Bayesian optimization of laser-plasma accelerators assisted by reduced physical models

A. Ferran Pousa,[1,*] S. Jalas,[1] M. Kirchen,[1] A. Martinez de la Ossa,[1] M. Thévenet,[1]
S. Hudson,[2] J. Larson,[2] A. Huebl,[3] J.-L. Vay,[3] and R. Lehe[3]

[1]*Deutsches Elektronen-Synchrotron DESY, Notkestrasse 85, 22607 Hamburg, Germany*
[2]*Argonne National Laboratory, Lemont, Illinois 60439, USA*
[3]*Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*

(Received 3 January 2023; accepted 10 July 2023; published 15 August 2023)

Particle-in-cell simulations are among the most essential tools for the modeling and optimization of laser-plasma accelerators, since they reproduce the physics from first principles. However, the high computational cost associated with them can severely limit the scope of parameter and design optimization studies. Here, we show that a multitask Bayesian optimization algorithm can be used to mitigate the need for such high-fidelity simulations by incorporating information from inexpensive evaluations of reduced physical models. In a proof-of-principle study, where a high-fidelity optimization with FBPIC is assisted by reduced-model simulations with Wake-T, the algorithm demonstrates an order-of-magnitude speedup. This opens a path for the cost-effective optimization of laser-plasma accelerators in large parameter spaces, an important step toward fulfilling the high beam quality requirements of future applications.

DOI: 10.1103/PhysRevAccelBeams.26.084601

I. INTRODUCTION

Laser-plasma accelerators (LPAs) make use of a plasma medium to transform the energy of a laser pulse into large longitudinal electric fields capable of accelerating particles to high energies in a short distance [1]. This process depends on a complex interplay of nonlinear physical phenomena that determine the final performance of the accelerator. The laser-plasma interaction (manifested as laser self-focusing, dephasing, and depletion [2]), the injection of electrons into the plasma wake [3–11], the beam-plasma interaction (especially beam loading [12–18]), and the dynamics of the injected electrons in the resulting plasma fields [19–21] dictate the final properties of the generated beams. These processes can be controlled, up to a certain extent, by the parameters and design properties of the setup. Typical examples include the plasma density profile (e.g., [11,17,18,22–25]), the properties of the laser pulse [26], or the use of external laser guiding [27–29]. Careful tuning and optimization of these parameters is critical for realizing LPAs that are capable of delivering the high beam quality and stability demanded by applications, particularly for free-electron lasers [30], storage ring injectors [31,32], and future colliders [33].

Due to the complexity of the physical processes involved, the optimization of an LPA design requires the use of high-fidelity particle-in-cell (PIC) simulations [34] where the self-consistent interaction between particles and electromagnetic fields is computed with minimal assumptions. However, the high computational cost associated with these simulations makes optimizing over a large set of parameters practically unfeasible. This limits the number of configurations that can be explored for achieving optimal performance.

Developing more efficient techniques for optimizing the design of LPAs is therefore an important step toward realizing the full potential of these novel accelerators. Besides the continued growth of available computing power, two approaches for more affordable optimization can be identified: reducing the number of simulations required to find the best-performing configuration, and reducing the cost per simulation.

The number of required simulations can be minimized by utilizing advanced algorithms that predict and evaluate only the most promising configurations throughout an optimization run. An example of this is Bayesian optimization [35], a machine learning-based technique that has gained popularity within the accelerator community [18,36–40]. This method generates a surrogate model of the simulation outcome (typically using Gaussian processes [41]) and suggests the most promising candidates for evaluation based on a balance between exploration (evaluating unmapped regions of the

[*] angel.ferran.pousa@desy.de

parameter space where new optima could be located) and exploitation (further sampling around known optima). The underlying model is continuously updated with the results from new evaluations, allowing for more promising and accurate suggestions in successive iterations. With this approach, the method can identify global optima with a reduced number of evaluations.

The computational cost of the individual simulations can be mitigated by making use of reduced models that sacrifice generality or accuracy by introducing physical approximations. This can involve both reducing the dimensionality (e.g., assuming quasicylindrical symmetry [42]) or neglecting certain physical properties of the laser-plasma interaction that are not dominant in the problem at hand. Common examples of the latter include the use of a laser envelope model [43] or assuming the wakefield to be quasistatic [43,44]. In principle, simulations with such reduced models can fully replace a complete PIC description if they accurately capture all the relevant physics involved. In other cases, they provide an approximate solution from which useful information might still be extracted.

In this paper, we show that the computational cost of Bayesian optimization can be further reduced with the assistance of inexpensive reduced-model evaluations that are performed in tandem with costly high-fidelity simulations. The inexpensive evaluations are used to dynamically probe regions of high interest and gather information that improves the predictions of the most promising configurations to evaluate at high fidelity. This strategy is enabled by the use of a multitask Gaussian process model [45–47], whereby the correlation between the outputs of different tasks (i.e., the two levels of fidelity in the proposed method) is learned so that information gained on one task results in an improved model of the other. In this way, the need for high-fidelity simulations is further reduced, leading to a faster and cheaper optimization. This is demonstrated here by a proof-of-principle study combining the simulation codes FBPIC [48] and Wake-T [49], which provide a full PIC description in quasi-3D geometry and inexpensive reduced models, respectively.

II. MULTITASK BAYESIAN OPTIMIZATION

Bayesian optimization is an efficient technique for the global optimization of black-box functions that are noisy and expensive to evaluate. It operates by building a probabilistic *surrogate model* of the *objective function* f (the function to minimize or maximize) that is cheaper to evaluate than f and from which the most promising points to query can be determined by maximizing an *acquisition function*. The surrogate model is typically obtained by performing Gaussian process regression over the available data. This provides an estimate of f and its associated uncertainty at any point of the parameter space. Determining which points to evaluate next depends on a balance between querying around known optima or exploring regions of high uncertainty where new optima could be identified. This balance is quantified by the acquisition function, and the points that maximize it are deemed as most promising for future evaluation. A typical choice for the acquisition function is the expected improvement [35], which quantifies how much a new evaluation is expected to improve over the current optimum. Once the new evaluations are completed, the Gaussian process model is updated with the obtained data and the same procedure is repeated. This continuously improves the accuracy of the model and of the suggested evaluations.

With the use of a multitask Gaussian process (MTGP) [45], Bayesian optimization can be extended to a collection of objective functions $f_1, \ldots, f_{N_t}$ from N_t different *tasks*. The MTGP learns the correlations between them and provides a surrogate model of each objective that features a reduced uncertainty by incorporating information from highly correlated tasks. This exchange of information was originally proposed as a way of transferring the knowledge of previous optimizations to new tasks in order to optimize them more efficiently [46]. Here, we make use of the approach described in Ref. [47], where an inexpensive task f_R (the reduced physical models) is used to assist in the optimization of a costly function f_H (the high-fidelity PIC simulations) so that the number of required evaluations of f_H is reduced. This strategy is a special case of multifidelity optimization where only two discrete levels of fidelity are considered. Alternative multifidelity algorithms adapted to multiple objectives have also been explored in the context of particle accelerators [50,51].

In this two-task approach, the covariance function—or *kernel*—that enables the MTGP to transfer information between tasks t and t' with inputs x and x' is defined as $k[(t, \mathbf{x}), (t', \mathbf{x}')] = B_{tt'}\kappa(\mathbf{x}, \mathbf{x}')$ [47]. This expression determines the covariance between data points from different tasks by assuming that both tasks share the same kernel $\kappa(\mathbf{x}, \mathbf{x}')$ for the input parameters (here, a Matérn $5/2$ kernel [41] is used) and that the task covariance can be captured separately by a 2×2 matrix B, where element $B_{tt'}$ is the covariance between t and t'. The coefficients of B as well as the parameters of $\kappa(\mathbf{x}, \mathbf{x}')$ are kernel hyperparameters that are inferred from the available data by maximizing marginal likelihood [47]. The degree of inter-task correlation can be quantified by $\rho^2 = B_{tt'}^2/(B_{tt}B_{t't'})$, which ranges between $\rho^2 = 0$ (no correlation) to $\rho^2 = 1$ (maximum correlation).

Using this MTGP model, the Bayesian optimization loop described in Ref. [47] is performed. As summarized in Fig. 1, batches of N_R reduced-model simulations and N_H high-fidelity simulations (with $N_H \leq N_R$) are executed in tandem. At each iteration, the optimizer (i) fits an MTGP to the available data, (ii) determines a set $\{\mathbf{x}_i\}_{i=1,\ldots,N_R}$ of the N_R most promising points to query by maximizing noisy expected improvement [52] on the MTGP model for the *high-fidelity* output $f_H(\mathbf{x})$, (iii) evaluates these N_R points using *reduced-model* simulations, (iv) updates the MTGP

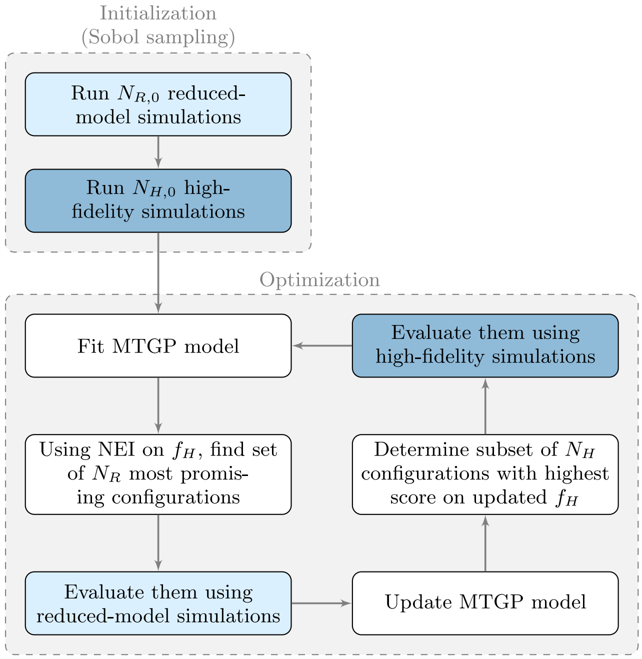
Initialization
(Sobol sampling)
Run $N_{R,0}$ reduced-model simulations
Run $N_{H,0}$ high-fidelity simulations
Optimization
Fit MTGP model
Evaluate them using high-fidelity simulations
Using NEI on f_H, find set of N_R most promising configurations
Determine subset of N_H configurations with highest score on updated f_H
Evaluate them using reduced-model simulations
Update MTGP model

FIG. 1. Workflow of the implemented algorithm for multitask Btextayesian optimization. Batches of reduced-model and high-fidelity simulations are highlighted in light and dark blue, respectively.

with the obtained results, (v) evaluates the N_R points in the updated surrogate model of $f_H(\mathbf{x})$ to select the N_H points with the most promising outcome, and (vi) evaluates the reduced sample of N_H points with *high-fidelity* simulations. To start the optimization, initial quasirandom samples of $N_{R,0}$ reduced-model simulations and $N_{H,0}$ high-fidelity simulations are generated by using two separate scrambled Sobol sequences [53] of input parameters.

This workflow has been implemented in OPTIMAS [54,55], a Python package that allows for scalable optimization on high-performance computing systems. The Bayesian optimization functionality relies on the AX library [56] and can be executed on both CPUs and GPUs. The allocation of computing resources for the optimizer and the simulations, as well as the coordination, execution, and communication between them, is orchestrated internally by the LIBENSEMBLE library [57]. This allows for the concurrent evaluation of multiple simulations that can make use of variable resources (number of CPUs and GPUs) across the available computing nodes.

The use of noisy expected improvement as acquisition function allows for the generation of large simulation batches while taking into account the noise of the observations and, if needed, of any optimization constraints. However, it is also the reason why N_H and N_R stay fixed throughout the optimization, as it does not directly provide a way of determining the fidelity and size of the simulation batches. Alternative acquisition functions, such as knowledge gradient [58] or predictive entropy search [59,60], could allow for a dynamic selection of the fidelity, but their implementation in a multitask optimization workflow is not trivial and remains an active area of research [47].

III. PROOF-OF-PRINCIPLE STUDY

The effectiveness of the proposed algorithm is demonstrated here by a proof-of-principle optimization study combining the simulation codes FBPIC [48] and Wake-T [49]. While FBPIC provides a high-fidelity, fully electromagnetic PIC description of the LPA physics in quasi-3D geometry [42], Wake-T allows for inexpensive simulations by using a reduced quasistatic wakefield model with 2D cylindrical symmetry [61] and a laser envelope model [62].

The setup to be optimized is an LPA booster stage for an externally injected electron bunch. Given a fixed laser driver and plasma profile, the goal is to determine the bunch current profile that results in the lowest energy spread with the highest possible charge and energy. This involves optimizing the net beam loading effect [12–14,17,18] throughout the LPA, a nontrivial process affected by laser dephasing, depletion, and diffraction for which no analytical theory is available and that must therefore be addressed with simulations.

To simultaneously achieve low energy spread, high charge, and high energy, these quantities are combined into a single objective to maximize:

$$f = \frac{k_Q E_{\mathrm{MED}}[\mathrm{GeV}]}{k_{\mathrm{MAD}}}, \tag{1}$$

where $k_Q = Q_{\mathrm{tot}}/Q_{\mathrm{ref}}$ is the ratio between the total Q_{tot} and a reference $Q_{\mathrm{ref}} = 10\mathrm{pC}$ charge, $k_{\mathrm{MAD}} = \Delta E_{\mathrm{MAD}}/\Delta E_{\mathrm{MAD,ref}}$ is the ratio between the relative energy spread ΔE_{MAD} and a reference value $\Delta E_{\mathrm{MAD,ref}} = 10^{-2}$, and E_{MED} is the median energy. The use of the median absolute deviation (MAD) energy spread and median energy provides a robust characterization of the energy spectrum in distributions with outliers, as typically observed in LPAs [17,18]. The value of f given by Eq. (1) can span over several orders of magnitude and feature sharp extremes that are not ideal for Gaussian process modeling. To alleviate this, the objective is internally treated by the optimizer as $\log(f)$.

The parameters of the laser driver are an energy $E_L = 10$ J, an FWHM duration $\tau_{\mathrm{FWHM}} = 25\mathrm{fs}$, a focal spot size $w_0 = 40$ μm, a wavelength $\lambda_0 = 800\mathrm{nm}$, and a peak normalized vector potential $a_0 \simeq 2.6$. The plasma density profile is a simple 10 cm-long flat-top with an on-axis electron density $n_{e,0} = 2\times 10^{17}$ cm^{-3} and a parabolic radial profile for laser guiding $n_e(r) = n_{e,0} + r^2/(\pi r_e w_0^4)$ [63]. The externally injected electron bunch has an initial energy $E_{b,0} = 200$ MeV with an rms energy spread of 0.1%. It features a normalized emittance of $\epsilon_{n,x} = 3$ μm in the horizontal direction and of $\epsilon_{n,y} = 0.5$ μm in the vertical plane. This difference between the x and y emittances typically arises in LPAs based on ionization injection as a result of the laser polarization [18]. Here, it is included in the externally injected bunch in order to ensure a bias between the two simulation codes, as this asymmetry can only be fully captured by the high-fidelity FBPIC simulations. The initial transverse size is matched to the focusing strength in the plasma, allowing for emittance preservation [19,64]. The longitudinal profile of the bunch is assumed to be trapezoidal, as it is known to be well suited for beam loading [13], and features smooth Gaussian ramps (1 μm rms) at the head and tail. The parameters exposed to the optimizer are the current at the head I_h and tail I_t, the bunch length L_b, and its longitudinal position in the wake, parameterized by the distance $\Delta z_{l,h} = z_l - z_h$ between the head of the bunch z_h and the center of the laser driver z_l. They are allowed to vary in the following ranges: $I_h \in [0.1, 10]$ kA, $I_t \in [0.1, 10]$ kA, $L_b \in [1, 20]$ μm, and $\Delta z_{l,h} \in [40, 60]$ μm.

The FBPIC simulations are performed using the boosted frame technique [65,66] with a Lorentz boost factor of 25. The longitudinal and radial resolutions are $dz = \lambda_0/80$ and $dr = k_p^{-1}/20$, respectively, where $k_p = (n_{e,0}e^2/m_e\epsilon_0 c^2)^{1/2}$ is the plasma wave number, e is the elementary charge, ϵ_0 is the vacuum permittivity and c is the speed of light. The number of particles per cell is 2 in both z and r, and 8 in the azimuthal direction. Three azimuthal modes are used to properly describe the ellipticity of the electron bunch. The simulations are performed on a single NVIDIA A100 GPU and have a typical execution time of ∼40 min. The Wake-T simulations have a resolution of $dz = c\tau_\mathrm{FWHM}/40$ and $dr = k_p^{-1}/20$ with two particles per cell. Each simulation is performed on a single core of an AMD EPYC 7643 CPU, with a typical execution time of ∼3 min. The entire optimization is carried out in one compute node with 96 CPU cores and 4 GPUs. One of the GPUs is dedicated to the optimizer (fitting the MTGP, acquisition function, etc.), and the remaining resources are available for simulations. With this setup, batches of either $N_R = 96$ concurrent Wake-T simulations or $N_H = 3$ concurrent FBPIC simulations can be performed. The scripts used to define and carry out this optimization workflow are available on Ref. [67].

To quantify the performance gain from the multitask approach, the same physical setup is also optimized solely with batches of three FBPIC simulations using a Bayesian algorithm based on a standard single-task Gaussian process model. This optimization is carried out in the same hardware and uses the same initialization routine and acquisition function as the multitask case. Since each optimization run evolves differently—both the initial sample of points and the optimization of the acquisition function include a certain degree of randomness—a total of six independent multitask and single-task optimizations have been carried out to determine the average evolution and its variance.

The results of this optimization study, shown in Fig. 2, indicate that incorporating information from reduced-model simulations using the multitask algorithm leads, on average, to an order-of-magnitude speedup in terms of the time to converge to a solution, and to a reduced variability in the convergence rate. For example, an average objective value of $f_H = 280$ is reached after ∼6 h when the optimization is assisted with Wake-T simulations, while this number grows to ∼45 h when only FBPIC is used. This boost in performance is achieved despite the outcome of both codes not being in full agreement with each other, as evidenced in Fig. 2(b). However, owing to the high degree of correlation between them ($\langle\rho^2\rangle \simeq 0.82$, where $\langle\rangle$ denotes average over the 6 runs), the multitask algorithm can capture the bias of the reduced model with respect to the high-fidelity simulations and extract useful information from it.

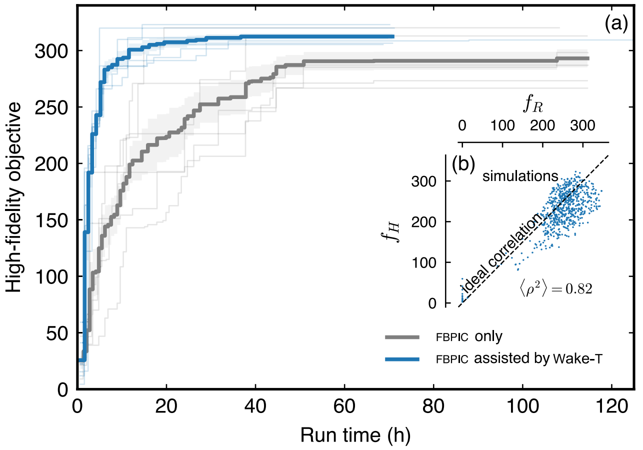

FIG. 2. (a) Average (thick line) and standard error (shaded area) of the evolution of the high-fidelity FBPIC objective with and without the assistance of reduced-model simulations with Wake-T. Six runs (thin lines) were performed for each case. (b) Correlation between the FBPIC (f_H) and Wake-T (f_R) objectives in the multitask optimization as obtained from individual simulations.

This approach successfully manages to optimize the given setup. The highest scoring FBPIC simulation ($f_H \simeq 323$) from the 6 multitask optimizations corresponds to a configuration with $I_h = 4.26$ kA, $I_t = 3.50$ kA, $L_b = 6.35$ μm and $\Delta z_{l,h} = 55.2$ μm, which results in a total charge of 114.6 pC, a mean energy of 2.9 GeV, and a relative energy spread of 0.1% (MAD). Figure 3 shows the outcome of the FBPIC and Wake-T simulations for this working point. Certain differences can be observed in the plasma wake, particularly toward the back, where highly relativistic plasma electrons cannot be accurately modeled within the quasistatic approximation. The evolution of the longitudinal phase space seen in Figs. 3(b)–3(d) shows that, as originally intended, an optimal net beamloading is achieved at the end of the LPA. Even though the energy spread can be locally high at some points during acceleration, the laser evolution and the subsequent changes to the plasma wake along the LPA end up resulting in a flattened energy distribution.

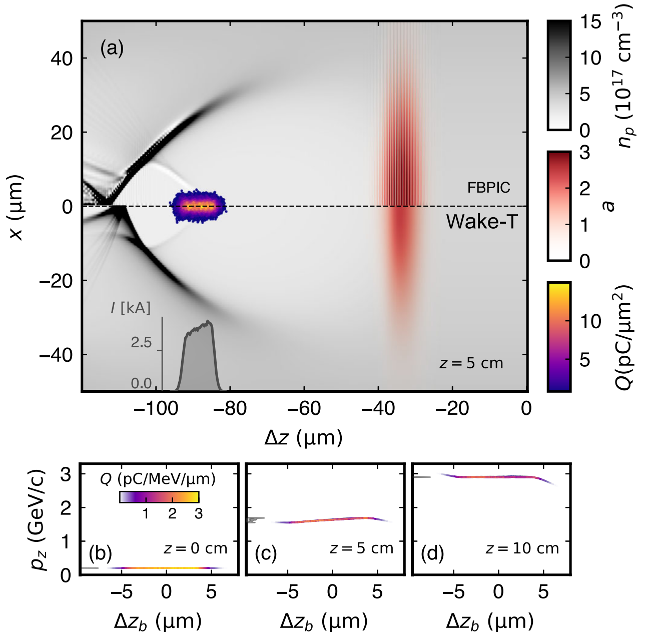

FIG. 3. Outcome of the highest scoring FBPIC simulation. (a) Plasma wakefields at the center of the LPA as obtained from FBPIC (top) and Wake-T (bottom). Longitudinal phase space at the (b) start, (c) middle, and (d) end of the FBPIC simulation. Δz and Δz_b are the longitudinal positions relative to the front of the simulation box and the beam center, respectively.

A detailed view of the sequence of simulation batches and the evolution of f_R and f_H in a multitask optimization is included in Fig. 4. The Wake-T simulations have a negligible cost compared to the FBPIC batches, allowing for broad and inexpensive exploration so that only the most promising configurations are evaluated at high fidelity. This allows for a much faster convergence of f_H, which evolves at virtually the same rate as f_R despite the reduced number of simulations. However, one potential drawback of performing large batches of reduced-model simulations is a rapid increase in the cost of suggesting new configurations. This is a result of the $\mathcal{O}(N^3)$ cost scaling of fitting the MTGP [41], where N is the total amount of evaluations, together with the also increasing costs of evaluating the model and optimizing the acquisition function. Clear evidence of this can be seen in Fig. 4, where the intervals between simulation batches progressively widen as the total number of evaluations increases. Therefore, determining an adequate ratio between N_R and N_H is of high relevance for a well-performing optimization. Otherwise, the cost savings from the increased convergence rate could be counterbalanced, at least in part, by the growing cost of the multitask optimizer.

The influence of the ratio between N_R and N_H is investigated here with a series of optimizations where the number of Wake-T simulations per batch is varied. In addition to the original study with $N_R = 96$, three more cases (each of them consisting of six independent runs) with $N_R = 48, 24$, and 12 are included. For each case, the evolution of f_H as well as the fraction of time that is spent purely in the optimizer, t_{opt}, are quantified. The general outcome of this scan is that reducing N_R leads to a slower convergence in terms of the number of iterations but to a faster optimizer (i.e., smaller t_{opt}). These two effects partially compensate each other in terms of total run time, leading to no significant differences between the cases with $N_R = 96, 48$, and 24, as seen in Fig. 5. To achieve an objective $f_H \geq 250$, which is reached by all runs, the

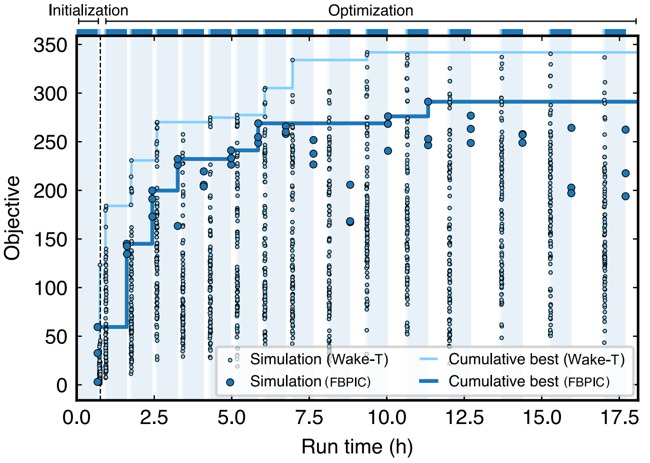

FIG. 4. Evolution of a multitask optimization with alternating batches of Wake-T and FBPIC simulations. The run time of each batch is indicated by the light blue (Wake-T) and dark blue (FBPIC) shaded areas. Intervals between batches correspond to the time when the optimizer is computing the next set of configurations to evaluate. Outcomes of each simulation and the evolution of the cumulative best objective are also included.

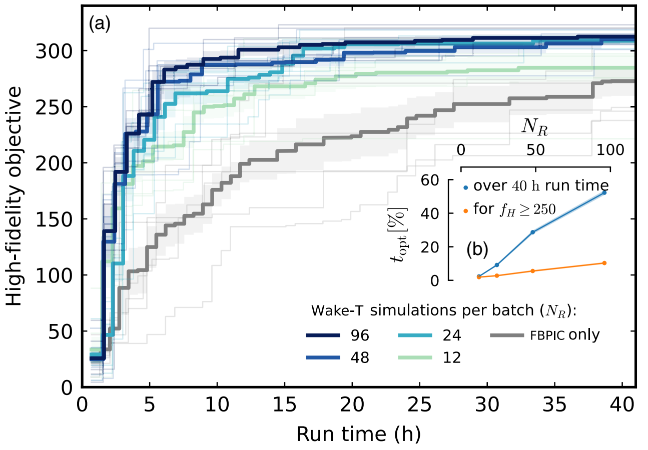

FIG. 5. (a) Evolution of the high-fidelity objective in multitask optimizations using a different number of Wake-T simulations per batch compared against a single-task (FBPIC only) benchmark. (b) Percentage of total run time consumed by the optimizer as a function of N_R. This percentage is measured both over the 40 h period shown and, alternatively, over the time needed to reach an objective $f_H \geq 250$.

time consumed by the optimizer is moderate in all cases, ranging from $t_{\mathrm{opt}} \simeq 10\%$ when $N_R = 96$ to $t_{\mathrm{opt}} \simeq 2\%$ when $N_R = 12$. However, when quantifying t_{opt} over the 40 h period displayed in Fig. 5, the optimizer becomes the dominant contribution to the total run time ($t_{\mathrm{opt}} \simeq 52\%$) when $N_R = 96$ while remaining negligible ($t_{\mathrm{opt}} \simeq 2\%$) when $N_R = 12$. This is not particularly concerning here, as the case with $N_R = 96$ reaches a close-to-optimal objective well before the 40 h threshold. However, if a higher number of iterations were required, such as in a case where the reduced models and the high-fidelity simulations are not as well correlated, it could lead to a significant loss in performance. Based on the results from this scan, $N_R = 24$ appears to be an adequate choice that leads to virtually the same rate of convergence as cases with higher N_R while allowing, if needed, for a larger number of iterations.

Potential strategies that could be pursued in the future for reducing t_{opt} include, for example, the use of dynamic batch sizes, so that the number of reduced-model simulations can be decreased when they no longer provide useful information, or establishing a criterion for removing the least valuable reduced-model observations at each iteration.

In general, depending on the physical problem to optimize, different reduced models of varying fidelity and cost might be available. As such, studying the behavior of the multitask method under varying degrees of inter-task correlation is of high relevance for its general applicability. In particular, it is important to ensure that the method converges to a meaningful optimum despite any degradation of the information gained from the reduced model and a potential increase in overall costs. To test this, an additional set of optimizations has been performed where the fidelity of f_R is reduced by decreasing the resolution of the Wake-T simulations. In addition to the original setup, 3 cases with a factor of 2, 4, and 8 lower resolution in both z and r are included. Due to the expected reduction in convergence rate, all optimizations are performed with $N_R = 24$ to allow for a larger number of iterations without significantly increasing t_{opt}. The results from this study, summarized in Fig. 6, clearly indicate that a loss in correlation directly translates into a slower convergence rate. However, even with a moderate correlation ($\langle \rho^2 \rangle \simeq 0.57$), the multitask algorithm can still provide a significant performance gain. Only when the two tasks are essentially independent (i.e., $\langle \rho^2 \rangle \sim 10^{-3}$ in the lowest resolution case) does the rate of convergence decrease below the single-task benchmark. This is because even though the MTGP reduces to a single task when $\rho^2 = 0$ [47], the inaccurate data from f_R can still influence the surrogate model of f_H until sufficient evaluations to infer the lack of correlation have been gathered. As such, the multitask technique converges toward the optimum even with unreliable or misleading reduced-model data, and provides a performance boost over single-task optimization as long as a meaningful inter-task correlation can be recognized.

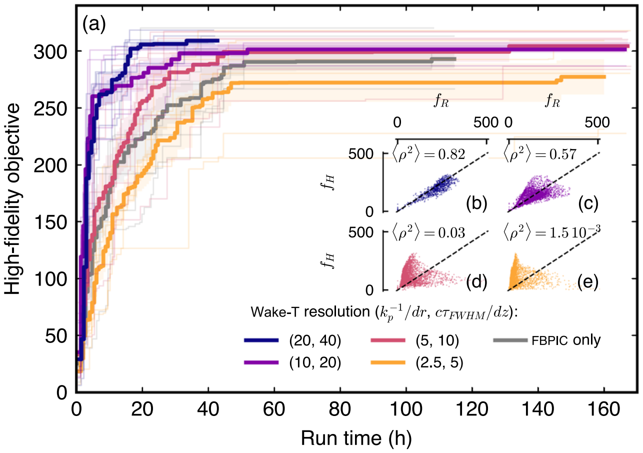

FIG. 6. (a) Evolution of the high-fidelity objective in multitask optimizations assisted by Wake-T simulations of different resolutions compared against a single-task (FBPIC only) benchmark. (b), (c), (d), and (e) show the correlation between Wake-T (f_R) and FBPIC (f_H) results in each case.

Multitask optimization could also prove to be useful beyond laser-plasma acceleration. For example, the design of conventional accelerator components, such as bunch compressors or transfer lines, typically involves multiple levels of fidelity (from simple transfer matrix calculations to full particle-tracking simulations with 3D effects such as space-charge or synchrotron radiation) that are optimized separately (see, e.g., [68,69]). A multitask algorithm would be able to combine these different levels of fidelity and cost into a single optimization.

IV. CONCLUSION

The proposed multitask method introduces the capability of leveraging reduced physical models for assisting in the Bayesian optimization of LPAs and lowering the need for costly high-fidelity simulations. In a proof-of-principle study combining the simulation codes FBPIC (high fidelity) and Wake-T (reduced models), this technique demonstrates an order-of-magnitude speedup over an equivalent single-task Bayesian optimization consisting solely of FBPIC simulations. This improvement in performance depends on the ratio of reduced-model to high-fidelity simulations, the cost difference between them, and their degree of correlation. An excessive number of reduced model simulations can increase the computational cost of suggesting new configurations, thus partially counterbalancing the gain in performance, while carrying out too few can slow down the convergence rate. Batches of $N_R = 24$ Wake-T simulations and $N_H = 3$ FBPIC simulations were found to be an adequate balance in the presented study. The choice of a reduced model that correlates well with the high-fidelity simulations is essential for achieving a significant speedup, although the algorithm converges toward the optimum even if no information is gained from the inexpensive simulations. The high computational efficiency of this method allows for the cost-effective optimization of LPAs in large parameter spaces. This is a critical step toward unlocking the full potential of these devices and fulfilling the high beam quality requirements of applications such as free-electron lasers, storage-ring injectors, and future particle colliders.

ACKNOWLEDGMENTS

We acknowledge support from DESY (Hamburg, Germany), a member of the Helmholtz Association HGF, and funding by the Helmholtz Matter and Technologies Accelerator Research and Development Program. This research was supported in part through the Maxwell computational resources operated at DESY, Hamburg, Germany. This material is based upon work supported by the CAMPA collaboration, a project of the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research and Office of High Energy Physics, Scientific Discovery through Advanced Computing (SciDAC) program under Contract No. DE-AC02-06CH11357 and No. DE-AC02-05CH11231 and by the Exascale Computing Project (17-SC-20-SC).

[1] T. Tajima and J. M. Dawson, Laser Electron Accelerator, Phys. Rev. Lett. **43**, 267 (1979).

[2] E. Esarey, C. B. Schroeder, and W. P. Leemans, Physics of laser-driven plasma-based electron accelerators, Rev. Mod. Phys. **81**, 1229 (2009).

[3] A. Modena, Z. Najmudin, A. Dangor, C. Clayton, K. Marsh, C. Joshi, V. Malka, C. Darrow, C. Danson, D. Neely, and F. N. Walsh, Electron acceleration from the breaking of relativistic plasma waves, Nature (London) **377**, 606 (1995).

[4] S. Bulanov, F. Pegoraro, A. Pukhov, and A. Sakharov, Transverse-Wake Wave Breaking, Phys. Rev. Lett. **78**, 4205 (1997).

[5] S. Bulanov, N. Naumova, F. Pegoraro, and J. Sakai, Particle injection into the wave acceleration phase due to nonlinear wake wave breaking, Phys. Rev. E **58**, R5257 (1998).

[6] D. Umstadter, J. K. Kim, and E. Dodd, Laser Injection of Ultrashort Electron Pulses into Wakefield Plasma Waves, Phys. Rev. Lett. **76**, 2073 (1996).

[7] A. Pak, K. A. Marsh, S. F. Martins, W. Lu, W. B. Mori, and C. Joshi, Injection and Trapping of Tunnel-Ionized Electrons into Laser-Produced Wakes, Phys. Rev. Lett. **104**, 025003 (2010).

[8] C. McGuffey, A. G. R. Thomas, W. Schumaker, T. Matsuoka, V. Chvykov, F. J. Dollar, G. Kalintchenko, V. Yanovsky, A. Maksimchuk, K. Krushelnick, V. Y. Bychenkov, I. V. Glazyrin, and A. V. Karpeev, Ionization Induced Trapping in a Laser Wakefield Accelerator, Phys. Rev. Lett. **104**, 025004 (2010).

[9] J. Faure, C. Rechatin, A. Norlin, A. Lifschitz, Y. Glinec, and V. Malka, Controlled injection and acceleration of electrons in plasma wakefields by colliding laser pulses, Nature (London) **444**, 737 (2006).

[10] A. J. Gonsalves, K. Nakamura, C. Lin, D. Panasenko, S. Shiraishi, T. Sokollik, C. Benedetti, C. B. Schroeder, C. G. R. Geddes, J. van Tilborg, J. Osterhoff, E. Esarey, C. Toth, and W. P. Leemans, Tunable laser plasma accelerator based on longitudinal density tailoring, Nat. Phys. **7**, 862 (2011).

[11] A. Buck, J. Wenz, J. Xu, K. Khrennikov, K. Schmid, M. Heigoldt, J. M. Mikhailova, M. Geissler, B. Shen, F. Krausz, S. Karsch, and L. Veisz, Shock-front Injector for High-Quality Laser-Plasma Acceleration, Phys. Rev. Lett. **110**, 185006 (2013).

[12] S. van der Meer, Improving the power efficiency of the plasma wakefield accelerator, CERN, Geneva, Tech. Report No. CERN-PS-85-65-AA, CLIC-Note-3, 1985, https://cds.cern.ch/record/163918.

[13] M. Tzoufras, W. Lu, F. S. Tsung, C. Huang, W. B. Mori, T. Katsouleas, J. Vieira, R. A. Fonseca, and L. O. Silva, Beam Loading in the Nonlinear Regime of Plasma-based Acceleration, Phys. Rev. Lett. **101**, 145002 (2008).

[14] K. V. Lotov, Efficient operating mode of the plasma wakefield accelerator, Phys. Plasmas **12**, 053105 (2005).

[15] C. Rechatin, X. Davoine, A. Lifschitz, A. B. Ismail, J. Lim, E. Lefebvre, J. Faure, and V. Malka, Observation of Beam Loading in a Laser-Plasma Accelerator, Phys. Rev. Lett. **103**, 194804 (2009).

[16] J. P. Couperus, R. Pausch, A. Köhler, O. Zarini, J. M. Krämer, M. Garten, A. Huebl, R. Gebhardt, U. Helbig, S. Bock, K. Zeil, A. Debus, M. Bussmann, U. Schramm, and A. Irman, Demonstration of a beam loaded nanocoulomb-class laser wakefield accelerator, Nat. Commun. **8**, 487 (2017).

[17] M. Kirchen, S. Jalas, P. Messner, P. Winkler, T. Eichner, L. Hübner, T. Hülsenbusch, L. Jeppe, T. Parikh, M. Schnepp, and A. R. Maier, Optimal Beam Loading in a Laser-Plasma Accelerator, Phys. Rev. Lett. **126**, 174801 (2021).

[18] S. Jalas, M. Kirchen, P. Messner, P. Winkler, L. Hübner, J. Dirkwinkel, M. Schnepp, R. Lehe, and A. R. Maier, Bayesian Optimization of a Laser-Plasma Accelerator, Phys. Rev. Lett. **126**, 104801 (2021).

[19] R. Assmann and K. Yokoya, Transverse beam dynamics in plasma-based linacs, Nucl. Instrum. Methods Phys. Res., Sect. A **410**, 544 (1998).

[20] P. Michel, C. B. Schroeder, B. A. Shadwick, E. Esarey, and W. P. Leemans, Radiative damping and electron beam dynamics in plasma-based accelerators, Phys. Rev. E **74**, 026501 (2006).

[21] A. Ferran Pousa, A. Martinez de la Ossa, and R. W. Assmann, Intrinsic energy spread and bunch length growth in plasma-based accelerators due to betatron motion, Sci. Rep. **9**, 17690 (2019).

[22] A. Döpp, E. Guillaume, C. Thaury, J. Gautier, K. Ta Phuoc, and V. Malka, 3d printing of gas jet nozzles for laser-plasma accelerators, Rev. Sci. Instrum. **87**, 073505 (2016).

[23] K. Schmid, A. Buck, C. M. S. Sears, J. M. Mikhailova, R. Tautz, D. Herrmann, M. Geissler, F. Krausz, and L. Veisz, Density-transition based electron injector for laser driven wakefield accelerators, Phys. Rev. ST Accel. Beams **13**, 091301 (2010).

[24] E. Guillaume, A. Döpp, C. Thaury, K. Ta Phuoc, A. Lifschitz, G. Grittani, J.-P. Goddet, A. Tafzi, S. W. Chou, L. Veisz, and V. Malka, Electron Rephasing in a Laser-Wakefield Accelerator, Phys. Rev. Lett. **115**, 155002 (2015).

[25] G. A. Bagdasarov, P. V. Sasorov, V. A. Gasilov, A. S. Boldarev, O. G. Olkhovskaya, C. Benedetti, S. S. Bulanov, A. Gonsalves, H.-S. Mao, C. B. Schroeder, J. van Tilborg, E. Esarey, W. P. Leemans, T. Levato, D. Margarone, and G. Korn, Laser beam coupling with capillary discharge plasma for laser wakefield acceleration applications, Phys. Plasmas **24**, 083109 (2017).

[26] A. R. Maier, N. M. Delbos, T. Eichner, L. Hübner, S. Jalas, L. Jeppe, S. W. Jolly, M. Kirchen, V. Leroux, P. Messner, M. Schnepp, M. Trunk, P. A. Walker, C. Werle, and P. Winkler, Decoding Sources of Energy Variability in a Laser-Plasma Accelerator, Phys. Rev. X **10**, 031039 (2020).

[27] A. J. Gonsalves, K. Nakamura, J. Daniels, C. Benedetti, C. Pieronek, T. C. H. de Raadt, S. Steinke, J. H. Bin, S. S. Bulanov, J. van Tilborg, C. G. R. Geddes, C. B. Schroeder, C. Tóth, E. Esarey, K. Swanson, L. Fan-Chiang, G. Bagdasarov, N. Bobrova, V. Gasilov, G. Korn, P. Sasorov, and W. P. Leemans, Petawatt Laser Guiding and Electron Beam Acceleration to 8 GeV in a Laser-Heated Capillary Discharge Waveguide, Phys. Rev. Lett. **122**, 084801 (2019).

[28] B. Miao, J. E. Shrock, L. Feder, R. C. Hollinger, J. Morrison, R. Nedbailo, A. Picksley, H. Song, S. Wang, J. J. Rocca, and H. M. Milchberg, Multi-GeV Electron Bunches from an All-Optical Laser Wakefield Accelerator, Phys. Rev. X **12**, 031038 (2022).

[29] K. Oubrerie, A. Leblanc, O. Kononenko, R. Lahaye, I. A. Andriyash, J. Gautier, J.-P. Goddet, L. Martelli, A. Tafzi, K. T. Phuoc *et al.*, Controlled acceleration of gev electron beams in an all-optical plasma waveguide, Light. Sci. Appl. **11**, 180 (2022)..

[30] W. Wang, K. Feng, L. Ke, C. Yu, Y. Xu, R. Qi, Y. Chen, Z. Qin, Z. Zhang, M. Fang, J. Liu, K. Jiang, H. Wang, C. Wang, X. Yang, F. Wu, Y. Leng, J. Liu, R. Li, and Z. Xu, Free-electron lasing at 27 nanometres based on a laser wakefield accelerator, Nature (London) **595**, 516 (2021).

[31] S. Hillenbrand, R. Assmann, A.-S. Müller, O. Jansen, V. Judin, and A. Pukhov, Study of laser wakefield accelerators as injectors for synchrotron light sources, Nucl. Instrum. Methods Phys. Res. A **740**, 153 (2014), proceedings of the first European Advanced Accelerator Concepts Workshop 2013.

[32] S. A. Antipov, A. Ferran Pousa, I. Agapov, R. Brinkmann, A. R. Maier, S. Jalas, L. Jeppe, M. Kirchen, W. P. Leemans, A. M. de la Ossa, J. Osterhoff, M. Thévenet, and P. Winkler, Design of a prototype laser-plasma injector for an electron synchrotron, Phys. Rev. Accel. Beams **24**, 111301 (2021).

[33] C. B. Schroeder, E. Esarey, C. G. R. Geddes, C. Benedetti, and W. P. Leemans, Physics considerations for laser-plasma linear colliders, Phys. Rev. ST Accel. Beams **13**, 101301 (2010).

[34] J.-L. Vay, A. Almgren, J. Bell, R. Lehe, A. Myers, J. Park, O. Shapoval, M. Thévenet, W. Zhang, D. P. Grote, M. Hogan, L. Ge, and C. Ng, Toward plasma wakefield simulations at exascale, in *Proceedings of the 2018 IEEE Advanced Accelerator Concepts Workshop (AAC)* (IEEE, New York, 2018), pp. 1–5, https://doi.org/10.1109/AAC.2018.8659392.

[35] D. R. Jones, M. Schonlau, and W. J. Welch, Efficient global optimization of expensive black-box functions, J Glob Optim **13**, 455 (1998).

[36] A. Hanuka, J. Duris, J. Shtalenkova, D. Kennedy, A. Edelen, D. Ratner, and X. Huang, Online tuning and light source control using a physics-informed Gaussian process, in *Machine Learning for the Physical Sciences Workshop, NeurIPS 2019, Vancouver, Canada* (2019), https://arxiv.org/abs/1911.01538.

[37] J. Duris, D. Kennedy, A. Hanuka, J. Shtalenkova, A. Edelen, P. Baxevanis, A. Egger, T. Cope, M. McIntire, S. Ermon, and D. Ratner, Bayesian Optimization of a Free-Electron Laser, Phys. Rev. Lett. **124**, 124801 (2020).

[38] R. Shalloo, S. Dann, J.-N. Gruse, C. Underwood, A. Antoine, C. Arran, M. Backhouse, C. Baird, M. Balcazar, N. Bourgeois *et al.*, Automation and control of laser wakefield accelerators using Bayesian optimization, Nat. Commun. **11**, 6355 (2020).

[39] A. Hanuka, X. Huang, J. Shtalenkova, D. Kennedy, A. Edelen, Z. Zhang, V. R. Lalchand, D. Ratner, and J. Duris, Physics model-informed gaussian process for online optimization of particle accelerators, Phys. Rev. Accel. Beams **24**, 072802 (2021).

[40] R. Roussel, J. P. Gonzalez-Aguilera, Y.-K. Kim, E. Wisniewski, W. Liu, P. Piot, J. Power, A. Hanuka, and A. Edelen, Turn-key constrained parameter space exploration for particle accelerators using bayesian active learning, Nat. Commun. **12**, 5612 (2021).

[41] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, Adaptive computation and machine learning (MIT Press, 2006), https://mitpress.mit.edu/9780262182539/gaussian-processes-for-machine-learning/.

[42] A. Lifschitz, X. Davoine, E. Lefebvre, J. Faure, C. Rechatin, and V. Malka, Particle-in-cell modelling of laser–plasma interaction using Fourier decomposition, J. Comput. Phys. **228**, 1803 (2009).

[43] P. Mora and T. M. Antonsen, Jr., Kinetic modeling of intense, short laser pulses propagating in tenuous plasmas, Phys. Plasmas **4**, 217 (1997).

[44] P. Sprangle, E. Esarey, and A. Ting, Nonlinear interaction of intense laser pulses in plasmas, Phys. Rev. A **41**, 4463 (1990).

[45] E. V. Bonilla, K. Chai, and C. Williams, Multi-task Gaussian process prediction, in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, Vol. 20, edited by J. Platt, D. Koller, Y. Singer, and S. Roweis (Curran Associates, Inc., Red Hook, NY, 2007), https://proceedings.neurips.cc/paper_files/paper/2007/file/66368270ffd51418ec58bd793f2d9b1b-Paper.pdf.

[46] K. Swersky, J. Snoek, and R. P. Adams, Multi-task Bayesian optimization, in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, Vol. 26, edited by C. Burges, L. Bottou,

M. Welling, Z. Ghahramani, and K. Weinberger (Curran Associates, Inc., Red Hook, NY, 2013), https://proceedings.neurips.cc/paper_files/paper/2013/file/f33ba15effa5c10e873bf3842afb46a6-Paper.pdf.

[47] B. Letham and E. Bakshy, Bayesian optimization for policy search via online-offline experimentation, J. Mach. Learn. Res. **20**, 1 (2019), https://jmlr.org/papers/v20/18-225.html.

[48] R. Lehe, M. Kirchen, I. A. Andriyash, B. B. Godfrey, and J.-L. Vay, A spectral, quasi-cylindrical and dispersion-free particle-in-cell algorithm, Comput. Phys. Commun. **203**, 66 (2016).

[49] A. Ferran Pousa, R. Assmann, and A. Martinez de la Ossa, Wake-T: a fast particle tracking code for plasma-based accelerators, J. Phys. Conf. Ser. **1350**, 012056 (2019).

[50] F. Irshad, S. Karsch, and A. Döpp, Expected hypervolume improvement for simultaneous multi-objective and multi-fidelity optimization, arXiv:2112.13901.

[51] F. Irshad, S. Karsch, and A. Döpp, Multi-objective and multi-fidelity bayesian optimization of laser-plasma acceleration, Phys. Rev. Res. **5**, 013063 (2023).

[52] B. Letham, B. Karrer, G. Ottoni, and E. Bakshy, Constrained Bayesian optimization with noisy experiments, Bayesian Anal. **14**, 495 (2019).

[53] A. B. Owen, Scrambling Sobol' and Niederreiter–Xing points, J. Complex. **14**, 466 (1998).

[54] https://github.com/optimas-org/optimas.

[55] A. Ferran Pousa, S. Jalas, M. Kirchen, A. Martinez de la Ossa, M. Thévenet, J. Larson, S. Hudson, A. Huebl, J.-L. Vay, and R. Lehe, optimas-org/optimas: v0.1.0 (2023).

[56] E. Bakshy, L. Dworkin, B. Karrer, K. Kashin, B. Letham, A. Murthy, and S. Singh, AE: A domain-agnostic platform for adaptive experimentation, in *Proceedings of the Advances in Neural Information Processing System, Montréal, Canada* (2018), https://eytan.github.io/papers/ae_workshop.pdf.

[57] S. Hudson, J. Larson, J.-L. Navarro, and S. M. Wild, libEnsemble: A library to coordinate the concurrent evaluation of dynamic ensembles of calculations, IEEE Trans. Parallel Distrib. Syst. **33**, 977 (2022).

[58] M. Poloczek, J. Wang, and P. Frazier, Multi-information source optimization, in *Advances in Neural Information Processing Systems*, Vol. 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., Long Beach, CA, 2017).

[59] J. M. Hernandez-Lobato, M. Gelbart, M. Hoffman, R. Adams, and Z. Ghahramani, Predictive entropy search for bayesian optimization with unknown constraints, in *Proceedings of the 32nd International Conference on Machine Learning, Proceedings of Machine Learning Research*, Vol. 37, edited by F. Bach and D. Blei (PMLR, Lille, France, 2015) pp. 1699–1707.

[60] M. McLeod, M. A. Osborne, and S. J. Roberts, Practical bayesian optimization for variable cost objectives, arXiv:1703.04335.

[61] P. Baxevanis and G. Stupakov, Novel fast simulation technique for axisymmetric plasma wakefield acceleration configurations in the blowout regime, Phys. Rev. Accel. Beams **21**, 071301 (2018).

[62] C. Benedetti, C. B. Schroeder, C. G. R. Geddes, E. Esarey, and W. P. Leemans, An accurate and efficient laser-envelope solver for the modeling of laser-plasma accelerators, Plasma Phys. Controlled Fusion **60**, 014002 (2017).

[63] E. Esarey, J. Krall, and P. Sprangle, Envelope Analysis of Intense Laser Pulse Self-Modulation in Plasmas, Phys. Rev. Lett. **72**, 2887 (1994).

[64] T. Mehrling, J. Grebenyuk, F. S. Tsung, K. Floettmann, and J. Osterhoff, Transverse emittance growth in staged laser-wakefield acceleration, Phys. Rev. ST Accel. Beams **15**, 111303 (2012).

[65] J.-L. Vay, Noninvariance of Space- and Timescale Ranges under a Lorentz Transformation and the Implications for the Study of Relativistic Interactions, Phys. Rev. Lett. **98**, 130405 (2007).

[66] J.-L. Vay, C. G. R. Geddes, E. Esarey, C. B. Schroeder, W. P. Leemans, E. Cormier-Michel, and D. P. Grote, Modeling of 10 GeV-1 TeV laser-plasma accelerators using Lorentz boosted simulations, Phys. Plasmas **18**, 123103 (2011).

[67] A. Ferran Pousa, S. Jalas, M. Kirchen, A. Martinez de la Ossa, M. Thévenet, J. Larson, S. Hudson, A. Huebl, J.-L. Vay, and R. Lehe, Optimization scripts used for "Bayesian optimization of laser-plasma accelerators assisted by reduced physical models", 10.5281/zenodo.7997698 (2023).

[68] J. Zhu, R. W. Assmann, M. Dohlus, U. Dorda, and B. Marchetti, Sub-fs electron bunch generation with sub-10-fs bunch arrival-time jitter via bunch slicing in a magnetic chicane, Phys. Rev. Accel. Beams **19**, 054401 (2016).

[69] S. Yamin, R. Assmann, F. Burkart, A. Ferran Pousa, F. Lemery, E. Panofski, W. Hillert, and B. Marchetti, Final focus system for injection into a laser plasma accelerator, Phys. Rev. Accel. Beams **24**, 091602 (2021).