Rapid Generation of Fully Relativistic Extreme-Mass-Ratio-Inspiral Waveform Templates for LISA Data Analysis

Alvin J. K. Chua,[1] Michael L. Katz,[2,3] Niels Warburton,[4] and Scott A. Hughes[5]

[1]*Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA*
[2]*Department of Physics and Astronomy, Northwestern University, Evanston, Illinois 60208, USA*
[3]*Center for Interdisciplinary Exploration and Research in Astrophysics (CIERA), Evanston, Illinois 60208, USA*
[4]*School of Mathematics and Statistics, University College Dublin, Belfield, Dublin 4, Ireland*
[5]*Department of Physics and MIT Kavli Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

(Received 23 August 2020; accepted 20 November 2020; published 4 February 2021)

The future space mission LISA will observe a wealth of gravitational-wave sources at millihertz frequencies. Of these, the extreme-mass-ratio inspirals of compact objects into massive black holes are the only sources that combine the challenges of strong-field complexity with that of long-lived signals. Such signals are found and characterized by comparing them against a large number of accurate waveform templates during data analysis, but the rapid generation of templates is hindered by computing the $\sim 10^3$–10^5 harmonic modes in a fully relativistic waveform. We use order-reduction and deep-learning techniques to derive a global fit for the ≈ 4000 modes in the special case of an eccentric Schwarzschild orbit, and implement the fit in a complete waveform framework with hardware acceleration. Our high-fidelity waveforms can be generated in under 1 s, and achieve a mismatch of $\lesssim 5 \times 10^{-4}$ against reference waveforms that take $\gtrsim 10^4$ times longer. This marks the first time that analysis-length waveforms with full harmonic content can be produced on timescales useful for direct implementation in LISA analysis algorithms.

DOI: 10.1103/PhysRevLett.126.051102

Introduction.—As gravitational-wave (GW) astronomy continues to bear fruit [1], preparatory work is underway for a future generation of ground- and space-based observatories that span the astrophysical GW spectrum [2–4], and whose success will depend on further advancements in the technology and methods of GW detection. The theorist's contribution to this endeavor lies primarily in the construction of *waveform* models to describe GW signals from astrophysical phenomena, as well as the application of statistical analysis to infer their presence in noisy data and their source properties. For highly relativistic sources, the computational burden of solving Einstein's equations in numerical modeling is fundamentally at odds with the Monte Carlo nature of modern signal processing and Bayesian-inference techniques.

Extreme-mass-ratio inspirals (EMRIs) are the most conspicuous example of such dissonance. These are the late capture orbits of stellar-mass ($\mu \sim 1$–$100\ M_\odot$) compact objects into the massive ($M \sim 10^5$–$10^7\ M_\odot$) black holes in galactic nuclei. They radiate millihertz GWs, and will be a key source class for the space mission LISA [4] upon its launch in the next decade. An EMRI signal typically has $\sim 10^5$ observable cycles carrying the imprint of the compact object's complex dynamical motion deep in the central black hole's gravitational field. This wealth of information is double edged: it will allow probes of galactic-nuclei astrophysics and strong-field gravity to unprecedented precision [5,6], but it places exacting constraints on the accuracy and efficiency of both modeling and data analysis for EMRIs—to a combined extent far surpassing that for other important LISA sources.

Calculations from black-hole perturbation theory, and in particular from the ongoing gravitational self-force program [7], are on target to produce EMRI waveforms that meet the accuracy requirements of LISA science [8,9]. Such models are computationally intensive, and hence ill suited for direct use in analysis algorithms that are tailored to the EMRI problem [10–14]. As in the case of numerical-relativity waveforms for comparable-mass binaries, self-force waveforms must be supplemented and approximated by *template* models that are (i) efficiency oriented, (ii) extensive in their description of both intrinsic and extrinsic effects, and (iii) end to end from source parameters to detector response. The challenge is to achieve this with a controlled and tolerable loss of accuracy. Strategies developed for the comparable-mass case, such as the standard construction of reduced-order-modeling (ROM) surrogates (e.g., Ref. [15]), are less likely to scale feasibly to the signal duration, harmonic complexity, and information volume of the full EMRI problem.

The semirelativistic "kludges" [16–20] are the only existing examples of EMRI template models. Kludges trade accuracy for efficiency by means of a modular build and various computational approximations. Their common distinguishing feature is a reliance on some weak-field assumption at one or more stages of their construction. The speed and generality of kludge models has greatly facilitated numerous LISA studies on mission performance, data analysis approaches, and potential scientific applications. However, kludges incur significant error with respect to fully relativistic models for many sources in the observable space of EMRIs [17], and have little room for improvement due to the limitations of the weak-field assumption. This inherent cap on accuracy may count against the continued development and adoption of kludge models, at least in the long term.

In this Letter, we report promising headway against the main obstacle to the rapid generation of fully relativistic EMRI waveforms: efficiently computing the slowly evolving *amplitudes* of the $\sim 10^3$–10^5 harmonic modes that comprise a single waveform in the canonical angular and frequency-based decomposition [21,22]. Through the integration of ROM and deep-learning techniques [23], an analytic model for these amplitudes is fitted to numerical data from a frequency-domain Teukolsky solver [24]. The key to our approach is the use of regression rather than interpolation, resulting in a less precise but global fit that returns the full set of amplitudes simultaneously. This allows the inclusion of relativistic amplitudes in template models, where they are combined with existing fast methods for generating the phasing *trajectories* to varying levels of accuracy.

As the mode-amplitude model is a neural network, it is composed of simple linear-algebra operations and hence amenable to acceleration through a highly parallelized implementation for graphics processing units (GPUs). We exploit this to construct the first EMRI waveform model with subsecond run-times in a realistic setting, i.e., analysis-length signals [$\sim 10^7\ M$ at sampling rate $1/(2M)$], and full harmonic content (retaining up to 1–10^{-9} of total power at initial orbital eccentricities of up to 0.7 [25]). The present model describes the source-frame GW field for eccentric orbits in Schwarzschild, with inspiral trajectories that are accurate at adiabatic order. Our code infrastructure is designed with the end goal of providing analysis-ready template models; specifically, it will readily accommodate postadiabatic trajectories informed by future self-force calculations, as well as the eventual extension to generic Kerr orbits and the integration of a compatible LISA response model.

Adiabatic waveforms.—An EMRI's disparate masses $(M, \mu \ll M)$ create a wide separation between its orbital and radiation-reaction timescales. This allows EMRIs to be modeled through a two-timescale expansion [26]. In the leading adiabatic part of this expansion, the equations of motion follow from flux balance laws. Though a purely adiabatic treatment of waveform phasing will be insufficiently accurate to describe a typical EMRI signal over its full duration [27], adiabatic trajectories can still be used for data analysis within a hierarchical semicoherent *search* scheme [10,19], or for more slowly evolving binaries with $\mu/M < 10^{-6}$ [28,29]. Postadiabatic corrections (once known) can be easily added by including additional phase corrections [9,30,31]. The computation of mode amplitudes is also only required at adiabatic order [9], even for the most stringent analysis task of *inference*. This is due to the disproportionate dependence of GW matched filtering on waveform phasing, rather than its amplitude.

For an EMRI with a nonrotating central black hole, the adiabatic evolution of the orbital energy $\mathcal{E}$ and angular momentum $\mathcal{L}$ is given by $(\dot{\mathcal{E}}, \dot{\mathcal{L}}) = -(\dot{E}, \dot{L})$, where an overdot denotes differentiation with respect to coordinate time t, and $(\dot{E}, \dot{L})$ is the total flux of energy and angular momentum radiated through null infinity and the event horizon. It is useful to parametrize the system by an equivalent set of quasi-Keplerian orbital elements: the semilatus rectum (henceforth "separation") p and eccentricity e, with $\mathcal{E}^2 = p^{-1}(p-2-2e)(p-2+2e)/(p-3-e^2)$ and $\mathcal{L}^2 = p^2M^2/(p-3-e^2)$ [32]. In this parametrization, stable bound orbits exist for $p > p_s = 6+2e$ and $0 \le e < 1$, where p_s denotes the separatrix [33]. Each instantaneous orbit (p, e) is associated with a radial and azimuthal frequency, denoted by Ω_r and Ω_φ, respectively.

In the Newman-Penrose formalism [34], the GW field h at null infinity is related to the Weyl curvature scalar ψ_4 via $\ddot{h} = 2\psi_4$, where $h = h_+ - ih_\times$ with the usual transverse traceless polarizations $h_{+,\times}$. For each orbit (p, e), ψ_4 may be obtained by solving the Teukolsky equation [35] in the frequency domain; this requires a decomposition of the form $\psi_4 = \sum_{lmn} R_{lmn}(r) Y_{lm}(\theta, \varphi) e^{-i\omega_{mn} t}$, where Y_{lm} are spherical harmonics with spin weight -2, and $\omega_{mn} = m\Omega_\varphi + n\Omega_r$ are the mode frequencies. It is convenient to define and solve for the complex Teukolsky amplitudes $Z^{\infty,H}_{lmn}$, which describe the limiting behavior of R_{lmn} as $r \to \infty$ and $r \to 2M$, respectively [21].

The GW strain for a detector at some suitably distant coordinates (t, r, θ, φ) is then given by [22]

$$h = \frac{1}{r} \sum_{lmn} A_{lmn}(t-r) Y_{lm}(\theta, \varphi) e^{-i\phi_{mn}(t-r)}, \qquad (1)$$

where $A_{lmn} = -2Z^\infty_{lmn}/\omega^2_{mn}$, and $\phi_{mn} = m\Phi_\varphi + n\Phi_r$ with $\Phi_{r,\varphi}(t) = \int_0^t d\tau \Omega_{r,\varphi}(p(\tau), e(\tau))$. In the results we show, we put $t = \phi = 0$ at periastron; other initial conditions are easily accommodated by adjusting the phase of A_{lmn} [27,30]. The sum over modes spans the indices $2 \le l \le l_{\max}$, $|m| \le l$ and $|n| \le n_{\max}$, with $l_{\max}$ and $n_{\max}$ determined by some convergence criterion (e.g., Ref. [36]). For the present work, we set $(l_{\max}, n_{\max}) = (10, 30)$, resulting in the sum of 7137 modes (but the explicit evaluation of only 3843, by exploiting mode symmetry [21]).

Fast trajectories.—To generate fast inspiral trajectories $(p(t), e(t), \Phi_{r,\varphi}(t))$ for use in template models, we need to rapidly evaluate $(\dot{p}, \dot{e})$ across the domain of (p, e). In a *flux-driven* trajectory for adiabatic waveforms, $(\dot{p}, \dot{e})$ is given in terms of the flux $(\dot{E}, \dot{L})$ through null infinity and the horizon, which can be calculated directly from the Teukolsky amplitudes [21]. However, numerical solutions for the amplitudes are computationally costly and can only be precomputed at a limited number of points in (p, e) space. Fast flux-driven trajectories must thus rely on an accurate and efficient interpolation scheme for the fluxes derived from this numerical data.

In this work, we first introduce a new parameter $u = \ln(p - p_s + 3.9)$, then calculate Teukolsky amplitudes and fluxes on a uniform grid in (u, e), where $1.37 \le u \le 3.82$ with spacing 0.05 and $0 \le e \le 0.8$ with spacing 0.025. The grid in u gives $p \in [p_s + 0.03, p_s + 41.6]$ and places more points near the separatrix, where the data vary more rapidly. Before interpolating the flux data, we factor out the leading post-Newtonian (PN) behavior $(\dot{E}_{\mathrm{PN}}, \dot{L}_{\mathrm{PN}})$ [37] to reduce the impact of interpolation error. We then create bicubic splines for $(\dot{E}/\dot{E}_{\mathrm{PN}}, \dot{L}/\dot{L}_{\mathrm{PN}})$ over (u, e), with PN factors restored after evaluating the splines. The inspiral trajectory is computed at run-time for initial values (p_0, e_0), by numerically integrating (for $p > p_s + 0.1$) the coupled ordinary differential equations $\{\dot{p}, \dot{e}, \dot{\Phi}_{r,\varphi}\}$ with an adaptive eighth-order Runge-Kutta method. As the flux varies on the radiation-reaction timescale M^2/μ, the solution is very smooth. This permits large integration steps, so generating each trajectory typically takes only a few milliseconds.

Going beyond flux-driven trajectories to make postadiabatic waveforms requires the inclusion of gravitational self-force corrections [7]. This introduces orbital-timescale variations into the equations of motion, which slows the calculation of a *self-forced* trajectory to minutes or even hours [38]. Recently, this barrier was overcome using near-identity transformations [31], allowing the transformed equations of motion to be evaluated in milliseconds. Key postadiabatic corrections at second order in the mass ratio [8] are being calculated in the two-timescale framework [9], which will incorporate a similar averaging procedure. Thus, the generation of the inspiral trajectory is unlikely to constitute a computational bottleneck for postadiabatic models either.

Neural-network amplitudes.—With the inspiral trajectory on hand, the remaining computationally nontrivial operation in Eq. (1) (besides the sum over modes at high resolution in time) is the evaluation of the mode amplitudes $A_{lmn}(t)$. (We fit A_{lmn} directly rather than Z^{∞}_{lmn}, to avoid numerical divergences due to fitting error whenever ω_{mn} approaches zero.) Although these are very slowly evolving and can be down sampled significantly in time, a conventional spline-interpolation approach requires the creation and evaluation of ≈4000 splines over the (p, e) space. Furthermore, future waveforms for generic Kerr orbits would involve $\sim 10^5$ splines over the four-dimensional space of separation, eccentricity, orbital inclination, and primary spin. This is problematic, as the ability of most interpolation schemes to simultaneously maintain accuracy and efficiency rapidly degrades for $\gtrsim 3$ variables.

To address the issue of high dimensionality (in both the space of modes and the space of orbits), we propose the approach of precomputing an analytic global fit for the mode amplitudes. The particular method we use is `Roman` [23], which combines the compressive power of ROM with the high-dimensional regression capabilities of deep neural networks. `Roman` was developed within the paradigm of ROM in GW modeling and analysis [39], and provides an alternative to the combination of surrogate waveforms [40] with the inference technique of reduced-order quadrature [41] (albeit at the expense of a more difficult initial fit). However, one open problem with the direct usage of ROM to fit full waveforms is accuracy. While the errors incurred by leading models (e.g., Ref. [42]) are sufficiently small for present ground-based applications, waveform templates for LISA data analysis will require far more stringent modeling [43].

In this work, we apply `Roman` to the fitting of mode amplitudes instead. A greedy algorithm [44] is first used to construct a reduced basis B for (the span of) the Teukolsky amplitude data on the uniform grid in (u, e). This allows the vectorized amplitudes $A_i = \mathrm{vec}(A_{lmn}) \in \mathbb{C}^{3843} \cong \mathbb{R}^{7686}$ to be represented in the reduced form

$$A_i(u, e) = \sum_j \alpha_j(u, e) B_{ji} \equiv \alpha_j(u, e), \qquad (2)$$

where $\alpha_j \in \mathbb{C}^{99} \cong \mathbb{R}^{198}$ for an effective compression factor of around 40. A deep neural network is then trained on the reduced dataset $\{u, e, \alpha_{\mathrm{num}}\}$ as a regression model for $\alpha(u, e)$. The architecture and training of the network is identical to the main example in Ref. [23], with the following exceptions: (i) Our network contains 20 hidden layers a_ℓ, where the first six comprise $2^{\ell+1}$ nodes and the remaining layers have 256 nodes each. (ii) As the training-set size of 1640 is small, Monte Carlo validation [45] is used to prevent overfitting, with 20 random examples held out at each epoch. (iii) The minibatch size is 810. (iv) The loss function is the standard L^2 loss $|\alpha - \alpha_{\mathrm{num}}|^2$ averaged over each minibatch, where $|\cdot|$ is Hermitian. Our network is trained over 3×10^4 epochs (4 h on one CPU core), after which it is evaluated at run-time for a set of input points $\{(p, e)\}$ to simultaneously output the corresponding set of mode amplitudes $\{\alpha \cdot B\}$. Finally, we renormalize each amplitude vector by a more accurate estimate for the vector norm, which is obtained through bicubic interpolation of the numerical norms.

Parallelized implementation.—The long-lived nature of EMRI signals will necessitate parallel implementations of template models and analysis algorithms, which can then be fully capitalized on through hardware acceleration. To date, accelerator hardware such as GPUs are very underutilized in GW astronomy. However, there are a few examples of GPU usage for both modeling and analysis: the generation of EMRI waveforms with time-domain Teukolsky solvers [46,47]; binary-black-hole waveform modeling and population inference for ground-based observing [48]; as well as massive-black-hole-binary waveform creation and parameter estimation for LISA [49].

In this work, our flux-driven trajectory and Roman amplitudes are combined in Eq. (1) to form an efficient adiabatic waveform model $h_{+,\times}(t)$ for eccentric Schwarzchild orbits, parametrized by the set $\{M, \mu, p_0, e_0, r, \theta, \varphi\}$. This model is implemented natively for GPUs, with an otherwise-equivalent counterpart implementation for CPUs. The source code is written in PYTHON (interface), C++, and CUDA, and is publicly available online [50].

GPU acceleration is crucial for relieving the main computational bottleneck in the construction of a time-domain EMRI waveform: the combination and summation of amplitude and phase information at a sufficiently high sampling rate for fully coherent analysis [defined here as $1/(2M)$ for concreteness]. In our model, this bottleneck is dealt with through a large-scale cubic-spline interpolation of $A_{lmn}(t)$ and $\Phi_{r,\varphi}(t)$ at a sparse ($\sim 10^2$) set of points in time. The number of considered modes is first reduced significantly (to $\sim 10^2$–10^3) by a run-time selection routine, where all modes at each point in time are sorted by power and removed if they do not contribute cumulatively up to some specified fraction of their total power (typically $\gtrsim 1$–10^{-5} for satisfactory waveform accuracy). Specific sets of modes can also be chosen for particular analysis purposes, e.g., $l_{\max} = 2$ to search for EMRIs at large separation. The selected amplitude (and phase) splines are then fed into a summation kernel, where they are evaluated and summed at full resolution.

Results.—The domain of validity for our waveform model is defined as $p_{\min} \leq p \leq p_s + 10$ and $0 \leq e \leq 0.7$, where $p_{\min} = \max\{p_s + 0.1, 7p_s - 41.9\}$. Orbits at small p and large e are excluded as they lack astrophysical relevance, and are also difficult to fit due to their high degree of variability. The large-p boundary is justified by the reduced sensitivity of LISA at frequencies corresponding to $p \gtrsim 20$. We assess the individual accuracies of the trajectory and amplitude modules against numerical Teukolsky flux and amplitude calculations, using a test dataset of 232 orbits that spans the domain of validity (but has no orbit in common with the training set). For the relative flux error $(\Delta\dot{E}/\dot{E}_{\mathrm{num}}, \Delta\dot{L}/\dot{L}_{\mathrm{num}})$, both components have a median value of 3×10^{-7}. As the vectorized Roman amplitudes are renormalized to similar accuracy, we consider their "mode-distribution" error $1 - \Re(A^\dagger A_{\mathrm{num}})/(|A||A_{\mathrm{num}}|)$, which reduces to (half of) the relative L^2 error when $|A| = |A_{\mathrm{num}}|$. The mode-distribution error has a median value of 3×10^{-5}.

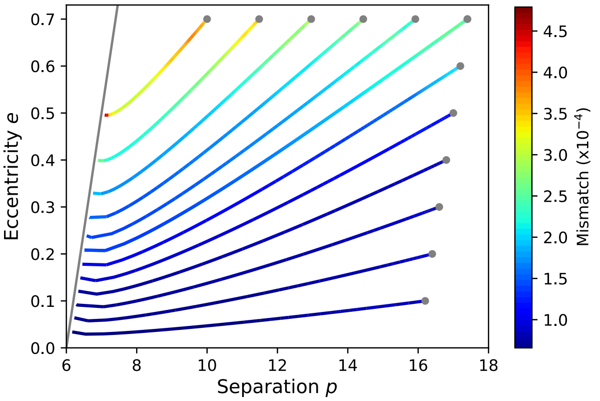

FIG. 1. Evolution of mismatch between fast and fiducial waveforms from (p_0, e_0) to (p, e), for 12 EMRIs with $M = 10^6\, M_\odot$, $\mu \in [15, 304] M_\odot$, and (p_0, e_0) along the model domain boundary. Each small mass is chosen such that the EMRI plunges after a year. These results are for $(\theta, \varphi) = (\pi/2, 0)$, but do not depend strongly on the viewing angle. In the worst case (top-left curve), the final 0.01% of the waveform causes the mismatch to increase from under 4×10^{-4} to 5×10^{-4}.

Our fast model is then benchmarked against a slower fiducial model that uses standard bicubic-spline interpolation for the amplitude of each mode (without mode selection), as well as integration steps at full time resolution for the inspiral trajectory. The bicubic amplitudes in the slow model are significantly more faithful to the numerical test data than the Roman amplitudes, with a median mode-distribution error of 3×10^{-11}. To quantify the overall error in the fast waveform with respect to the slow fiducial waveform, we examine their mismatch: $1 - \Re(h^\dagger h_{\mathrm{fid}})/(|h||h_{\mathrm{fid}}|)$, defined here without noise weighting for simplicity. The mismatch is dominated by amplitude error, as the phase difference $\Delta\phi$ between the fast and slow trajectories typically has a maximal value of $\sim 10^{-3}$ over the full duration of a waveform.

Figure 1 shows how the mismatch from (p_0, e_0) up to (p, e) changes for a representative set of EMRIs in the domain of validity, as they evolve towards the separatrix over a duration of $\sim 10^7\, M$. The EMRI with the largest e_0 and the smallest p_0 plunges at high eccentricity $e \approx 0.5$, and has the "worst-case" full mismatch of 5×10^{-4}; snapshots of this waveform at initial and plunge time are shown in Fig. 2. In general, the signal-to-noise ratio ρ of a GW source determines the required level of mismatch $\sim 1/\rho^2$ for inference purposes [43], and so the accuracies achieved by a waveform with Roman amplitudes should be adequate for LISA EMRIs (where $\rho \lesssim 10^2$). In terms of efficiency, wall times for the slow model ($\sim$1h for the worst-case waveform

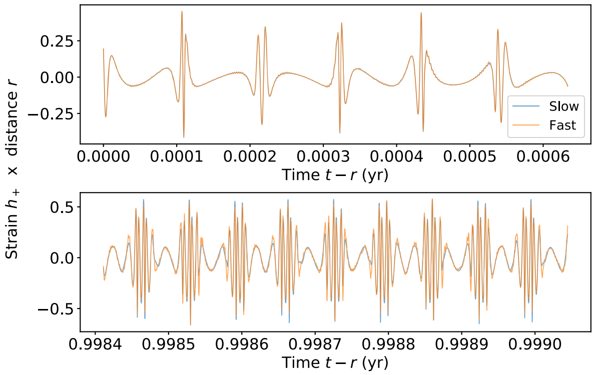

FIG. 2. Six hour snapshots of fast (orange) and fiducial (blue) waveforms, 1 yr before plunge (top) and just before plunge (bottom). Waveforms are for the worst-case EMRI $(M, \mu, p_0, e_0) = (10^6\ M_\odot, 15\ M_\odot, 10, 0.7)$, with a 1 yr mismatch of 5×10^{-4}. Small amplitude deviations are visible just before plunge at $(p, e) \approx (7, 0.5)$, where the mode-distribution error approaches its maximum across the domain of validity.

with $\sim 10^3$ modes) are dominated not just by mode summation, but also the evaluation of mode amplitudes (see Fig. 3). This is not the case for our fast model, where the bottleneck is reduced solely to summation, and wall times are reduced to ~ 1 min on a CPU and further to $\sim 10^2$ ms on a GPU.

Conclusion.—The efficient computation of fully relativistic EMRI waveform templates has yet to be achieved under the constraints of LISA data analysis, as a significant bottleneck is posed by the interpolation and evaluation of the $\sim 10^3$–10^5 mode amplitudes. In this Letter, we propose that the bottleneck can first be relieved by combining order-reduction and deep-learning techniques in the amplitude fit [23], and then virtually removed through the use of GPU acceleration. We demonstrate this by introducing the first EMRI waveform model with subsecond run-times for analysis-length signals with full harmonic content. Access to higher modes during analysis is important not just for precise inference, but also for finding signals in the first place: using our model, we find that a quadrupolar waveform with $l_{\max} = 2$ typically has a mismatch of ≈ 0.1 against a fiducial waveform, which may be suboptimal even for search [19].

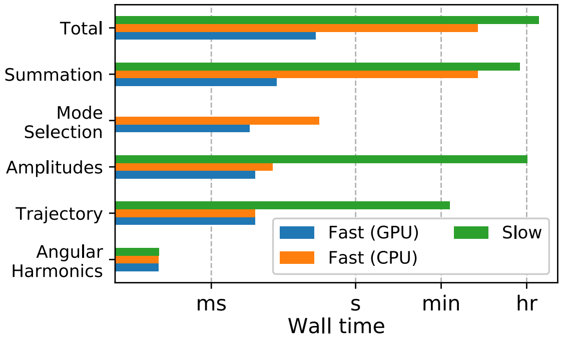

FIG. 3. Computational wall time for fast and fiducial waveforms, broken down into individual modules. All times are averaged over ≥ 5 evaluations of the worst-case waveform on a single CPU core (and GPU), where the CPU is an Intel Xeon Gold 6132 and the GPU is an NVIDIA Tesla V100.

Our present waveform model is accurate at adiabatic order for eccentric Schwarzschild orbits, and thus can already be used to construct search templates for EMRIs with a nonrotating large mass. However, LISA data analysis needs template models that describe generic Kerr EMRIs at sufficient accuracy for inference. The framework presented in this Letter is designed to accommodate the increased accuracy and extensiveness of such models while retaining efficiency. Postadiabatic waveforms require the replacement of flux-driven trajectories with self-forced trajectories, which will be equally efficient [9,31]. Practical schemes for dealing with transient resonances [51–53] can be included as well. Although the mode amplitudes are required only at leading order [9], they must be extended to cover the space of Kerr orbits; our fitting technique is promising for dealing with the increased dimensionality. The source-end waveform also has to be integrated with a realistic LISA response, which could be done through a frequency-domain approximation for both waveform [30] and response [11,14], or by developing accelerated versions of more accurate time-domain simulators [54,55]. Finally, the modular nature of the framework allows the incorporation of additional physics as well. This could include environmental effects, e.g., accretion disks [56] and massive perturbers [57,58], or new physics, e.g., beyond-general-relativity corrections [59] and beyond-standard-model physics [60,61].

A. J. K. C. acknowledges support from the Jet Propulsion Laboratory (JPL) Research and Technology Development program, and from the NASA Grant No. 18-LPS18-0027. M. L. K. acknowledges support from the National Science Foundation under Grant No. DGE-0948017. N. W. gratefully acknowledges support from a Royal Society–Science Foundation Ireland University Research Fellowship. S. A. H. is supported by NASA ATP Grant No. 80NSSC18K1091 and NSF Grant No. PHY-1707549. We thank Chad Galley for the RomPy package [44] and discussions on reduced-order modeling, as well as Ian Hinder and Barry Wardell for the SimulationTools analysis package [62]. Final thanks go out to Soichiro Isoyama and Adam Pound for helpful discussions. This work makes use of the Black Hole Perturbation Toolkit [24]. Our research was supported in part through the computational resources and staff contributions provided for the Quest/Grail high performance computing facility at Northwestern University. Parts of this work were carried out at JPL, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

[1] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. X **9**, 031040 (2019).

[2] LVK Collaboration, Living Rev. Relativity **21**, 3 (2018).

[3] S. Burke-Spolaor *et al.*, Astron. Astrophys. Rev. **27**, 5 (2019).

[4] P. Amaro-Seoane *et al.*, arXiv:1702.00786.

[5] S. Babak, J. Gair, A. Sesana, E. Barausse, C. F. Sopuerta, C. P. L. Berry, E. Berti, P. Amaro-Seoane, A. Petiteau, and A. Klein, Phys. Rev. D **95**, 103012 (2017).

[6] C. P. Berry, S. A. Hughes, C. F. Sopuerta, A. J. Chua, A. Heffernan, K. Holley-Bockelmann, D. P. Mihaylov, M. C. Miller, and A. Sesana, arXiv:1903.03686.

[7] L. Barack and A. Pound, Rep. Prog. Phys. **82**, 016904 (2019).

[8] A. Pound, B. Wardell, N. Warburton, and J. Miller, Phys. Rev. Lett. **124**, 021101 (2020).

[9] J. Miller and A. Pound, arXiv:2006.11263.

[10] J. R. Gair, L. Barack, T. Creighton, C. Cutler, S. L. Larson, E. S. Phinney, and M. Vallisneri, Classical Quantum Gravity **21**, S1595 (2004).

[11] S. Babak, J. R. Gair, and E. K. Porter, Classical Quantum Gravity **26**, 135004 (2009).

[12] N. J. Cornish, Classical Quantum Gravity **28**, 094016 (2011).

[13] Y. Wang, Y. Shang, and S. Babak, Phys. Rev. D **86**, 104050 (2012).

[14] A. J. K. Chua, N. Korsakova, C. J. Moore, J. R. Gair, and S. Babak, Phys. Rev. D **101**, 044027 (2020).

[15] N. E. M. Rifat, S. E. Field, G. Khanna, and V. Varma, Phys. Rev. D **101**, 081502 (2020).

[16] L. Barack and C. Cutler, Phys. Rev. D **69**, 082005 (2004).

[17] S. Babak, H. Fang, J. R. Gair, K. Glampedakis, and S. A. Hughes, Phys. Rev. D **75**, 024005 (2007).

[18] A. J. K. Chua and J. R. Gair, Classical Quantum Gravity **32**, 232002 (2015).

[19] A. J. K. Chua, C. J. Moore, and J. R. Gair, Phys. Rev. D **96**, 044005 (2017).

[20] EMRI Kludge Suite, http://github.com/alvincjk/EMRI_Kludge_Suite.

[21] S. A. Hughes, Phys. Rev. D **61**, 084004 (2000).

[22] S. Drasco and S. A. Hughes, Phys. Rev. D **73**, 024027 (2006).

[23] A. J. K. Chua, C. R. Galley, and M. Vallisneri, Phys. Rev. Lett. **122**, 211101 (2019).

[24] Black Hole Perturbation Toolkit, http://bhptoolkit.org.

[25] C. Hopman and T. Alexander, Astrophys. J. **629**, 362 (2005).

[26] T. Hinderer and E. E. Flanagan, Phys. Rev. D **78**, 064028 (2008).

[27] S. Drasco, E. E. Flanagan, and S. A. Hughes, Classical Quantum Gravity **22**, S801 (2005).

[28] E. Gourgoulhon, A. Le Tiec, F. H. Vincent, and N. Warburton, Astron. Astrophys. **627**, A92 (2019).

[29] P. Amaro-Seoane, Phys. Rev. D **99**, 123025 (2019).

[30] S. A. Hughes *et al.* (to be published).

[31] M. Van De Meent and N. Warburton, Classical Quantum Gravity **35**, 144003 (2018).

[32] C. Cutler, D. Kennefick, and E. Poisson, Phys. Rev. D **50**, 3816 (1994).

[33] L. C. Stein and N. Warburton, Phys. Rev. D **101**, 064007 (2020).

[34] E. Newman and R. Penrose, J. Math. Phys. (N.Y.) **3**, 566 (1962).

[35] S. A. Teukolsky, Astrophys. J. **185**, 635 (1973).

[36] R. Fujita, W. Hikida, and H. Tagoshi, Prog. Theor. Phys. **121**, 843 (2009).

[37] C. Munna, C. R. Evans, S. Hopper, and E. Forseth, Phys. Rev. D **102**, 024047 (2020).

[38] T. Osburn, N. Warburton, and C. R. Evans, Phys. Rev. D **93**, 064024 (2016).

[39] S. E. Field, C. R. Galley, F. Herrmann, J. S. Hesthaven, E. Ochsner, and M. Tiglio, Phys. Rev. Lett. **106**, 221102 (2011).

[40] S. E. Field, C. R. Galley, J. S. Hesthaven, J. Kaye, and M. Tiglio, Phys. Rev. X **4**, 031006 (2014).

[41] P. Canizares, S. E. Field, J. R. Gair, and M. Tiglio, Phys. Rev. D **87**, 124005 (2013).

[42] J. Blackman, S. E. Field, M. A. Scheel, C. R. Galley, C. D. Ott, M. Boyle, L. E. Kidder, H. P. Pfeiffer, and B. Szilágyi, Phys. Rev. D **96**, 024058 (2017).

[43] C. Cutler and M. Vallisneri, Phys. Rev. D **76**, 104018 (2007).

[44] RomPy, http://bitbucket.org/chadgalley/rompy.

[45] M. Kuhn and K. Johnson, *Applied Predictive Modeling* (Springer, New York, 2013).

[46] G. Khanna and J. McKennon, Comput. Phys. Commun. **181**, 1605 (2010).

[47] J. McKennon, G. Forrester, and G. Khanna, in *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the EXtreme to the Campus and Beyond, XSEDE 12* (Association for Computing Machinery, New York, NY, USA, 2012).

[48] C. Talbot, R. Smith, E. Thrane, and G. B. Poole, Phys. Rev. D **100**, 043030 (2019).

[49] M. L. Katz, S. Marsat, A. J. K. Chua, S. Babak, and S. L. Larson, Phys. Rev. D **102**, 023033 (2020).

[50] FastEMRIWaveforms, Code available at https://bhptoolkit.org/FastEMRIWaveforms, along with detailed documentation and example notebooks.

[51] J. R. Gair, E. E. Flanagan, S. Drasco, T. Hinderer, and S. Babak, Phys. Rev. D **83**, 044037 (2011).

[52] E. E. Flanagan, S. A. Hughes, and U. Ruangsri, Phys. Rev. D **89**, 084028 (2014).

[53] J. Brink, M. Geyer, and T. Hinderer, Phys. Rev. Lett. **114**, 081102 (2015).

[54] M. Vallisneri, Phys. Rev. D **71**, 022001 (2005).

[55] A. Petiteau, G. Auger, H. Halloin, O. Jeannin, E. Plagnol, S. Pireaux, T. Regimbau, and J.-Y. Vinet, Phys. Rev. D **77**, 023002 (2008).

[56] B. Kocsis, N. Yunes, and A. Loeb, Phys. Rev. D **84**, 024032 (2011).

[57] H. Yang and M. Casals, Phys. Rev. D **96**, 083015 (2017).

[58] B. Bonga, H. Yang, and S. A. Hughes, Phys. Rev. Lett. **123**, 101103 (2019).

[59] A. Maselli, N. Franchini, L. Gualtieri, and T. P. Sotiriou, Phys. Rev. Lett. **125**, 141101 (2020).

[60] O. A. Hannuksela, K. W. Wong, R. Brito, E. Berti, and T. G. Li, Nat. Astron. **3**, 447 (2019).

[61] O. A. Hannuksela, K. C. Ng, and T. G. Li, Phys. Rev. D **102**, 103022 (2020).

[62] SimulationsTools, https://simulationtools.org.