Draft version June 21, 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 present a simple and general method of recalibrating gravitational wave (GW) phenomenological waveform models jointly. By using `jax`and `ripple`, we can perform automatic differentiationto functions, which allows us to use gradient-based optimization methods to recalibrate waveform coefficients in IMRPhenomD model. This method reduces systematic bias previously introduced to the model and generally can improve waveform accuracy. With recalibrated coefficients , we found that the typical *mismatch* has a 50% decrease. Furthermore, we analyze the accuracy base on the waveform 's intrinsic parametersinvestigate the possibility of improving the accuracy of a 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 of the waveform were calibrated separately. Using `ripple`, a library of waveform models compatible with automatic differentiation, we can perform gradients-based optimization to all the waveform coefficients the same time, which should improve the quality of waveform by capturing correlations between previous separately fitted parts. We found that waveform accuracy has significant dependence on black hole spin. Reduced spin approximation introduces degeneracy in spin, which prevented further improvement. We isolated regions in the parameter space that does not fit the waveform ansatz. These results allow us to understand more about how to develop newer phenomenological after recalibration, the median mismatch between the model and NR waveforms decrease by 50%. We further explore how different parts 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

In the future, the Laser Interferometer Gravitational-wave Observatory (LIGO) will finish its maintenance and start observing new Many data analysis tasks in today's gravitational wave (GW) results. This new O4 run is expected to double the rate of current binary black hole (BBH) observations (Abbott et al. 2020). Additionally, the sensitivity of interferometers will be increased to capture more details of GW. Having instruments with higher sensitivity, GW modelsof equal or higher accuracy than observations should be used to extract GW information. Otherwise, the extracted information would be affected more by GW models instead of interferometer sensitivity, resulting in a bottleneck in GW analyses. Although GW models are accurate enough for current analyses, the accuracy of current models will no longer suffice for future data analyses (Pürrer & Haster 2020). Hence, it is necessary for us to develop and improve GW models.

kelvin33550336@gmail.com

Currently, astrophysics, such as match filtering search (?) and parameter estimation(?), rely on accurate waveform models. Because using numerical relativity (NR) waveforms in these tasks is prohibitively expensive, the community has constructed waveform approximants that can be evaluated much faster. There are three families of GW models are commonly used . They are the commonly used GW approximants, which are effective-one-body (EOB) (Ossokine et al. 2020; Cotesta et al. 2020; Taracchini et al. 2014), Numerical Relativity (NR) surrogate (Islam et al. 2022; 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). EOB models are constructed by mapping two masses onto an effective body under an effective metric; NR surrogate models construct waveforms using combinations of NR waveforms; Phenom models are formulated using specific ansatz and inspiral approximations. While EOB and NR surrogate models give better waveform approximants, Phenom waveforms can be produced much faster, hence it is used mostly in data analysis tasks that requires many waveform generations. This advantage scales up in data analysis tasks such as matched filtering and

parameter estimation, where many waveforms are required in each run. This motivates us to improve upon the current framework of Phenom models, thus can retain the advantage of fast waveform generation while improving the model 's accuracy. While the detail of construction of these models are different, they all have a set of internal parameters that can be calibrated to NR waveforms. The quality of a waveform model is determined by the ansatz used and the accuracy of the calibrated parameters.

Automatic The LIGO-VIRGO-KAGRA (LVK) collaboration ((?????).) has recently started their fourth observational run on May 26, 2023. Because of the sensitivity improvement, the new run is expected to double the size of current binary black hole (BBH) observations (Abbott et al. 2020). The improved sensitivity also implies we expect to detect individual events with a higher signal-to-noise ratio (SNR). This means we can resolve more features in the signal, at the same time put a more stringent requirement on the accuracy of our waveform model(?).

Because of the large number of calibration parameters (often in a couple hundreds if not more), waveform models is usually calibrated in pieces (?). This ignores the correlation between different parts of the waveform model and limits it quality. Recently, there has been efforts in rebuilding waveform models (?) that supports automatic differentiation (AD)is a method to calculate (?), which is a technique used to compute derivatives of functions up down to machine precision . In traditional numerical calculations, derivatives are usually obtained through numerical derivatives. Symbolic derivatives were available but it was less efficient. Both methods were not viable in machine learning, where back-propagation requires precise and rapid derivative calculations. In without issues of scaling up to high dimension or expression swelling. In particular, `python`**`ripple`** , packages including `pytorch` (Paszke et al. 2019), `tensorflow` (Abadi et al. 2015), etc. utilizes AD to train machine learningmodels. AD's algorithm is intuitive in nature. Functions defined are decomposed into tree structures of primitives(Edwards et al. 2023) exposes the calibration parameters to the user. This allows us to make use of infrastructures that are heavily used in machine learning, such as addition or function evaluations. Since these operations are fundamental, they were saved as pairs internally. Differentiation proceeds forward following the tree structure, with the application of the chain rule in each step to evaluate its derivative. Analytic derivatives of such operations are applied in each step and the desired derivative can then be obtained by composing back the original function according to gradient descent and back propagation?, to improve the calibration of the original structure. `ripple`(Edwards et al. 2023) was a new implementation of IMRPhenomD, one of the models within the Phenom family. It was first implemented in `lalsuite` using `C`. In order to make use of AD, it was rewritten using `jax`, a `python`package that supports AD. Using `ripple`, one can apply AD to GW models to obtain precise derivatives, thus allowing one to freely use derivative-based algorithms to perform data analyses. waveform models.

In this paper, we investigate the possibility of further improving the accuracy of IMRPhenomD a waveform model, IMRPhenomD(?), by jointly optimizing all the fitting coefficients given NR waveforms, and what constraints one may face when trying to further improve. We find that simply by applying gradient descent algorithm, one can obtain a better set of waveform coefficients, thus improving the accuracy of the model.Furthermore, by comparing the accuracy of optimized and original waveforms, we find that model-generated waveforms are very sensitive to their intrinsic parameters. Specifically, IMRPhenomD favors certain parts calibration coefficients given a set of NR waveforms. Training on a few NR waveforms, we demonstrate one can improve the match between IMRPhenomD and NR waveforms over a decently sized parameter space, up to mass ratio $q = 8$ We also 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 of the parameter space. This means IMRPhenomD introduces systematic bias to other GW analysis tasks. This showcases the flaws of the ansatz and allows us to have a deeper understanding of Phenom modelsgives us hints on whether the waveform model 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 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 optimized waveforms with original waveforms. We also show how the optimization result differs with waveforms of different intrinsic source parameters. In Sec. 4, we address the difference between our calibrating procedure with (Khan et al. 2016). We also explain how 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 ϕ respectively):

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

where f is the frequnecy, θ are the intrinsic parameters of the binary, and Λ 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 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. 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 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 Λ 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 parameters back to the intrinsic parameter space, IMRhenomD 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 stiching 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.

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

Importantly, since the fitting was performed on the individual segments, the final waveform is not guarenteed 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 significatnly 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 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 (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), $f_{\rm max}$ and $f_{\rm min}$ are the relevant maximum and minimum frequencies for the integration. We note here that the mismatch can be viewed as the mean square error (MSE) between the two waveforms.

[TE: Up to here]

Since we wish to optimize the model over the whole parameter space, we need to compare multiple model generated waveforms with NR waveforms. However, mismatch only quantifies the difference between IMRPhenomD and NR waveform for one particular set of intrinsic parameters. To take into account of various different waveforms in the parameter space, we pick waveforms from the different parts of the parameter space. We define the loss function as an average of training waveforms in two ways, the simple average of mismatches and the normalize average of mismatches,

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

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

where $\mathcal{M}_i$ represents the mismatch of an individual training waveform, $\mathcal{M}_{i,\mathrm{ini}}$ represents the initial mismatch of the individual training waveform, and N is the total number of individual training waveforms. Note that we choose to use two different averages, since they have different preferences in optimization base on waveform mismatches. For the first choice, simple average serves as the simplest choice of loss function, but is prone to be dominated by a single waveform with a large mismatch. Other waveforms with smaller mismatches would be insignificant comparatively, and might not be able to improve under such optimization. Alternatively, the second choice, normalized average eliminates the aforementioned issue. Nevertheless, it excludes the information on initial mismatches. $\mathcal{L}_{\mathrm{fl}}$ restricts every training waveform to decrease at similar rates, hence it is hard to obtain optimized waveforms with mismatches in the same order of magnitude. Instead, their ratios in mismatches would remain approximately the same. Conversely, $\mathcal{L}_{\mathrm{mean}}$ allows the loss function to automatically adjust and individual mismatches would be in a similar order of magnitude after optimization. In this paper, we showcase 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 take NR waveforms for calculating the mismatch. We choose 11-16 NR waveforms from the set of waveforms used in the original calibration process as training waveforms. Originally, 19 waveforms are taken from NR simulations for calibrating IMRPhenomD (Khan et al. 2016), which are waveforms from the SXS catalog (Boyle et al. 2019) or BAM simulation. As BAM waveforms are not publicly available, we cannot take the identical training set as them. Instead, we take the available waveforms from the SXS catalog to construct our loss function. Training waveforms used are listed in Tab. 1 and 2. The training waveforms chosen has maximum mass ratio to be 8. This is because SXS catalog does not have NR waveforms with extremely high mass ratio. 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 from LIGO have $q \leq 8$. Hence, we calibrate IMRPhenomD with waveforms of $q \leq 8$.

In the SXS catalog, NR waveforms are in the form of time-series strain. Since time-series data is oscillatory, performing optimization in the time-domain is not ideal. Hence, we transform NR waveforms to frequency-domain to compare with IMRPhenomD waveforms with the same intrinsic parameters. We taper the time-series using Tukey window. [2] Then, the frequency spectra can be obtained by taking the Fourier transform of the time-series.

Other than NR waveforms, one need to choose a relevant noise PSD for mismatch. We have opted to use a flat PSD for this purpose, 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 and piece-wise fitting, but not due to the use of a different mass scale. Furthermore, 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. 2015). 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. We selected an arbitrary mass scale of $M = 50M_\odot$ for all waveforms for demonstration, as this is a commonly observed mass scale in LIGO observations.

We point out that our treatment to NR waveforms is different from that of (Husa et al. 2016; Khan et al. 2016). In the original calibration process, training waveforms are hybrid waveforms of NR and SpinAlignedEOB (SEOB) waveforms. The low frequency inspiral part is taken from the SEOB waveforms while the rest of the waveforms are taken from NR simulations. Instead, we solely use NR waveforms for comparison, since we are only exploring the possibility of optimizing waveform models. Thus, for simplicity sake, we ignore this procedure. On the other hand, most NR waveforms used (for both training and testing) have long enough time series data, i.e. > 15 orbits (Boyle et al. 2019), in which they are long enough to contain part of the inspiral segment and all merger and ringdown frequency information. We take the frequency limits as $f_{\mathrm{min}} = 0.1 f_{\mathrm{RD}}$ and $f_{\mathrm{max}} = 1.2 f_{\mathrm{RD}}$, where f_{RD} is the frequency at ringdown. This range covers most of the IMRPhenomD's frequency range, except the minimum frequency is set higher than that in the original calibration due to NR length. When compared with IMRPhenomC, the frequency range is slightly extended to have a higher maximum frequency. We have the dimensionless frequency spacing $M\Delta f = 2.5\times10^{-6}$, which is sufficient to capture all features of GW strain.

With the loss function evaluated, we apply gradient descent to optimize the tunable coefficients as shown in

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

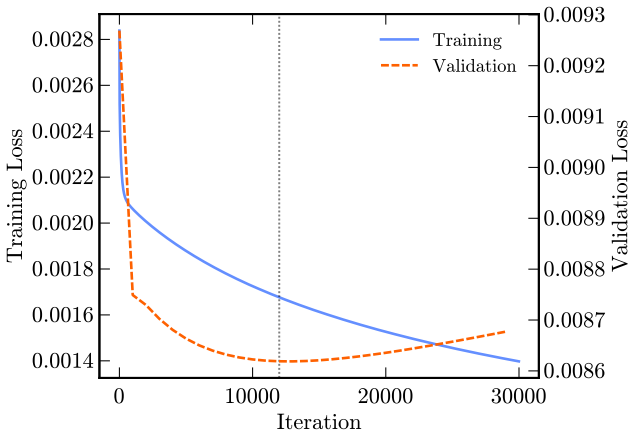

Figure 1. Loss functions against number of iterations. We take the set of coefficients at the minimum of the validation loss.

Algorithm 1: Gradient descent pseudocode

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

$\lambda \leftarrow \lambda_i$
/* 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 λ

Algorithm 1. We take λ_i to be the original coefficients given in (Khan et al. 2016). We take them as the initial waveform coefficients because they lie in the neighborhood of the minimum that we wish to find. Then, by taking $\alpha = 10^{-6}$, we see the validation loss in Fig. 1 plateau at around $N = 12000$, hence we end the optimization here.

3. RESULT AND COMPARISON WITH ORIGINAL MODEL

To evaluate the effectiveness of the optimization , we analyze how well the optimization procedure generalize to waveforms that are not in the training set, we evaluate the mismatch between the fine-tuned model and the data provided in the SXS catalog for 536 NR waveforms from the SXS catalog in this study. We specifically select waveforms with NR waveforms. We select waveforms that share the same part of the parameter space with the training waveform, i.e. the waveforms with negligible eccentricity ($e < 2 \times 10^{-3}$) and precession ($\chi_{x,y} < 5 \times 10^{-3}$)that are consistent with the constraints of the waveform model. As shown in . Fig. 2 , the intrinsic parameters of the chosen test waveforms fall within the parameter spacedefined by the training waveforms. This suggests that these test waveforms can be used for direct comparison with the original model. shows how the training and testing waveforms are distributed in the $q - \chi_{PN}$ space.

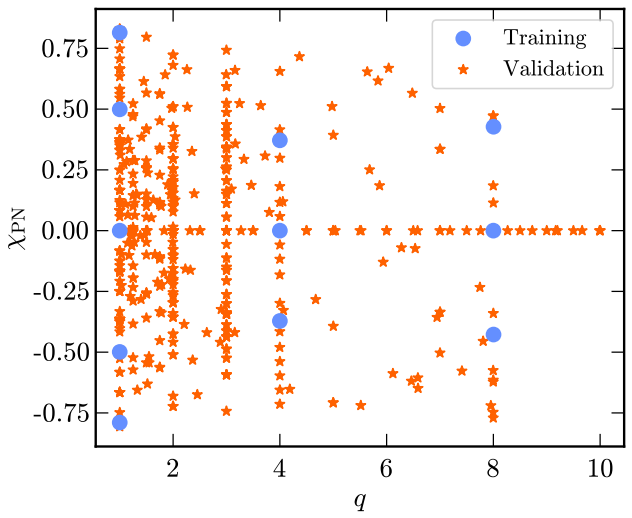

Figure 2. Parameter Distribution of training and testing waveforms in $q - \chi_{PN}$ spacewith mass ratio q against post-newtonian spin parameter $\chi_{\rm PN}$. Orange: Training The blue dots represent the waveforms ; Blue: Testing used in training and the orange stars represent the waveforms used in testing.[KW: Change the legend, and include the extra points. Change the training point to be star and testing point to be dot.]

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 waveforms used to recalibrate the model. The mass ratio $q = m_1/m_2 \geq 1$ with spins $\chi_{1,2}$. Out of the 11 waveforms listed here, 9 of them are also used in the original IMRPhenomD calibration. (Khan et al. 2016) The two remaining waveforms were from BAM simulation, to which we do not have access.

To illustrate the improvement effect of optimiziation on an individual waveform level, we compare the mismatch of a

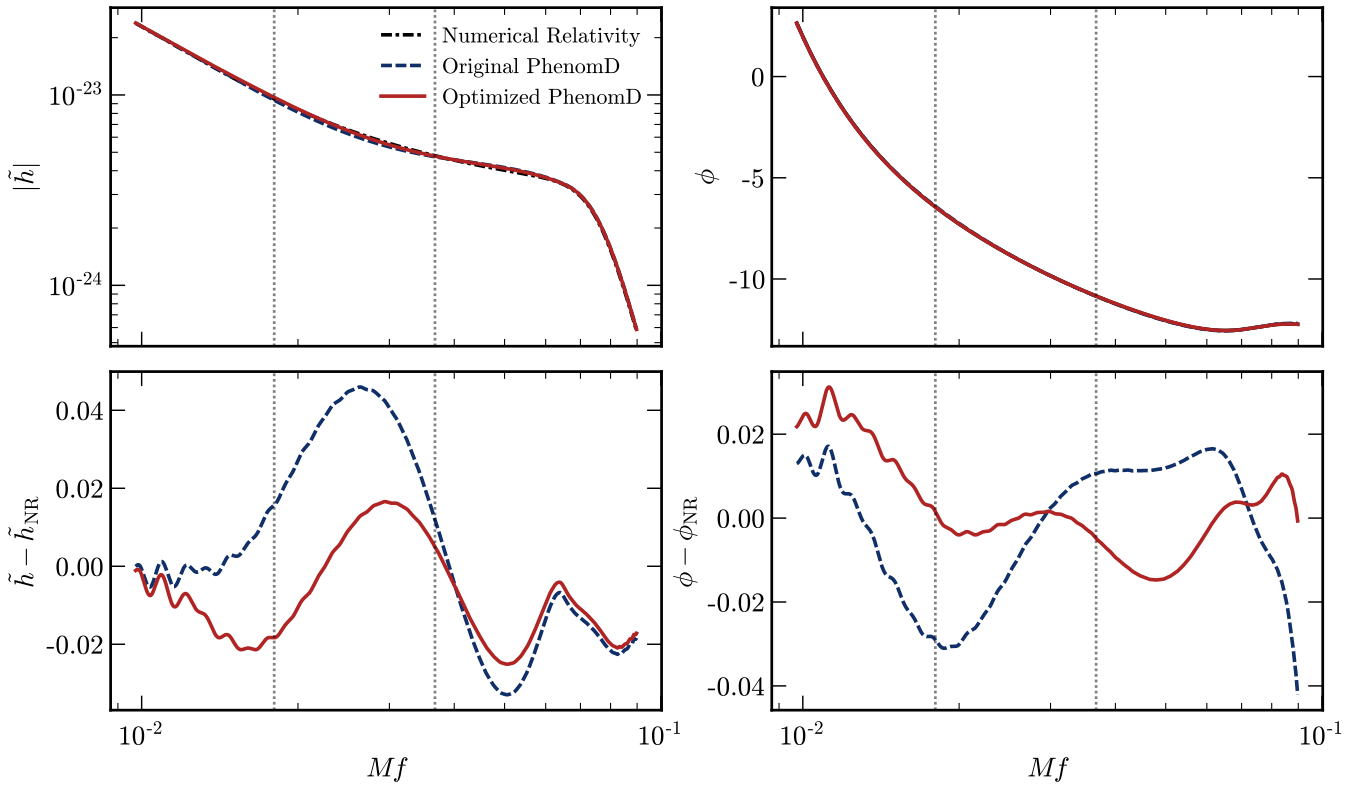

Figure 3. Comparison between original and optimized IMRPhenomD waveforms. Here shows the SXS:BBH:0154 NR waveform, which has mass ratio $q = 1$ and $\chi_1 = \chi_2 = -0.8$. The original mismatch is around 2.8×10^{-4} and the optimized mismatch is around 5.3×10^{-5}. Top: It shows the amplitude and phase of NR, original IMRPhenomD and optimized IMRPhenomD waveform. Bottom: It shows the relative error of amplitudes between NR and IMRPhenomD waveforms, and the absolute error of phases between NR and IMRPhenomD waveforms

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 waveforms used in further recalibration.

particular waveform before and after optimization with respect to the NR waveform taken directly from the SXS catalog in Fig. 3. Compared to the original IMRPhenomD waveform, we can see the optimized waveform has smaller residual from NR waveform 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 one of the testing waveforms from the catalog presented in (Khan et al. 2016).

To quantify the effect of optimization 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 dataset, We evaluate the mismatch of all testing waveforms using training dataset. Fig. 4 shows the distribution of log-mismatch for the testing waveforms before and after the optimization procedure. For generality, we use a constant PSD weighted loss function, $\mathcal{L}_{\rm mean}$. We present the resulting distribution in Fig. 4. The in our loss function. One can see the distribution have been skewed toward lower mismatch, where the peak of the distribution has shifted towards a lower mismatch, with a decrease of almost one is shifted by approximately an order of magnitudeand a , and a median mismatch is reduced by 50%reduction in the median. When using $\mathcal{L}_{\rm fit}$, we observe a similar improvement with a 22.9% decrease in the median. Even though the distribution lacks a clear peak due to the problem discussed in Section 2.3, both distributions show observable improvement.

Applying the same methods, we will see that the distributions of mismatches calculated using the `zdethp`PSD exhibit superior improvement compared to the unweighted mismatch. The shape of the distribution is similar to that shown in Fig. 4. This result is expected, as the Note that the IMRPhenomD model was initially constructed and fitted using the `zdethp` weighted mismatch. Consequently, the model is anticipated to closely fit the NR waveforms while incorporating the influence of the To ensure a fair

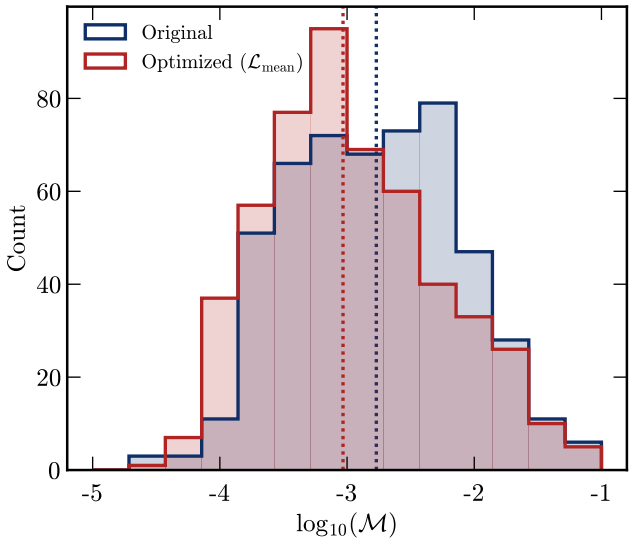

Figure 4. Distributions of waveform mismatches calculated using $\mathcal{L}_{\rm mean}$. We use training waveforms listed in Tab. 1 and mismatches are weighted with the constant PSD. Dashed lines represent the median of the distributions, which has decreased by 45.3%.

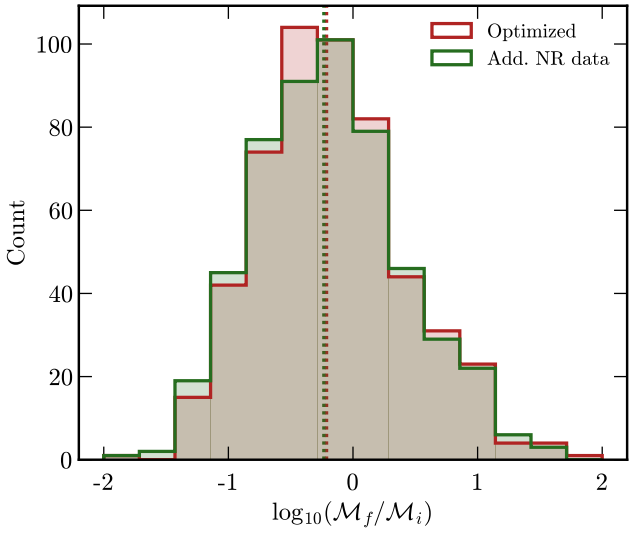

Figure 5. Distributions of $\log_{10}$ difference in mismatch. The distribution labeled `q148` uses training waveforms listed in Tab. 1 while the `q1248` distribution uses waveforms listed in Tab. 1 and 2. Mismatches are calculated using the constant PSD with the loss function $\mathcal{L}_{\rm mean}$. Dashed lines represent the median of the distributions, which has decreased by 10.4%.

comparison, we apply the same methods using the `zdethp` PSD , rather than a constant PSD. exhibit superior improvement compared to the unweighted mismatch. We find no significant qualitative and quantitative difference between the two PSDs in the resulting distribution of mismatches.

Motivated by the successful improvement of the waveforms, we expand the training dataset to optimize additional waveforms listed To understand whether additional training data can further improve the performance of the model, we include waveforms that are not presence in the original IMRPhenomD calibration in our training dataset with parameters tabulated 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 optimized distribution remains comparable in length and endpoint to the original distribution, indicating that our procedure is incapable of improving these waveformsthe original dataset is sufficient for this task. Similarly, utilizing the `zdethp` PSD to optimize the loss function with additional waveforms leads to similar improvement in the resulting distributionresults in similar level of improvement.

Although most of the waveforms show improvement in figure 4, the high mismatch tail of the distribution remains unaffected. 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 intrinsic parameters, it is conceivable that some trade-offs in accuracy exist between different parts To investigate the performance of recalibration over the source parameter space, we plot the improvement of mismatch in log projected over the parameter space of q vs. $\chi_{\rm PN}$ in Fig. 6. On the mass ratio axis, we can see the waveforms with $q \leq 4$ show most consistent average improvement. This is because that part 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 problemis better covered by the training waveform, as compared to testing waveforms from $q = 4$ to $q = 8$, where some testing waveforms lie outside the source parameter space covered by the training set of waveforms. On the other hand, if the ansatz lacks the correct parameterized form to capture the NR waveforms' behavior as a function of the intrinsic parameters, the results will always be biased, and we should not expect any improvement, even if we segment the parameter space during training.

Since we know intrinsic parameters play an important role in the ansatz, we would like to investigate how intrinsic parameters affect the recalibration process. First, we plot the parameter space of q vs. $\chi_{\rm PN}$ in Fig. 6. We can see near non-spinning waveforms demonstrate spin axis, we can see the waveform with $\chi_{\rm PN}$ close to zero show the most consistent improvement, and we start deviate from 0, there could be fluctuations in the average improvement. On top of that, in the $q \leq 4$ region, there seems to be more consistent improvement , since the ansatz are developed base on non-spinning waveforms . Also, the original

coefficients were fitted using NR waveforms with equal or similar spin , hence the model prefers waveforms with similar spin. On the other hand, spinning waveforms can either improve or worsen in terms of mismatch base on their intrinsic parameters. To further discuss the behavior of non-spinning of waveforms with $\chi_{\rm PN} < 0$. This can be understood as since the original waveform was developed for aligned spin system, the waveforms with $\chi_{\rm PN}$ is less well-fitted [KW: Varify this.], hence there could be more room for improvement. To show this, we plot the parameter space of χ_1 vs. χ_2 in Fig. 7. Waveforms along the diagonal axis, i.e. $\chi_1 \approx \chi_2$, 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 optimization differently. In the top-left region, waveforms generally improve slightly with along optimization. However, waveforms in the bottom-right region do not improve after optimization. Some waveforms even turned worse 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, mean there could be some 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. 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 has great improvement. Above the diagonal , the improvement is significant, except for a few defects The improvement above the diagonal is also substantial compared to recalibrate using all waveforms at once, barring a few caused by some testing waveforms having $q > 4$. This can be seen clearer in Fig. 8, which we see that the optimized histogram shifts downward uniformly. Below the diagonal, waveforms have negligible improvement (Fig. 8 and 9), indicating the original set of coefficients is the optimal set of coefficients, and cannot be further improved. HenceOn the other hand, the improvement below the diagonal is negligible, indicating our method struggles to further improve on top of the original IMRPhenomD waveform. Interestingly, we can deduce IMRPhenomD has better match with the top-left region and does not fit well with waveforms lying in the bottom-right region. Thus, the ansatz restricts additional improvements and further optimizing in a smaller region does not generally improve the resultssee in the top-right and bottom-left quadrant, the lower right corner shows systematic worsening similar to the case of all waveforms. This hinted in general, the IMRPhenomD waveform ansatz does not fit well with waveforms with anti-aligned spin where the primary spin is positive.

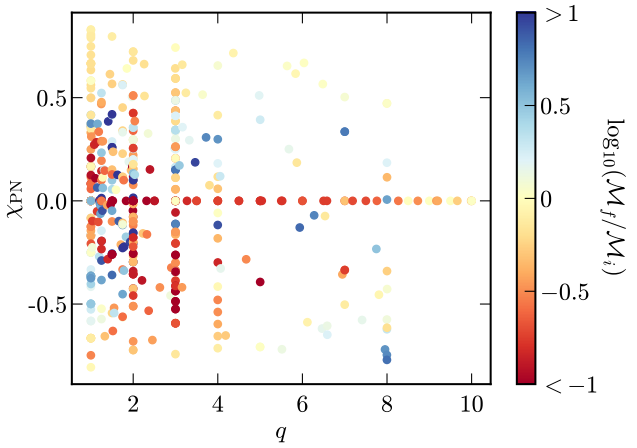

Figure 6. Parameter space of testing waveforms with q against $\chi_{\rm PN}$. We show the result from optimizing $\mathcal{L}_{\rm mean}$ with constant PSD and training waveforms in Tab. 1. Here, the colorbar represents the $\log_{10}$ difference between optimized and original mismatches.

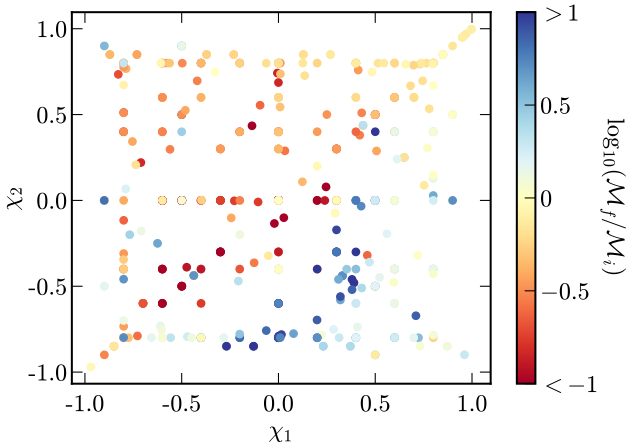

Figure 7. Parameter space of testing waveforms with χ_1 against χ_2. We show the result from optimizing $\mathcal{L}_{\rm mean}$ with the constant PSD and training waveforms listed in Tab. 1.

4. DISCUSSION

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 waveforms used in recalibrating coefficients in 4 regions. From top to down are 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 $\chi_1 \approx -\chi_2$.

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 waveforms, frequency range, etc. Nonetheless, as the decrease in mismatch is rather significant, this optimization procedure should be able to improve the accuracy

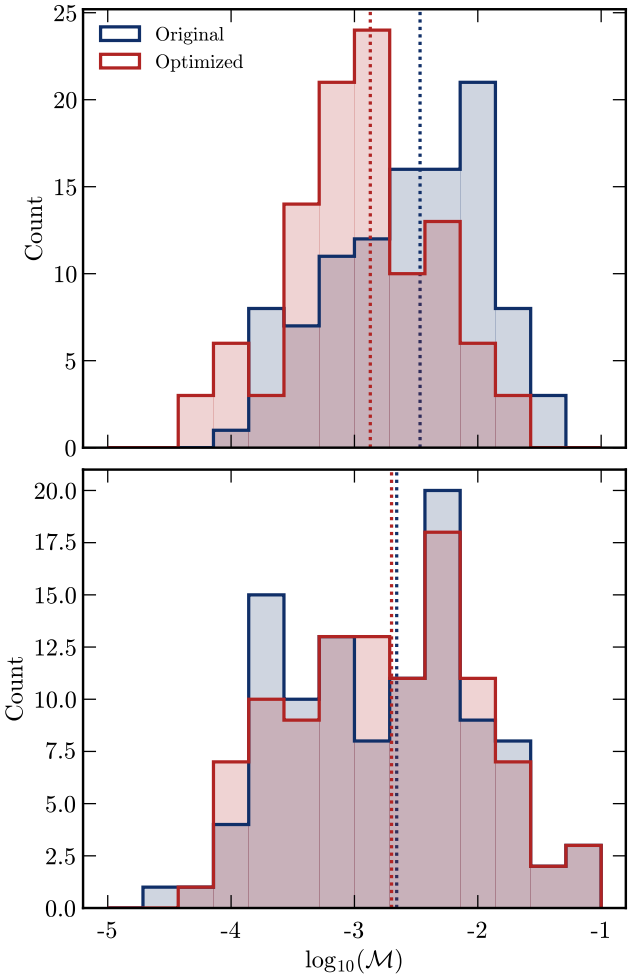

Figure 8. Distributions of mismatches in the top-left (Top) and bottom-right (Bottom) regions. We use a constant noise spectrum to calculate mismatch and $\mathcal{L}_{\rm mean}$ as the loss function. Waveforms in the top-left region generally improves while waveforms in the bottom-right region has very little improvement, as indicated by the median (dashed lines). [KW: Is the original referring to original IMRPhenomD or optimized using everything? Can we have a reference for optimizing using everything?]

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 result in any significant improvement of the model's accuracy. These observations suggest that the parameterized ansatz employed in our study may not be suitable

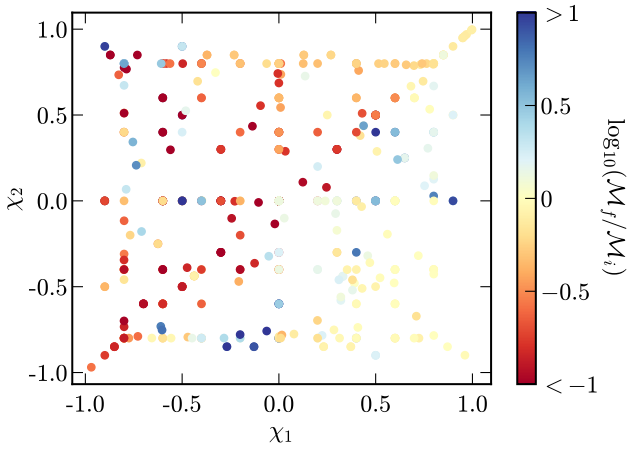

Figure 9. Parameter space of testing waveforms. Each region is fitted independently with waveforms listed in Tab. 3.

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 general, 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, $\mathcal{L}_{\rm mean}$ outperforms $\mathcal{L}_{\rm fl}$. By increasing the number of training waveforms, we see a slight improvement increase in Fig. 5.

Furthermore, we investigated how the intrinsic 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 χ_1 and χ_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., Abbott, B., Abbott, R., et al. 2015, 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., Abbott, R., Abbott, T., et al. 2020, Living reviews in relativity, 23, 1

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

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

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

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

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

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

Islam, T., Field, S. E., Hughes, S. A., et al. 2022, 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

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

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