforms for the cases that result in the largest mismatch for NRHybSur3dq8 (top) and SEOBNRv4HM (bottom) in the left panel of Fig. 6. We also show the corresponding hybrid waveforms (labeled as NR as only the late part is shown). Each waveform is projected using all available modes for that model, along the direction which results in the largest mismatch for NRHybSur3dq8 (SEOBuNRv4HM) in the top (bottom) panel. Note that NRHybSur3dq8 is evaluated using trial surrogates that are not trained using these cases. The binary parameters and the direction in the source frame are indicated in the inset text. All waveforms are time shifted such that the peak of the total waveform amplitude occurs at \( t = 0 \) (using all available modes, according to Eq. (38)). Then the waveform modes are rotated about the \( z \)-axis such that the orbital phase is zero at \( t = -3500M \).

using a flat noise curve over the NR part of the hybrid waveforms (to do this, we truncate the waveforms and begin tapering at \( t = -3500M \)). We see that the mismatches for NRHybSur3dq8 are about two orders of magnitude lower than that of SEOBNRv4HM. We compare this with the truncation error in the NR waveforms themselves, by computing the mismatch between the two highest available resolutions of each NR waveform. The errors in the surrogate model are well within the truncation error of the NR simulations. Note that NR error estimated in this manner is a conservative estimate; if we treat the high resolution simulation as the fiducial case, the NR curve in Fig. 6 can be thought of as the error in the lower resolution simulation. This explains why the errors in the surrogate are lower than the NR errors. We suspect that the error of the high resolution simulations is close to the surrogate model’s error.

The right panel of Fig. 6 shows mismatches computed using the Advanced LIGO design sensitivity noise curve. The mismatches are now dependent on the total mass of the system, so we show mismatches for masses starting at the lower limit of the range of validity of the surrogate: \( M \geq 2.25M_\odot \). 95th-percentile mismatches for NRHybSur3dq8, are always below \( \sim 3 \times 10^{-4} \) in the mass range \( 2.25M_\odot \leq M \leq 300M_\odot \). At high masses (\( M \geq 40M_\odot \)), where the merger-ringdown is more prominent, our model is more accurate than SEOBNRv4HM by roughly two orders of magnitude, in agreement with the left panel of Fig. 6.

For high masses only the last few orbits of the hybrid waveforms are in the LIGO band, and the hybrid waveforms are effectively the same as the NR waveforms. For low masses, the errors in the right panel of Fig. 6 quantify how well different models reproduce the hybrid waveforms. However, this comparison cannot account for the errors in the hybridization procedure itself. We provide some evidence for the fidelity of the hybrid waveforms in Sec. VII B, by comparing against some long NR waveforms.

Fig. 7 shows NRHybSur3dq8 and SEOBNRv4HM waveforms for the cases leading to the largest errors in the left panel of Fig. 6. The surrogate shows very good agreement with the NR waveform, even for its worst case. SEOBNRv4HM shows a noticeably larger deviation that cannot all be accounted for with a time and/or phase shift. Note that we align the time and orbital phase of the waveforms in Fig. 7.

We note the main improvement over SEOBNRv4HM is not due to the inclusion of more modes. We find that the agreement between SEOBNRv4HM and the NR/hybrid waveforms in Figs. 6 and 7 improves only marginally when restricting the NR/hybrid waveforms to the same
set of modes as SEOBNRv4HM.

B. Hybridization errors

The errors described in Sec. VII A are computed by comparing the surrogate against hybrid waveforms, hence they do not include the errors in the hybridization procedure or the errors from EOB-corrected-PN waveforms (cf. Sec. IV B) we use for the early inspiral. To estimate these errors, we compare the surrogate against a few very long NR simulations. We perform five new simulations that are \( \sim 10^5 M \) long and two that are \( \sim 3 \times 10^4 M \) long. These have been assigned the identifiers SXS:BBH:1412 - SXS:BBH:1418, and will be made publicly available in the upcoming update of the SXS public catalog. In addition, we use two simulations of length \( \sim 3 \times 10^4 M \) from Ref. [106]. These nine simulations are represented as square markers in Fig. 2, and have not been used in training the surrogate. The surrogate was trained against hybrid waveforms whose NR duration varied between 3270M and 4227M. Therefore, comparing against long NR waveforms, which include the early inspiral, is a good way to estimate the hybridization error.

We begin by repeating the mismatch computation from the right panel of Fig. 6, using the \( 10^5 M \) long NR waveforms. This is shown in Fig. 8. We also show the errors in the NR simulations, estimated by comparing the two highest available NR resolutions. We find that the mismatches between the surrogate and the long NR waveforms for \( M > 30 M_\odot \) are below \( 10^{-4} \), in agreement with Fig. 6. For lower masses, the mismatches quickly increase and can be as high as \( \sim 10^{-2} \). However, this increase in mismatch is accompanied by an increase in the error of the NR waveforms. This is expected, since for very long NR waveforms the accumulated phase error is a dominant source of numerical error, which becomes increasingly relevant for low mass systems as more of the waveform moves in-band. Therefore, in Fig. 8, at low masses, the comparison between the surrogate and NR waveforms is largely dominated by the numerical resolution error of the long NR waveforms themselves.

We find that a better test of the hybridization procedure, one that is less sensitive to NR phase accumulation errors, is to compare against different segments of the NR waveform. Since the phase errors accumulate over a large number of cycles, by looking at smaller segments we ensure that this contribution is not the dominant error. To be precise, we compare the surrogate and the NR data, using segments of length \( \Delta t = 5 \times 10^3 M \) ending at a particular number of orbits before the peak of the waveform. For each segment we compute mismatches

\[ q=1.6, \chi_1=0.4, \chi_2=-0.3 \]

\[ q=1.4, \chi_1=0.5, \chi_2=0.4 \]

\[ q=1.8, \chi_1=-0.5, \chi_2=0.4 \]

\[ q=1.5, \chi_1=0.5, \chi_2=0.5 \]

\[ q=1.8, \chi_1=-0.4, \chi_2=-0.4 \]

FIG. 8. Comparisons between the NRHybSur3dq8 surrogate model and a few NR waveforms of \( \sim 10^5 M \) in duration. We also show the NR resolution error. 95th-percentile mismatches (over points in the sky) are shown as a function of total mass. The inset text indicates the mass ratio and component spins. Mismatches are computed using the Advanced LIGO design sensitivity noise curve. To best assess the error introduced by the hybridization procedure we use the same set of modes for the NR waveforms as the surrogate. At low masses, the hybridization errors (red circles) become less reliable measures of accuracy due to the large NR resolution error (black circles) itself. Fig. 9 describes a refined comparison to improve the assessment at low masses.

at several points in the sky using a flat noise curve. By varying the number of orbits to the peak, we can cover the entire NR waveform including the early inspiral region where the surrogate depends on the hybridization procedure. These errors are shown in Fig. 9. We find that in each segment, the mismatch between the surrogate and the NR data is, in general, lower or comparable to the NR resolution error. Therefore, the surrogate reproduces the NR data accurately in the early inspiral and the hybridization errors are smaller than or comparable to the NR resolution error for these cases. We note that the surrogate errors in Fig. 9 depend on the length of the segment considered and are only meaningful when compared to the NR errors in the same segment.

Note that for these long NR simulations, the outer boundary location is chosen based on the length of the simulations so as to avoid unphysical center-of-mass accelerations seen in earlier long-duration runs [106].
Unfortunately, long NR simulations such as these are not available at regions of the parameter space where both mass ratio and spin magnitudes are large. These are the cases where PN is expected to perform poorly, so we expect larger hybridization errors for these cases.

C. Extrapolation outside the training range

We now investigate the efficacy of NRHybSur3dq8 to extrapolate beyond its training parameter range by comparing against SpEC NR simulations [106–110] at larger mass ratios ($8 < q \leq 10$) and/or larger spin magnitudes ($|\chi_{1z}| > 0.8$ or $|\chi_{2z}| > 0.8$). These NR simulations are represented as triangle markers in Fig. 2.

Fig. 10 shows mismatches for NRHybSur3dq8 when compared against these simulations. We find that surrogate extrapolates remarkably well, with the mismatch always below $4 \times 10^{-4}$ for all cases, which include mass ratios up to $q = 10$ and spin magnitudes up to $|\chi| = 0.998$. However, the extrapolation errors can be about half an order of magnitude larger than errors within the training range. Note that NR simulations with both high mass ratios and high spin magnitudes are not currently available, and the ones used here represent the most extreme cases found in the SXS Catalog. We do not hybridize these simulations before comparing to NRHybSur3dq8 because several of them are too short. In Fig. 10, the minimum mass for each case is chosen to be the lowest mass at which all used modes of the NR simulation lie fully in the LIGO band with a low frequency cut-off of 20 Hz.

At much higher mass ratios than those tested here, such as $q = 15$, we find that the waveforms generated
by the surrogate can have “glitches” in the time series. Therefore, we recommend the surrogate be used for $q \leq 10$ and $|\chi_{1z}|, |\chi_{2z}| \leq 1$. However, we advise caution with any extrapolation in general.

D. Mode mixing

Numerical relativity waveforms are extracted as spin-weighted spherical harmonic modes [32, 33]. However, in the ringdown regime, the natural basis to use is the spin-weighted spheroidal harmonic basis [111, 112]. A spherical harmonic mode $h_{\ell m}$ can be written as a linear combination of all spheroidal harmonic modes $h_{\ell m}^S$ with the same $m$ index [113]. Therefore, during the ringdown, we expect leakage of power between different spherical harmonic modes with the same $m$. This is referred to as mode mixing.

Since the surrogate accurately reproduces the spherical harmonic modes from the NR simulations, it also captures this mode mixing. We demonstrate this for an example case in Fig. 11. Here we compute the Fourier transform of different spherical harmonic modes in the ringdown stage of the waveform. Before computing the Fourier transform, we first drop all data before $t = 20M$, where $t = 0$ corresponds to the peak of the waveform amplitude (cf. Eq. (38)). Then, we taper the data between $t = 20M$ and $t = 40M$, as well as the last 10$M$ of the time series, using a Planck window [103]. The tapering width at the start is chosen such that the remaining signal is dominated by the fundamental quasi-normal mode (QNM) overtone. Fig. 11 shows the absolute value of these Fourier transforms for different modes, for both the surrogate and the NR waveform. In addition, we show the frequency of the fundamental QNM overtone for each mode [114].

Note that the $(2, 2)$ mode and the $(3, 2)$ mode have the same $m$ index, the condition required for mode mixing. We see that the peak of the $(2, 2)$ mode agrees with the QNM frequency as expected. For the $(3, 2)$ mode however, while there are features of a peak at the expected QNM frequency, there is a much larger peak at the frequency of the $(2, 2)$ mode. This is because some of the energy of the stronger $(2, 2)$ mode has leaked into the $(3, 2)$ mode due to mode mixing. Mode mixing can also be seen for the $(3, 3)$ and $(4, 3)$ modes, which also have the same $m$ index. Fig. 11 shows that not only does the surrogate agree with NR in the ringdown, it also reproduces the mode mixing present in the NR data.

E. Evaluation cost

Figure 12 shows the evaluation cost for NRHybSur3dq8, at different total masses, starting at 20Hz, and using a sampling rate of 4096Hz. This suggests that NRHybSur3dq8 is fast enough for direct use in parameter estimation. We also show the evaluation cost per mode. This is compared to the evaluation cost of SEOBNRv4_ROM [17], a Fourier domain Reduced Order Model (ROM) version of SEOBNRv4. Note that...
We present NRHybSur3dq8, the first NR-based surrogate waveform model that spans the entire LIGO bandwidth, valid for stellar mass binaries with total masses \( M \geq 2.25M_\odot \). This model is trained on 104 NR-PN/EOB hybrid waveforms of nonprecessing quasicircular BBH systems with mass ratios \( q \leq 8 \), and spin magnitudes \( |\chi_{1z}|,|\chi_{2z}| \leq 0.8 \). The parametric fits for this model are performed using Gaussian process regression. This model includes the following spin-weighted spherical harmonic modes: \( l \leq 4 \) and \((5,5)\), but not \((4,1)\) or \((4,0)\). We make our model available publicly through the easy-to-use Python package gwsurrogate [47], therefore further speed-up could be possible with a C implementation.

**VIII. CONCLUSION**

We present NRHybSur3dq8, the first NR-based surrogate waveform model that spans the entire LIGO bandwidth, valid for stellar mass binaries with total masses \( M \geq 2.25M_\odot \). This model is trained on 104 NR-PN/EOB hybrid waveforms of nonprecessing quasicircular BBH systems with mass ratios \( q \leq 8 \), and spin magnitudes \( |\chi_{1z}|,|\chi_{2z}| \leq 0.8 \). The parametric fits for this model are performed using Gaussian process regression. This model includes the following spin-weighted spherical harmonic modes: \( l \leq 4 \) and \((5,5)\), but not \((4,1)\) or \((4,0)\). We make our model available publicly through the easy-to-use Python package gwsurrogate [47].

Through a cross-validation study, we show that the surrogate accurately reproduces the hybrid waveforms. The mismatch between them is always less than \( \sim 3 \times 10^{-4} \) for total masses \( 2.25M_\odot \leq M \leq 300M_\odot \). For high masses \( (M \geq 40M_\odot) \), where the merger-ringdown is more prominent, we show roughly a 2 orders of magnitude improvement over the current state-of-the-art model with nonquadrupole modes, SEOBNRv4HM [15].

By comparing against several long NR simulations, we show that the errors in our hybridization procedure are comparable or lower than the resolution error in current NR simulations. In addition, by comparing against available NR simulations at higher mass ratios and spins, we show that our model extrapolates reasonably well outside its training range. Based on these tests, we are cautiously optimistic that the surrogate can be used for \( q \leq 10 \) and \( |\chi_{1z}|,|\chi_{2z}| \leq 1 \), and we leave a more detailed investigation for future work.

**A. Future work**

While our tests of the hybridization procedure are encouraging, long NR simulations are available only for low mass ratios and low spin magnitudes. Therefore, we have no means to test hybridization at high mass ratios and/or high spins, where PN is expected to perform poorly. An improved surrogate model and refined study of the hybridization errors will require longer inspiral waveforms with greater coverage of the parameter space.

Another extension of interest is towards larger mass ratios and spin magnitudes. While the surrogate extrapolates very well when compared to available simulations at larger mass ratios and spins, no NR simulations are available with both large mass ratios \( (q > 8) \) and large spins \( (\chi > 0.8) \). Therefore, our model is untested in that parameter space and it might be necessary to add training points there. The model could also be extended to include precession and/or eccentricity, however this is more challenging because of the enlarged parameter space as well as more complicated hybridization.

Finally, as mentioned in Sec. VII E, the evaluation time of NRHybSur3dq8 can likely be reduced by constructing a Fourier domain ROM [23] of the time-domain model.

We leave these explorations to future work.

**ACKNOWLEDGMENTS**

We thank Matt Giesler for helping carry out the new SpEC simulations used in this work. We thank Michael Boyle, Kevin Barkett, Matt Giesler, Yanbei Chen, and Saul Teukolsky for useful discussions. We thank Patricia Schmidt for careful and detailed feedback on an earlier draft of this manuscript. V.V. and M.S. are supported by the Sherman Fairchild Foundation, and NSF grants PHY–170212 and PHY–1708213 at Caltech. S.E.F is partially supported by NSF grant PHY-1806665. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. This research is part of the Blue Waters sustained-petascale computing project, which is sup-
ported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. Simulations were performed on NSF/NCSA Blue Waters under allocation NSF PRAC–1713694; on the Wheeler cluster at Caltech, which is supported by the Sherman Fairchild Foundation and by Caltech; and on XSEDE resources Bridges at the Pittsburgh Supercomputing Center, Comet at the San Diego Supercomputer Center, and Stampede and Stampede2 at the Texas Advanced Computing Center, through allocation TG-PHY990007N. Computations for building the model were performed on Wheeler and Stampede2.

REFERENCES


[45] R. O’Shaughnessy, Benjamin Farr, E. Ochsner, Hec-


qc/9608012 [gr-qc].


naries,” Class. Quantum Grav. 27, 084020 (2010), arXiv:1003.2939 [gr-qc].


