Draft version April 18, 2024
Typeset using LaTeX **twocolumn** style in AASTeX631

Recalibrating Gravitational Wave Phenomenological Waveform Model

Kelvin K. H. Lam,[1] Kaze W. K. Wong,[2] and Thomas D. P. Edwards[3]

[1] *Department of Physics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong*
[2] *Center for Computational Astrophysics, Flatiron Institute, New York, NY 10010, USA*
[3] *William H. Miller III Department of Physics and Astronomy, Johns Hopkins University, Baltimore, Maryland 21218, USA*

ABSTRACT

We investigate the possibility of improving the accuracy of the phenomenological waveform model, IMRPhenomD, by jointly optimizing all the calibration coefficients at once, given a set of numerical relativity (NR) waveforms. When IMRPhenomD was first calibrated to NR waveforms, different parts (i.e., the inspiral, merger, and ringdown) of the waveform were calibrated separately. Using `ripple`, a library of waveform models compatible with automatic differentiation, we can, for the first time, perform gradient-based optimization on all the waveform coefficients at the same time. This joint optimization process allows us to capture previously ignored correlations between separate parts of the waveform. We found that after recalibrating with a slightly restricted parameter subspace ($q \leq 8$), despite the tail of the mismatch distribution remains similar, the median mismatch between the model and NR waveforms decreases by 50%. We further explore how different regions of the source parameter space respond to the optimization procedure. We find that the degree of improvement correlates with the spins of the source. This work shows a promising avenue to help understand and treat systematic error in waveform models.

1. INTRODUCTION

Many data analysis tasks in gravitational wave (GW) astrophysics, such as matched filtering (Owen 1996; Owen & Sathyaprakash 1999) and parameter estimation (Dax et al. 2021; Christensen & Meyer 2022; Koposov et al. 2022; Islam et al. 2022a; Romero-Shaw et al. 2020; Zackay et al. 2018; Ashton et al. 2019), rely upon accurate waveform models. Since the generation of numerical relativity (NR) waveforms is prohibitively expensive, the community has constructed waveform approximants that can be evaluated much faster. There are three families of commonly used GW approximants: the effective-one-body (EOB) (Ossokine et al. 2020; Cotesta et al. 2020; Taracchini et al. 2014), NR surrogate (Islam et al. 2022b; Varma et al. 2019b,a), and phenomenological (Phenom) models (Husa et al. 2016; Khan et al. 2016; García-Quirós et al. 2020; Pratten et al. 2021). While the detailed construction of each model is different, they all have a set of internal parameters that can be calibrated to NR waveforms. The quality of the waveform model is therefore determined by the ansatz used and the accuracy of the calibrated parameters.

kelvin33550336@gmail.com

The LIGO-VIRGO-KAGRA (LVK) collaboration (Aasi et al. 2015a; Abbott et al. 2021a,b; Acernese et al. 2015; Akutsu et al. 2021) recently started their fourth observational run on May 26, 2023. Impressively, they are expected to double the total number of observed binary black holes (BBHs) (Abbott et al. 2020). Moreover, the improved sensitivity also implies that we expect to detect individual events with a higher signal-to-noise ratio (SNR) than ever before. This means we can resolve more features in the signal, therefore putting more stringent requirements on the accuracy of our waveform model (Pürrer & Haster 2020; Hu & Veitch 2022).

Because of the large number of calibration parameters (often a few hundred if not more), waveform models are usually calibrated separately for the inspiral, merger, and ringdown parts of the waveform (Khan et al. 2016; Santamaria et al. 2010; Pratten et al. 2021). This ignores the correlation between different parts of the waveform model and limits it quality. Recently, there has been an effort to rebuild waveform models (Khan et al. 2016) using programming languages that support automatic differentiation (AD) (Edwards et al. 2023; Iacovelli et al. 2022a,b; Coogan et al. 2022); AD is a technique used to compute machine precision derivatives of func-

tions without the issues of scaling up to high dimension or expression swelling. In particular, `ripple` (Edwards et al. 2023) exposes the calibration parameters to the user. This allows us to make use of common techniques from machine learning, such as gradient descent and back propagation (Bradbury et al. 2018; Paszke et al. 2019; Abadi et al. 2015), to improve the calibration of the waveform models.

In this paper, we investigate the possibility of further improving the accuracy of a waveform model, IMRPhenomD (Khan et al. 2016; Husa et al. 2016), by jointly optimizing all the calibration coefficients for given a set of NR waveforms. Using a similar set of NR waveforms to those used in (Khan et al. 2016; Husa et al. 2016), we demonstrate that one can improve the match between IMRPhenomD and NR waveforms over a decently sized parameter space, up to a mass ratio $q = 8$. We additionally explore how different parts of the source parameter space (e.g. the primary and secondary spins) respond to the optimization procedure by optimizing the waveform separately for different regions. This can help in understanding whether the waveform model ansatz performs equally well in different regions of the parameter space.

The rest of the paper is structured as follows: In Sec. 2, we review the parameterization of the IMRPhenomD model and the mismatch function that is used as an objective/loss function for the calibration, followed by outlining the specific optimization scheme used for recalibration. In Sec. 3, we give the optimization result by comparing mismatches of the optimized waveforms with the original waveforms. We also show how the optimization result differs as a function of the source parameters. Finally, in Sec. 4, we discuss the differences between our calibration procedure and the procedure used in (Khan et al. 2016). Additionally, we discuss some caveats (such as the difference in training waveforms) that limit our conclusions on the mismatch improvement. We also explain how the reduced spin parameterization affects the accuracy of the model. Note that throughout this paper we use the terms recalibration and optimization interchangeably.

2. OPTIMIZATION METHOD

In this section, we first briefly review the construction of the IMRPhenomD model and discuss how the calibration parameters enter the waveform. We then outline the mismatch and how it can be used as a loss function. Finally we discuss the gradient descent algorithm and our stopping criterion.

2.1. *Waveform Model*

We start by giving a succinct summary of the IMRPhenomD model and the relevant parameters. Interested readers should refer to (Khan et al. 2016) for more details.

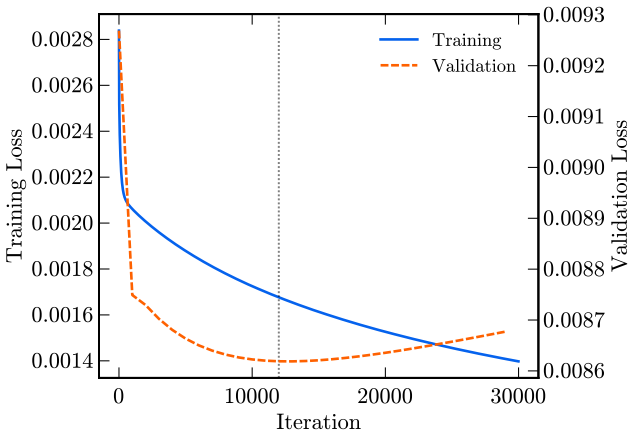

Figure 1. Average Loss $\mathcal{L}_{\rm ave}$ against the number of iterations. The vertical dotted line indicates the minimum of the validation loss, which is where we stop the optimization.

Aligned-spin, frequency-domain waveform models (such as IMRPhenomD) can be written as a combination of amplitude and phase functions (A and ϕ respectively):

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

where f is the frequency, θ are the intrinsic parameters of the binary, and Λ^i ($i = 1, 2, \cdots, 19$) is a set of 19 additional parameters which will be discussed below. The phase and amplitude functions are then split into three sections which represent the inspiral, intermediate, and merger-ringdown (MR) parts of the waveform. During inspiral, A and ϕ are known analytically from post-Newtonian (PN) theory; IMRPhenomD uses the TaylorF2 model (Buonanno et al. 2009; Arun et al. 2005) which is known up to 3.5PN order augmented with the next four higher order PN terms. To model the intermediate and MR regions, IMRPhenomD (and all waveforms in the IMRPhenom family) uses a series of parameterizations [1] which depend purely on Λ^i and can be calibrated to numerical relativity (NR) simulations. The three sections are then *stitched* together using step functions. Importantly, the parameterizations are chosen such that they can be made $\mathcal{C}^1$ continuous at the boundary between each section.

In practice the Λ^i parameters are fit for each section independently, i.e., intermediate coefficients are fit whilst ignoring the MR region. Finally, to map the grid of tuned Λ^i parameters back to the intrinsic parameter

[1] The parameterizations for both the amplitude and phase functions can be found in (Khan et al. 2016).

space, IMRPhenomD uses the polynomial function:

$$\begin{aligned}\Lambda^i = &\lambda^i_{00} + \lambda^i_{10}\eta \\ &+ (\chi_{\rm PN} - 1)(\lambda^i_{01} + \lambda^i_{11}\eta + \lambda^i_{21}\eta^2) \\ &+ (\chi_{\rm PN} - 1)^2(\lambda^i_{02} + \lambda^i_{12}\eta + \lambda^i_{22}\eta^2) \\ &+ (\chi_{\rm PN} - 1)^3(\lambda^i_{03} + \lambda^i_{13}\eta + \lambda^i_{23}\eta^2)\,, \end{aligned} \tag{2}$$

where the λ's are the fitting coefficients we are going to optimize below, η is the symmetric mass ratio, and $\chi_{\rm PN}$ is the post-Newtonian spin parameter, which is defined as

$$\chi_{\rm PN} = \frac{m_1\chi_1 + m_2\chi_2}{m_1 + m_2} - \frac{38\eta}{113}(\chi_1 + \chi_2)\,. \tag{3}$$

Here, $m_{1,2}$ and $\chi_{1,2}$ are the primary and secondary mass and spin, respectively.

Although initially independent, the stitching procedure means that each section of the waveform intrinsically depends on the full set of λ's. A slightly inaccurate set of λ's can therefore lead to inaccuracies in the generated waveforms. Thus, the calibration of these coefficients is crucial to the accuracy of IMRPhenom GW models. Importantly, since the fitting was performed on the individual segments, the final waveform is not guaranteed to have λ's close to global minima.

At the time of construction this piece-wise approach was necessary since λ has 209 components, making the fitting to NR simulations computationally prohibitive. Here, for the first time we recalibrate the λ coefficients jointly. This is made possible by the use of gradient-based optimization algorithms, enabled by AD from `jax` and `ripple`, which are significantly more efficient in high dimensions.

2.2. *Loss Function*

In order to optimize the coefficients, we need to define a loss function that quantifies the difference between the waveform model and the target NR simulations which we want to match. Here we adopt a quantity commonly used in GW physics called the *mismatch* function (Owen 1996; Husa et al. 2016). It is defined as

$$\mathcal{M}(h_1, h_2) = 1 - \max_{t_0,\phi_0}\langle \hat{h}_1, \hat{h}_2\rangle, \tag{4}$$

where $h_{1,2}$ are the two GW waveforms we are comparing, and t_0 and ϕ_0 are a relative time and phase shift respectively. The inner product, $\langle h_1, h_2\rangle$, is defined as

$$\langle h_1, h_2\rangle = 4\,\mathrm{Re}\int_{f_{\rm min}}^{f_{\rm max}} \frac{h_1(f)h_2^*(f)}{S_n(f)}\,df, \tag{5}$$

where $\hat{h} = h/\sqrt{\langle h, h\rangle}$ is the normalized GW strain, $S_n(f)$ is the power spectral density (PSD), and $f_{\rm max}$ ($f_{\rm min}$) is the maximum (minimum) frequency for the integration. We note here that the mismatch can be viewed as a mean square error (MSE) between the two waveforms.

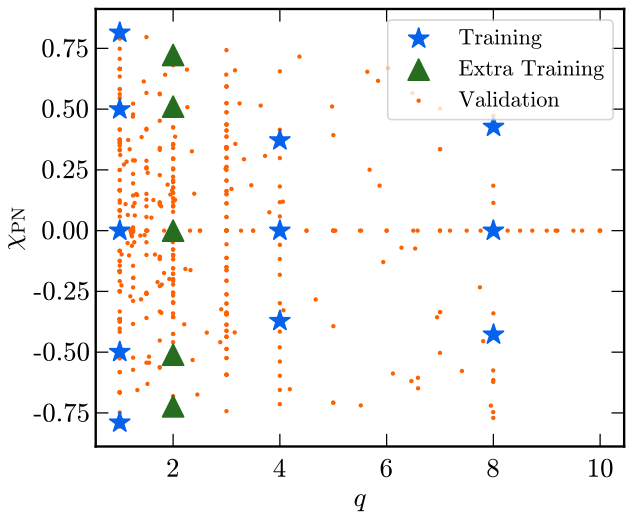

Figure 2. Distribution of training and validation NR waveforms in the $q - \chi_{\rm PN}$ space. The blue stars represent the waveforms used to compute the training loss, green triangles represent extra training data, and the orange dots represent the waveforms used for validation.

Since we wish to optimize the model over the whole parameter space, we need to compare multiple model generated waveforms with NR waveforms. However, the mismatch is only defined for two input waveforms at a particular set of intrinsic parameters. We therefore define the loss function as an average of training waveforms in two ways, the simple average of mismatches and the normalized average of mismatches,

$$\mathcal{L}_{\rm ave} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{M}_i \tag{6}$$

$$\mathcal{L}_{\rm norm} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathcal{M}_i}{\mathcal{M}_{i,\rm ini}}, \tag{7}$$

where $\mathcal{M}_i$ represents the mismatch of an individual training waveform, $\mathcal{M}_{i,\rm ini}$ represents the mismatch between original IMRPhenomD waveforms and NR waveforms, and N is the total number of training waveforms.

The two loss functions are chosen to give different behavior during the optimization process. The simple average, $\mathcal{L}_{\rm ave}$, serves as the simplest choice of loss function, but is prone to be dominated by a single point in parameter space with a large mismatch. Other points with smaller mismatches would be insignificant comparatively, and might not be able to improve under such a loss function. Alternatively, the normalized average,

$\mathcal{L}_{\rm norm}$, eliminates the aforementioned issue by encouraging the waveform to improve at each training point at a similar rate. The ratio in $\mathcal{L}_{\rm norm}$ will therefore remain approximately the same for each training point. Conversely, $\mathcal{L}_{\rm ave}$ allows the loss function to automatically adjust and preferentially optimize the largest mismatches, encouraging the waveform to have similar mismatches everywhere. In this paper, we show the results of using both loss functions and examine the differences between them.

2.3. *Optimization Scheme*

To compute the loss functions, we have to choose NR waveforms for calculating the mismatch. Originally, 19 NR simulations were used to calibrate IMRPhenomD (Khan et al. 2016); nine of these are from the SXS catalog (Boyle et al. 2019) and 10 BAM simulations. [2] As BAM waveforms are not publicly available, we cannot use a training set identical to the original work. Instead, we take the same nine waveforms from the SXS catalog plus two additional waveforms which closely mimic two of the low mass ratio BAM simulations. The remaining BAM simulations have no close SXS counterparts, and are therefore not included in the training set. Our main results therefore uses 11 NR waveforms for calibration which are listed in Tab. 1. Additional NR waveforms which are used for further calibration are listed in Tab. 2.

The training set has a maximum mass ratio of eight due to the lack of high mass ratio simulations in the SXS catalog. In fact, the SXS catalog only has NR waveforms with $q \leq 10$. Nevertheless, we are interested in the behavior of the IMRPhenomD model with small q, as most BBH events observed by LIGO and Virgo have $q \leq 8$.

The SXS NR waveforms are given as time-series strain. Since IMRPhenomD is modeled in the frequency domain, we need to Fourier transform the simulation results in order to compute the mismatch, (4). For this, we taper the time-series using Tukey window (Usman et al. 2016) [3] before using standard FFT routines to compute the fast Fourier transform.

In addition to choosing the NR waveforms for the training set, one needs to choose a relevant PSD for the mismatch. We have opted to use a flat PSD for the mismatch calculation, as it provides results that are independent of the detector sensitivity and mass scale. The use of a flat PSD ensures that the improvement in accuracy is due mainly to the difference in high-dimensional fitting. Additionally, we are interested in examining the effect of introducing a detector PSD on the optimization process. For this, we have chosen the zero-detuned high-power (`zdethp`) noise PSD (Aasi et al. 2015b). Since the total mass of the system scales with the frequency of the waveform, we must choose a corresponding mass scale to match the frequency range of our noise PSD. To demonstrate the effect of introducing a detector specific PSD, we selected an arbitrary mass scale of $M = 50M_{\odot}$, as binaries of this mass are commonly observed by the LIGO and Virgo detectors (Abbott et al. 2019, 2021c,d,e).

We point out that our treatment of NR waveforms is different from that of (Husa et al. 2016; Khan et al. 2016). In the original calibration process, the training waveforms are hybrids of NR and SpinAlignedEOB (SEOB) waveforms. The low frequency inspiral part is taken from the SEOB waveforms while the rest is taken from NR simulations. Instead, we solely use NR waveforms for calibration since most NR waveforms used (for both training and validation) have long enough time series data, i.e. > 15 orbits (Boyle et al. 2019), to contain part of the inspiral segment and all merger and ringdown frequency information. We use the frequency limits $f_{\rm min} = 0.1 f_{\rm RD}$ and $f_{\rm max} = 1.2 f_{\rm RD}$, where $f_{\rm RD}$ is the frequency at ringdown. This range covers most of the IMRPhenomD's frequency range, except the minimum frequency is set slightly higher than in the original calibration due to the NR simulation length. When compared with IMRPhenomC, the frequency range is slightly extended to have a higher maximum frequency (Santamaria et al. 2010). We use the dimensionless frequency spacing $M\Delta f = 2.5 \times 10^{-6}$, which is sufficient to capture all features of the GW strain.

With the loss function evaluated, we apply gradient descent to optimize the tunable coefficients as shown in Algorithm 1. We take λ_i to be the original coefficients given in (Khan et al. 2016) because they likely lie in the neighborhood of the minimum that we wish to find. We fix our learning rate, α, to be 10^{-6}, which is small enough to ensure we don't move far from the minimum. Finally, we stop the optimization when the validation loss stops decreasing (see Sec. 3 for description of the validation set). This can be seen in Fig. 1 at around 12000 iterations.

3. RESULT AND COMPARISON WITH ORIGINAL MODEL

To evaluate how well the optimization procedure generalizes to waveforms that are not in the training set, we evaluate the mismatch between the fine-tuned model

[2] We took the CoM-corrected waveforms with 4th-order extrapolation.

[3] Specifically, we choose $\alpha = 2t_{\rm RD}/T$, where $t_{\rm RD}$ is the duration of ringdown (maximum amplitude to the end of the strain) and T is the duration of the entire GW strain.

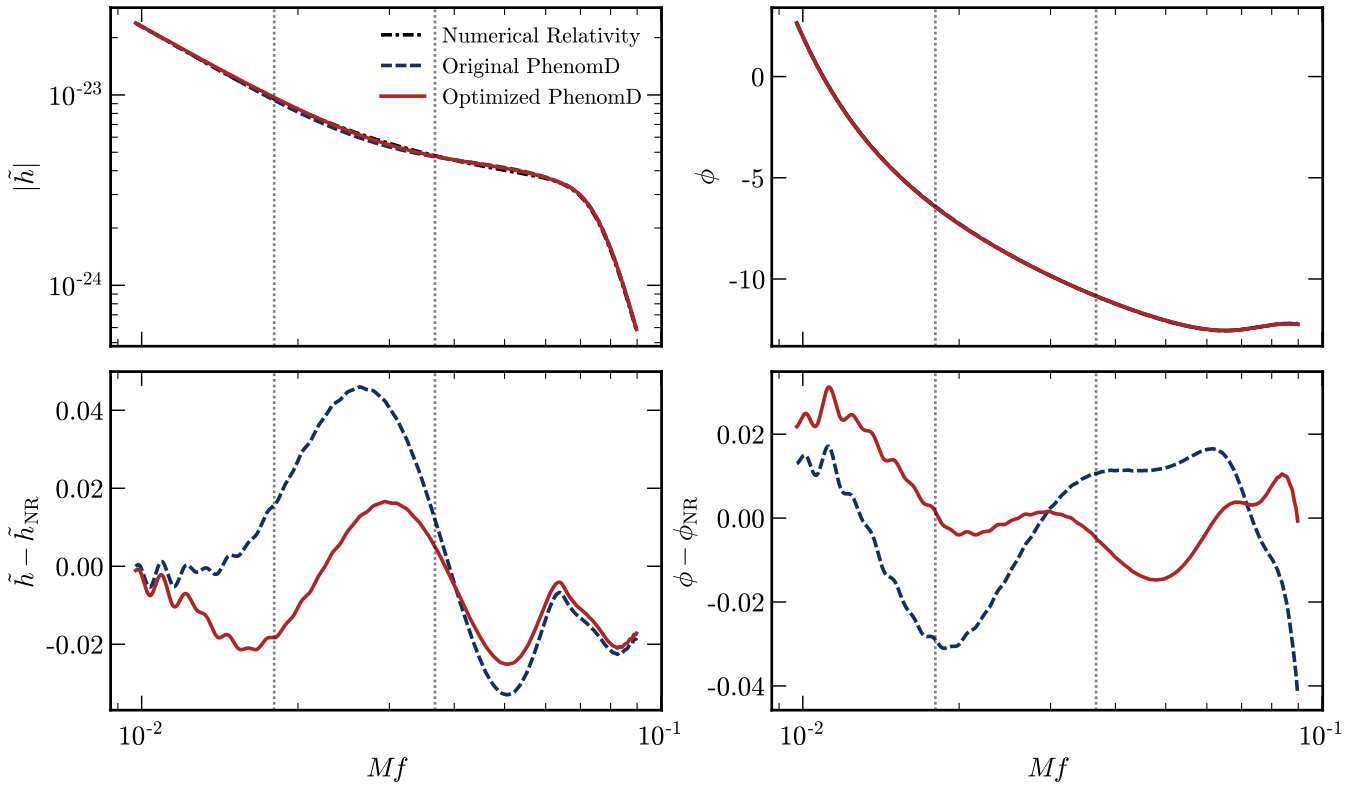

Figure 3. Comparison between original and optimized IMRPhenomD waveforms. Here we show the SXS:BBH:0154 NR waveform, which has a mass ratio of $q = 1$ and spins $\chi_1 = \chi_2 = -0.8$. The original mismatch is 2.8×10^{-4} and the optimized mismatch is 5.3×10^{-5}. *Top*: Here we show the amplitude (left) and phase (right) of the NR, original IMRPhenomD, and optimized IMRPhenomD waveforms. *Bottom*: Here we show the relative error between the NR and IMRPhenomD waveform ampltiudes (left) as well as the absolute error of the phases between the NR and IMRPhenomD waveforms (right).

Algorithm 1: Gradient descent pseudocode

Input: initial coefficients λ_i
Parameters: number of iterations M, learning rate α
Variables: current coefficients λ, mismatch gradient $\nabla\mathcal{L}$
Result: output coefficients λ

$\lambda \leftarrow \lambda_i$
`/* Gradient Descent */`
for $i < M$ **do**
  $\mathcal{L} \leftarrow Mismatch(\lambda)$
  $\nabla\mathcal{L} \leftarrow AutoDiff(\mathcal{L})$
  $\lambda \leftarrow \lambda - \alpha\nabla\mathcal{L}$
return λ

and an additional 526 NR waveforms in the SXS catalog i.e., the validation set. We select waveforms that share the same part of the parameter space with the training set, i.e., waveforms with negligible eccentricity ($e < 2 \times 10^{-3}$) and precession ($\chi_{x,y} < 5 \times 10^{-3}$). Figure 2 shows how the training and validation waveforms are distributed in the $q - \chi_{\rm PN}$ space.

To illustrate the effect of optimization on an individual waveform level, in Fig. 3 we plot the phase and am-

Code	q	χ_1	χ_2
SXS:BBH:0156	1.0	-0.95	-0.95
SXS:BBH:0151	1.0	-0.60	-0.60
SXS:BBH:0001	1.0	0.00	0.00
SXS:BBH:0152	1.0	0.60	0.60
SXS:BBH:0172	1.0	0.98	0.98
SXS:BBH:1418	4.0	-0.40	-0.50
SXS:BBH:0167	4.0	0.00	0.00
SXS:BBH:1417	4.0	0.40	0.50
SXS:BBH:0064	8.0	-0.50	-0.46
SXS:BBH:0063	8.0	0.00	0.00
SXS:BBH:0065	8.0	0.50	0.46

Table 1. List of NR waveforms used to recalibrate the model. The mass ratio is defined as $q = m_1/m_2 \geq 1$ and the spins are denoted by $\chi_{1,2}$. Out of the 11 waveforms listed here, two SXS waveforms are analogues to two of the BAM NR simulations used in the original calibration (SXS:BBH:1417 as A7 and SXS:BBH:1418 as A9 in (Khan et al. 2016)). The rest of the SXS waveforms are also used in the original IMRPhenomD calibration.

Code	q	χ_1	χ_2
SXS:BBH:0234	2.0	-0.85	-0.85
SXS:BBH:0235	2.0	-0.60	-0.60
SXS:BBH:0169	2.0	0.00	0.00
SXS:BBH:0256	2.0	0.60	0.60
SXS:BBH:0257	2.0	0.85	0.85

Table 2. Additional NR waveforms used in further recalibration.

plitude of a particular waveform before and after optimization together with the NR waveform taken directly from the SXS catalog. In the bottom panel one can see that, compared to the original IMRPhenomD waveform, the optimized waveform has smaller residuals both in amplitude and phase, particularly in the inspiral region where the amplitude displays a 50% reduction in error. For a fair comparison, we selected the SXS:BBH:0154 NR waveform which was also used in (Khan et al. 2016) to validate the original waveform.

With the purpose of improving downstream tasks such as parameter estimation in mind, the more relevant metric of improvement is the distribution of improvement in mismatch over the entire validation dataset. Figure 4 shows the distribution of log-mismatches for the validation waveforms before and after the optimization procedure. Here we show results using a constant PSD in our loss function. One can see the distribution of the optimized waveform is skewed toward lower mismatches, with the peak of the distribution being shifted by approximately an order of magnitude. The median mismatch is reduced by 50% (see vertical dotted lines) while the tail hardly changes. When using $\mathcal{L}_{\rm norm}$, we observe a less pronounced improvement with a 22.9% decrease in the median.

Note that the performance of the IMRPhenomD model was initially tested using the `zdethp` weighted mismatch.[4] We would therefore like to examine whether using the `zdethp` PSD in our loss function could lead to an improved mismatch. Performing the optimization again, we find no significant difference between the results using the two PSDs in the distribution of mismatches.

To understand whether additional training data can further improve the performance of the model, we include waveforms that are not present in the original IMRPhenomD calibration in our training dataset; the

[4] Note, however, that the mismatch was never directly used during the calibration process (Khan et al. 2016, 2019).

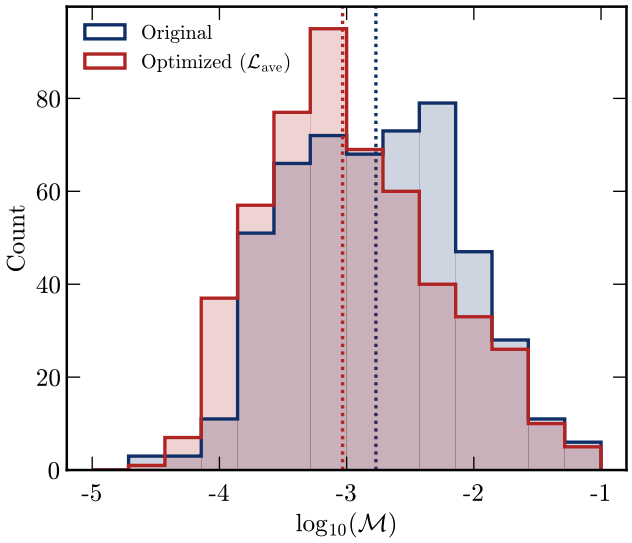

Figure 4. Distributions of mismatches before and after optimization using the $\mathcal{L}_{\rm ave}$ loss function. Mismatches are calculated using the training waveforms listed in Tab. 1 and are weighted with a constant PSD. The dotted lines represent the median of the distributions, which decreased by 45.3% during optimization.

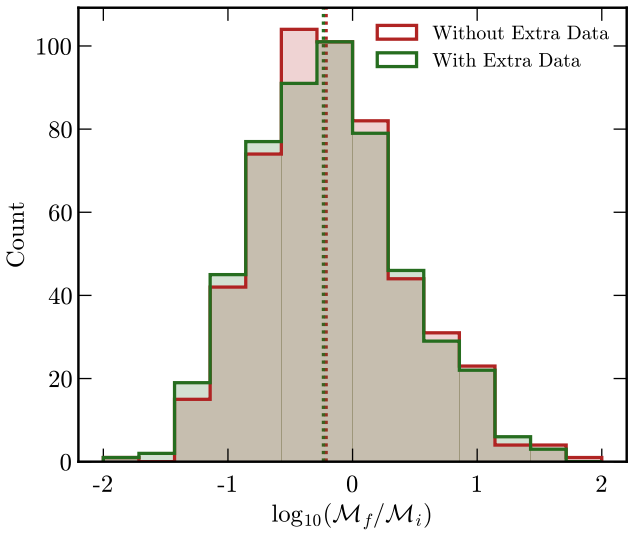

Figure 5. Distributions of $\log_{10}$ difference in mismatch. The red distribution uses the training waveforms listed in Tab. 1 while the green distribution uses waveforms listed in Tab. 1 and 2. Mismatches are calculated using a constant PSD with the loss function $\mathcal{L}_{\rm ave}$. The dotted lines represent the median of the distributions, which decreased by an additional 10.8% during optimization with the additional data.

parameters can be found in Tab. 2. We specifically choose to use $q = 2$ events since we have abundant $q = 2$ NR waveforms to validate the final result. The new set of coefficients generated from this optimization process

yields only marginal improvements in the newly produced waveforms, as seen in Fig. 5. The high mismatch tail of the newly optimized distribution remains comparable to the distribution from the first optimization, indicating that the original dataset is sufficient for this task. Similarly, utilizing the `zdethp` PSD in the loss function together with additional waveforms results in a similar level of improvement.

To investigate the performance of recalibration over the source parameter space, we plot the improvement of the log-mismatch as a function of the parameter space $q-\chi_{\rm PN}$ in Fig. 6. Red points indicate that the waveform is improved by the optimization procedure, while blue points indicate that the waveform mismatch *increases* during optimization. We can see that waveforms with $q \leq 4$ show the most consistent average improvement. This is likely due to the better coverage of training waveforms in that part of the parameter space (see Fig. 2). On the spin axis, we can see that the waveforms with $\chi_{\rm PN} \sim 0$ show the most consistent improvement. When we move away from $\chi_{\rm PN} \sim 0$, the improvement fluctuates but exhibits an overall trend. This is particularly true in the $q \leq 4$ region, where we see a consistent improvement of the waveform for $\chi_{\rm PN} < 0$. We also plot the parameter space $\chi_1 - \chi_2$ in Fig. 7. Points along the diagonal axis, $\chi_1 \sim \chi_2$, show good mismatch improvements as discussed above. Meanwhile, the top-left and bottom-right regions respond to the optimization differently. In the top-left region, the waveform generally improves after optimization. However, in the bottom-right region, the waveform does not improve after optimization.

Given that the waveform model's ansatz may not be entirely compatible with NR, and the optimization procedure is carried out over a distribution of waveforms with varying source parameters, it is conceivable that different parts of the source parameter space may not share the same set of optimal IMRPhenomD parameters. This would mean that there are trade-offs in accuracy between different parts of the parameter space. If this is the cause of the lack of improvement in the high mismatch tail of the distribution, segmenting the parameter space into smaller subspaces should alleviate this problem. On the other hand, if the ansatz lacks the correct parameterized form to capture the NR waveforms' behavior as a function of the source parameters, the results will always be biased, and we should not expect any improvement, even if we segment the parameter space during training.

We divided the parameter space into four regions to analyze the effect of the recalibration procedure on each region separately (Fig. 8 and 9). The training waveforms

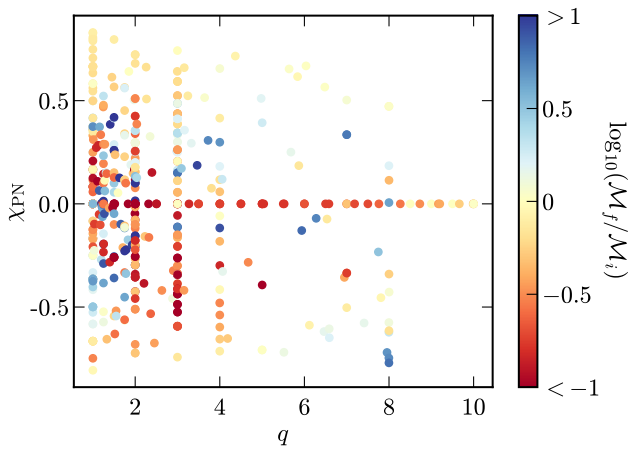

Figure 6. Fractional mismatch change of the validation waveforms in the $q-\chi_{\rm PN}$ plane. We show results for the $\mathcal{L}_{\rm ave}$ loss function with a constant PSD and training waveforms in Tab. 1. Here, the colorbar represents the $\log_{10}$ difference between optimized and original mismatches. Red points indicate that the waveform is improved by the optimization procedure, while blue points indicate that the waveform mismatch *increases* during optimization.

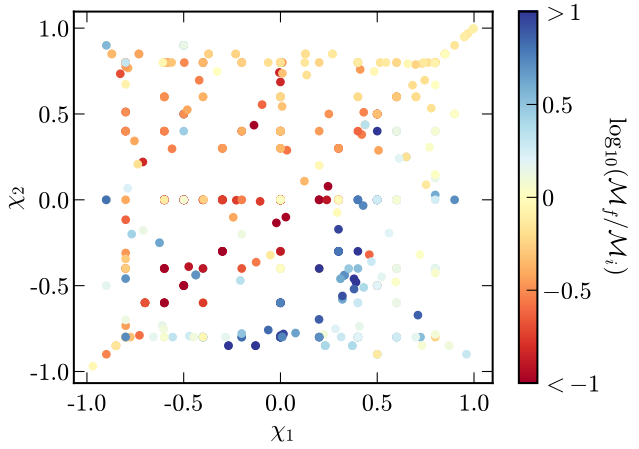

Figure 7. Fractional mismatch change of the validation waveforms in the $\chi_1-\chi_2$ plane. We show results for the $\mathcal{L}_{\rm ave}$ loss function with a constant PSD and training waveforms listed in Tab. 1.

used for fitting in this scenario are listed in Tab. 3 and the loss functions are calculated using a simple average of the mismatches, $\mathcal{L}_{\rm ave}$. The top-left and bottom-right regions have limited data for $q > 4$, hence the result is only valid up to $q \leq 4$ and we only use waveforms with $q \leq 4$ as validation waveforms for these two regions. From Fig. 8, we observe that both the top-right and bottom-left regions improve significantly over the original model. This is especially pronounced for the bottom-left region, where the improvement is significantly better than optimizing all regions simultaneously.

Code	q	χ_1	χ_2
SXS:BBH:0172	1.0	0.98	0.98
SXS:BBH:0152	1.0	0.60	0.60
SXS:BBH:0001	1.0	0.00	0.00
SXS:BBH:1417	4.0	0.40	0.50
SXS:BBH:0167	4.0	0.00	0.00
SXS:BBH:1426	8.0	0.48	0.75
SXS:BBH:0063	8.0	0.00	0.00
SXS:BBH:0370	1.0	-0.20	0.40
SXS:BBH:2092	1.0	-0.50	0.50
SXS:BBH:0330	1.0	-0.80	0.80
SXS:BBH:2116	2.0	-0.30	0.30
SXS:BBH:2111	2.0	-0.60	0.60
SXS:BBH:0335	2.0	-0.80	0.80
SXS:BBH:0263	3.0	-0.60	0.60
SXS:BBH:2133	3.0	-0.73	0.85
SXS:BBH:0263	4.0	-0.80	0.80
SXS:BBH:0156	1.0	-0.95	-0.95
SXS:BBH:0151	1.0	-0.60	-0.60
SXS:BBH:0001	1.0	0.00	0.00
SXS:BBH:1418	4.0	-0.40	-0.50
SXS:BBH:0167	4.0	0.00	0.00
SXS:BBH:1419	8.0	-0.80	-0.80
SXS:BBH:0063	8.0	0.00	0.00
SXS:BBH:0304	1.0	0.50	-0.50
SXS:BBH:0327	1.0	0.80	-0.80
SXS:BBH:2123	2.0	0.30	-0.30
SXS:BBH:2128	2.0	0.60	-0.60
SXS:BBH:2132	2.0	0.87	-0.85
SXS:BBH:2153	3.0	0.30	-0.30
SXS:BBH:0045	3.0	0.50	-0.50
SXS:BBH:0292	3.0	0.73	-0.85

Table 3. List of NR waveforms used in recalibrating the coefficients in the four $\chi_1 - \chi_2$ regions. From top to bottom the lines denote the top-right ($\chi_1, \chi_2 > 0$), top-left ($\chi_1 < 0 < \chi_2$), bottom-left ($\chi_1, \chi_2 < 0$), and bottom-right ($\chi_1 > 0 > \chi_2$) regions respectively. Note that for the top-right and bottom-left regions, waveforms are chosen to have $\chi_1 \approx \chi_2$, while the training waveforms for the other two regions are chosen to have $\chi_1 \approx -\chi_2$.

This suggests the ansatz fits this part of the parameter space well. The top-left region also improves over the original model although it is similar to when optimizing all waveforms at once. On the other hand, the bottom-right region does not improve over the original model, indicating that a change to the ansatz is required to fit the NR data better.

Overall, the improvement for both optimization schemes are similar (Fig. 9). Although split region optimization uses more training waveforms, some of the waveforms, e.g. opposite spins, hinders the optimization procedure and hence gives a slightly worse result when comparing with all region optimization.

4. DISCUSSION AND CONCLUSION

The results in this work show a promising way to understand and improve the accuracy of phenomenological waveform models by jointly optimizing all coefficients at once. However, there are a number of caveats as well as multiple ways to improve upon the current work.

While this study focuses on the IMRPhenomD model due to its availability as an automatically differentiable waveform in `ripple`, other more sophisticated waveforms such as IMRPhenomP (Hannam et al. 2014; Khan et al. 2019) and IMRPhenomXAS (Pratten et al. 2020, 2021) can also potentially benefit from recalibration. In fact, since these waveforms have a larger number of calibration parameters and are fit to a larger number of NR simulations, it is possible that the improvement will be more dramatic than for IMRPhenomD. `ripple` is being regularly updated with new waveforms, and future work should carefully examine whether these more modern waveforms can be improved.

The correlation we observe between the mismatches and the waveform's parameters can also be used to design more physical ansatz. For example, we see the recalibration process struggles to improve in the regime where χ_1 is positive and χ_2 is negative. This means there is a dependency between the two spins that is not captured in the current ansatz. In fact, this issue can be explained by PN theory, where the leading spin-orbit coupling term depends on the effective spin while higher order terms depends on both χ_1 and χ_2. Nevertheless, we hope to show a systematic scheme that could reveal hidden degeneracies, as illustrated by the effective spin degeneracy. Hence, such systematic scheme that incorporates our calibration method with waveform design will be handy in constructing new waveform models in the future.

We note that the parameter space we use is restricted to $q \leq 8$, while waveforms of $q = 18$ are used in the original calibration trial. Also, we have directly used NR waveforms instead of hybridized SEOB-NR waveforms for training and validation purposes. These choices may put the original IMRPhenomD result at a disadvantage when assessing accuracy, as both the parameter space and frequency range are restricted. We tested our result by training the model using a set of training waveforms of the same intrinsic parameters, but with BAM wave-

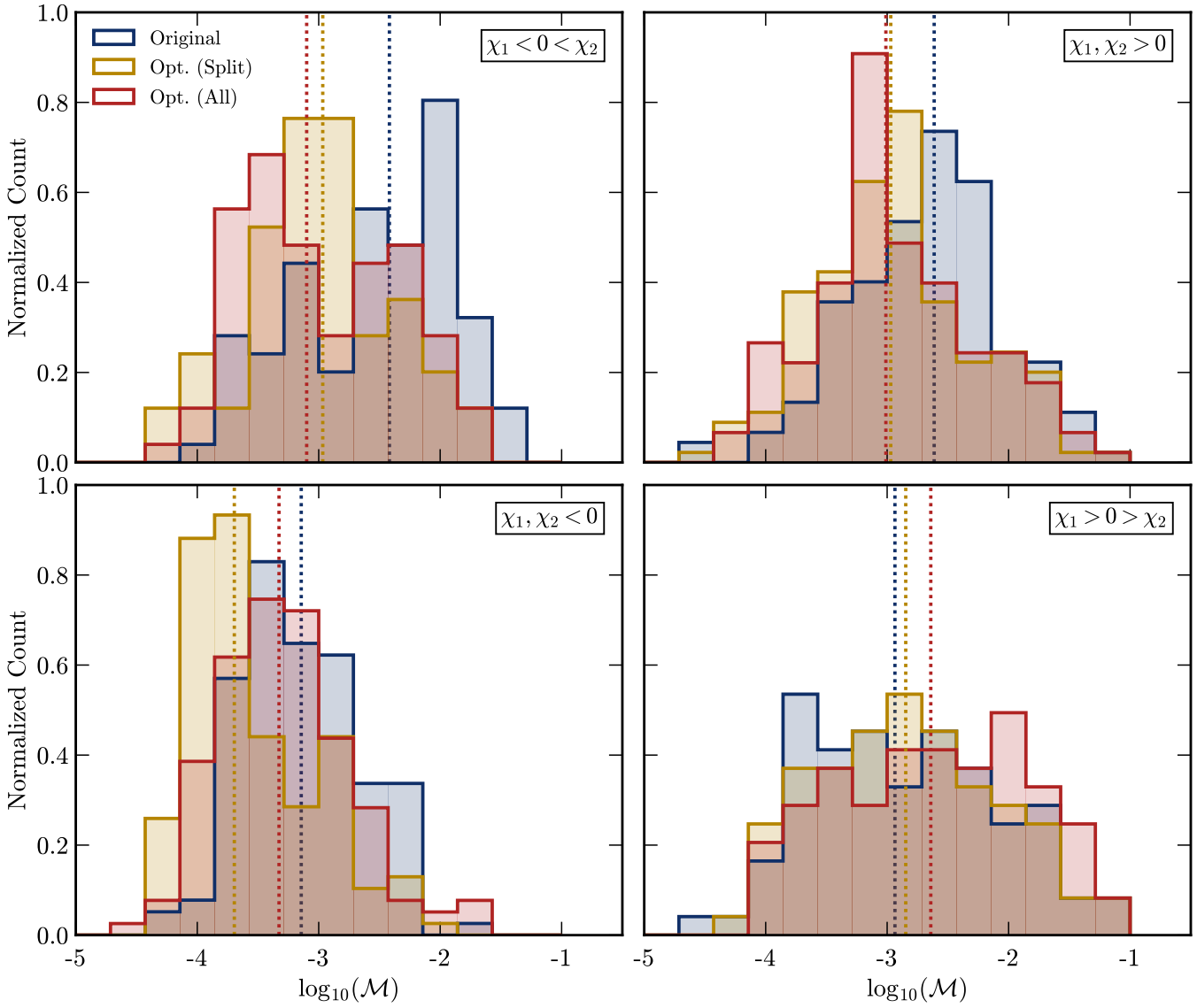

Figure 8. Distributions of mismatches for both split region optimization and all region optimization in all 4 regions. We use a constant PSD to calculate the mismatch and $\mathcal{L}_{\rm ave}$ as the loss function. The medians of each distribution are indicated by dotted lines.

forms replaced with SEOBNR waveforms. We can see the result in Fig. 10, where the mismatch median decreases by about 8%. Nevertheless, it is known that IMRPhenomD has a large inaccuracy in the large mass-ratio regime, and is rarely used in this region of parameter space. Hence, we focus on showing that AD with gradient descent, as a general and robust method, can be used to further optimize GW waveform models with arbitrary sets of training waveforms used. With more publicly available data in the future, more systematic studies should be made to precisely assess the improvement one can get out of further optimization.

In making Fig. 8 and 9, we split the parameter space into four regions purely for simplicity. These simple cuts demonstrate the accuracy of the waveform can be further improved, but they are almost certainly not the optimal way to incorporate the extra information we have about the waveform. A more general way such as adding new functional dependency between the parameters should be explored in the future.

In this work, we use the mismatch to quantify the accuracy of the waveform model. A natural extension of this work is to investigate how the recalibration process affects downstream analyses such as parameter estimation and population modeling. Differentiable samplers such as (Wong et al. 2023), which have been recently introduced to the community, potentially allow one to optimize the waveform directly using metrics from pa-

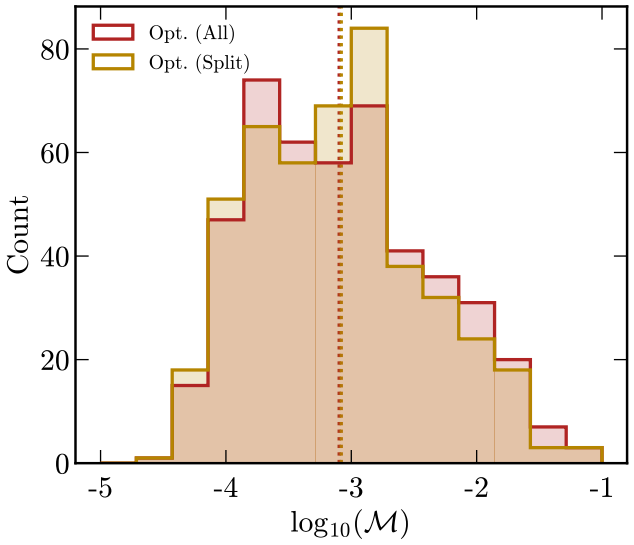

Figure 9. Combined mismatch distributions for both optimization schemes shown in Fig. 8. The medians of both distributions are indicated by dotted lines.

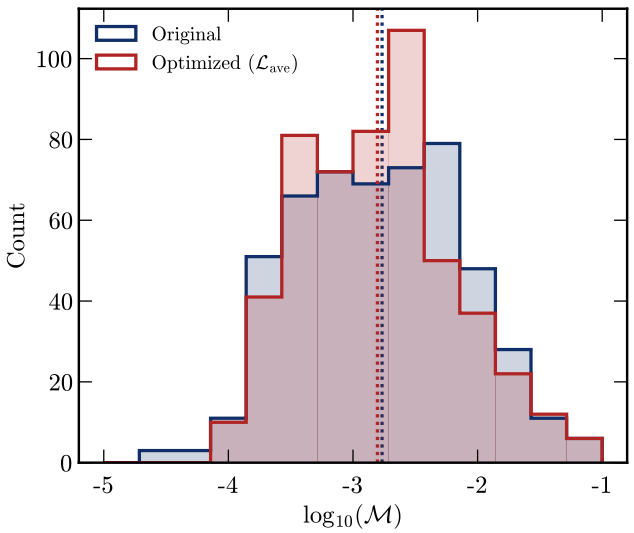

Figure 10. Distributions of mismatches before and after optimization using the $\mathcal{L}_{\rm ave}$ loss function. Mismatches are calculated using the training waveforms listed in Khan et al. (2016), but BAM waveforms are replaced by SEOBNRv4 waveforms of the same intrinsic parameters. Mismatches are calculated with a constant PSD. The dotted lines represent the median of the distributions, which decreased by 8.2% during optimization.

rameter estimation. For example, minimizing the bias in an injection-recovery run could be used as a loss function. Overall, this approach could help reduce systematic waveform error in parameter estimation simply through recalibrated waveforms. We plan to investigate these avenues in the future.

The results presented in Fig. 5 indicate that increasing the number of training NR waveforms used in the waveform optimization yields only a marginal increase in accuracy. This observation suggests that the parameterized ansatz employed in IMRPhenomD struggles to capture the full complexity of the NR waveforms. However, as the accuracy requirements of waveform models increases with more sensitive detectors, more NR waveforms will be required to train more flexible Phenom models. Accurate NR simulations must therefore be developed in parallel with waveform models to ensure we meet future detector accuracy requirements.

Overall, the development of accurate waveform models is crucial for the success of GW astronomy. In this work, we have explored how modern computational tools (automatic differentiation and gradient descent) can help with this task. Our method is general and can be applied to any waveform model that is differentiable. We therefore encourage the waveform development community to utilize these tools and hope that this work can effectively contribute to the development of accurate waveform models.

5. ACKNOWLEDGMENTS

We thank Will M. Farr, Max Isi, and Mark Hannam for helpful discussions; we also thank Carl-Johan Haster, Neil J. Cornish and Thomas Dent for comments on the draft. The Flatiron Institute is a division of the Simons Foundation. T.E. is supported by the Horizon Postdoctoral Fellowship.

REFERENCES

Aasi, J., et al. 2015a, Class. Quant. Grav., 32, 074001, doi: 10.1088/0264-9381/32/7/074001

Aasi, J., Abbott, B., Abbott, R., et al. 2015b, Classical and quantum gravity, 32, 074001

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/

Abbott, B., Abbott, R., Abbott, T., et al. 2019, Physical Review X, 9, 031040

Abbott, B. P., Abbott, R., Abbott, T., et al. 2020, Living reviews in relativity, 23, 1

Abbott, R., et al. 2021a. https://arxiv.org/abs/2108.01045

—. 2021b. https://arxiv.org/abs/2111.03606

Abbott, R., Abbott, T., Abraham, S., et al. 2021c, Physical Review X, 11, 021053

Abbott, R., Abbott, T., Acernese, F., et al. 2021d, arXiv preprint arXiv:2108.01045

—. 2021e, arXiv preprint arXiv:2111.03606

Acernese, F., et al. 2015, Class. Quant. Grav., 32, 024001, doi: 10.1088/0264-9381/32/2/024001

Akutsu, T., et al. 2021, PTEP, 2021, 05A101, doi: 10.1093/ptep/ptaa125

Arun, K. G., Iyer, B. R., Sathyaprakash, B. S., & Sundararajan, P. A. 2005, Phys. Rev. D, 71, 084008, doi: 10.1103/PhysRevD.71.084008

Ashton, G., et al. 2019, Astrophys. J. Suppl., 241, 27, doi: 10.3847/1538-4365/ab06fc

Boyle, M., Hemberger, D., Iozzo, D. A., et al. 2019, Classical and Quantum Gravity, 36, 195006

Bradbury, J., Frostig, R., Hawkins, P., et al. 2018, JAX: composable transformations of Python+NumPy programs, 0.2.5. http://github.com/google/jax

Buonanno, A., Iyer, B., Ochsner, E., Pan, Y., & Sathyaprakash, B. S. 2009, Phys. Rev. D, 80, 084043, doi: 10.1103/PhysRevD.80.084043

Christensen, N., & Meyer, R. 2022, Rev. Mod. Phys., 94, 025001, doi: 10.1103/RevModPhys.94.025001

Coogan, A., Edwards, T. D. P., Chia, H. S., et al. 2022, Phys. Rev. D, 106, 122001, doi: 10.1103/PhysRevD.106.122001

Cotesta, R., Marsat, S., & Pürrer, M. 2020, Physical Review D, 101, 124040

Dax, M., Green, S. R., Gair, J., et al. 2021, Phys. Rev. Lett., 127, 241103, doi: 10.1103/PhysRevLett.127.241103

Edwards, T. D. P., Wong, K. W. K., Lam, K. K. H., et al. 2023, RIPPLE: Differentiable and Hardware-Accelerated Waveforms for Gravitational Wave Data Analysis. https://github.com/tedwards2412/ripple

García-Quirós, C., Colleoni, M., Husa, S., et al. 2020, Phys. Rev. D, 102, 064002, doi: 10.1103/PhysRevD.102.064002

Hannam, M., Schmidt, P., Bohé, A., et al. 2014, Physical review letters, 113, 151101

Hu, Q., & Veitch, J. 2022, Physical Review D, 106, 044042

Husa, S., Khan, S., Hannam, M., et al. 2016, Physical Review D, 93, 044006

Iacovelli, F., Mancarella, M., Foffa, S., & Maggiore, M. 2022a, Astrophys. J., 941, 208, doi: 10.3847/1538-4357/ac9cd4

—. 2022b, Astrophys. J. Supp., 263, 2, doi: 10.3847/1538-4365/ac9129

Islam, T., Roulet, J., & Venumadhav, T. 2022a. https://arxiv.org/abs/2210.16278

Islam, T., Field, S. E., Hughes, S. A., et al. 2022b, Physical Review D, 106, 104025

Khan, S., Chatziioannou, K., Hannam, M., & Ohme, F. 2019, Physical Review D, 100, 024059

Khan, S., Husa, S., Hannam, M., et al. 2016, Physical Review D, 93, 044007

Koposov, S., Speagle, J., Barbary, K., et al. 2022, joshspeagle/dynesty: v2.0.3, v2.0.3, Zenodo, doi: 10.5281/zenodo.7388523

Ossokine, S., Buonanno, A., Marsat, S., et al. 2020, Physical Review D, 102, 044055

Owen, B. J. 1996, Physical Review D, 53, 6749

Owen, B. J., & Sathyaprakash, B. S. 1999, Physical Review D, 60, 022002

Paszke, A., Gross, S., Massa, F., et al. 2019, in Advances in Neural Information Processing Systems 32, ed. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Curran Associates, Inc.), 8024–8035. http://arxiv.org/abs/1912.01703

Pratten, G., Husa, S., Garcia-Quiros, C., et al. 2020, Physical Review D, 102, 064001

Pratten, G., García-Quirós, C., Colleoni, M., et al. 2021, Physical Review D, 103, 104056

Pürrer, M., & Haster, C.-J. 2020, Physical Review Research, 2, 023151

Romero-Shaw, I. M., et al. 2020, Mon. Not. Roy. Astron. Soc., 499, 3295, doi: 10.1093/mnras/staa2850

Santamaria, L., Ohme, F., Ajith, P., et al. 2010, Physical Review D, 82, 064016

Taracchini, A., Buonanno, A., Pan, Y., et al. 2014, Physical Review D, 89, 061502

Usman, S. A., Nitz, A. H., Harry, I. W., et al. 2016, Classical and Quantum Gravity, 33, 215004

Varma, V., Field, S. E., Scheel, M. A., et al. 2019a, Physical Review Research, 1, 033015

—. 2019b, Physical Review D, 99, 064045

Wong, K. W. k., Gabrié, M., & Foreman-Mackey, D. 2023, J. Open Source Softw., 8, 5021, doi: 10.21105/joss.05021

Zackay, B., Dai, L., & Venumadhav, T. 2018, arXiv preprint arXiv:1806.08792