

Recalibrating Gravitational Wave Phenomenological Waveform Model

KELVIN K. H. LAM,¹ KAZE W. K. WONG,² AND THOMAS D. P. EDWARDS³

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 differentiation to 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 parameters. 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 models.

1. INTRODUCTION

In the future, the Laser Interferometer Gravitational-wave Observatory (LIGO) will finish its maintenance and start observing new gravitational wave (GW) results. This new O4 run is expected to double the rate of current binary black hole (BBH) observations (?). Additionally, the sensitivity of interferometers will be increased to capture more details of GW. Having instruments with higher sensitivity, GW models of 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 (?). Hence, it is necessary for us to develop and improve GW models.

Currently, three families of GW models are commonly used. They are the effective-one-body (EOB) (???), Numerical Relativity (NR) surrogate (???) and phenomenological (Phenom) models (????). 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 wave-

forms; 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.

Automatic differentiation (AD) is a method to calculate derivatives of functions up 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 `python`, packages including `pytorch` (?), `tensorflow` (?), etc. utilizes AD to train machine learning models. AD's algorithm is intuitive in nature. Functions defined are decomposed into tree structures of primitives, 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 the original structure. `ripple` (?) 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.

In this paper, we investigate the possibility of further improving the accuracy of 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 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 models.

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 parameters. In Sec. 4, we address the difference between our calibrating procedure with (?). We also explain how reduced spin parameterization affects the accuracy of the model.

2. METHOD

2.1. Waveform Model

In order to recalibrate the model, we have to understand what parameters the model has. Here we give a succinct summary of the IMRPhenomD model and the relevant parameters. For interested readers, please refer to (?) for more details on construction of the model.

The IMRPhenomD model is constructed by combining three individually fitted parts into one coherent waveform model, which consists of the inspiral, intermediate, and merger-ringdown part, [KW: Check if the following equation is correct. I think it is off by some smoothing factors.] [KL: I don't remember there is any smoothing

factors. They are just connected using step functions, then the entire waveform is made to have continuous first derivative by fixing some of the coefficients for the two connections.]

$$h(f, \theta, \Lambda^i) = h_{\text{ins}}(f, \theta, \Lambda^i) + h_{\text{int}}(f, \theta, \Lambda^i) + h_{\text{rd}}(f, \theta, \Lambda^i). \quad (1)$$

Instead of fitting the strain, which is a highly oscillatory function that is difficult to fit, the amplitude and phase are fitted since they are smoother functions. In each part, the amplitude and phase are made using simple functions of frequency such as polynomials or lorentzians. Specifically, the merger-ringdown amplitude is fitted by a lorentzian and the other parts are fitted using polynomials.

$$\begin{aligned} A_0 &\equiv \sqrt{\frac{2\eta}{3\pi^{1/3}}} f^{-7/6} \\ A_{\text{ins}}(f; \theta) &= A_{\text{PN}}(f; \theta) + A_0 \sum_{i=1}^3 \rho_i f^{(6+i)/3} \\ A_{\text{int}}(f; \theta) &= A_0 (\delta_0 + \delta_1 f + \delta_2 f^2 + \delta_3 f^3 + \delta_4 f^4) \\ A_{\text{rd}}(f; \theta) &= A_0 \left[\gamma_1 \frac{\gamma_3 f_{\text{damp}}}{(f - f_{\text{RD}})^2 + (\gamma_3 f_{\text{damp}})^2} e^{-\frac{\gamma_2 (f - f_{\text{RD}})}{\gamma_3 f_{\text{damp}}}} \right], \end{aligned} \quad (2)$$

where A_{PN} is the post-newtonian expansion of the inspiral amplitude up to order $A_0 f^2$, f_{damp} is the damping frequency, and f_{RD} is the frequency at ringdown. The ansatzes for phase can be found in (?). These simple analytic functions consists of parameters $\Lambda^i = \{\rho_i, \delta_i, \gamma_i\}$, which are defined as follows.

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

where λ are fitting coefficients obtained during calibration, η is the symmetric mass ratio, and χ_{PN} is the post-Newtonian spin parameter, which is defined as

$$\chi_{\text{PN}} = \frac{m_1 \chi_1 + m_2 \chi_2}{m_1 + m_2} - \frac{38\eta}{113} (\chi_1 + \chi_2). \quad (4)$$

Here, $m_{1,2}$ and $\chi_{1,2}$ are the primary and secondary mass and spin, respectively. Finally, the individual segments are first connected directly using step functions. Then, by fixing coefficients in the intermediate segment, one can make the final waveform is continuous in its first derivative.

Combining Eq. 1, 2 and 3, we can see that the entire waveform is non-linear in λ . A slightly inaccurate set

of λ can significantly affect the shape of the generated waveforms. Thus, having a set of accurate waveform coefficients is important and fundamental to having accurate GW models. Generally, waveform coefficients are obtained by calibrating with NR waveforms, which are waveforms computed using NR simulations. In the case of (?), they first obtain a set of Λ by fitting model generated waveforms to Eq. 2 and the phase ansatzes. Repeating with different NR waveforms, they obtain multiple sets of Λ , and λ are subsequently found by fitting against Eq. 3. Since the fitting procedure is done in a piece-wise manner, the correlations between different segments are omitted, which could limit the accuracy of the model. Also, since fitting was performed before connecting individual segments, the final waveform does not guarantee to achieve the optimal waveform. The connecting procedure alters the previously fitted waveform. Hence, the model generated waveforms contains additional inaccuracies.

Instead, we recalibrate coefficients jointly, which we can remove inaccuracies and biases discussed above, and can improve model accuracy. In the past, due to the complex nature of GW strains and piece-wise formalism of IMRPhenomD, non-linear fitting was difficult to be performed in optimizing coefficients. Hence, piece-wise optimization was done to obtain coefficients. However, with `ripple` and AD from `jax`, gradients of IMRPhenomD can be easily obtained, thus allowing the use of gradient-based algorithms for us to recalibrate the model.

2.2. Loss Function

In order to recalibrate the model, we need to define an objective function that quantify the performance of the model. A common choice for such metric is the *mismatch* function (?). It is defined as

$$\mathcal{M}(h_1, h_2) = 1 - \max_{t_0, \phi_0} \langle \hat{h}_1, \hat{h}_2 \rangle, \quad (5)$$

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

$$\langle h_1, h_2 \rangle = 4 \text{Re} \int_{f_{\min}}^{f_{\max}} \frac{h_1(f) h_2^*(f)}{S_n(f)} df, \quad (6)$$

where $\hat{h} = h/\sqrt{\langle h, h \rangle}$ is the normalized GW strain, $S_n(f)$ is the power spectral density(PSD) of noise from the instrument, f_{\max} and f_{\min} are the relevant maximum and minimum frequencies for the integration. The mismatch is a quantity that is closely related to the mean square error (MSE) between the two waveforms, as it is often considered a weighted version of the MSE.

[KL: The maximum is calculated by comparing the phase of IMRPhenomD and NR waveforms. We know that they are off by a linear relation of f ,

$$\phi_{NR} - \phi_{IMR} = (2\pi t_0)f + \phi_0.$$

Since I have ϕ_{NR} , ϕ_{IMR} and the frequency array, I did a linear regression to get $2\pi t_0$ and ϕ_0 .]

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,

$$\mathcal{L}_{\text{mean}} = \frac{1}{N} \sum_{i=1}^N \mathcal{M}_i \quad (7)$$

$$\mathcal{L}_{\text{fl}} = \frac{1}{N} \sum_{i=1}^N \frac{\mathcal{M}_i}{\mathcal{M}_{i,\text{ini}}}, \quad (8)$$

where \mathcal{M}_i represents the mismatch of an individual training waveform, $\mathcal{M}_{i,\text{ini}}$ 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}_{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, $\mathcal{L}_{\text{mean}}$ 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 (?), which are waveforms from the SXS catalog (?) 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.¹ Then, the frequency spectra can be obtained by taking the Fourier transform of the time-series.

[TE: I think this discussion needs to be much more precise. Having a waveform model calibrated to particular noise curve could make sense, but its inherently different to just fitting the phase and the amplitude separately like they do in the original paper. So I would recommend that our default is to use a flat PSD and then in addition discuss the one with the PSD. This way, I think we demonstrate more clearly that its the high dimensional fitting that helps, not the just changing the metric used when fitting.] [TE: This choice seems pretty arbitrary. Looking at Fig 8 of 2111.03606, they seem to cluster closer to 60-70. This choice is important and should be more informed.]

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 solely to the difference in high-dimensional fitting and piece-wise fitting. 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 (?). 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. Therefore, we have selected a mass scale of $M = 50M_{\odot}$ for all waveforms, as this is a commonly observed mass scale in LIGO observations. We will then examine the differences in optimization.

We point out that our treatment to NR waveforms is different from that of (??). 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 (?), 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 $f_{\min} = 0.1f_{\text{RD}}$ and $f_{\max} = 1.2f_{\text{RD}}$, 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 $M\Delta f = 2.5 \times 10^{-6}$, 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 Algorithm 1. We take λ_i to be the original coefficients given in (?). 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 $\alpha = 10^{-6}$ and $N = 30000$, we can optimize the coefficients systematically. [TE: I think we need a bit more discussion about the stopping criterion here. What is N, and why did we choose it to be this number? Ideally we would actually have a plot of the loss function during training. I don't have a good intuition for if its noisy. Does it plateau? If we wanted to ensure that people don't complain about overfitting we could also plot the loss of the validation set to show its not going up.]

3. RESULT AND COMPARISON WITH ORIGINAL MODEL

To evaluate the effectiveness of our optimization approach, we analyze approximately 530 NR waveforms from the SXS catalog in this study. We specifically select

¹ Specifically, we choose $\alpha = 2t_{\text{RD}}/T$, where t_{RD} is the duration of ringdown and T is the duration of the entire GW strain.

figures/loss.pdf

Figure 1. The value of loss function against number of iterations. The code is terminated after 30000 steps

figures/intrin_space.pdf

Figure 2. Parameter space with mass ratio against normalized reduced spin. Orange: Training waveforms; Blue: Testing waveforms

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

```

1  $\lambda \leftarrow \lambda_i$ 
/* Gradient Descent */
2 for  $i < N$  do
3    $\mathcal{L} \leftarrow \text{Mismatch}(\lambda)$ 
4    $\nabla \mathcal{L} \leftarrow \text{AutoDiff}(\mathcal{L})$ 
5    $\lambda \leftarrow \lambda - \alpha \nabla \mathcal{L}$ 
6 return  $\lambda$ 

```

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 space defined by the training waveforms. This suggests that these test waveforms can be used for direct comparison with the original model.

We begin by investigating the impact of joint optimization on an individual waveform. Fig. 3 illustrates that the optimized waveform outperforms the original waveform, particularly in the inspiral region, where the amplitude displays a 50% reduction in error. To ensure fairness, we selected one of the testing waveforms from the catalog presented in (?).

| 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. (?) The two remaining waveforms were from BAM simulation, to which we do not have access.

We use a constant noise spectrum, $\mathcal{L}_{\text{mean}}$, to evaluate the mismatch of all testing waveforms, and present the resulting distribution in Fig. 4. The peak of the distribution has shifted towards a lower mismatch, with a decrease of almost one order of magnitude and a 50% reduction in the median. When using \mathcal{L}_{fl} , we observe a similar improvement with a 22.9% decrease in the me-

figures/0154.pdf

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.

figures/q148.pdf

Figure 4. Distributions of waveform mismatches calculated using both $\mathcal{L}_{\text{mean}}$ and \mathcal{L}_{fl} in recalibration. We use training waveforms listed in Tab. 1 and mismatches are weighted with the constant noise spectrum. For the $\mathcal{L}_{\text{mean}}$ distributions, the median decreased by 50.0% while the median of \mathcal{L}_{fl} distribution decreased by 22.9%.

dian. However, the distribution lacks a clear peak due to the problem discussed in Section 2.3. Despite this, both distributions exhibit significant improvement, and the optimization scheme remains unaffected.

Applying the same methods, we observe that the distributions of mismatches calculated using the **zdet**hp noise spectrum exhibit superior improvement compared to the weighted mismatch. The shape of the distribution is similar to that shown in Fig. 4. This result is expected, as the IMRPhenomD model was initially constructed and fitted using the **zdet**hp weighted mismatch. Consequently, the model is anticipated to closely fit the NR waveforms while incorporating the influence of the **zdet**hp noise spectrum, rather than a constant spectrum.

figures/q148_q1248_compare.pdf

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 noise spectrum with the loss function $\mathcal{L}_{\text{mean}}$.

Motivated by the successful improvement of the waveforms, we expand the training dataset to optimize additional waveforms listed in Tab. 2. 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 waveforms. Similarly, utilizing the **zdet**hp spectrum to optimize the loss function with additional waveforms leads to minimal improvement in the resulting distribution.

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 some trade-offs in accuracy exist between different parts of the parameter space. If this is the cause of the high mismatch tail’s failure to improve during joint-optimization, 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 intrinsic parameters, the results will always be biased, and we should not expect any improvement, even if we segment the parameter space during training.

figures/ps_q148_qchi.pdf

Figure 6. Parameter space of testing waveforms of q vs. χ_{PN} . We use the recalibrated result from $\mathcal{L}_{\text{mean}}$ with the constant noise spectrum and training waveforms in Tab. 1. Here, the colorbar represents the \log_{10} difference between optimized and original unweighted mismatches.

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. χ_{PN} in Fig. 6. We can see near non-spinning waveforms demonstrate more consistent improvement, whereas spinning waveforms can either improve or worsen in terms of mismatch base on their intrinsic parameters.

Furthermore, 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. We expect to see this feature since the original coefficients were fitted using NR waveforms with equal or similar spin, hence the model prefers waveforms with similar spin. One interesting feature is how the top-left ($\chi_1 < 0$ and $\chi_2 > 0$) and bottom-right ($\chi_1 > 0$ and $\chi_2 < 0$) regions respond to optimization. In the top-left region, waveforms generally improve with along optimization. However, mismatches in the bottom-right region do not improve after optimization. Most waveforms even turned worse after optimization. These waveforms correspond to the waveforms in the high mismatch tail in Fig. 4. [KW: Give justification.]

We divided the parameter space into four regions to analyze the effect of the recalibration procedure on each region separately (Fig.7). The training waveforms used for fitting are listed in Table 3. The top-left and bottom-

figures/ps_q148_chi1chi2.pdf

Figure 7. Parameter space of testing waveforms of χ_1 vs. χ_2 . We use the recalibrated result from $\mathcal{L}_{\text{mean}}$ with the constant noise spectrum and training waveforms in Tab. 1.

right regions have limited data for $q > 4$, hence the results are only valid up to $q \leq 4$. From Fig.9, we observe improvements in waveform mismatch for all waveforms, except those in the bottom-right region. For equal-spin waveforms lying in the proximity of the diagonal axis, the mismatch improves significantly due to most training waveforms being equal-spin. In the top-left region, the improvement is noticeable, except for a few defects caused by some testing waveforms having $q > 4$. The optimized waveforms in the bottom-right region have a higher mismatch than the original ones, indicating poor fitting. Further optimization in a smaller subspace does not improve the results.

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 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, this 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 im-

| 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:0167 | 8.0 | 0.00 | 0.00 |
| <hr/> | | | |
| 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 |
| <hr/> | | | |
| 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 |
| <hr/> | | | |
| 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 equal or similar spins, while the training waveforms for the other two regions are chosen to have opposite spins.

provement 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 for certain regions in the parameter space, leading to mismatches for some waveforms while other waveforms re-

figures/all_quadrants.pdf

Figure 8. Distributions of mismatches after optimizing in separate regions. We use a constant noise spectrum to calculate mismatch and $\mathcal{L}_{\text{mean}}$ for the loss function. Generally, most waveforms with mismatch $> 10^{-2}$ lies in the bottom-right region.

figures/ps_q148_quadrants.pdf

Figure 9. [KW: The color scheme makes the plot a bit hard to read.] Parameter space of testing waveforms. Each region is fitted independently. Colorbar represents log difference of mismatches before and after optimization.

main 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, χ_{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 χ_{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, as shown in Fig. 7. Notably, the ansatz performs better in the top left region than the bottom right along a degeneracy line. In an attempt to address this issue, we partitioned the parameter space into four regions, as described in Sec. 3. However, even with separate optimizations for each region, Fig. 9 indicates that the bottom-right region exhibits similar mismatches as before, while the top-left region’s performance has improved. This observation suggests that the ansatz is specific to certain regions of the parameter space, with a preference for BBH events featuring $\chi_1 < 0$ and $\chi_2 > 0$.

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.

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 (?) or IMRPhenomX (?) 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. Interestingly, while recalibrating the IMRPhenomX model (?), which is parameterized by an additional anti-symmetric spin parameter, it is expected not

to exhibit the degeneracy previously described. 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 (?). 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 `zdeathp` 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, where the result can be seen in Fig. 4. By increasing the number of training waveforms, we see a slight improvement increase in Fig. 5.

Furthermore, we investigated how the intrinsic 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 fourth quadrant cannot be improved. To further test this result, we recalibrate waveforms in separate regions in parameter space. As shown in Fig. 9, this recalibration process gives further improvement to the top-left region while the bottom-right region shows similar result. This indicates that the model’s ansatz does not fit waveforms in this region. This phenomenon is 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