Draft version June 23, 2023
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 recalibration, 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 match filtering (Owen 1996a; Owen & Sathyaprakash 1999) and parameter estimation (Dax et al. 2021; Islam et al. 2022a; 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.

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). Morover, 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 functions 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

kelvin33550336@gmail.com

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 second 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. In Sec. 4, we discuss the difference between our calibration procedure and the procedure used in (Khan et al. 2016). We also explain how the reduced spin parameterization affects the accuracy of the model.

## 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 fucntion. 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.

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

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

where $f$ is the frequnecy, $\theta$ are the intrinsic parameters of the binary, and $\Lambda$ is a set of 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 $\phi$ 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. 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 $\Lambda$ and can be calibrated to numerical relativity (NR) simulations. The three sections are then *stictched* 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 $\Lambda$ 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 $\Lambda$ parameters back to the intrinsic parameter 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 $\lambda$'s are the fitting coefficients we are going to optimize below, $\eta$ 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 $\lambda$'s. A slightly inaccurate set of $\lambda$'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 $\lambda$'s close to global minima.

At the time of construction this piece-wise approach was necessary since $\lambda$ has 209 components, making the fitting to NR simulations computationally prohibitive. Here, for the first time we recalibrate the $\lambda$ coefficients

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

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 1996b; 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 $\phi_0$ are a relative time and phase shift respectively. The inner product, $\langle h_1, h_2 \rangle$, is defined as

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

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

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,

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

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

where $\mathcal{M}_i$ represents the mismatch of an individual training waveform, $\mathcal{M}_{i,\mathrm{ini}}$ represents the initial mismatch of the individual training waveform, and $N$ is the total number of training waveforms.

The two loss functions are chosen to give different behaviour during the optimization process. The simple average, $\mathcal{L}_{\mathrm{mean}}$, 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, $\mathcal{L}_{\mathrm{norm}}$, eliminates the aforementioned issue by encouraging the waveform to improve at each training point at a similar rate. The ratio in $\mathcal{L}_{\mathrm{norm}}$ will therefore remain approximately the same for each training point. Conversely, $\mathcal{L}_{\mathrm{mean}}$ allows the loss function to automatically adjust and preferencially 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); 9 of these are from the SXS catalog (Boyle et al. 2019) and 10 BAM simulations. 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 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 (Harris 1978)[2] before using standard FFT routines to compute the Fourier transform.

In addition to choosing the NR waveforms for the training set, one needs to choose a relevant PSD for 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 accu-

[2] Specifically, we choose $\alpha = 2t_{\mathrm{RD}}/T$, where $t_{\mathrm{RD}}$ is the duration of ringdown and $T$ is the duration of the entire GW strain. [TE: What is $t_{\mathrm{RD}}$? and how is it computed? ]

racy 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 testing) 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 $\lambda_i$ 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, $\alpha$, 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 and an additional 536 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 optimiziation on an individual waveform level, in Fig. 3 we plot the phase and amplitude 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 met-

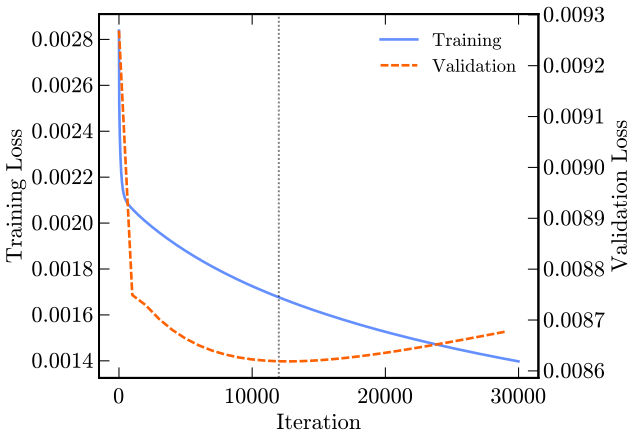


**Figure 1.** Loss functions against the number of iterations. The vertical dotted line indicates minimum of the validation loss, where we stop the optimization and obtain a new set of waveform coefficients, $\lambda$.

**Algorithm 1:** Gradient descent pseudocode

**Input:** initial coefficients $\lambda_i$
**Parameters:** number of iterations $N$, learning rate $\alpha$
**Variables:** current coefficients $\lambda$, mismatch gradient $\nabla\mathcal{L}$
**Result:** output coefficients $\lambda$

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

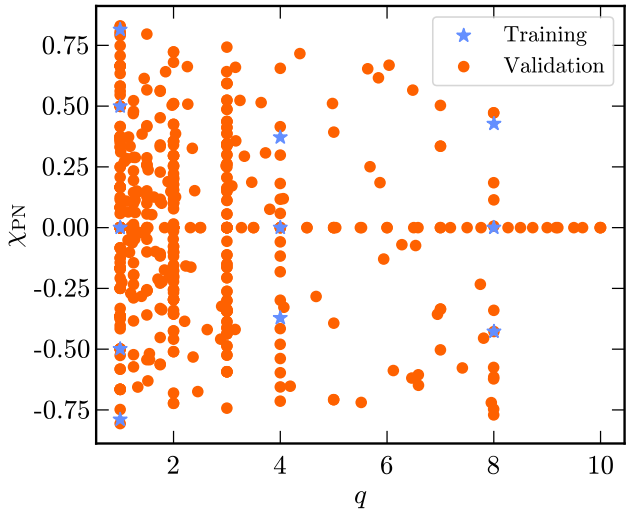


**Figure 2.** Distribution of training and validation NR waveforms in the $q - \chi_{PN}$ space. The blue stars represent the waveforms used to compute the training loss and the orange dots represent the waveforms used for validation.[KW: Change the legend, and include the extra points. Change the training point to be star and testing point to be dot. ]

| Code | $q$ | $\chi_1$ | $\chi_2$ |
|---|---|---|---|
| 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 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, nine of them are also used in the original IMRPhenomD calibration. The remaining two are SXS analogues to two of the BAM NR simulations used in the original calibration (Khan et al. 2016).

ric 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,

| Code | $q$ | $\chi_1$ | $\chi_2$ |
|---|---|---|---|
| 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.

with the peak of the distribution being shifted by approximately an order of magnitude, and the median mismatch being reduced by 50% (see vertical dotted lines). When using $\mathcal{L}_{\rm nl}$, we observe a less pronounced improvement with a 22.9% decrease in the median.

Note that the performace of the IMRPhenomD model was initally tested using the `zdethp` weighted mismatch.[3] We would therefore like to examine whether using the `zdethp` PSD in our loss function could lead to an improved to mismatch. Peforming 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 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.

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

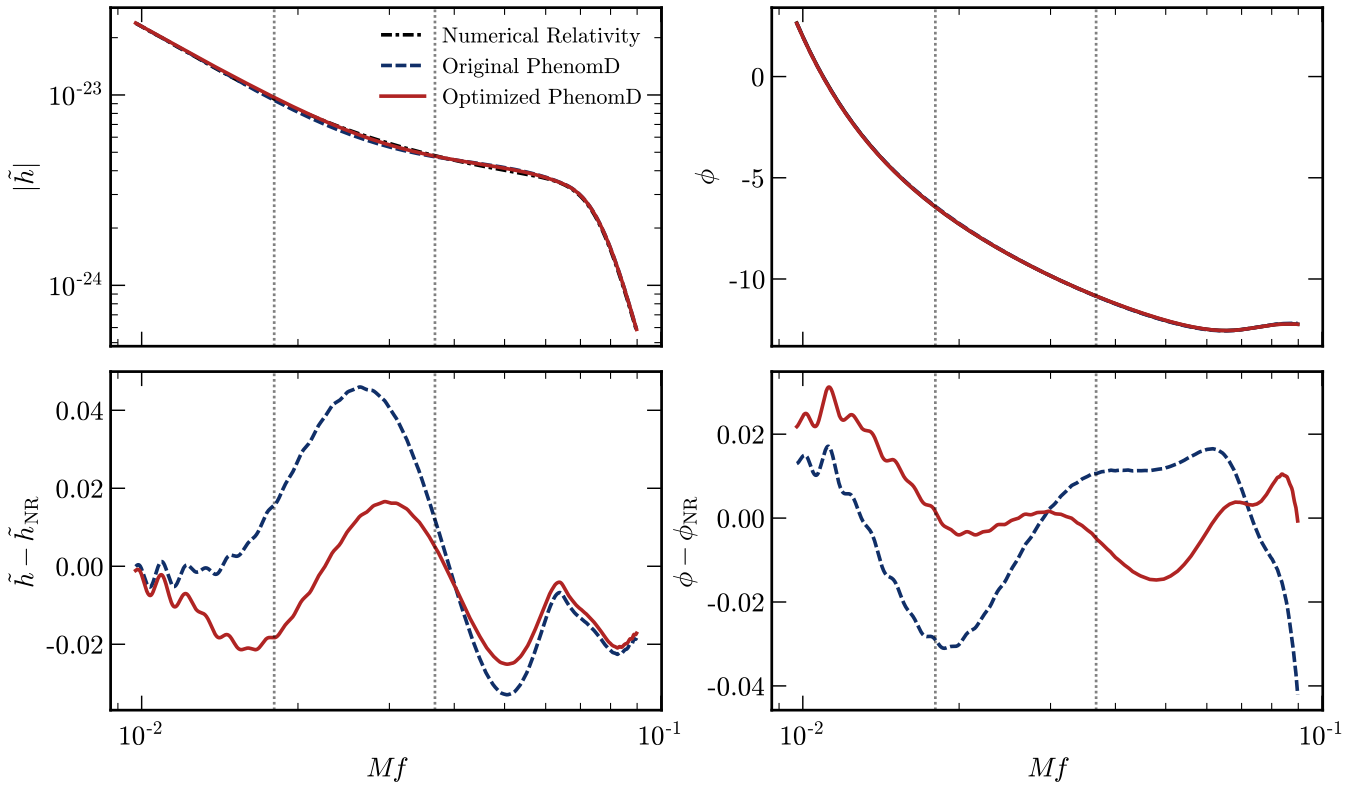


**Figure 3.** Comparison between original and optimized IMRPhenomD waveforms. Here we show the SXS:BBH:0154 NR waveform, which has a mass ratio of $q = 1$ and spins $\chi_1 = \chi_2 = -0.8$. The original mismatch is $2.8 \times 10^{-4}$ and the optimized mismatch is $5.3 \times 10^{-5}$. *Top*: Here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

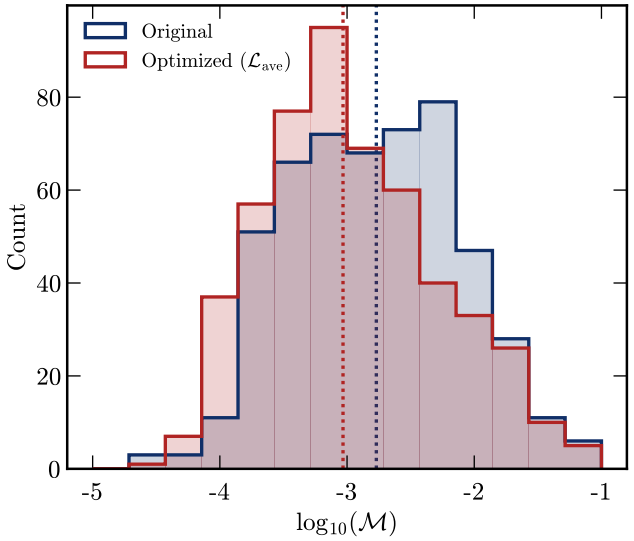


**Figure 4.** Distributions of waveform mismatches calculated using $\mathcal{L}_{\rm ave}$. We use the training waveforms listed in Tab. 1 and mismatches are weighted with a constant PSD. The dashed lines represent the median of the distributions, which decreased by 45.3% during optimization.

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 $\chi_{\rm PN} < 0$. Since the original model was developed for aligned-spin systems, NR waveforms with opposite sign $\chi_{1,2}$ are likely less well-fit [KW: Verify this. ]. To show this, we 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 ($\chi_1 < 0 < \chi_2$) and bottom-right ($\chi_1 > 0 > \chi_2$) 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

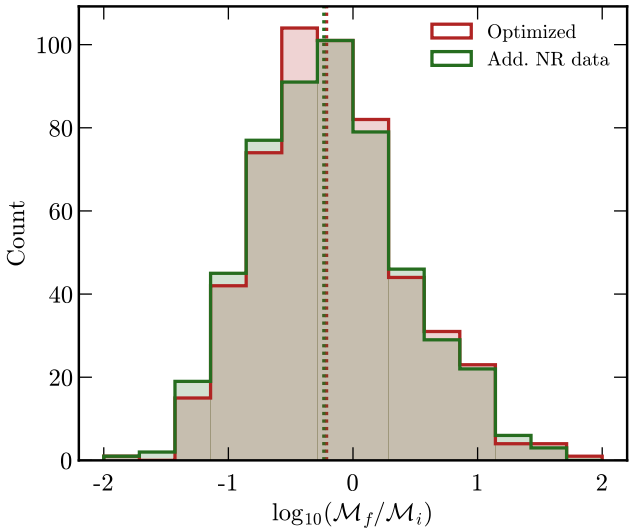


**Figure 5.** Distributions of $\log_{10}$ difference in mismatch. The red distribution uses the training waveforms listed in Tab. 1 while the green distribution uses waveforms listed in Tab. 1 and 2. Mismatches are calculated using a constant PSD with the loss function $\mathcal{L}_{\rm ave}$. The dashed lines represent the median of the distributions, which decreased by 10.4% during optimization.

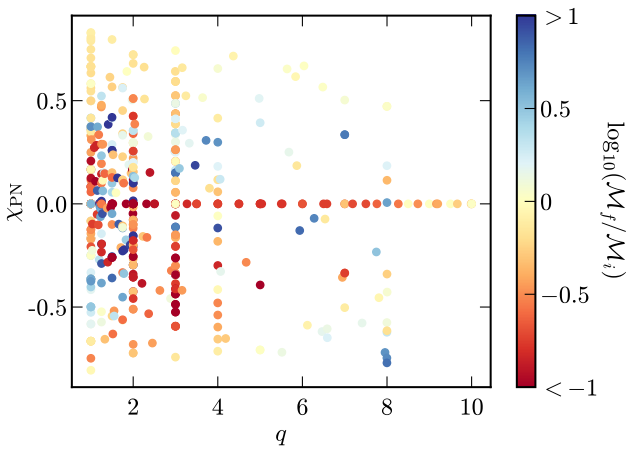


**Figure 6.** Fractional mismatch change of the validation waveforms in the $q-\chi_{\rm PN}$ plane. We show results for the $\mathcal{L}_{\rm ave}$ loss function with a constant PSD and training waveforms in Tab. 1. Here, the colorbar represents the $\log_{10}$ difference between optimized and original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

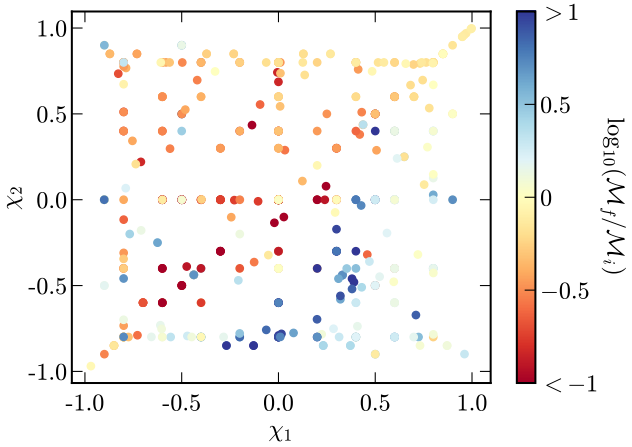


**Figure 7.** Fractional mismatch change of the validation waveforms in the $\chi_1-\chi_2$ plane. We show results for the $\mathcal{L}_{\rm ave}$ loss function with a constant PSD and training waveforms listed in Tab. 1.

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. 9). The training waveforms 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 results are only valid up to $q \lesssim 4$. From Fig. 9, we observe that equal-spin waveforms lying on the diagonal improve the most. The improvement above the diagonal is also significantly more than when optimizing using all waveforms at once. This can be seen clearly in the top panel of Fig. 8, where we see that the optimized histogram shifts to lower mismatches.[4] On the other hand, the improvement below the diagonal is negligible, indicating that our method struggles to further improve over the original IMRPhenomD waveform. Interestingly, we can see in the bottom-right quadrant that the waveform does not show a significant change to the overall mismatch (see bottom panel of Fig. 8). This suggests that

[4] Note that the top panel of Fig. 8 specifically shows the distribution of mismatches for the top-left quadrant of Fig. 9.

| Code | $q$ | $\chi_1$ | $\chi_2$ |
|---|---|---|---|
| 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 $\chi_1 - \chi_2$ quadrants. 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. 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 $\chi_1 \approx -\chi_2$.

the IMRPhenomD waveform ansatz fits waveforms with $\chi_1 > 0$ and $\chi_2 < 0$ less well.

## 4. DISCUSSION

We have shown the result of recalibrating waveform coefficients. One thing to note is that our recalibration procedure is not exactly the same as the original calibration. For instance, we use a different set of NR

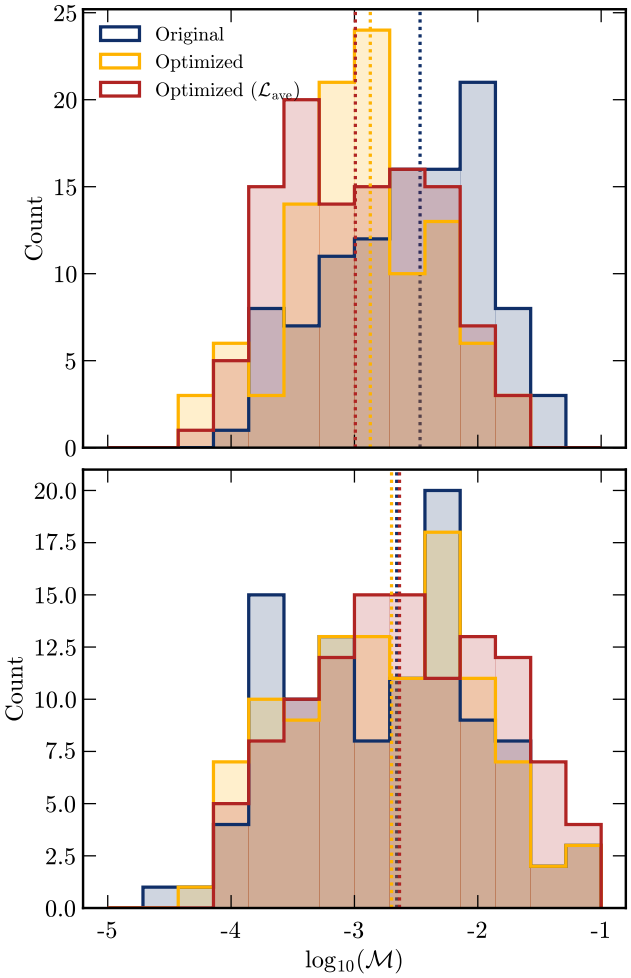


**Figure 8.** Distributions of mismatches in the top-left (top) and bottom-right (bottom) regions of Fig. 9. We use a constant PSD to calculate the mismatch and $\mathcal{L}_{\rm ave}$ as the loss function. The waveform generally improves in the top-left region while it shows very little improvement in the bottom-right, as indicated by the change of the medians of each distribution (dashed lines). [KW: Is the original referring to original IMRPhenomD or optimized using everything? Can we have a reference for optimizing using everything? ]

waveforms, frequency range, etc. Nonetheless, as the decrease in mismatch is rather significant, this optimization procedure should be able to improve the accuracy of IMRPhenomD on a similar scale regardless of the differences. Here, the result serves as a demonstration of the general method used.

The results presented in Fig. 5 demonstrate that increasing the number of training waveforms used in waveform optimization yields only a marginal increase in accuracy. Our analysis suggests that this marginal improvement is a consequence of over-determination of the waveform coefficients. Consequently, increasing the number of calibration NR waveforms is unlikely to re-

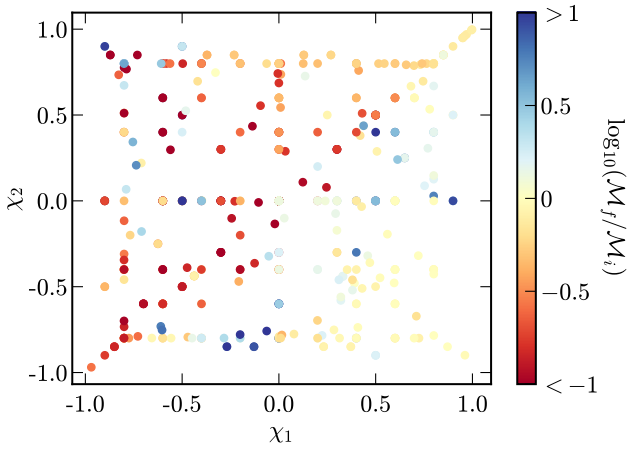


**Figure 9.** Fractional mismatch change of the validation waveforms. Each quadrant is fitted independently with to the NR waveforms listed in Tab. 3.

sult in any significant improvement of the model's accuracy. These observations suggest that the parameterized ansatz employed in our study may not be suitable for certain regions in the parameter space, leading to low mismatches for some waveforms while other waveforms remain at the high mismatch tail with negligible changes. This highlights the constraints of the model's flexibility that ultimately limit its performance.

The reduced spin approximation is a major contributor to the inaccuracies observed in the ansatz. In IMRPhenomD, this approximation employs a single spin parameter, $\chi_{\rm PN}$, as described in Sec. 2. The parameterization of BBH mergers using a single spin parameter results in a degeneracy within the parameter space. Specifically, black hole events with different spins may generate the same waveform due to identical values of $\chi_{\rm PN}$, leading to erroneous results, particularly for highly unequal spin events. This degeneracy produces straight lines in the parameter space with negative slopes that depend on the mass ratio, which can be seen in Fig. 7. Notably, the ansatz performs better in the top-left region than the bottom-right region. In an attempt to address this issue, we partitioned the parameter space into four regions, as described in Sec. 3. With separate optimizations for each regions, Fig. 9 indicates that the top-left region's performance has improved, while the bottom-right region hardly improves. This observation suggests the ansatz is specific to certain regions of the parameter space, with a preference for BBH events lying above the diagonal, and it has limited enhancement for events lying below the diagonal.

The division of the parameter space into four regions was a simple approach taken for practical reasons. A more systematic approach would involve the use of level set estimation algorithms to identify regions of interest within the parameter space. Such an algorithm can reveal additional degeneracies or issues that may exist within the ansatz. One possible strategy is to recalibrate individual regions of interest to achieve better results. An alternative approach is to select regions based on degeneracy lines. However, due to the limited number of NR waveforms available, we were unable to implement this approach. With more NR waveforms available in the future that cover the entire parameter space, we can perform optimization with fewer restrictions and select regions more systematically. Other than how to divide regions of interest, the choice of training waveforms also affects the final results greatly. Note that the choice of training waveforms listed in Tab. 3 were taken arbitrarily to test the effectiveness of separate fitting. Hence, if one takes a different set of training waveforms over the parameter space, the result might give additional features that can test and explain IMRPhenomD better.

Although our study primarily focused on the IMRPhenomD model, this simple yet versatile approach can be applied to other differentiable GW models, such as the IMRPhenomP (Hannam et al. 2014; Khan et al. 2019) or IMRPhenomX (Pratten et al. 2020, 2021) models within the same family. By jointly optimizing a new set of coefficients, it is expected that both models can be enhanced since they share similar construction principles to the IMRPhenomD model. For instance, they also use PN approximants as part of the ansatz in the inspiral segment. It will be interesting to recalibratethe IMRPhenomXAS model (Pratten et al. 2020). Because it is parameterized by an additional anti-symmetric spin parameter, it is expected not to exhibit the degeneracy previously described. With the currently developing `jax` IMRPhenomXAS model in `ripple`, A more detailed investigation may provide valuable insights into the systematics of the Phenom models. Furthermore, this approach is applicable to other GW model families, such as NR surrogate models or EOB models introduced in Sec. 1. Such an approach could simplify NR waveform calibration procedures and lead to the improvement of existing models.

## 5. CONCLUSION

In this paper, we have presented a systematic method to recalibrate GW models. This method utilizes `jax`'s automatic differentiation to apply derivative-based optimization to recalibrate GW models jointly. Using the new implementation of the IMRPhenomD model, `ripple`, which is written in `jax`, in conjunction with NR waveforms from the SXS catalog, we recalibrate waveform coefficients of the IMRPhenomD model. In gen-

eral, the waveform accuracy can be improved by 50%. Comparing `zdethp` weighted and unweighted mismatch, weighted mismatches have a slightly better improvement. In contrast, different types of loss function result in significantly different final mismatch distributions. As seen in Fig. 4, $\mathcal{L}_{\rm ave}$ outperforms $\mathcal{L}_{\rm nl}$. By increasing the number of training waveforms, we see a slight improvement increase in Fig. 5.

Furthermore, we investigated how the source parameters affect the improvement. Fig. 7 shows that the optimization procedure has a certain preference for waveforms lying in the top-left region while the bottom-right region hardly improved. To further test this result, we recalibrate waveforms in separate regions in parameter space. From Fig. 9, we can see that this recalibration process gives further improvement to the top-left region while the bottom-right region only have little improvement. This indicates that the ansatz has limited match in this region and does not fit most waveforms in this region, hence introduces bias to downstream GW analysis. This phenomenon is mainly due to the reduced spin approximation used in parameterizing the ansatz, where degeneracies between $\chi_1$ and $\chi_2$ are introduced.

While we naively separate the optimization process into 4 regions, one can perform systematic region-selection. In principle, we can apply this general method to other newer and more accurate models such as IMRPhenomX or IMRPhenomP models. Then, we can perform all the above analyses to understand how to construct better GW Phenom models in the future.

## 6. 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. P., et al. 2019, Phys. Rev. X, 9, 031040, doi: 10.1103/PhysRevX.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

—. 2021c, Phys. Rev. X, 11, 021053, doi: 10.1103/PhysRevX.11.021053

—. 2021d. https://arxiv.org/abs/2108.01045

—. 2021e. https://arxiv.org/abs/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

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

Harris, F. 1978, Proceedings of the IEEE, 66, 51, doi: 10.1109/PROC.1978.10837

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

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

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

—. 1996b, Phys. Rev. D, 53, 6749, doi: 10.1103/PhysRevD.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

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

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

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

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