Actively learning numerical relativity

Tomas Andrade[©],[1] Rossella Gamba[©],[2,3,4] and Juan Trenado[©][1]

[1]*Departament de Física Quàntica i Astrofísica, Institut de Ciències del Cosmos, Universitat de Barcelona, Martí i Franquès 1, E-08028 Barcelona, Spain*
[2]*Theoretisch-Physikalisches Institut, Friedrich-Schiller-Universität Jena, 07743, Jena, Germany*
[3]*Institute for Gravitation and the Cosmos, The Pennsylvania State University, University Park, Pennsylvania 16802, USA*
[4]*Department of Physics, University of California, Berkeley, California 94720, USA*

Data analysis of gravitational waves detected by LIGO–Virgo–KAGRA collaboration and future observatories relies on precise modeling of the sources. In order to build, calibrate, and validate current models, we resort to expensive simulations in numerical relativity (NR), the fully-fledged simulation of Einstein's equations. Since simulation costs and the dimensionality of parameter space are prohibitive to perform a dense coverage, approximate models interpolate among the available simulation data. We put forward the technique of Gaussian process active learning (GPAL), an adaptive, data-driven protocol for parameter space exploration and training of gravitational wave approximants. We evaluate this proposal by studying a computationally inexpensive scenario, in which we calibrate the approximant TEOBResumS using the NR-informed model as a proxy for NR. In this case study, we find that GPAL reduces the computational cost of training by a factor of 4 with respect to uniform or randomly distributed simulations. Moreover, we consider a parallel implementation which reduces computational time, and hybrid strategies which improve precalibrated models. The Gaussian process regression employed in this approach naturally endows the algorithm with notion of model uncertainty. We comment on the implications of this feature for data analysis.

I. INTRODUCTION

The detection of gravitational waves by LIGO–Virgo–KAGRA collaboration in 2015 opened a new scientific era [1]. This emerging multidisciplinary endeavour combines detectors of exquisite precision and sophisticated data analysis techniques [2]. Due to the large amount of experimental noise, the most sensitive searches involve matched filtering, in which one compares the data stream to a large template bank of candidate signals [3]. Moreover, the characterization of the physical sources is conducted via parameter estimation, a Bayesian inference technique which also compares the detected signal with theoretical templates [4]. Both aspects highlight the crucial role of the precise mathematical modeling of gravitational waves.

For concreteness, henceforth we restrict ourselves to the case of binary black holes (BBHs) described by general relativity. These are by far the most numerous events detected to date, with 90 BBHs events recorded in the observational campaign O3 [5], a number which will significantly increase with the data currently being collected by O4 [6].

The fully fledged simulation of general relativity is termed numerical relativity (NR), and it is by now a mature field, at least in the absence of matter sources. The first successful simulations of a black hole binary in 2005 [7–9]

paved the way for many developments which have materialized in a series of end-to-end codes [10–13] which allow researchers to carry out these simulations routinely. The complexity of the problem at hand is that of exploring a large parameter space—ten dimensions of intrinsic parameters corresponding to the mass ratio, three relative momenta, and six independent spin components—with highly expensive simulations. As a conservative estimate, we take the average cost for short quasicircular signals of about 10^4 CPU hours [14]. The most common strategy to alleviate this technical challenge is to resort to *approximants*, i.e., analytic or semianalytic[1] approximate models which in turn require calibration by a reduced set of NR simulations. Once properly trained, these models are fast to evaluate and provide a good basis to build template banks. We can distinguish three main families of approximants: (i) phenomenological models [15–31]; (ii) effecive one body [32–43], implemented via two subfamilies,

---

[1]Here by semianalytic we mean models that require solving ordinary differential equations (ODEs). While requiring some numerical computation, the cost of this is several orders of magnitude smaller than that of NR which requires solving partial differential equations (PDEs).

SEOBNRv* [44–52] and TEOBResumS [53–64]; and (iii) NR surrogates [65–69]. The details of the training strategy depend on the family of approximants under consideration. The typical procedure is to construct the training set from preexisting NR databases, which are built relying on domain expert knowledge, e.g., range of parameters believed to be astrophysically relevant, technical feasibility of carrying out the simulations for certain parameters, etc. The test set is also usually chosen from existing NR simulation data, but sometimes other approximants are used for validation in certain regimes. The main NR open databases are the SXS [70–72], RIT [73–75], CoRe [76,77], SACRA-MPI [78,79], Cardiff [80], and MAYA [81] catalogues.

Our goal here is to describe a systematic protocol to choose a series of NR simulations for GW model calibration in order to make them as informative as possible, thus reducing computational costs and alleviating human bias. The key observation is that the problem of exploring parameter space with expensive simulations is precisely the goal of Experimental Design [82,83], a field of statistics which is concerned with the determination of a series of "experiments" (here NR simulations) which can best inform and validate a given model (here the waveform approximants). As we will describe, the methodology of Gaussian process active learning (GPAL), developed in Cognitive Science [84], can be readily adapted to the problem of interest. This approach builds upon the extensive literature in Bayesian optimization experimental design [85] and active learning [86,87] within machine learning.[2] The main feature of this approach is to take a set of existing simulations as Bayesian priors, and suggest the next experiment as a posterior maximizing a certain acquisition function, which captures the degree of new information gained by performing this new experiment. In this context, the natural acquisition function is the variance of the Gaussian process (GP) regression [88–90] used to fit the data, which intuitively captures the idea of maximizing information at the regions of parameter space where the variance is highest. This and related *adaptive* methods have proven useful to increase model accuracy and reduce experimental costs in a variety of contexts [91–96] as we also observe in our problem of interest.

As an interesting byproduct of this approach, the GP variance can be used to construct a measure of model uncertainty. We shall leverage this feature to endow the approximant with a precise notion of local (input specific) waveform uncertainty and error estimates in the time domain. This has been previously noted in [69], which modeled the strain time series as GP and used the variance to provide error estimates in the time domain. The relevance of this model uncertainty for parameter

estimation has been put forward in [97,98], and recently considered in [99,100].

As a concrete example, we will calibrate the EOB model TEOBResumS, taking as a proxy for NR simulations a previously calibrated version of TEOBResumS itself. This will serve as a proof of concept revealing the potential and challenges of the approach, while reducing the computational costs to a minimum. Based on this case study, we estimate that the active approach to build a training set of NR simulations can reduce the computational cost by a factor of 4 with respect to uniform or randomly chosen NR simulations. We also discuss a parallel implementation of the model which helps reduce computational time, and hybrid strategies which could be used to improve previously calibrated models.

This paper is organized as follows. In Sec. II we review the waveform model TEOBResumS, paying special attention to the internal parameters which require NR information. We lay out the methodology in Sec. III, focusing on the general use of the GPAL algorithm for time series, and its applications to GW modeling. We present our results in Sec. IV, showcasing the calibration improvements and introducing the notion of waveform uncertainty. We conclude in Sec. V, summarizing the advantages of our approach. We discuss some limitations of the scope of our case study, and some potential limitations of the algorithm itself in Sec. VI. Finally, we put forward some concrete open problems suggested by the current study in Sec. VII. We collect some relevant details of GP regression and additional numerical results in the Appendices.

II. WAVEFORM MODEL

Our main tool in this study will be the EOB model TEOBResumS, a state-of-the-art approximant for spinning compact binaries [BBHs, binary neutron stars (BNSs), or black hole–neutron star (BHNSs)] coalescing along generic orbits. The EOB approach is based on the mapping of the dynamics of the general relativistic two-body problem into the motion of a test mass in an effective background metric. By further resumming high order post-Newtonian results, and augmenting them via NR information, TEOBResumS can robustly and efficiently generate waveforms throughout the entire inspiral-merger-postmerger phases of a coalescence. There are several versions of this model publicly available. We will base our work on v3.0.0 of TEOBResumS-GIOTTO [55,57,101]. Although this version of the model is not the most developed for the case of quasicircular (QC) orbits, all results obtained below can be straightforwardly applied to the more recent versions of the model. From now on, we shall refer to this model simply as TEOBResumS. Within the QC case, we shall set the precession to zero, which implies that the black hole spins are aligned with the orbital angular momentum. Therefore, the main physical input parameters for TEOBResumS are the mass ratio $q = m_1/m_2$ (where m_1, m_2 are the black hole masses), and the two dimensionless spin parameters χ_1, χ_2,

---

[2]Note that this is unrelated to the Bayesian inference approach employed in parameter estimation for gravitational waves.
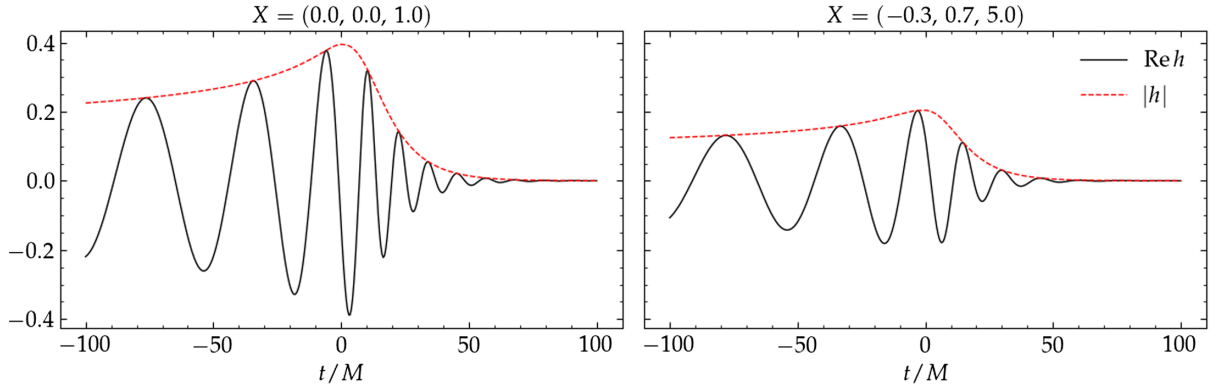
FIG. 1.   Selected waveforms with initial data $X = (\chi_1, \chi_2, q)$ for a spinless binary with $q = 1$ (left) and a spinning binary with $q = 5$ (right).

which take values between -1 and 1, where the sign accounts for the orientation of the spin with respect to the orbital angular momentum. The code can generate waveforms for any value of mass ratio $q \geq 1$ and spins $|\chi_{1,2}| \leq 1$, and has been validated against NR simulations up to values of $q = 128$ [102]. In addition, we can specify an initial frequency f_0—or alternatively an initial radial distance r_0—which controls the starting point of the dynamics. As such, this quantity is not an intrinsic parameter of the source and does not stand on the same footing as the mass ratio or spins. We set the starting point to $r_0 = 14M$ in geometric units without loss of generality, corresponding to an initial frequency of approximately 0.006. For given values of the initial data (χ_1, χ_2, q), the code outputs a set of complex time series, denoted as modes, labeled by (ℓ, m) referring to the spherical harmonic decomposition. Here we will focus on the leading mode $(\ell, m) = (2, 2)$, and omit the mode indices henceforth.[3] We show some selected examples of the waveforms in Fig. 1.

As mentioned above, the model depends on some analytically known parameters (e.g., the EOB potentials expressed as resummed post-Newtonian expansions), and some free parameters which are fixed by calibrating (or informing) the model by comparing the resulting waveforms to a set of NR simulations taken as "ground truth," i.e., strictly accurate and precise.[4] More precisely, denoting the initial data collectively as X, NR produces the ground truth signals

$$h_{gt}(X, t) = h_{NR}(X; t). \tag{1}$$

Denoting the internal model parameters which require calibration as $c_i(X)$, the EOB model produces an approximate waveform $\tilde{h}(X, t)$,

$$\tilde{h}(X, t) = h_{EOB}(c_i(X); t). \tag{2}$$

The internal parameters of the model are given by smooth functions of the mass ratio and spins, which typically take the form of ratios of polynomials. In this work we will focus on six internal parameters $c_i = \{A_{mrg}, \omega_{mrg}, A_{NQC}, \omega_{NQC}, \dot{A}_{NQC}, \dot{\omega}_{NQC}\}$ (note that we suppress mode indices since we focus on the 22 mode). The parameters A_{mrg}, ω_{mrg} correspond to the amplitude and frequency at merger, while the remaining parameters ensure that the waveforms are smooth near the peaks; see, e.g., Refs. [58,101] for a precise definition. The dependence of the model parameters with the initial data is given by smooth, slowly varying functions, which can be written as products of ratios of polynomials that contain arbitrary coefficients to be fixed upon informing the model with NR; see Appendix A.

Calibrating or informing the model amounts to

(i)   Choosing a set of points $\{X\}^{train}$, the *training set*, representative of some region of parameter space the phenomenology of which we want to capture.

(ii) Extracting the values of the internal parameters $c_i(\{X\}^{train})$.

(iii) Providing global fits $c_i(X)$ which interpolate among the $c_i(\{X\}^{train})$.

While minimizing one or more target metrics which attempt to capture how similar $\tilde{h}(X, t)$ and $h_{gt}(X, t)$ are evaluated over some test set $\{X\}^{test}$. A common choice of such metrics is the unfaithfulness or mismatch (see Appendix C). In addition to the unfaithfulness, TEOBResumS aims to minimize the phase and amplitude differences at merger [64], since it has proven useful to

---

[3]The relative importance of the subleading modes depends on the intrinsic parameters, being higher for high mass ratios. For the highest mass ratio considered here, $q = 6$, the maximum value of the amplitude of the (2,2) mode is 3.5 times larger than the first subleading mode (3,3). Extra information about the extrinsic parameters such as binary orientation is needed to assess the impact of the higher modes in the full strain computed as the sum of modes weighted by the corresponding spherical harmonics. In any case, all methods developed here equally apply to subleading modes, with little modification.

[4]There are of course sources of inaccuracy and imprecision in NR waveforms, but these are usually assumed to be under control and much smaller than those of the approximants.

incrementally improve different aspects of the model. The version of the model used here has been informed with about 130 simulations; see [58] for details.

III. METHODOLOGY

As stated above, calibration of a given model requires carrying out NR simulations in a judiciously chosen training set $\{X\}^{\text{train}}$. Here we consider GPAL, an implementation of Bayesian optimal experimental design which deploys active learning in combination with Gaussian process regression. In particular, this strategy selects the training set $\{X\}^{\text{train}}$ iteratively, maximizing the variance of the Gaussian process fits performed on the parameter space after each iteration. This captures the intuitive idea that every new simulation is the one that provides more "information" about the system.

A. GPAL for time series

Let us briefly describe the GPAL algorithm and the relevant modifications we introduce for our study. For a given target $f(X)$—the function to be fitted or "learned"—defined on a domain $D(X)$, the GPAL algorithm can be summarized as follows. As the initial step, give prior values $\tilde{f}(X)$ and an initial variance $\sigma(X)$. Then, iterate over the following:

(1) Evaluate f at the maximum of $\sigma(X)$, X_0.

(2) Taking into account the newly extracted value $f(X_0)$, compute the posterior of $f_{\text{post}}(X)$ as the Gaussian process regression of the function on $D(X)$, with new variance $\sigma_{\text{post}}(X)$.

(3) Update $\tilde{f}(X) = f_{\text{post}}(X)$ and $\sigma(X) = \sigma_{\text{post}}(X)$.

Thus, after each iteration, we obtain a training set $\{X\}^{\text{train}}$, and approximate function $\tilde{f}(X)$ and a variance $\sigma(X)$. We have used GP regression with kernel (covariance function) Matern52, with $\sigma = 1$, implemented in the library GPY [103]; see Appendix B for details.

In the application at hand, we use GPAL to evaluate the internal model parameters $c_i(X)$ which play the role of the function $f(X)$ in the discussion above.[5] This has allowed us to deploy this strategy to the case of time series. Note that, as opposed to [69,97,98], we are not describing the time series as GP, but as given by the fixed (time) functional forms specified by the model at hand from the onset.

The algorithm easily allows for other ways of selecting the evaluation points, or *acquisition functions*, instead of maximizing $\sigma(X)$. One that appears particularly useful in the case where there is high variability in a particular region of parameter space is the norm of the gradient of f (or even

$\tilde{f}$, to reduce evaluation costs). We leave this investigation for future work.

B. EOB as proxy for NR

In order to carry out an in-depth study of the calibration process with minimal computational cost, we shall take EOB as a proxy for NR. More specifically, we take the NR informed approximant TEOBResumS as the ground truth, i.e.,

$$h_{gt}(X, t) = h_{\text{EOB}}(c_i(X); t), \tag{3}$$

and use it to calibrate an uninformed version of the model

$$\tilde{h}(X, t) = h_{\text{EOB}}(\tilde{c}_i(X); t). \tag{4}$$

Note that the ground truth and approximate models in (3) and (4) only differ by the values for the internal parameters, which are chosen as the calibrated values $c_i(X)$, and initialized to random values $\tilde{c}_i(X)$. As mentioned above, we only change the calibrated values of the internal parameters corresponding to $c_i = \{A_{\text{mrg}}, \omega_{\text{mrg}}, A_{\text{NQC}}, \omega_{\text{NQC}}, \dot{A}_{\text{NQC}}, \dot{\omega}_{\text{NQC}}\}$, keeping the remaining parameters fixed to the previously calibrated values. Since these affect more strongly the late inspiral and ring down of the waveforms, we restrict our study to a time interval $(-100M, 100M)$ centered at the peak of the absolute value of each signal.

We extract the internal parameters for the EOB approximant simply by calling the output of the code. This eliminates uncertainties in the extraction, which are present in a realistic case.[6] We will discuss some of the limitations of our approach in Sec. VI.

C. Calibration strategies

To benchmark the GPAL strategy, we consider two alternative approaches consisting of uniformly spaced grids and random grids in parameter space. We will compare the results provided by these different choices using the same number of training points in each approach. We choose the input parameters in the range $-0.85 \leq \chi_{1,2} \leq 0.85$, $1 \leq q \leq q_{\text{max}}$, and study how the algorithm behaves as we vary $2 \leq q_{\text{max}} \leq 6$. We discretize the space of parameters by choosing a uniform grid with 64^3 points in the directions χ_1, χ_2, q.

We initialize the active algorithm by choosing flat priors for all internal parameters $c_i(X) = 1$ and their variance $\sigma(X) = 1$, and picking the first iteration to be a random point. We observe that the subsequent iterations tend to lie at the boundaries of the computational domain. After a

---

[5]We have done so by fitting each parameter individually as a function of X. Other "multioutput" choices are possible in which one can take into account correlations among the various functions to be approximated [104].

[6]For example, in order to extract the amplitude of the waveform at merger, we would have to evaluate the value of the waveform at the amplitude peak. It turns out that there are minor differences between the direct output value and the extracted value, which introduce extra difficulties in the calibration.

TABLE I. Subsample step size in each direction of the grid, number of training points, and number of test points for each epoch of our numerical experiments.

$1d$ subsample	32	22	16	13	11	9	8
N_{train}	27	64	125	216	343	512	729
N_{test}	64	125	216	343	512	729	1000

certain number of iterations, internal points get selected.[7] We choose the uniform training grids by specifying points on the boundary of the domain and subsampling each direction by the same number of points, e.g., for a 1d grid of 64 points, a subsampling of 32 yields three equally spaced nodes in each direction, so that the number of training points is 3^3. The points of the random grid are selected without restrictions, so they may or may not cover the regions near the edges of the domain. The test set for each numerical experiment is chosen to be the reciprocal lattice of the corresponding homogeneous training grid, including boundary points equidistant from it.

In this setup, we monitor how the performance and uncertainty of the algorithm changes as we vary the number of training points. For our numerical experiments we take the subsampling of the uniform grid, number of training points, and size of test set as given in Table I. We depict the resulting training sets for $N = 125$ in Fig. 2.

D. Model assessment

At the end of each calibration round, we compute various performance and uncertainty metrics, as we now describe. For a given point in the test set X, we extract the ground truth model parameters $c_i(X)$ and the fitted values $\tilde{c}_i(X)$, and evaluate the ground truth and approximate signals using (3) and (4). We then align—adjust the relative time and phases—the so-obtained signals with respect to one another using the Python package PyCBC [105].

The main performance metric we consider is the unfaithfulness between the ground truth and approximate signals

$$\bar{\mathcal{F}}(X) = \bar{\mathcal{F}}(h_{gt}(X), \tilde{h}(X)). \quad (5)$$

This is a standard quantity to characterize the notion of "distance" between two waveforms in the context of GW data analysis. In addition, we consider the amplitude $\delta A(X)$ and phase differences $\delta\phi(X)$ at merger between the ground truth and approximate signals at a given point in parameter space X, similarly to (5). As argued in [64], these merger

---

[7]Since we take $\sigma(X)$ to be the acquisition function and initialize the algorithm with flat priors, the selected training points and fit variance are independent of the functions being fitted; see Appendix B. In particular, this means that the variance is the same for each internal parameter c_i.
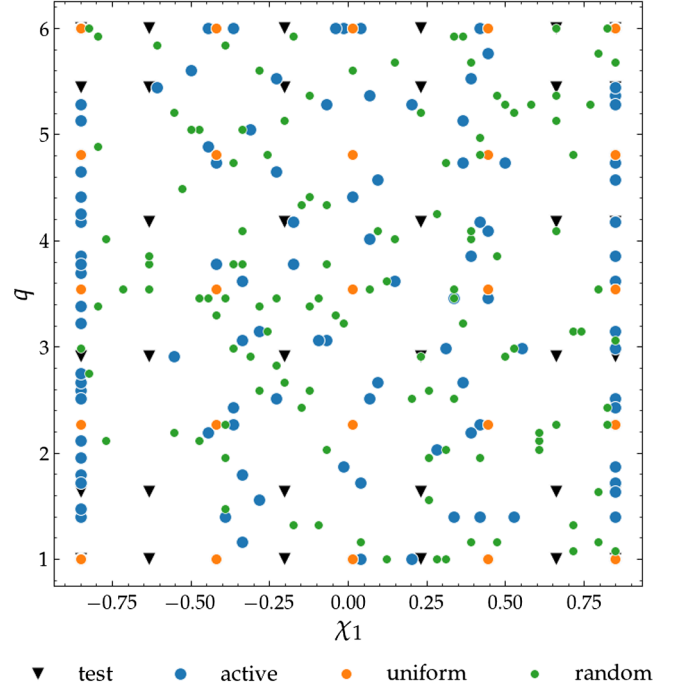

FIG. 2. Training and test sets for $N = 125$, projected on the 2D slice (χ_1, q).

differences provide useful guidance in waveform modeling. See Appendix C for the precise definition of these quantities. These metrics are chosen such that a model with perfect performance has $\bar{\mathcal{F}}(X) = \delta A(X) = \delta\phi(X) = 0$ across the entire test set.

The variance of the Gaussian fits, $\sigma(X)$ endows the algorithm with a natural measure of uncertainty at each point in parameter space. In order to make this more easily interpretable in the context of GW data analysis, we introduce the notion of *waveform uncertainty* $u(X)$ by considering the unfaithfulness between the approximate signal at zero variance and the most uncertain signal given the variance of all internal parameters, $\tilde{h}^{(\alpha_0)}(X)$; see Appendix C for a precise definition. The waveform uncertainty can be expressed as

$$u(X) = \bar{\mathcal{F}}(\tilde{h}(X), \tilde{h}^{(\alpha_0)}(X)). \quad (6)$$

Finally we consider the fit accuracy, defined as the Euclidean norm of the difference between the ground truth values of the parameters and the fitted ones, which we denote as $\delta c(X)$.

We will typically be interested in the average values of these metrics on the test set, which we denote as

$$\langle M \rangle = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} M(X_i) \quad (7)$$

with M being any of the aforementioned metrics, and i running over the N_{test} points in the test set $\{X\}^{\text{test}}$.
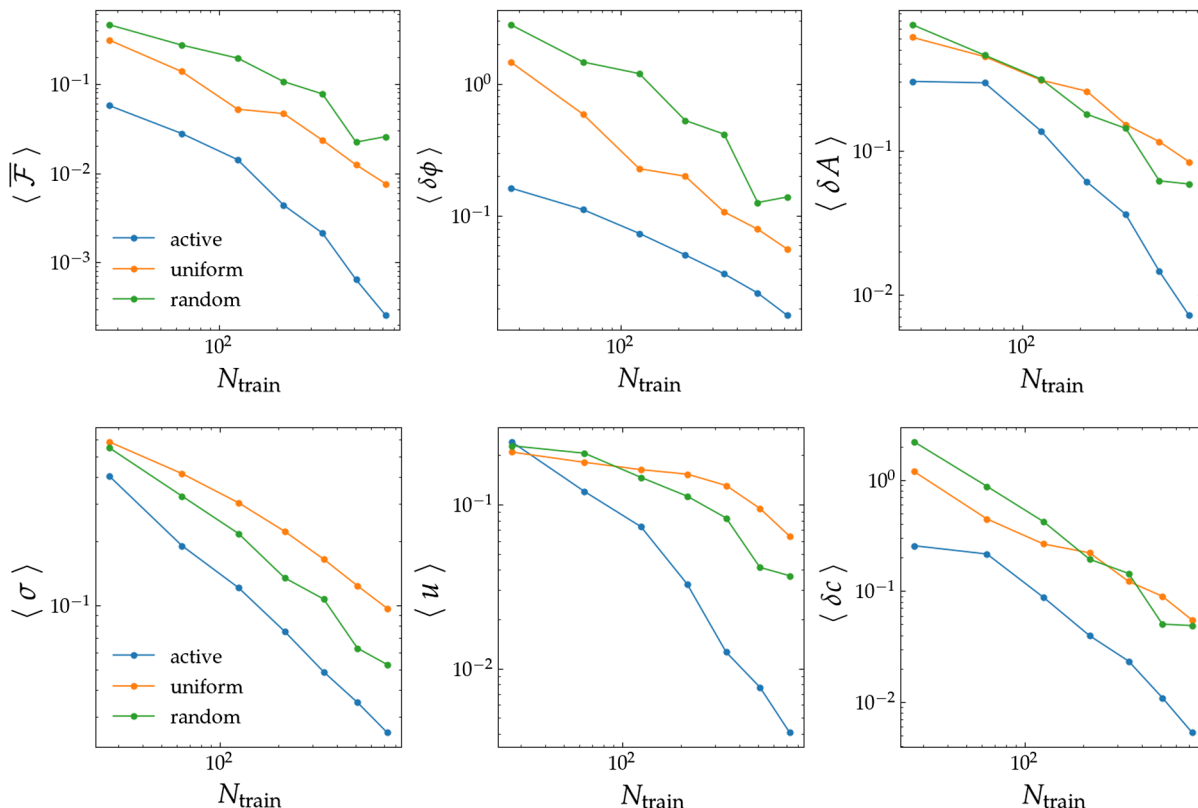
FIG. 3. Dependence of the average performance (top) and uncertainty metrics (bottom) with number of training points. All plots are in log-log scale.

IV. RESULTS

A. Performance and calibration cost

We compare the three training strategies, active, uniform, and random, by carrying out each training with a given number of points, and evaluating the performance and uncertainty metrics.

We begin discussing the average values of the metrics over the test set, constructed as explained in Sec. III D. In Fig. 3 we show the dependence of all metrics with the number of training points for $q_{max} = 6$. We clearly appreciate that the active training strategy outperforms the uniform and random strategies. Interestingly, we observe an approximate power-law dependence with N_{train} in most of these quantities, with the exception of those associated to random training. We expect that averaging over realizations would bring the dependence of these quantities closer to power laws. We obtain qualitatively similar results for smaller values of q_{max}, although the performance difference between the active and other training protocols increases with q_{max}.

Taking as a reference target value $\bar{\mathcal{F}} = 0.01$, we interpolate the performance curves and extract the number of training points $N_{train}^{(type)*}$ required to achieve this performance target, with type = active, uniform, random. We can then estimate the cost reduction provided by the active strategy

with respect to the uniform/random ones, by computing the number of training points required to achieve the corresponding targets, and taking their ratio.[8] Carrying out the same analysis with the amplitude and phase differences at merger with reference values $\delta A = 0.1$, $\delta \phi = 0.06$ yields the results in Fig. 4. We observe a maximum cost reduction of around 4 for all the metrics considered.

We have also studied the dependence of the various metrics as a function of the location in parameter space. We observe that performance and uncertainty metrics deteriorate significantly near the edges of the domain; see Appendix E.

We consider extra alternative test and training sets in Appendix D. In particular, we monitor the effect of altering the test set by removing boundary points and making it fully random. Moreover, we consider a training set given by Chebyshev nodes along all directions, allowing a more dense coverage of the regions near the boundary. In all cases we see robust results for the supremacy of the active training, with a cost reduction of roughly 4 times in most cases.

---

[8]Note that this estimate assumes that the cost of all simulations is the same, which is not the case in a realistic scenario; see Sec. VI.
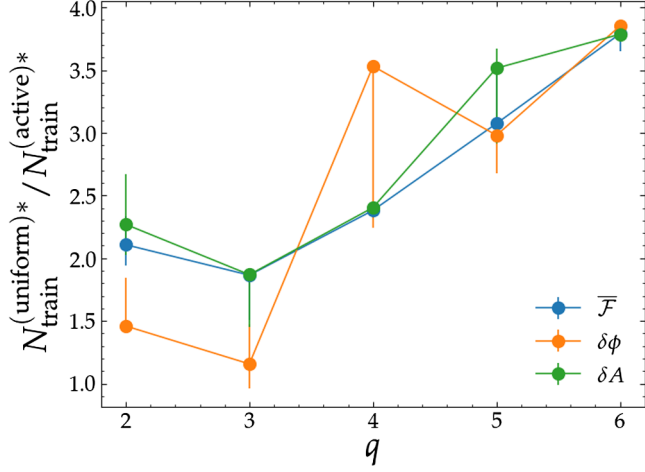
FIG. 4. Average performance gain for all performance metrics, as a function of q_{max}. We obtain the error bars for each data point by varying the baseline values by $\pm 30\%$, computing all intercepts, and keeping the highest and lowest values.

B. Parallelization and calibration time

In principle, the GPAL strategy as described above should be followed serially. However, it is possible to parallelize it by partitioning the computational domain to obtain the training points, and performing a global fit at the end. For concreteness, we partition the domain in $N_q = 17$ two-dimensional slices of constant q, corresponding to approximately $\{q^i\} = 1 + 5/16 i$, with $i = 0 \ldots 16$, explore each slice *in parallel* with $N_{train}^{(2d)} = 6, 8, 10, 12, 14,$ 16, 18 simulations, and finally perform a global GP regression in the 3D domain. We then study the performance of the algorithm as we vary the total number of simulations $N_{train}^{\parallel} = N_q N_{train}^{(2d)}$ keeping N_q fixed. We show our results in Fig. 5. We observe that in order to achieve the
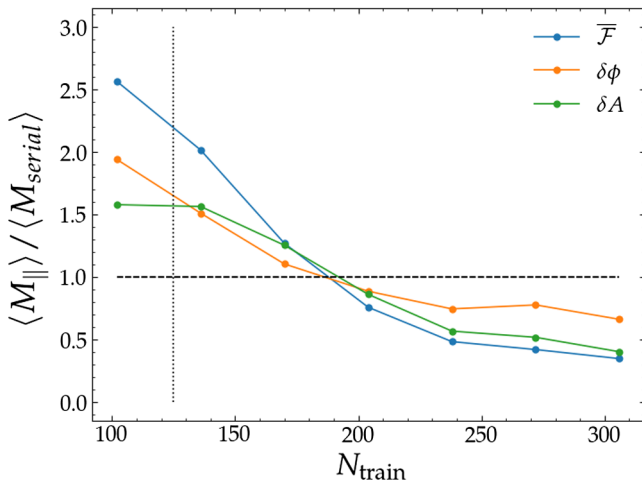

FIG. 5. Relative performance between the parallel and serial versions of our active training strategy, expressed as ratios of the performance metrics versus number of training points.

same performance as the one with $N_{train} = 125$ we require about $N_{train} = 187$ (corresponding to $N_{train}^{(2d)} = 11$), i.e., around 50% more simulations to reach the same performance considering all metrics.

Let us call the computation time per simulation t_{sim}. In Sec. IV A we found that the performance achieved with $N_{train}^{active} = 125$ is comparable to the one resulting from $N_{train}^{uniform} = 512$. Assuming that we can carry out $N_{train}^{(2d)} = 17$ simulations in parallel, the time it takes to calibrate the model with a uniform grid is $T_{uniform} = 512/17 t_{sim} \sim 30 t_{sim}$. On the other hand, since we found that $N_{train}^{(2d)} = 11$, building the training set in the parallel implementation takes $T_{active}^{\parallel} = N_{train}^{(2d)} t_{sim} = 11 t_{sim}$, approximately one third of the time required by the uniform grid calibration.

C. Hybrid strategy and model improvement

The active approach can be also deployed to improve a model previously calibrated with a different strategy, as we will now illustrate.

For concreteness, consider the case $q_{max} = 6$ discussed in Sec. IV A, focusing on the active and uniform grid cases. Instead of utilizing the active strategy from the onset, we can take as a starting point for the active algorithm a model calibrated with $N_{train} = 125$ using a uniform grid, and then switching to active training.

We see the outcome of this hybrid procedure in Fig. 6, which also shows for comparison the purely active and purely uniform results. We observe that the hybrid training quickly catches up with the active one, and even outperforms it for large training sets. For the average unfaithfulness, we see an order of magnitude performance gain for $N_{train} = 216$ of the hybrid protocol over the uniform one, and almost a two order of magnitude gain for $N_{train} = 729$. The phase and amplitude differences at merger show a significant but slightly lower increase.

D. Waveform uncertainty

As explained above, the variance of the Gaussian fits provides a natural way of capturing the model uncertainties across parameter space. Moreover, this can be translated into waveform uncertainties as explained in Sec. III D. We have seen in Fig. 3 that the average waveform uncertainty as a function of training points decreases much faster with the active approach compared to uniform and random grids. In particular, we see that only for the active approach this uncertainty goes below the target unfaithfulness $\bar{\mathcal{F}} = 0.01$ in the ranges of parameters we have considered in this study.

In Appendix E we show a histogram of the waveform uncertainties resulting from all training strategies at the maximum number of training points $N_{train} = 729$. This again shows the better accuracy of the active approach, with only very few cases exceeding $\bar{\mathcal{F}} = 0.01$.
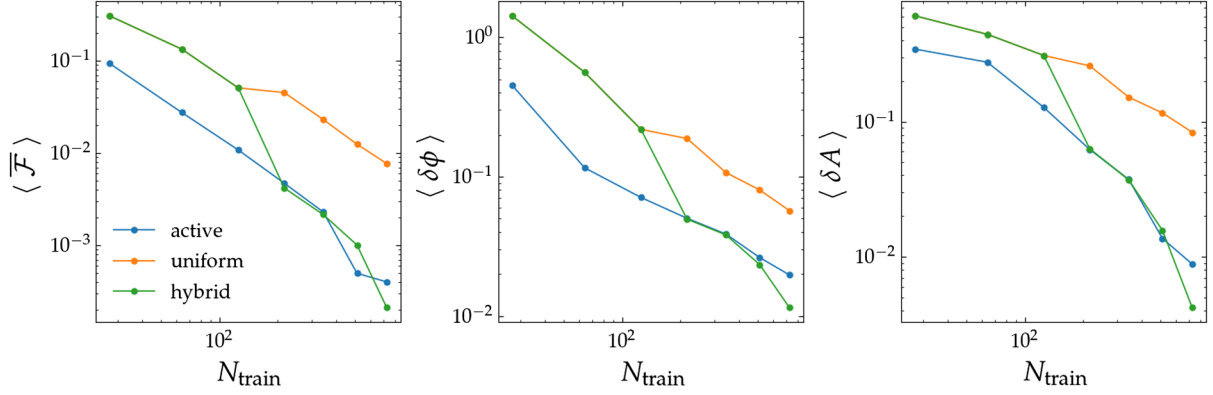
FIG. 6.   Average performance metrics as a function of the number of training points for active, uniform, and hybrid strategies. In the hybrid case, we initialize the training with the same prior as for uniform training with $N_{\text{train}} = 125$ (hence the curves coincide up to that point), and switch to active training.
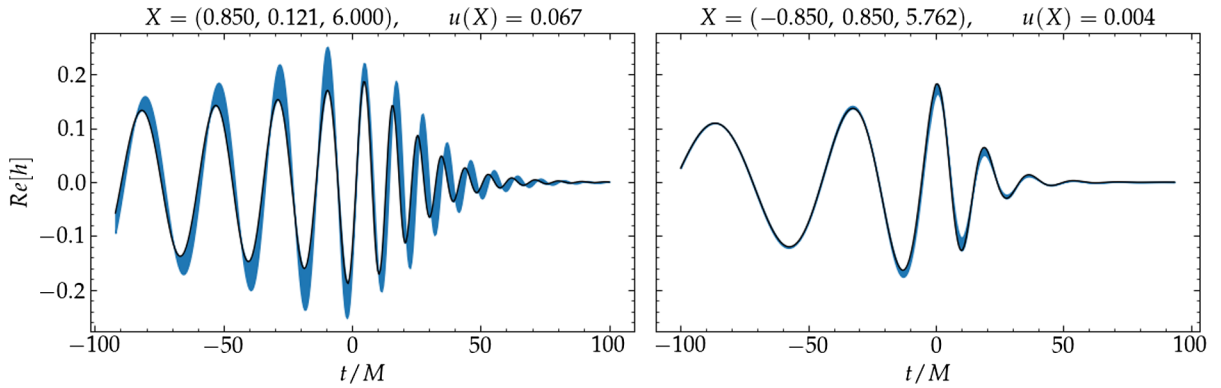

FIG. 7.   Predicted waveform and the corresponding most uncertain waveform (as defined in Appendix C) for the maximum (left) and average value (right) of the waveform uncertainty $u(X)$ in the case $q_{\text{max}} = 6$. We shade the region between both waveforms to ease visualization.

Using this notion of uncertainty, we can provide error bars for the waveforms in the time domain, as shown in Fig. 7. We leave for future work a careful consideration of the implications of this feature for data analysis.

V. CONCLUSIONS

We propose the use active learning to systematically construct the training sets upon which waveform approximants are built. More precisely, we put forward the idea of adapting the technique GPAL to time series modeling. This involves performing a GP regression to fit the internal parameters from available NR data, and using the variance of the GP as the acquisition function to explore NR parameter space maximizing the information gained from each simulation. Furthermore, we introduce the notion of waveform uncertainty, which is built out of the variance of the GP and provides a local, quantitative, and intuitive measure to endow the approximants with error estimates.

This stands in contrast with the standard techniques used for model calibration, in which a fixed set of NR simulations is chosen based on domain knowledge (e.g., priors of astrophysical relevance, ease to carry out NR simulations), and global fits of the internal parameters are performed by postulating fixed functional forms. Moreover, with the notable exception of some surrogate models which also employ GP regression, e.g., [69], most of the available approximants typically only provide estimates of global inaccuracy (e.g., maximum unfaithfulness over a chosen test set) and do not provide local uncertainty estimates of the waveforms.

Our strategy appears to be advantageous in the following ways:

(i) For a fixed performance target in various metrics, it requires fewer simulations compared to nonadaptive methods. We observe a cost reduction by a factor of up to 4 in the simple model considered here. Moreover, larger gains occur for larger parameter spaces.

(ii) It is amenable to parallelization. In our case study discussed in Sec. IV B, we observe that in order to achieve the same performance as the serial active

algorithm, the parallelized version requires more simulations (about 50% in our examples), but this can still accomplish a speed up with respect to a uniform grid calibration similarly parallelized (up to a factor of 3 in our examples).

(iii) It can be used to improve models initially calibrated using a different strategy, say, with a uniform grid. In our example in Sec. IV C, we observe an order of magnitude performance improvement when switching from uniform to active training.

(iv) It provides a quantitative, local, and easy to interpret the measure of uncertainty for the waveforms across parameter space. We moreover find that the average uncertainty over test set decreases as a power law with the number of simulations used for training and is significantly smaller for active training than for uniform or random grids.

(v) By making the construction of the training set systematic, it reduces human bias in the construction of the approximants. While it is true that there is certain freedom in the choice of the acquisition function and other implementation details, the criteria for doing so can be succinctly expressed in an intuitive way.

(vi) The interpolation algorithm easily accommodates complex data patterns as it is based on GP regression, a nonparametric algorithm. In the simple case considered, we see a power-law decrease of the fit inaccuracy with the number of simulations. We expect this feature to have the highest impact in cases which display intricate data patterns. Note that in order to fully exploit this feature, the acquisition function used should favor regions of higher parameter variation. We leave a careful study of this aspect for future work.

(vii) Our framework can be seamlessly integrated with different GW approximants. More generally, our approach can be deployed in conjunction with any model of temporal series.

VI. DISCUSSION

Only after the application of the proposed technique to more realistic scenarios we shall be able to properly evaluate its merits. With this in mind, let us discuss some potential limitations we identify in our study:

(i) We assume that the model can perfectly describe the simulated data. We have made this choice since it allowed us to disentangle the effect of calibration from model improvement. However, in more realistic applications, there could be difficulties arising from model incompleteness or incorrectness, which are known to affect Bayesian adaptive experimental designs [106]. This limitation appears to have the lowest impact in the case of NR surrogates, since they are the most agnostic models, although they require larger training sets.

(ii) We use GP regression to fit the data for all choices of training sets. It is possible that for some specific models, using parametric fitting procedures (say, based on ratios of polynomials) could yield more accurate waveforms in some regions of parameter space. In particular, it might be that these simpler fits allow for extrapolation outside of the training region. However, we do not expect this to be a robust feature, since it is in particular contingent to the choice of specific model parameters.

(iii) Our computational cost and time estimates are based on the total number of simulations required by each training strategy. It is well known, however, that some simulations are more costly than others (e.g., quasicirculars with higher mass ratios). Since our main results involve relative costs, we expect that our conclusions should hold even when more precise cost estimates are taken into account.

(iv) We are assuming that the extraction of the model parameters from NR data is exact, since we read it off from the code output. In practice, this process involves some extra manipulations such as computing time derivatives of the signals, performing fits in the time domain, etc. Having available NR data of sufficient quality should reduce the impact of this limitation.

Let us also consider some possible scenarios in which GPAL may not be most efficient when applied to GW modeling:

(i) Given enough resources to parallelize the data collection of a large training set, our algorithm will eventually require more computational time than other nonadaptive approaches. This is an unlikely scenario given the high processing and memory requirements of the simulations. Moreover, as explained in Sec. IV B, this can be largely overcome by parallelizing the GPAL algorithm.

(ii) We have observed that using the GP variance as acquisition function tends to select points at the edges of the computational domain, which is typically where simulations are most challenging to carry out. For instance, the cases of large q and large spins $|\chi| \sim 1$ typically become more computationally expensive that those for moderate values of these parameters. We thus recommend the use of this algorithm in regions of parameter space in which all simulations are feasible to carry out. Furthermore, our setting does not allow for extrapolation outside of the training domain.

VII. OUTLOOK

We identify several research questions which require further investigation:

(i) Our results suggest that it is possible to efficiently improve existing models of GW by enlarging their

calibration set of NR simulations as suggested by the active learning algorithm. It would be interesting to quantify this potential improvement for the available GW approximants for quasicircular binary black hole (BH).

(ii) The case of BH dynamical captures (or highly eccentric binaries) [107–109] has been much less developed than that quasicircular case, but recent progress has been made [60,110,111]. In this scenario, the two BHs undergo a series of close encounters before merger, resulting in a much richer phenomenology which involves a larger parameter space, and displays significant complexity in the model parameters. GPAL could be useful to guide parameter exploration in this and similar scenarios.

(iii) The notion of waveform uncertainty has been argued to play an important role in parameter estimation [97,98]. It would be interesting to extend these analyses to other models in which this notion was previously not incorporated, e.g., EOB approximants. Moreover, it could serve as a statistical weight to build model ensembles or meta models combining several approximants, see, e.g., [112,113].

We hope to address some of these open problems in the near future.

ACKNOWLEDGMENTS

APPENDIX A: EOB PARAMETERS

In the model TEOBResumS, the EOB parameters discussed in the main text $c_i = \{A_{\mathrm{mrg}}, \omega_{\mathrm{mrg}}, A_{\mathrm{NQC}}, \omega_{\mathrm{NQC}}, \dot{A}_{\mathrm{NQC}}, \dot{\omega}_{\mathrm{NQC}}\}$ are taken to be functions of the following form

$$Y = Y^0 Y^{\mathrm{orb}}(\nu) Y^{\hat{S}}(X_{12}, \hat{S}) \tag{A1}$$

for $Y = \{A_{\mathrm{mrg}}, \omega_{\mathrm{mrg}}, A_{\mathrm{NQC}}, \omega_{\mathrm{NQC}}\}$, and

$$Z = Z^{\mathrm{orb}}(\nu) + Z^{\hat{S}}(X_{12}, \hat{S}) \tag{A2}$$

for $Z = \{\dot{A}_{\mathrm{NQC}}, \dot{\omega}_{\mathrm{NQC}}\}$.[9] In the expressions above $X_{12} = (m_1 - m_2)/M$ and $\hat{S} = (S_1 + S_2)/M^2$, and we denote with a "0" superscript the test-particle limit, while the "orb" and "$\hat{S}$" superscripts indicate the orbital and spin part respectively. We further have that

$$Y^{\mathrm{orb}} = 1 + a_1^Y \nu + a_2^Y \nu^2, \tag{A3}$$

$$Y^{\hat{S}} = \frac{1 + b_1^Y \hat{S} + b_2^Y \hat{S}^2}{1 + b_3^Y \hat{S}}, \tag{A4}$$

$$b_i^Y = \frac{c_{i,0}^Y + c_{i,1}^Y X_{12}}{1 + c_{i,2}^Y X_{12}}, \tag{A5}$$

$$Z^{\mathrm{orb}} = 1 + a_1^Z \nu + a_2^Z \nu^2, \tag{A6}$$

$$Z^{\hat{S}} = (b_1^Z + c_1^Z X_{12}) \hat{S} + (b_2^Z + c_2^Z X_{12}) \hat{S}^2. \tag{A7}$$

The set of quantities fit to NR are therefore the $\{a^Y, c^Y, a^Z, b^Z, c^Z\}$ coefficients, where subindices have been dropped for brevity.

APPENDIX B: GP REGRESSION

In this appendix we collect some useful information about GP regression closely following [89]. A GP is a set of random variables, any of which have a joint Gaussian (multivariate, normal) distribution. More concretely, if we consider a set of inputs $\{X\}$, we assume that a function f takes values $f(X)$ which are normally distributed with some mean $m(X)$ (which for notational simplicity we take to be 0) and a covariance function (also known as kernel) $K(X, X)$ which is symmetric and positive definite.

$$f \sim \mathcal{N}(0, K(X, X)) \tag{B1}$$

where $\mathcal{N}$ is the normal distribution.

Similarly, for unobserved points X_* (say, in the test set), the values $f_* := f(X_*)$ are distributed normally with the same kernel. The joint distribution of the observed (training) and unobserved (test) points is given by the prior

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K & K_* \\ K_* & K_{**} \end{bmatrix}\right) \tag{B2}$$

where $K = K(X, X)$, $K_* = K(X_*, X)$, $K_{**} = K(X_*, X_*)$. To get the posterior distribution used for regression, we condition over the observed points, which after some manipulations, yields

$$f_* | X_*, X, f \sim \mathcal{N}(K_* K^{-1} f, K_{**} - K_* K^{-1} K_*). \tag{B3}$$

---

[9]In practice, the leading order multipolar behavior of the amplitude is factorized out before fitting; see Appendix 2 of [58]. Similarly, it was found that it is more convenient to fit $Z/(\nu\omega_{22}^{\mathrm{NQC}})$, rather than Z directly.

This equation expresses that the distribution of the function on the test set values f^*, conditioned over the test set values X_*, the training set points X, and the values of the function f, is normally distributed with the mean and variance indicated on the right hand side. Thus, the predictions and variance of the GP regression are the mean and variance of the distribution in (B3). Note that the variance only depends on the training and test sets, but not on the function being sampled.

In our numerical experiments, we have used the kernel denoted Matern52 with auto relevance determination, which is given by the expression

$$k(r) = \sigma^2 \left(1 + \sqrt{5}r + \frac{5}{3}r^2 \right) \exp(-\sqrt{5}r) \quad \text{(B4)}$$

with

$$r^2(x, x') = \sum_{q=1}^{Q} \frac{(x_q - x'_q)^2}{\ell_q^2} \quad \text{(B5)}$$

where the ℓ_q are relative length scales automatically determined for each direction and Q the number of dimensions in parameter space. We fix $\sigma = 1$ throughout. Note that the correlation decays exponentially with the distance between inputs, which ensures the smoothness of the output functions. We have used the implementation provided by the Python library GPy [103].

APPENDIX C: PERFORMANCE AND UNCERTAINTY METRICS

Here we provide details regarding the performance and uncertainty metrics used in the main text. We begin with the faithfulness, or match, which is defined as

$$\mathcal{F} = \frac{\langle h_1, h_2 \rangle}{\sqrt{\langle h_1, h_1 \rangle \langle h_2, h_2 \rangle}} \quad \text{(C1)}$$

where the inner product between two waveforms h_1, h_2 is given by

$$\langle h_1, h_2 \rangle = 4\Re \int \frac{\hat{h}_1(f)\hat{h}_2^*(f)}{S_n(f)} \quad \text{(C2)}$$

where $\hat{h}_{1,2}(f)$ are the Fourier transforms of the time domain signals. Throughout this paper we consider a uniform power spectral distribution $S_n(f) = 1$, but more generally this is where the noise curve of a particular GW detector enters the calculation.[10] The unfaithfulness is then given by $\bar{\mathcal{F}} = 1 - \mathcal{F}$.

---

[10]The case $S_n(f) = 1$ is usually considered in GW modeling since it allows us to evaluate the model in a way independent of details of different detectors.

We also consider the amplitude and phase difference at merger. To obtain these, we decompose the signal in amplitude and phase as

$$h_{gt}(X, t) = A_{gt}(X, t)e^{-i\phi_{gt}(X,t)} \quad \text{(C3)}$$

$$\tilde{h}(X, t) = \tilde{A}(X, t)e^{-i\tilde{\phi}(X,t)} \quad \text{(C4)}$$

and compute the difference at merger, defined as the amplitude peak of each signal,

$$\delta A(X) = (A_{gt}(X, t_{\text{peak}}) - \tilde{A}(X, t_{\text{peak}}))/A_{gt}(X, t_{\text{peak}}) \quad \text{(C5)}$$

$$\delta\phi(X) = \phi_{gt}(X, t_{\text{peak}}) - \tilde{\phi}(X, t_{\text{peak}}). \quad \text{(C6)}$$

The most uncertain signal for a given variance is defined by first computing the family of waveforms

$$\tilde{h}^{(\alpha)}(X, t) = h_{EOB}\left(\tilde{c}_i(X) + \frac{1}{2}\Delta_i^{(\alpha)}\sigma(X); t \right) \quad \text{(C7)}$$

where $\Delta_i^{(\alpha)}$ account for all possible sign combinations of internal parameters, e.g., $\{+, +, +, +, +, +\}$, $\{-, +, +, +, +, +\}$, etc. For the case considered here having six internal parameters, we need to compute $2^6 = 64$ different waveforms. Note that this captures the notion of directional derivative of the waveform with respect to the c_i. We then select the most uncertain signal as the representative $\tilde{h}^{(\alpha_0)}(X, t)$ that maximizes the norm $|\tilde{h}^{(\alpha)}(X, t) - \tilde{h}(X, t)|$, given by

$$|h_1(t) - h_2(t)| = \int dt |h_1(t) - h_2(t)|^2 \quad \text{(C8)}$$

within this set.

In order to quantify the accuracy of the fit, we also introduce the Euclidean norm of the difference between the ground truth values of the parameters and the fitted ones,

$$\delta c(X) = \|c_i(X) - \tilde{c}_i(X)\|. \quad \text{(C9)}$$

APPENDIX D: ADDITIONAL NUMERICAL RESULTS

1. Alternative test sets

In the main text we reported the results obtained for a test set consisting of the reciprocal lattice associated with the uniform training set for given N_{train}, with the boundary points added; see Fig. 2. In this appendix we explore other possibilities for test sets, for the same training strategies (active, uniform, random). As in the main text, we use as reference values for the performance metrics $\bar{\mathcal{F}} = 0.01$, $\delta A = 0.1$, $\delta\phi = 0.06$.

FIG. 9. Average performance gain of active training over Chebyshev training as a function of q_{max}. All test sets are random grids with the same number of points as the training sets.
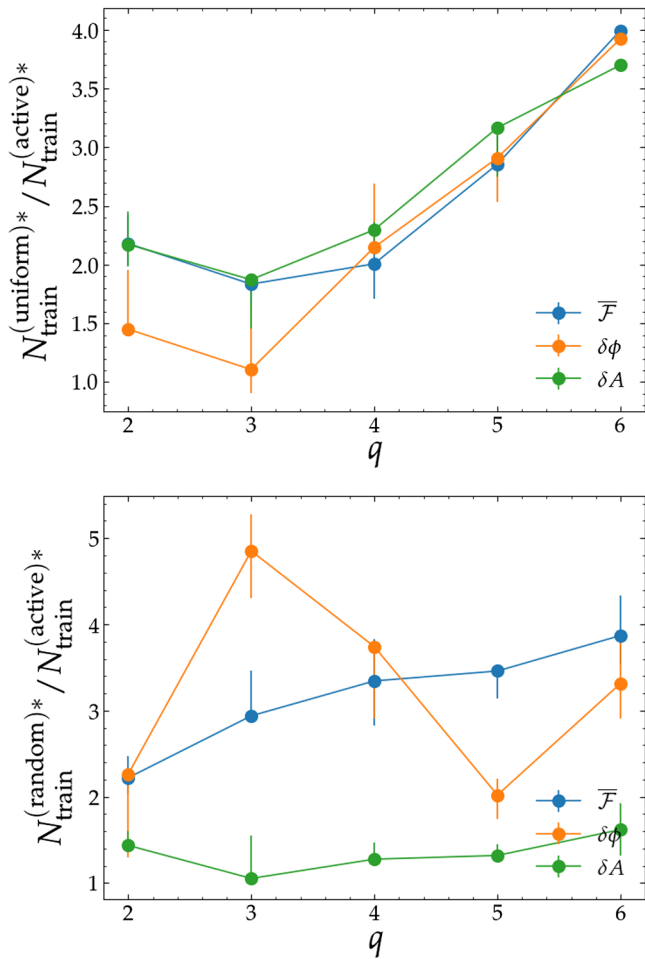
FIG. 8. Average performance gain of active training for test sets constructed as the reciprocal lattice used for uniform training (top) and for fully random test sets with $N_{rmtest} = 200$ (bottom). In the latter case, random training outperforms uniform training, so we report the performance gain with respect to that training strategy.

We first consider test sets such as the one discussed in the main text, but with the boundary points removed. This is motivated by the fact that boundary points are the most challenging to model (see Appendix E), which could lead to a misrepresentation of the performance if one is interested in "typical" points. We show the relative performance gain for the test set in the main text with the boundary points removed in Fig. 8, to be compared with Fig. 4. Besides the reduction of the error bars in the test set without boundary points, we observe essentially the same behavior, i.e., the performance gain grows with q_{max} reaching a factor of 4 for the largest value $q_{max} = 6$.

We have also considered test sets chosen at random with uniform distribution across parameter space. In such cases, we typically observe that random training outperforms uniform training, although active training continues to be superior. We show in Fig. 8 the average
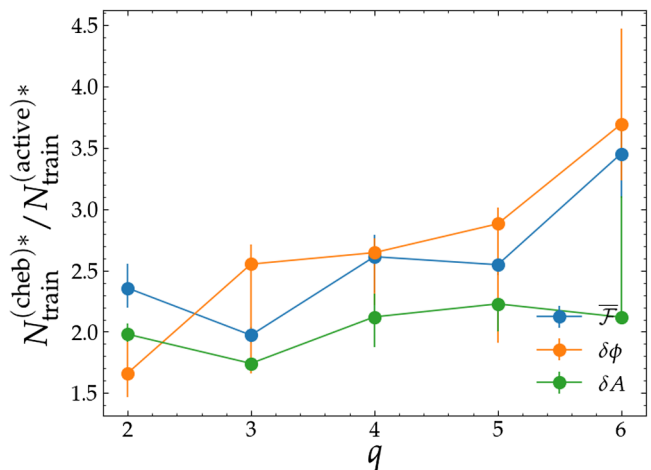
performance gain of active training over random training. In the case of the unfaithfulness, we observe a similar behavior to the reciprocal lattice test, with a monotonic growth of performance reaching nearly 4×. For the phase difference at merger, the performance gain is not monotonic and has larger error bars, which could be explained by the randomness of the test sets. For the amplitude difference at merger, we find a very mild performance gain slightly above 50%.

2. Alternative training sets

As discussed in Sec. III C, the active training strategy tends to select points near the boundary of the computational domain. This suggests that it might be possible to improve the performance of greedy approaches by replacing a uniform grid for a Chebyshev one, which is denser near the edges.[11] For the interval (a, b) the Chebyshev nodes are given by

$$x_k = \frac{1}{2}[(a+b) + (a-b)\cos\langle(k\pi/n)\rangle] \qquad k = 0 \ldots n.$$
(D1)

In addition to the active, uniform, and random training strategies discussed in detail in the main text, we consider numerical experiments with a (passive) Chebyshev training consisting of a 3d grid with Chebyshev nodes along each direction, having the same number of training points as all other approaches. Using the reciprocal lattice of the uniform training as test sets gives a biased evaluation of the different algorithms since the Chebyshev points tend

---

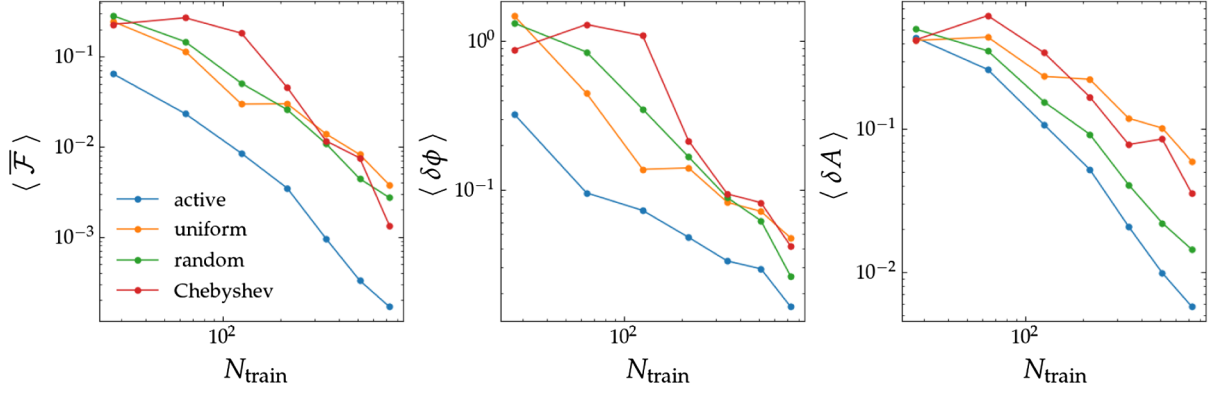[11]We thank Sebastiano Bernuzzi for suggesting this numerical experiment.

FIG. 10. Dependence of the average performance metrics with the number of training points, for active, uniform, random, and Chebyshev training with $q_{max} = 6$. The test sets are random grids the same number of points as the training sets. All plots are in log-log scale.

to be close to the reciprocal lattice. We therefore consider fully random test sets, with the same number of points as the corresponding training sets. We show our results for the average performance gain of active training over Chebyshev training in Fig. 9, and the average performance as a function of training points for $q_{max} = 6$ in Fig. 10. We observe that uniform, random, and Chebyshev training give similar results, all being superseded by active training. The maximum cost reduction obtained by using active training is again roughly a factor

of 4 for the mismatch, and slightly lower for other metrics.

APPENDIX E: LOCAL RESULTS

In the main text we focused on the averages of performance and uncertainty metrics. In this appendix we complement these results with the statistical and spatial distributions in parameter space. First, we show in Fig. 11 histograms of all metrics extracted for $N_{train} = 729$.
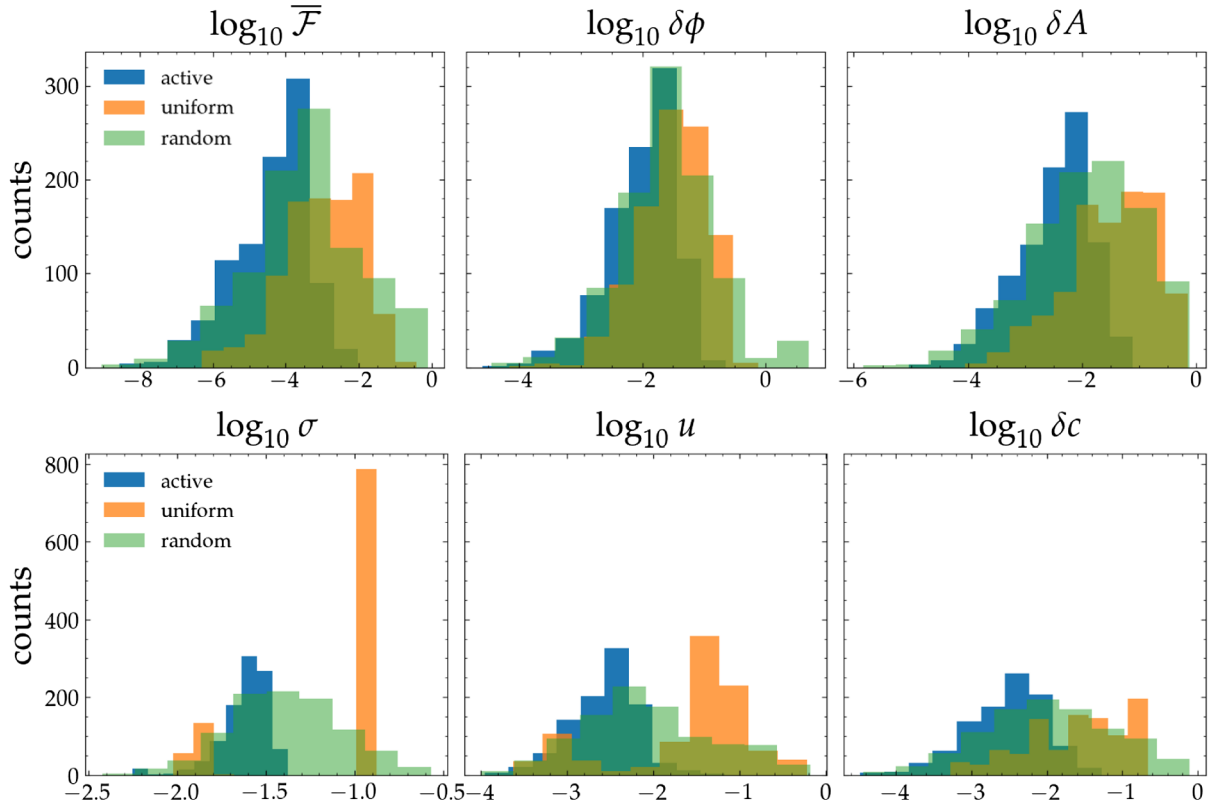

FIG. 11. Histograms of the performance (top) and uncertainty metrics (bottom) for $N_{train} = 729$ with $q_{max} = 6$.
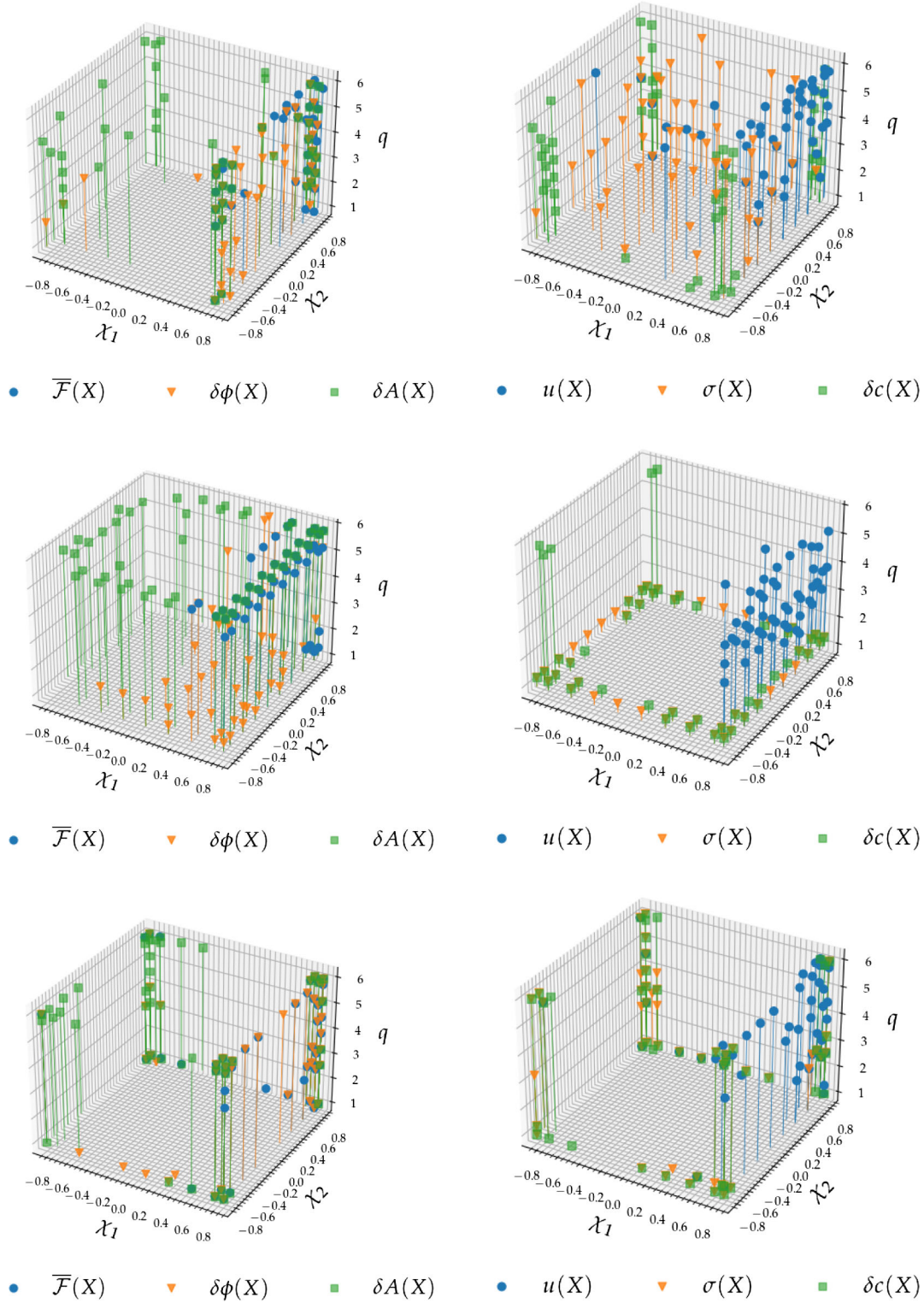
FIG. 12. Three-dimensional local dependence of the performance (left column) and uncertainty (right column) metrics for active (top row), uniform (center row), and random (bottom row) training with $N_{\text{train}} = 729$. We show the 50 worst points in each case.

We note that not only are the averages always smaller for active training, but the entire distributions show the superior performance of this protocol over uniform and random grid training. In Fig. 12 we display a visualization of the local dependence of all metrics as functions of the input parameters. We observe that, with the exception of the variance in active training, the most challenging cases tend to accumulate near the boundaries of the computational domain.

[1] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Observation of gravitational waves from a binary black hole merger, Phys. Rev. Lett. **116,** 061102 (2016).

[2] P. Jaranowski and A. Krolak, Gravitational-wave data analysis. Formalism and sample applications: The Gaussian case, Living Rev. Relativity **8,** 3 (2005).

[3] J. Abadie *et al.* (LIGO Scientific and Virgo Collaborations), Search for gravitational waves from binary black hole inspiral, merger and ringdown, Phys. Rev. D **83,** 122005 (2011); **86,** 069903(E) (2012).

[4] J. Aasi *et al.* (LIGO Scientific and Virgo Collaborations), Parameter estimation for compact binary coalescence signals with the first generation gravitational-wave detector network, Phys. Rev. D **88,** 062001 (2013).

[5] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run, Phys. Rev. X **13,** 041039 (2023).

[6] B. P. Abbott *et al.* (KAGRA, LIGO Scientific, and Virgo Collaborations), Prospects for observing and localizing gravitational-wave transients with Advanced LIGO, Advanced Virgo and KAGRA, Living Rev. Relativity **21,** 3 (2018).

[7] F. Pretorius, Evolution of binary black hole spacetimes, Phys. Rev. Lett. **95,** 121101 (2005).

[8] M. Campanelli, C. O. Lousto, P. Marronetti, and Y. Zlochower, Accurate evolutions of orbiting black-hole binaries without excision, Phys. Rev. Lett. **96,** 111101 (2006).

[9] J. G. Baker, J. Centrella, D.-I. Choi, M. Koppitz, and J. van Meter, Gravitational wave extraction from an inspiraling configuration of merging black holes, Phys. Rev. Lett. **96,** 111102 (2006).

[10] SPEC: Spectral Einstein code, https://www.black-holes.org/code/SpEC.html.

[11] Einstein Toolkit, https://einsteintoolkit.org.

[12] B. Daszuta, F. Zappa, W. Cook, D. Radice, S. Bernuzzi, and V. Morozova, GR-Athena++: Puncture evolutions on vertex-centered oct-tree adaptive mesh refinement, Astrophys. J. Suppl. Ser. **257,** 25 (2021).

[13] T. Andrade *et al.*, GRChombo: An adaptable numerical relativity code for fundamental physics, J. Open Source Software **6,** 3703 (2021).

[14] V. Vijay, Black hole simulations: From supercomputers to your laptop, Ph.D. thesis, 2019, https://thesis.library.caltech.edu/11544/.

[15] P. Ajith, S. Babak, Y. Chen, M. Hewitson, B. Krishnan *et al.*, Phenomenological template family for black-hole coalescence waveforms, Classical Quantum Gravity **24,** S689 (2007).

[16] P. Ajith, S. Babak, Y. Chen, M. Hewitson, B. Krishnan *et al.*, A template bank for gravitational waveforms from coalescing binary black holes. I. Non-spinning binaries, Phys. Rev. D **77,** 104017 (2008).

[17] P. Ajith, M. Hannam, S. Husa, Y. Chen, B. Brügmann *et al.*, Inspiral-merger-ringdown waveforms for black-hole binaries with non-precessing spins, Phys. Rev. Lett. **106,** 241101 (2011).

[18] L. Santamaria, F. Ohme, P. Ajith, B. Brügmann, N. Dorband *et al.*, Matching post-Newtonian and numerical relativity waveforms: Systematic errors and a new phenomenological model for non-precessing black hole binaries, Phys. Rev. D **82,** 064016 (2010).

[19] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. Jiménez Forteza, and A. Bohé, Frequency-domain gravitational waves from nonprecessing black-hole binaries. I. New numerical waveforms and anatomy of the signal, Phys. Rev. D **93,** 044006 (2016).

[20] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez Forteza, and A. Bohé, Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era, Phys. Rev. D **93,** 044007 (2016).

[21] G. Pratten, S. Husa, C. Garcia-Quiros, M. Colleoni, A. Ramos-Buades, H. Estelles, and R. Jaume, Setting the cornerstone for a family of models for gravitational waves from compact binaries: The dominant harmonic for non-precessing quasicircular black holes, Phys. Rev. D **102,** 064001 (2020).

[22] H. Estellés, A. Ramos-Buades, S. Husa, C. García-Quirós, M. Colleoni, L. Haegel, and R. Jaume, Phenomenological time domain model for dominant quadrupole gravitational wave signal of coalescing binary black holes, Phys. Rev. D **103,** 124060 (2021).

[23] E. Hamilton, L. London, J. E. Thompson, E. Fauchon-Jones, M. Hannam, C. Kalaghatgi, S. Khan, F. Pannarale, and A. Vano-Vinuales, Model of gravitational waves from precessing black-hole binaries through merger and ringdown, Phys. Rev. D **104,** 124027 (2021).

[24] H. Estellés, M. Colleoni, C. García-Quirós, S. Husa, D. Keitel, M. Mateu-Lucena, M. d. L. Planas, and A. Ramos-Buades, New twists in compact binary waveform modeling: A fast time-domain model for precession, Phys. Rev. D **105,** 084040 (2022).

[25] L. London, S. Khan, E. Fauchon-Jones, X. J. Forteza, M. Hannam, S. Husa, C. Kalaghatgi, F. Ohme, and F. Pannarale, First higher-multipole model of gravitational waves from spinning and coalescing black-hole binaries, Phys. Rev. Lett. **120,** 161102 (2018).

[26] C. García-Quirós, M. Colleoni, S. Husa, H. Estellés, G. Pratten, A. Ramos-Buades, M. Mateu-Lucena, and R. Jaume, Multimode frequency-domain model for the gravitational wave signal from nonprecessing black-hole binaries, Phys. Rev. D **102,** 064002 (2020).

[27] S. Khan, F. Ohme, K. Chatziioannou, and M. Hannam, Including higher order multipoles in gravitational-wave models for precessing binary black holes, Phys. Rev. D **101,** 024056 (2020).

[28] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Simple model of complete precessing black-hole-binary gravitational waveforms, Phys. Rev. Lett. **113,** 151101 (2014).

[29] P. Schmidt, F. Ohme, and M. Hannam, Towards models of gravitational waveforms from generic binaries II: Modelling precession effects with a single effective precession parameter, Phys. Rev. D **91,** 024043 (2015).

[30] S. Khan, K. Chatziioannou, M. Hannam, and F. Ohme, Phenomenological model for the gravitational-wave signal from precessing binary black holes with two-spin effects, Phys. Rev. D **100,** 024059 (2019).

[31] G. Pratten *et al.*, Computationally efficient models for the dominant and subdominant harmonic modes of precessing binary black holes, Phys. Rev. D **103**, 104056 (2021).

[32] A. Buonanno and T. Damour, Effective one-body approach to general relativistic two-body dynamics, Phys. Rev. D **59**, 084006 (1999).

[33] A. Buonanno and T. Damour, Transition from inspiral to plunge in binary black hole coalescences, Phys. Rev. D **62**, 064015 (2000).

[34] T. Damour, P. Jaranowski, and G. Schaefer, On the determination of the last stable orbit for circular general relativistic binaries at the third post-Newtonian approximation, Phys. Rev. D **62**, 084011 (2000).

[35] T. Damour, Coalescence of two spinning black holes: An effective one- body approach, Phys. Rev. D **64**, 124013 (2001).

[36] T. Damour, B. R. Iyer, and A. Nagar, Improved resummation of post-Newtonian multipolar waveforms from circularized compact binaries, Phys. Rev. D **79**, 064004 (2009).

[37] T. Damour, P. Jaranowski, and G. Schäfer, Effective one body approach to the dynamics of two spinning black holes with next-to-leading order spin-orbit coupling, Phys. Rev. D **78**, 024009 (2008).

[38] E. Barausse and A. Buonanno, An improved effective-one-body Hamiltonian for spinning black-hole binaries, Phys. Rev. D **81**, 084024 (2010).

[39] T. Damour, Gravitational self-force in a Schwarzschild background and the effective one body formalism, Phys. Rev. D **81**, 024017 (2010).

[40] T. Damour and A. Nagar, Effective one body description of tidal effects in inspiralling compact binaries, Phys. Rev. D **81**, 084016 (2010).

[41] T. Damour, Gravitational scattering, post-Minkowskian approximation and effective one-body theory, Phys. Rev. D **94**, 104015 (2016).

[42] J. Vines, Scattering of two spinning black holes in post-Minkowskian gravity, to all orders in spin, and effective-one-body mappings, Classical Quantum Gravity **35**, 084002 (2018).

[43] T. Damour, High-energy gravitational scattering and the general relativistic two-body problem, Phys. Rev. D **97**, 044038 (2018).

[44] A. Bohé *et al.*, Improved effective-one-body model of spinning, nonprecessing binary black holes for the era of gravitational-wave astrophysics with advanced detectors, Phys. Rev. D **95**, 044028 (2017).

[45] S. Babak, A. Taracchini, and A. Buonanno, Validating the effective-one-body model of spinning, precessing binary black holes against numerical relativity, Phys. Rev. D **95**, 024010 (2017).

[46] R. Cotesta, A. Buonanno, A. Bohé, A. Taracchini, I. Hinder, and S. Ossokine, Enriching the symphony of gravitational waves from binary black holes by tuning higher harmonics, Phys. Rev. D **98**, 084028 (2018).

[47] S. Ossokine *et al.*, Multipolar effective-one-body waveforms for precessing binary black holes: Construction and validation, Phys. Rev. D **102**, 044055 (2020).

[48] M. van de Meent, A. Buonanno, D. P. Mihaylov, S. Ossokine, L. Pompili, N. Warburton, A. Pound, B. Wardell, L. Durkan, and J. Miller, Enhancing the SEOBNRv5 effective-one-body waveform model with second-order gravitational self-force fluxes, Phys. Rev. D **108**, 124038 (2023).

[49] A. Ramos-Buades, A. Buonanno, M. Khalil, and S. Ossokine, Effective-one-body multipolar waveforms for eccentric binary black holes with nonprecessing spins, Phys. Rev. D **105**, 044035 (2022).

[50] L. Pompili *et al.*, Laying the foundation of the effective-one-body waveform models SEOBNRv5: Improved accuracy and efficiency for spinning non-precessing binary black holes, Phys. Rev. D **108**, 124035 (2023).

[51] M. Khalil, A. Buonanno, H. Estelles, D. P. Mihaylov, S. Ossokine, L. Pompili, and A. Ramos-Buades, Theoretical groundwork supporting the precessing-spin two-body dynamics of the effective-one-body waveform models SEOBNRv5, Phys. Rev. D **108**, 124036 (2023).

[52] A. Ramos-Buades, A. Buonanno, H. Estellés, M. Khalil, D. P. Mihaylov, S. Ossokine, L. Pompili, and M. Shiferaw, SEOBNRv5PHM: Next generation of accurate and efficient multipolar precessing-spin effective-one-body waveforms for binary black holes, Phys. Rev. D **108**, 124037 (2023).

[53] T. Damour and A. Nagar, A new analytic representation of the ringdown waveform of coalescing spinning black hole binaries, Phys. Rev. D **90**, 024054 (2014).

[54] A. Nagar, G. Riemenschneider, and G. Pratten, Impact of numerical relativity information on effective-one-body waveform models, Phys. Rev. D **96**, 084045 (2017).

[55] A. Nagar *et al.*, Time-domain effective-one-body gravitational waveforms for coalescing compact binaries with nonprecessing spins, tides and self-spin effects, Phys. Rev. D **98**, 104052 (2018).

[56] S. Akcay, R. Gamba, and S. Bernuzzi, A hybrid post-Newtonian—effective-one-body scheme for spin-precessing compact-binary waveforms, Phys. Rev. D **103**, 024014 (2021).

[57] R. Gamba, S. Akçay, S. Bernuzzi, and J. Williams, Effective-one-body waveforms for precessing coalescing compact binaries with post-Newtonian twist, Phys. Rev. D **106**, 024020 (2022).

[58] A. Nagar, G. Riemenschneider, G. Pratten, P. Rettegno, and F. Messina, Multipolar effective one body waveform model for spin-aligned black hole binaries, Phys. Rev. D **102**, 024077 (2020).

[59] D. Chiaramello and A. Nagar, Faithful analytical effective-one-body waveform model for spin-aligned, moderately eccentric, coalescing black hole binaries, Phys. Rev. D **101**, 101501 (2020).

[60] A. Nagar, P. Rettegno, R. Gamba, and S. Bernuzzi, Effective-one-body waveforms from dynamical captures in black hole binaries, Phys. Rev. D **103**, 064013 (2021).

[61] S. Albanesi, A. Nagar, and S. Bernuzzi, Effective one-body model for extreme-mass-ratio spinning binaries on eccentric equatorial orbits: Testing radiation reaction and waveform, Phys. Rev. D **104**, 024067 (2021).

[62] A. Nagar, A. Bonino, and P. Rettegno, Effective one-body multipolar waveform model for spin-aligned, quasicircular, eccentric, hyperbolic black hole binaries, Phys. Rev. D **103**, 104021 (2021).

[63] A. Nagar and P. Rettegno, Next generation: Impact of high-order analytical information on effective one body waveform models for noncircularized, spin-aligned black hole binaries, Phys. Rev. D **104,** 104004 (2021).

[64] A. Nagar, P. Rettegno, R. Gamba, A. Albertini, and S. Bernuzzi, Analytic systematics in next-generation of effective-one-body gravitational waveform models for future observations, Phys. Rev. D **108,** 124018 (2023).

[65] J. Blackman, B. Szilagyi, C. R. Galley, and M. Tiglio, Sparse representations of gravitational waves from precessing compact binaries, Phys. Rev. Lett. **113,** 021101 (2014).

[66] J. Blackman, S. E. Field, M. A. Scheel, C. R. Galley, D. A. Hemberger, P. Schmidt, and R. Smith, A surrogate model of gravitational waveforms from numerical relativity simulations of precessing binary black hole mergers, Phys. Rev. D **95,** 104023 (2017).

[67] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, L. E. Kidder, and H. P. Pfeiffer, Surrogate model of hybridized numerical relativity binary black hole waveforms, Phys. Rev. D **99,** 064045 (2019).

[68] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, D. Gerosa, L. C. Stein, L. E. Kidder, and H. P. Pfeiffer, Surrogate models for precessing binary black hole simulations with unequal masses, Phys. Rev. Res. **1,** 033015 (2019).

[69] D. Williams, I. S. Heng, J. Gair, J. A. Clark, and B. Khamesra, Precessing numerical relativity waveform surrogate model for binary black holes: A Gaussian process regression approach, Phys. Rev. D **101,** 063011 (2020).

[70] A. H. Mroue *et al.*, Catalog of 174 binary black hole simulations for gravitational wave astronomy, Phys. Rev. Lett. **111,** 241104 (2013).

[71] T. Chu, H. Fong, P. Kumar, H. P. Pfeiffer, M. Boyle, D. A. Hemberger, L. E. Kidder, M. A. Scheel, and B. Szilagyi, On the accuracy and precision of numerical waveforms: Effect of waveform extraction methodology, Classical Quantum Gravity **33,** 165001 (2016).

[72] M. Boyle *et al.*, The SXS Collaboration catalog of binary black hole simulations, Classical Quantum Gravity **36,** 195006 (2019).

[73] J. Healy, C. O. Lousto, Y. Zlochower, and M. Campanelli, The RIT binary black hole simulations catalog, Classical Quantum Gravity **34,** 224001 (2017).

[74] J. Healy, C. O. Lousto, J. Lange, R. O'Shaughnessy, Y. Zlochower, and M. Campanelli, Second RIT binary black hole simulations catalog and its application to gravitational waves parameter estimation, Phys. Rev. D **100,** 024021 (2019).

[75] J. Healy and C. O. Lousto, Fourth RIT binary black hole simulations catalog: Extension to eccentric orbits, Phys. Rev. D **105,** 124010 (2022).

[76] T. Dietrich, D. Radice, S. Bernuzzi, F. Zappa, A. Perego, B. Brügmann, S. V. Chaurasia, R. Dudi, W. Tichy, and M. Ujevic, CoRe database of binary neutron star merger waveforms, Classical Quantum Gravity **35,** 24LT01 (2018).

[77] A. Gonzalez *et al.*, Second release of the CoRe database of binary neutron star merger waveforms, Classical Quantum Gravity **40,** 085011 (2023).

[78] K. Kiuchi, K. Kawaguchi, K. Kyotoku, Y. Sekiguchi, M. Shibata, and K. Taniguchi, Sub-radian-accuracy gravitational waveforms of coalescing binary neutron stars in numerical relativity, Phys. Rev. D **96,** 084060 (2017).

[79] K. Kiuchi, K. Kawaguchi, K. Kyotoku, Y. Sekiguchi, and M. Shibata, Sub-radian-accuracy gravitational waves from coalescing binary neutron stars in numerical relativity. II. Systematic study on the equation of state, binary mass, and mass ratio, Phys. Rev. D **101,** 084006 (2020).

[80] E. Hamilton *et al.*, A catalogue of precessing black-hole-binary numerical-relativity simulations, Phys. Rev. D **109,** 044032 (2024).

[81] D. Ferguson *et al.*, Second MAYA catalog of binary black hole numerical relativity waveforms, arXiv:2309.00262.

[82] D. V. Lindley, On a measure of the information provided by an experiment, Ann. Math. Stat. **27,** 986 (1956).

[83] A. Atkinson, A. Donev, and R. Tobias, *Optimum experimental designs, with SAS* (oxford University Press, 2007).

[84] J. Chang, J. Kim, B.-T. Zhang, M. A. Pitt, and J. I. Myung, Data-driven experimental design and model development using Gaussian process with active learning, Cogn. Psychol. **125,** 101360 (2021).

[85] K. Chaloner and I. Verdinelli, Bayesian experimental design: A review, Stat. Sci. **10,** 273 (1995).

[86] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, Active learning with statistical models, arXiv:cs/9603104.

[87] A. Agnihotri and N. Batra, Exploring Bayesian optimization, Distill, 10.23915/distill.00026 (2020).

[88] D. Krige, *A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand* (Chemical, Metallurgical and Mining Society of South Africa, 1951).

[89] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (The MIT Press, Cambridge, MA, 2005).

[90] J. Wang, An intuitive tutorial to Gaussian processes regression, Comput. Sci. Eng. **25,** 4 (2023).

[91] D. R. Cavagnaro, M. A. Pitt, and J. I. Myung, Model discrimination through adaptive experimentation, Psychon. Bull. Rev. **18,** 204 (2011).

[92] D. R. Cavagnaro, M. A. Pitt, R. Gonzalez, and J. I. Myung, Discriminating among probability weighting functions using adaptive design optimization, J. Risk Uncertainty **47,** 255 (2013).

[93] H. Gu, W. Kim, F. Hou, L. A. Lesmes, M. A. Pitt, Z.-L. Lu, and J. I. Myung, A hierarchical Bayesian approach to adaptive vision testing: A case study with the contrast sensitivity function, J. Vis. **16,** 15 (2016).

[94] Z. Li, L. J. Segura, Y. Li, C. Zhou, and H. Sun, Multiclass reinforced active learning for droplet pinch-off behaviors identification in inkjet printing, J. Manuf. Sci. Eng. **145,** 071002 (2023).

[95] A. Foster, M. Jankowiak, E. Bingham, P. Horsfall, Y. W. Teh, T. Rainforth, and N. Goodman, Variational Bayesian optimal experimental design, arXiv:1903.05480.

[96] A. Foster, D. R. Ivanova, I. Malik, and T. Rainforth, Deep adaptive design: Amortizing sequential Bayesian experimental design, arXiv:2103.02438.

[97] C. J. Moore and J. R. Gair, Novel method for incorporating model uncertainties into gravitational wave parameter estimates, Phys. Rev. Lett. **113,** 251101 (2014).

[98] C. J. Moore, C. P. L. Berry, A. J. K. Chua, and J. R. Gair, Improving gravitational-wave parameter estimation using Gaussian process regression, Phys. Rev. D **93,** 064001 (2016).

[99] M. Breschi, S. Bernuzzi, K. Chakravarti, A. Camilletti, A. Prakash, and A. Perego, Kilohertz gravitational waves from binary neutron star mergers: Numerical-relativity informed postmerger model, Phys. Rev. D **109,** 064009 (2024).

[100] J. S. Read, Waveform uncertainty quantification and interpretation for gravitational-wave astronomy, Classical Quantum Gravity **40,** 135002 (2023).

[101] G. Riemenschneider, P. Rettegno, M. Breschi, A. Albertini, R. Gamba, S. Bernuzzi, and A. Nagar, Assessment of consistent next-to-quasicircular corrections and postadiabatic approximation in effective-one-body multipolar waveforms for binary black hole coalescences, Phys. Rev. D **104,** 104045 (2021).

[102] A. Nagar, J. Healy, C. O. Lousto, S. Bernuzzi, and A. Albertini, Numerical-relativity validation of effective-one-body waveforms in the intermediate-mass-ratio regime, Phys. Rev. D **105,** 124061 (2022).

[103] GPy a Gaussian processes framework in Python, Sheffield machine learning software, http://sheffieldml.github.io/GPy/.

[104] E. V. Bonilla, K. Chai, and C. Williams, Multi-task Gaussian process prediction, in *Advances in Neural Information Processing Systems*, edited by J. Platt, D. Koller, Y. Singer, and S. Roweis (NIPS Foundation, 2007), Vol. 20.

[105] A. Nitz *et al.*, GWASTRO/PyCBC: v2.0.5 release of PyCBC (Zenodo, 2022), 10.5281/zenodo.10473621.

[106] S. J. Sloman, D. M. Oppenheimer, S. B. Broomell, and C. R. Shalizi, Characterizing the robustness of Bayesian adaptive experimental designs to active learning bias, arXiv:2205.13698.

[107] W. E. East, S. T. McWilliams, J. Levin, and F. Pretorius, Observing complete gravitational wave signals from dynamical capture binaries, Phys. Rev. D **87,** 043004 (2013).

[108] A. Rasskazov and B. Kocsis, The rate of stellar mass black hole scattering in galactic nuclei, Astrophys. J. **881,** 20 (2019).

[109] H. Tagawa, Z. Haiman, and B. Kocsis, Formation and evolution of compact object binaries in AGN disks, Astrophys. J. **898,** 25 (2020).

[110] T. Andrade *et al.*, Towards numerical-relativity informed effective-one-body waveforms for dynamical capture black hole binaries, Phys. Rev. D **109,** 084025 (2024).

[111] G. Carullo, S. Albanesi, A. Nagar, R. Gamba, S. Bernuzzi, T. Andrade, and J. Trenado, Unveiling the merger structure of black hole binaries in generic planar orbits, Phys. Rev. Lett. **132,** 101401 (2024).

[112] L. Rokach, *Pattern Classification Using Ensemble Methods* (World Scientific Publishing Co., Inc., USA, 2010).

[113] J. Moreira, C. Soares, A. Jorge, and J. Sousa, Ensemble approaches for regression: A survey, ACM Comput. Surv. **45,** 1 (2012).

[114] https://bitbucket.org/eob_ihes/teobresums/.