Costs of Bayesian parameter estimation in third-generation gravitational wave detectors: An assessment of current acceleration methods

Qian Hu[*] and John Veitch[†]

Institute for Gravitational Research, School of Physics and Astronomy, University of Glasgow, Glasgow, G12 8QQ, United Kingdom

(Received 5 March 2025; accepted 29 September 2025; published 15 October 2025)

Bayesian inference with stochastic sampling has been widely used to obtain the properties of gravitational wave (GW) sources. Although computationally intensive, its cost remains manageable for current second-generation GW detectors because of the relatively low event rate and signal-to-noise ratio (SNR). The third-generation (3G) GW detectors are expected to detect hundreds of thousands of compact binary coalescence (CBC) events every year with substantially higher SNR and longer signal duration, presenting significant computational challenges. In this study, we systematically evaluate the computational costs of CBC source parameter estimation (PE) in the 3G era by modeling the PE time cost as a function of SNR and signal duration. We examine the standard PE method alongside acceleration methods including relative binning, multibanding, and reduced order quadrature. We predict that PE for a one-month-observation catalog with 3G detectors could require at least billions of CPU core hours with the standard PE method, whereas acceleration techniques can reduce this demand to less than millions of core hours, which is as high as the cost of analyzing GW events in the past 10 years. These findings highlight the necessity for more efficient PE methods to enable cost-effective and environmentally sustainable data analysis for 3G detectors. In addition, we assess the accuracy of accelerated PE methods, emphasizing the need for careful treatment in high-SNR scenarios.

DOI: 10.1103/dj7k-tk37

I. INTRODUCTION

Since the first direct detection of gravitational waves (GWs) by Advanced LIGO [1] in 2015 [2], over 200 GW events have been published by the LIGO-Virgo [3]-KAGRA [4] (LVK) collaboration [5–7], with hundreds more expected from the current observing run. All of these GW events are identified as compact binary coalescences (CBCs)—mergers of compact objects such as black holes, neutron stars, and potentially exotic stars [8,9]. Determining the properties of these sources is essential for various aspects of GW astronomy, including cosmology [10–12], astrophysical population [13–15], the equation of state of dense matter [16], stochastic background [17–19], among others.

CBC signals are characterized by the physical and geometric parameters of their sources, and the properties of the sources are inferred by estimating these parameters. A typical CBC system is described by at least 15 parameters, including two component masses, six component spin parameters, and seven extrinsic parameters describing the 3D sky location, inclination angle and polarization angle, and coalescence phase and time of the source. Additional parameters are required to account for specific phenomena, such as tidal deformability parameters for neutron stars [20], orbital eccentricity and eccentric anomaly for nonquasicircular orbits [21,22], and modification terms for alternative theories of gravity [23,24], etc.

Due to the transient nature of CBC signals, parameter estimation (PE) is solved under the Bayesian framework [25], in which the probability density distribution of the source parameters is given by the Bayes theorem, but not directly analyzable due to the complexity in GW waveform models. Stochastic sampling algorithms, Markov chain Monte Carlo (MCMC) or nested sampling, are used to draw samples from the probability distribution, which represent the estimated source parameters. These methods, however, are computationally intensive. Currently, a typical full PE run using the standard Bayesian framework [25] for a CBC source would cost hours to days, and the cost increases with the signal-to-noise ratio (SNR) and signal duration. A number of acceleration methods for GW source PE are proposed to mitigate the computational burden in stochastic sampling, including relative binning (RB, also referred to as heterodyning) [26–28], multibanding (MB) [29,30], and

[*]Contact author: Qian.Hu@glasgow.ac.uk
[†]Contact author: John.Veitch@glasgow.ac.uk

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

reduced order quadrature (ROQ) [31,32]. Nevertheless, little work has been done to systematically investigate and compare the performance of these algorithms. The LVK data release still extensively uses the standard PE method without acceleration [6,7,33,34], which is feasible given current detection rate, signal SNRs and durations.

However, there is a growing concern regarding the computational burden for the proposed third-generation (3G) GW detectors [35,36]. The 3G detectors, including Einstein Telescope (ET) [37–39] and Cosmic Explorer (CE) [40,41], are designed to improve the sensitivity by at least an order of magnitude over Advanced LIGO, and are expected to detect hundreds of thousands of CBC signals every year with significantly higher SNR and longer signal duration [38,42]. While these detectors promise exciting scientific discoveries, the computational cost of analyzing all the detected signals using current methods is anticipated to be enormous, though it is not yet fully quantified.

In this paper, we present a detailed estimate of the computational costs of PE for CBC sources in the 3G era, considering both the standard PE method and accelerated methods. By conducting 1200 PE experiments and fitting the PE time cost as a function of SNR and duration, we estimate the total costs of PE for 3G detectors based on the simulated catalog in ET Mock Data Challenge 1 (MDC-1) [43]. We find that billions of CPU core hours is required to analyze one month of observation with the 3G detectors using the standard PE method, with most of the computational load coming from long binary neutron star (BNS) and neutron star-black hole (NSBH) signals. With acceleration methods, the cost could be reduced to less than millions of CPU core hours. We compare the performance of different acceleration methods in terms of speed and accuracy. Our results identify ROQ as the fastest method overall, though it requires additional training before the PE. We find that, even with the accelerated methods, the one-month cost is sufficient to analyze all CBC events detected in the past ten years, highlighting the necessity of the use of more efficient methods such as machine learning. We also highlight accuracy issues in high-SNR scenarios, with RB showing the worst accuracy due to its reliance on fiducial parameters used in the heterodyning process.

This paper is organized as follows. Section II provides a brief introduction to Bayesian inference and stochastic sampling in Sec. II, including the standard methods in Sec. II A and acceleration methods in Sec. II B. The experiments are presented in Sec. III, with details on the experiment design in Sec. III A, fitting the PE cost in Sec. III B, and accuracy assessment in Sec. III C. Section IV applies the time cost fitting to the ET MDC-1 and estimates the total PE cost in the 3G era. The concluding remarks are provided in Sec. V. Throughout this paper, the phrase "GW sources" refers to CBCs.

II. OVERVIEW OF BAYESIAN PARAMETER ESTIMATION

A. Bayesian parameter estimation

1. Bayes theorem

Bayesian inference gives us means to quantify uncertainty about the source parameters using the concept of Bayesian updating. Here we give a brief overview of the Bayesian inference framework used in GW astronomy and in this work. More detailed reviews of this topic can be found in, e.g., Refs [25,44,45] Starting with generic prior expectations about the source (which may be informed by astrophysical or geometrical considerations), we observe the transient GW signal, and then update our knowledge based on the new data. Mathematically, given a hypothesis H (e.g., the waveform model) and a prior distribution of the parameters $p(\theta|H)$, we can update the probability distribution of θ with observation data d using Bayes' theorem

$$p(\theta|d,H) = \frac{p(\theta|H)p(d|\theta,H)}{p(d|H)}, \tag{1}$$

where

$$p(d|\theta,H) \propto \exp -\frac{1}{2}(d-h(\theta)|d-h(\theta)) \tag{2}$$

is the likelihood function, assuming the noise is stationary and Gaussian [46], $h(\theta)$ is the GW signal given parameter θ, and $(\cdot|\cdot)$ denotes the inner product between two strain datasets. The raw GW data is provided in the time domain, but the inner product is easier to estimate in the frequency domain using the windowed and Fourier-transformed data [46]

$$(a|b) = 4\ \mathrm{Re}\int_0^\infty \frac{a^*(f)b(f)}{S_n(f)}df, \tag{3}$$

where Re denotes real part and a^* denotes the complex conjugate of a. $S_n(f)$ is the one-sided power spectral density (PSD) of the detector noise. $p(d|H)$ is the evidence, a constant that is irrelevant for the shape of the posterior distribution, but is useful in model comparison [44]. The posterior probability distribution $p(\theta|d,H)$ provides the estimation of the parameter θ.

2. Stochastic sampling

Although the posterior probability density function is computable in the entire parameter space (as long as the waveform model is valid) via Eq. (1), it is impossible to compute it all over the space due to the high dimensionality. Instead, stochastic sampling algorithms are used to obtain a set of samples that follows the distribution given by Eq. (1).

MCMC and nested sampling are two families of sampling algorithms. MCMC constructs a Markov chain whose

equilibrium distribution is the target distribution. Over time, the chain wanders through the parameter space, producing samples with a density that approximates the target distribution. Nested sampling is designed to calculate the evidence integral, the denominator of Eq. (1) [47]. It works by initially drawing a set of live points from the prior distribution, then iteratively removing and replacing the sample with the lowest likelihood. As the algorithm progresses the live points occupy a geometrically shrinking volume of parameter space, whose volume can be statistically estimated. After the algorithm reaches the peak, the nested samples can be reweighted to draw or approximate the posterior distribution.

A number of variants of MCMC and nested sampling algorithms have been used in GW astronomy. DYNESTY [48] is the most widely used in recent LVK public data [6,7], although many novel samplers are being developed to achieve faster speeds. For example, NESSAI [49,50] is a nested sampling algorithm enhanced with machine learning and importance sampling, and it reduces the number of likelihood evaluations by an order of magnitude in GW source PE compared with DYNESTY. We find that DYNESTYv2.1.4 [51] with the configuration used in the LVK production in the third observing run performs extremely slowly in our 3G simulations, while NESSAIv0.8.1 [52] offers significantly improved speed and convergence without any fine-tuning. Therefore, we use NESSAI throughout this work.

B. Acceleration methods

Stochastic sampling for GW PE is computationally intensive. Sampling algorithms may struggle to identify the correct region in the parameter space, requiring long convergence times and numerous likelihood evaluations (waveform generation and vector manipulation). An increase in signal duration (which results in more observed GW cycles and a boarder bandwidth) or SNR results in a narrower posterior (see, e.g., Ref. [53,54]), making it more challenging for the sampler to locate and converge on the posterior. Additionally, longer signal durations slow down waveform generation and vector manipulation involved in the likelihood evaluation. Both of these factors increase the computational cost of performing PE.

Several acceleration methods have been proposed to speed up the PE. In addition to the general improvement of the sampler like NESSAI, there are methods specifically targeting the CBC PE problem, aiming to reduce the data size and simplify likelihood evaluations. In this section, we introduce three methods that we will examine in this work: relative binning (RB), multibanding (MB), and reduced order quadrature (ROQ). All of them have been implemented in the GW PE Python package BILBY [30,32,55,56]. We will also mention other fast PE methods in Sec. II B 4 but they are not involved in our assessment due to either having completely different mechanisms, being a mixture of many methods, only obtaining a subset of parameters, or having an incomplete software implementation.

1. Relative binning

Relative binning (RB) [26,27], first proposed as heterodyning [28], is based on the idea that many likelihood evaluations occur near the maximum likelihood point during PE, and that the GW waveform changes smoothly as a function of the source parameters. Therefore, if a fiducial source parameter that is close to the true value can be assigned before PE (which is possible given information from matched filtering detection, see, e.g., Ref. [57]), the PE can focus on in the nearby region where likelihood evaluations can be simplified by exploiting the smooth variations in the waveform as the fiducial parameter is varied. Since the smooth variation with respect to the fiducial waveform requires a lower bandwidth to describe, a coarser frequency resolution is sufficient to reconstruct waveforms near the fiducial waveform. This reduces the data size and therefore the computational cost of computing each likelihood. The new frequency bins with coarse frequency resolution can be chosen such that the GW signal only contains a few cycles within each bin, which is computable using the post-Newtonian (PN) phase expansion [58].

Considering frequency domain waveforms, we approximate the ratio between the fiducial waveform $h_0(f)$ and an arbitrary waveform $h(f)$ in a frequency bin b to the linear order

$$r(f) = \frac{h(f)}{h_0(f)} = r_0(h,\mathrm{b}) + r_1(h,\mathrm{b})(f - f_\mathrm{m}(\mathrm{b})) + \cdots, \tag{4}$$

where is f_m is the central frequency of the bin. r_0 and r_1 are bin-dependent coefficients that measure how h deviates from h_0, and can be efficiently derived from the values of $r(f)$ at the edges of bins. Since $(d|d)$ is a constant, the likelihood [Eq. (2)] only requires the computation of $(d|h)$ and $(h|h)$, which can be approximated using $r(f)$. Define

$$\begin{aligned} A_0(\mathrm{b}) &= (d|h_0)_\mathrm{b}, \qquad A_1(\mathrm{b}) = (d|(f - f_\mathrm{m})h_0)_\mathrm{b} \\ B_0(\mathrm{b}) &= (h_0|h_0)_\mathrm{b}, \qquad B_1(\mathrm{b}) = (h_0|(f - f_\mathrm{m})h_0)_\mathrm{b}, \end{aligned} \tag{5}$$

where $(\ldots|\ldots)_\mathrm{b}$ is the inner product inside the frequency bin b, i.e., the integral limits are the minimum and the maximum frequencies of the bin. $(d|h)$ and $(h|h)$ can be calculated as

$$\begin{aligned} (d|h)^\mathrm{RB} &= \sum_\mathrm{b} (d|h)_\mathrm{b} = \sum_\mathrm{b} (d|rh_0)_\mathrm{b} \\ &\approx \sum_\mathrm{b} (A_0(\mathrm{b}) r_0^*(h,\mathrm{b}) + A_1(\mathrm{b}) r_1^*(h,\mathrm{b})), \end{aligned} \tag{6}$$

and similarly

$$(h|h)^{\mathrm{RB}} \approx \sum_{\mathrm{b}} (B_0(\mathrm{b})|r_0(h,\mathrm{b})|^2 + 2B_1(\mathrm{b})\mathrm{Re}[r_0(h,\mathrm{b})r_1^*(h,\mathrm{b})]). \quad (7)$$

For any waveform h, it is only necessary to evaluate the waveform at the edges of the bins to obtain r_0 and r_1, which speeds up the likelihood evaluations. In addition, focusing on regions around the fiducial parameters significantly reduces the number of likelihood evaluations, although this may bring accuracy issues.

Examples of applications of RB can be found in Refs [59–63]. In particular, Dai *et al.* [61] showed that PE for GW170817 with aligned spin waveform model IMRPhenomD_NRTidal [64] and TaylorF2 [58] and RB can be done with 150 CPU core hours. Wong *et al.* [63], also using aligned spin waveform model IMRPhenomD(_NRTidal), demonstrated a combination of RB and a gradient-based sampler, and achieved 2 hours of sampling time for GW150914 and 1 day for GW170817 with 400 CPU cores.

2. Multibanding

The frequencies of signals from CBC sources increase during the binary inspiral. GW signals are present in the data from the low-frequency cutoff f_{min} of the GW detector (∼20 Hz for current detectors [1,3,4] and ∼5 Hz for 3G [39,40]) up to the maximum frequency f_{max} of hundreds of Hz up to ∼2 kHz, depending on the source masses. The full-bandwidth GW data (i.e., strain data with sampling rate of 4096 Hz or 16384 Hz) is excessive for the low-frequency early inspiral stage, therefore the data can be adaptively down-sampled as long as sampling rate remains above the Nyquist frequency. This approach is known as multibanding or adaptive frequency resolution, first proposed for source detection [65,66].

Several different MB schemes for PE have been proposed; see, e.g., Refs. [29,30]. In this work, we adopt the scheme used in Morisaki [30], in which the time-domain data is divided into bands with geometrically decreasing lengths $(T, T/2, T/4, \ldots, 4s)$. The starting and ending frequencies of each band are chosen such that (i) the data in the time-domain band can cover the signal within the frequency band, with time-frequency relation given by the 0PN equation for the lowest mass in the prior and (ii) the waveform smoothly vanishes at the left edge of each time-domain band with the window function applied. The frequency bands and relevant coefficients are determined once the prior is set, requiring no additional calculations before or during PE.

MB leads to coarser frequency grids on which the GW waveform needs to be evaluated and thus reduces the data size, accelerating waveform generation and likelihood evaluation. This requires waveform models that can be calculated at any frequency, typically with closed-form expressions in the frequency domain, such as the IMRPhenom waveforms [67,68]. However, MB does not reduce the number of likelihood evaluations. Morisaki [30] shows that MB can speed up PE by a factor of 20–50 for signals starting from 20 Hz, and by a factor of 100–500 for signals starting from 5 Hz, in line with the asymptotic scaling $5f_{\mathrm{max}}/3f_{\mathrm{min}}$ found in [29]. MB can be combined with ROQ [30,69] and RB [62,70] to further enhance the performance. In addition to PE and detection, the similar idea has been explored in source localization [71].

3. Reduced order quadrature

ROQ [31,32] utilizes the reduced order modeling (ROM) [72–74] of GW waveforms, which aims to accelerate waveform evaluation. In the frequency domain, the ROM of a waveform model can be written as the linear addition of N bases, where N is much smaller than the original length of the waveform

$$h(f;\theta) \approx h^{\mathrm{ROM}}(f;\theta) = \sum_{J=1}^{N} h(\mathcal{F}_J;\theta)B_J(f), \quad (8)$$

where $B_J(f)$ is a set of bases that span the signal space obtained via a greedy algorithm [72] or singular value decomposition [73]. $\{\mathcal{F}_J\}$ is a set of the most representative frequencies that can be used to reconstruct the entire waveform determined by empirical interpolation [72]. The $h^{\mathrm{ROM}}(f;\theta)$, as an approximation of the original waveform model $h(f;\theta)$, only requires waveform evaluation at $\{\mathcal{F}_J\}$ and a set of predetermined bases $\{B_J(f)\}$, and is therefore faster to calculate, at the cost of precomputing the bases. For low-dimensional parameter spaces, it is possible to build a surrogate model $h_J(\theta)$ to approximate the waveform value at a specific frequency node $h(\mathcal{F}_J;\theta)$. $h_J(\theta)$ can takes the form of polynomials [72], splines [73], or even neural networks [75]. However, PE for precessing systems has a high-dimensional parameter space in which surrogate models cannot be easily built, hence it is not suitable for our task. We therefore still need a waveform model that can be evaluated at any frequencies, like we do in MB.

The ROQ likelihood can be expressed using the ROM. The nonconstant terms in the likelihood are

$$(d|h)^{\mathrm{ROQ}} = \mathrm{Re}\sum_{J=1}^{N} h(F_J;\theta)\omega_J(t_c), \quad (9)$$

and

$$(h|h)^{\mathrm{ROQ}} = \sum_{I=1}^{N}\sum_{J=1}^{N} h^*(F_I;\theta)h(F_J;\theta)\psi_{IJ}, \quad (10)$$

where $\omega_J(t_c)$ and ψ_{IJ} are integration weights that depend on the basis functions, data, and PSD, and their definitions can be found in Ref. [76]. To perform ROQ, basis functions $\{B_J(f)\}$ need to be calculated under a prior distribution of the intrinsic parameters (which can be broader than the

prior used in PE) and the integration weights need to be tailored to the specific data. This precalculation stage can take anywhere from minutes to hours, depending on factors such as signal length and the prior range. However, the most computationally expensive step, basis construction, only needs to be done once for a wide range of sources, as long as the source falls within the prior used for basis construction. ROQ significantly accelerates the likelihood calculation by reducing it to linear and quadratic summations with waveforms evaluated at a reduced frequency grid.

ROQ can bring speed-up factors of more than ten times depending on the signal properties, and is widely used in GW PE, see, e.g., [31,69,76,77]. Notably, the speed up can goes up to thousands when the ROQ prior is fine-tuned using the information from detection [78]. Similarly, Ref. [76] illustrates ROQ's application to long signals from BNS in 3G detectors: using a restricted ROQ prior, PE for a 90-minute-long BNS signal of SNR of 2400 can be done with 20 CPU hours of precalculation and 1600 CPU hours of sampling. Bayesian inference for such signals would be prohibitively slow ($> 10^7$ CPU hours) using the standard PE method.

4. Other methods

There are a number of other approaches for fast PE. One common technique is marginalization, which reduces computational cost by integrating over certain parameters. The luminosity distance, coalescence time, and coalescence phase (for the dominant $\ell = m = 2$ mode of the signal) can be analytically marginalized in the posterior [25,44], allowing them to be excluded from the sampling process and reconstructed after the other parameters are sampled. Taking the matched filtering SNR timeseries as input, numerical marginalization of more parameters is demonstrated in COGWHEEL [60], achieving minutes to a few hours of PE time on a single CPU. RIFT [79,80] also makes use of likelihood marginalization, evaluating the marginal likelihood on a grid of intrinsic parameters, and then constructing a fast interpolator to explore the posterior. Marginalization techniques are also used for quickly obtaining subsets of parameters such as for source localization [81–83].

Machine learning methods based on generative models, such as DINGO [62,84–86] and other variants [87] based on normalizing flows [88,89] and `Vitamin` [90] based on variational autoencoders [91], have shown great potential in fast PE. Generative models learn the data distribution and can quickly draw samples from the posterior in seconds once trained. The samples can be further improved with importance sampling at an additional time cost of a few minutes [85]. DINGO has been applied to the inference of LVK events and presents fast speed and good accuracy, including the inference with slow waveforms [86] that could be extremely expensive for traditional Bayesian inference. There are also 3G-specific applications in the literature, such as for overlapping signals [92] and long BNS signals in 3G detectors [70]. In addition, methods that can make use of GPU, such as gradient-based samplers [63,93] and RIFT [94], could also bring significant acceleration compared to traditional samplers. The speed advantages of machine learning approaches and GPU hardware make them particularly well suited for catalog-level tasks, a critical need in the 3G era with the huge increase in signal numbers.

III. EXPERIMENTS

To estimate the cost of Bayesian inference in the 3G era, we need to understand the relationship between signal properties (duration, SNR, etc.) and the associated sampling cost. To do this, we perform 1200 PE runs with varying settings and fit the resulting relation. In this section, we describe the experiment setup, outline the process for obtaining the time cost fitting, and discuss the accuracy of different acceleration methods.

A. Setup

We consider a detector network contains one triangular 10 km-ET operating at the ET-D design sensitivity [95] at the Virgo site, and one 40 km CE detector with the CE2 design sensitivity [40] at the LIGO Hanford site. Detector noise is assumed to be stationary and Gaussian, and different events have different noise realizations. The starting frequency is 5 Hz and data is sampled at 2048 Hz.

We simulate three types of binary black hole (BBH) sources: GW151226-like ($14 + 8M_\odot$, duration = 256 s), GW150914-like ($36 + 30M_\odot$, duration = 64 s), and GW190521-like ($85 + 66M_\odot$, duration = 16 s). They represent typical low, medium, and high mass BBHs in the population and produce long, medium, and short duration signals, respectively. We fix their intrinsic parameters to the median values given in GWTC-3 [7] and GWTC-2.1 [6] results and randomly simulate extrinsic parameters except for coalescence time and luminosity distance. The coalescence time is fixed at GPS time $t_c = 0s$ as this choice should not influence the result. The luminosity distance is scaled to adjust the optimal network SNR to values in the set (12,30,100,500,1000), ranging from near-threshold detections to the high-SNR golden events. For each SNR and source type, 20 events with different extrinsic parameters are simulated. Each event is analyzed with four PE methods: the "standard" method (no acceleration method applied), RB, MB, and ROQ, resulting in a total of 1200 PE runs.

The signals are simulated and analyzed with the IMRPhenomPv2 [67,68] waveform model, which leads to 15 parameters to estimate. Although IMRPhenomPv2 is not the most accurate waveform model to date, it is very fast to evaluate and can be efficiently parallelized for calculation.

We may hope that by the time of 3G observing, advances in hardware and algorithms will reduce the cost of waveform evaluation, so this can be considered a conservative choice. The prior distributions for the parameters are as follows:

(i) Chirp mass $\mathcal{M}$: Uniform in $[8.5, 9.5]M_\odot$, $[26, 34]M_\odot$, and $[50, 100]M_\odot$, for GW151226-like, GW150914-like, and GW190521-like sources, respectively.
(ii) Mass ratio q: Uniform in [0.25, 1].
(iii) Spin dimensionless magnitudes a_1, a_2: Uniform in [0, 0.8].
(iv) Spin tilt θ_1, θ_2: Uniform in cosine (isotropic spin).
(v) Spin angles $\varphi_{12}, \varphi_{JL}$: Uniform in $[0, 2\pi]$ rad.
(vi) Inclination angle θ_{JN}: Uniform in cosine (isotropic orientation).
(vii) Right ascension α: Uniform in $[0, 2\pi]$ rad.
(viii) Declination angle δ: Uniform in sine (isotropic on the sky).
(ix) Polarization angle ψ: Uniform in $[0, \pi]$ rad.
(x) Coalescence phase ϕ_c: Uniform in $[0, 2\pi]$ rad.
(xi) Coalescence time t_c: Uniform in $[-0.1, 0.1]s$.
(xii) Luminosity distance d_L: Uniform in the volume between $[d_L/2, 2d_L]$, where d_L is the injected value.

We use the PE library BILBYv2.2.2 [96] for our analysis. We employ the sampler NESSAIv0.8.1 for faster convergence, though future improvements to the sampler may further enhance performance. We set the number of live points $\texttt{nlive} = 2000$. Larger nlive may be required for higher SNRs or narrower posteriors such as for BNS signals, which would increase the sampling time. Experiments are run with $\texttt{ncpu} = 32$ CPU cores with the AMD EPYC 7443 CPU model (base clock 2.85 GHz). The likelihood is evaluated in parallel but the training and sampling parts of NESSAI are not parallelized and only use one core. As a result, the total sampling CPU core time is calculated as total sampling wall time + $(\texttt{ncpu} - 1) \times$ likelihood evaluation wall time [97]. The time cost statistics are directly read from the NESSAI log files.

The ROQ bases are calculated using PYROQ package [98], with the same prior used in PE. Using reconstruction error tolerance of 10^{-8} for ROM bases and 10^{-10} for quadratic bases, we obtain 392, 183, 136 ROM bases, and 184, 96, 112 quadratic bases for GW151226-like, GW150914-like, GW190521-like sources, respectively. The data length is significantly compressed compared to the original lengths of 64k, 32k, and 16k data points. With a single CPU core, ROQ training takes less than one hour for GW150914-like and GW190521-like sources, but takes five days for the GW151226-like source. We note that the construction can be parallelized but we do not implement it in this work. Although it is possible to construct ROQs with a wider prior for more general applications, this would increase the training time and reduce the compression rate. Therefore, we focus the ROQ construction on the simulated signals. The precalculation time cost for building the ROQs is not included in the ROQ sampling time.

The fiducial parameter for RB is chosen to be close to the injection parameter, roughly 10 times more accurate than the preestimates from current matched filtering search [57]. The error in $\mathcal{M}$ is randomly sampled within $\pm 0.1\%$, while errors for q and d_L are within $\pm 5\%$. The errors for α, δ, and θ_{JN} are sampled within $\pm 10\%$, and the coalescence time has no error (0%). Parameters such as spins and angles, which are difficult to estimate accurately before PE, are set to zero in the fiducial parameters. All random errors are sampled uniformly within their ranges. Although we choose the error to be 10 times more accurate than current search pipelines can provide [57], they may not be accurate enough for the 3G detections as we observe convergence issues in PE with RB in our experiments, particularly in high-SNR events. RB is effective only within a region near the fiducial parameters, and it may fail to reach the narrow posterior region when the fiducial parameter is biased. In our experiments, five RB runs failed to converge and were excluded from the time cost estimate.

With our configuration, all methods produces ~12000 posterior samples when the sampler converges. The code for the experiments can be found in the GitHub repository 3gpemethods [99].

B. Bilinear fitting for time costs

We calculate the CPU core time for stochastic sampling in our experiments, and the results are shown in Fig. 1. Overall, ROQ exhibits the fastest speed among all PE methods, completing sampling within 2 CPU days for most events in our simulation. MB and RB typically finish within 5 CPU days. In contrast, the standard method can take tens to hundreds of CPU days. Acceleration methods provide a speed-up factor of up to several hundred times, especially in high-SNR and long-signal scenarios.

There is a clear trend that sampling time increases with both SNR and signal duration, with the scaling being approximately linear in the logarithmic scale. Therefore, we employ the following bilinear function to fit the relation between sampling CPU days D, signal duration T (in second) and network SNR

$$\log_{10} D = a \log_{10} T + b \log_{10} \mathrm{SNR} + c, \tag{11}$$

where the coefficients a, b, c for different PE methods are determined by linear regression using the scikit-learn package [100] and are listed in Table I. We also calculate the R^2 score, the coefficient of determination, a higher value of which indicates a better fit. The R^2 score exceeds 0.5 for all PE methods except RB, which may be caused by its convergence issues since the choice of fiducial parameter has impact on the convergence speed. The relatively low R^2 scores in the three acceleration methods implies their run time may depend on signal SNR and duration in a more complex way, but from the regression results in Fig. 1, the bilinear fitting is able to capture the overall tendency and

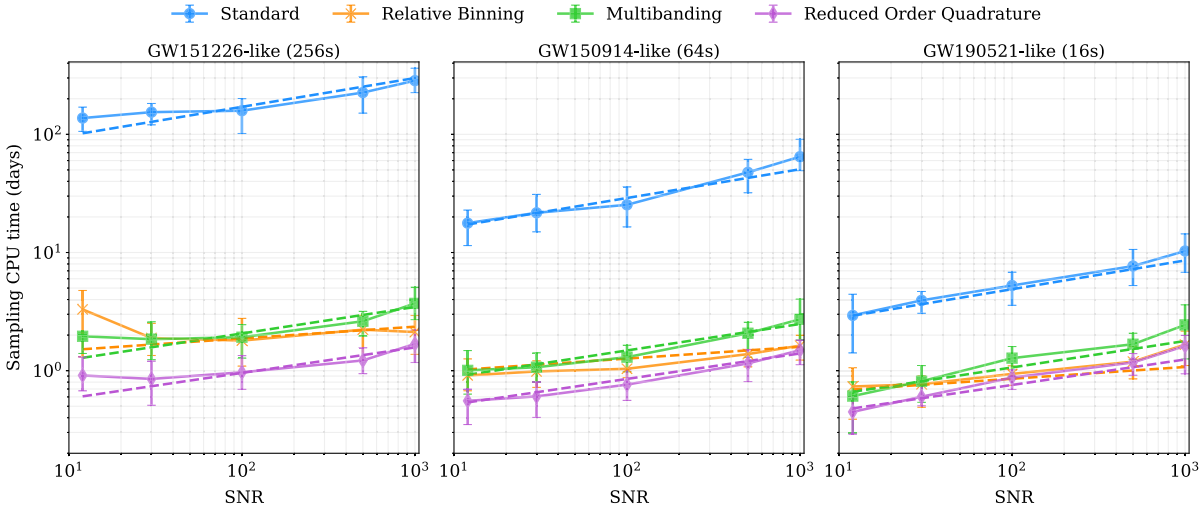


FIG. 1. The CPU core time of stochastic sampling in our experiments, along with the bilinear fitting. The error bars represent the 16% to 84% percentiles of the sampling time across 20 events for each configuration, with the mean indicated in the center. The dashed lines correspond to the bilinear fits of the sampling time cost.

the accuracy should be sufficient for a rough, order-of-magnitude level estimation for the PE configuration used in this work.

Looking into the coefficients, we notice the standard method has the largest scaling factors (a and b) with respect to the signal duration and SNR, indicating that its time cost grows most rapidly, which aligns with our expectations. ROQ has the lowest scaling factor (the coefficient a) with the signal duration, highlighting its good efficiency in speeding up likelihood evaluations. RB shows the lowest scaling factor (the coefficient b) with the signal SNR. For the runs that finish quickly (for example, low SNR and short-duration events), we note that the total cost is not dominated by the likelihood but by other operations of the NESSAI sampler.

It is worth mentioning that the effect of lower frequency cutoff $f_{\rm low}$ can be estimated roughly using a. If $f_{\rm low}$ changes to 2 Hz or 3 Hz, the SNR would not change significantly, but the duration of the signal would scale as $f_{\rm low}^{-8/3}$. Using Eq. (11), the sampling time D will be *multiplied* by a factor

$$D_{f_{\rm low}} = \left(\frac{5}{f_{\rm low}}\right)^{\frac{8a}{3}} D_{5\ \rm Hz}, \tag{12}$$

where $D_{5\ \rm Hz}$ is the result given by Eq. (11) and a is given in Table I. Substituting the values, the PE time cost of the standard method would increase roughly 22 and 6 times for $f_{\rm low}$ of 2 Hz, 3 Hz, respectively. For the accelerated methods, the time increase is between 10% and 100%. Note that the increase is underestimated, since the changes in SNR are not accounted for.

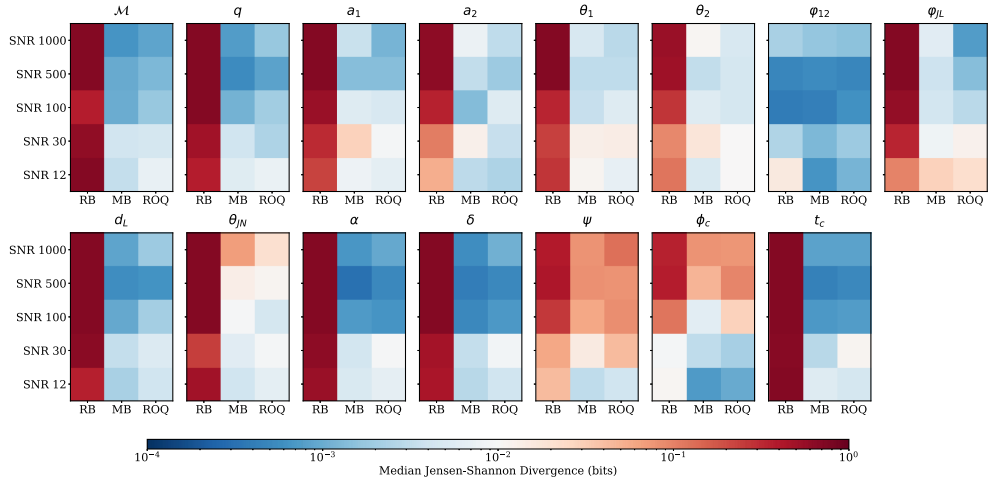
TABLE I. Coefficients a, b, c and their uncertainties from bilinear regression in Eq. (11) for sampling time. The last column gives the coefficient of determination R^2 score ranging from 0 to 1, with higher value indicating a better match between prediction and data.

Method	a	b	c	R^2
Standard	1.28 ± 0.02	0.24 ± 0.01	-1.34 ± 0.04	0.94
RB	0.28 ± 0.02	0.10 ± 0.02	-0.61 ± 0.05	0.40
MB	0.24 ± 0.02	0.22 ± 0.01	-0.71 ± 0.04	0.60
ROQ	0.09 ± 0.02	0.22 ± 0.01	-0.65 ± 0.04	0.51

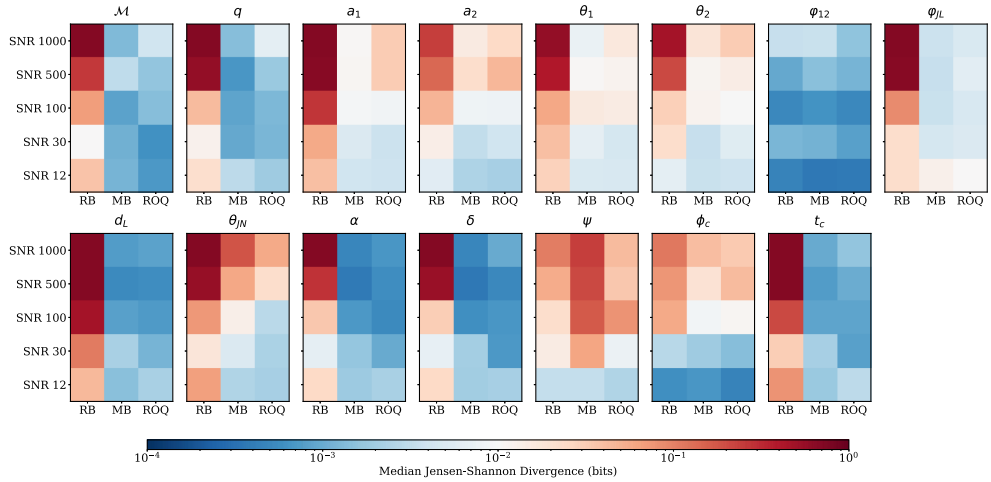
C. Accuracy

We assess the accuracy of the acceleration methods by comparing the posterior distributions obtained using each method to those obtained with the standard method. For each parameter in each event, we compute the Jensen-Shannon divergence (JSD) between the posterior distribution from an acceleration method and the standard method. The JSD is calculated with a base of 2, so it ranges from 0 to 1, with the unit of *bits*. A JSD value of 0 indicates that the two posterior distributions are identical, while a value of 1 means that they are completely different, with no overlap in the result. Reference [101] points out that the BILBY package produces posterior samples with $\lesssim 0.0015$ nat ($\approx$0.001 bit) of statistical fluctuations. Based on this, we consider two sets of posterior samples identical when their JSD is less than 10^{-3}. For two sets of samples with JSD above this threshold, it is possible to use importance sampling [85] to reweight the approximated posterior samples if they are not significantly biased. Therefore, we consider JSD at the order of 10^{-2} "acceptable" in this work.

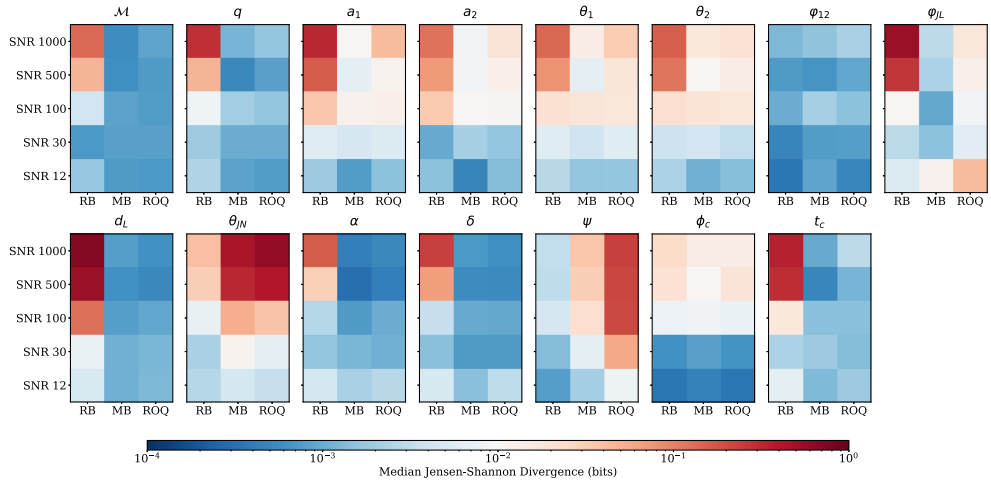
The JSD is shown in Fig. 2. For ROQ and MB, most parameters are estimated with good accuracy under SNR of



(a)GW151226-like



(b)GW150914-like



(c)GW190521-like

FIG. 2. The JSD between posteriors obtained using acceleration methods and the standard method. Each subfigure corresponds to a different source type, and each patch represents a different parameter. The color of each grid represent the value of JSD in *bits*. A JSD of 0 indicates that the two posterior distributions are identical, while a JSD of 1 indicates they are completely different.

100, with typical JSD below or slightly above 10^{-3} bits. We observe that the error increases with SNR: all acceleration methods fail to produce identical posterior samples in high-SNR scenarios, especially for spin-related parameters, polarization angle and the coalescence phase. This means that the likelihood needs to be approximated with a higher accuracy in high SNR cases. For instance, increasing the number of bases in ROQ may help address this issue. RB shows the poorest accuracy among all PE methods, despite having relatively accurate fiducial parameters in our experiments, suggesting more accurate fiducial parameters are required. Its accuracy deteriorates when the posterior is narrow, which happens in high SNR events and long signals (where the chirp mass posterior is narrow). Extra work is needed to determine the required accuracy of the fiducial parameter before performing a full PE. A comprehensive investigation of PE accuracy in the 3G era is beyond the scope of this paper and should be addressed in future work.

IV. COSTS ESTIMATES FOR 3G DETECTORS

With the estimated time cost given by Eq. (11), we now predict the total cost of Bayesian inference for 3G detectors using a realistic source population. We will first introduce the mock data catalog, ET MDC-1, and then demonstrate how the total cost is estimated.

A. Overview of ET MDC-1

ET MDC-1 contains one month of simulated observation with the 3G detector ET [102]. The population model used in ET MDC-1 includes 6812 BBH, 61217 BNS, and 2029 NSBH sources over the month, and varying numbers of detections possible depending on the detector network. We consider three types of detector networks: ET, ET-CE, and ET-CE-CEL (ET-2CE). Here, CE is assumed to be located at the LIGO Hanford site and CEL at the LIGO Livingston site, with the same PSD used in the PE experiments. We calculate the network SNR for each network configuration and present the cumulative distribution of the SNR in Fig. 3. Setting SNR $= 8$ as the detection threshold, a single ET could detect most BBH sources but would miss a certain fraction of BNS and NSBH sources. Adding CEs to the network significantly enlarges the detection range of BNS and NSBH sources.

We also show the cumulative distribution of the signal duration in Fig. 3. The signal duration is defined as the time it takes for the binary system to evolve from 5 Hz to the merger, computed using the 0PN equation. We add an additional 4s to the duration of each signal to account for the FFT corruption in real data analysis. Most BBH signals have durations less than 100s, while NSBH and BNS signals are mostly longer than 100s. Some BNS signals can last for over an hour.

Note that ET MDC-1 only provides the GW strain data for ET. However, in this study, we focus solely on the signal

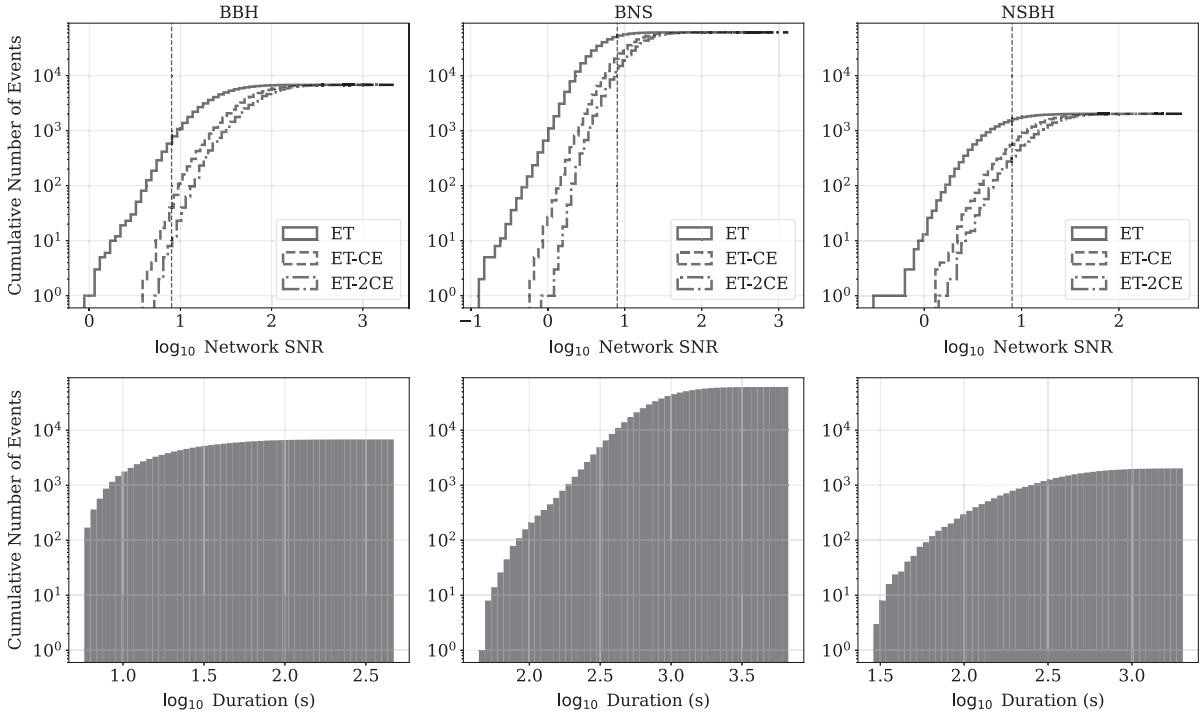


FIG. 3. Top: cumulative distributions of network SNR of different sources (in different columns) and different detector networks (in different line styles). The black dashed lines mark the SNR $= 8$ threshold. Bottom: cumulative distributions of signal durations, measured from 5 Hz, for different source types.

duration and the (optimal) SNR, which are independent of the actual noise realization.

B. Total PE cost estimates

Instead of directly using GW strain data, we simply use the signal durations and SNRs based on the MDC catalog to predict the sampling time with Eq. (11). However, we have overlooked some factors that also affect the sampling time. We discuss their expected contributions here

(i) Parameter space volume: Our experiments are based on 15-D PE for BBH systems. For NSBH and BNS systems, there are one and two more parameters for the tidal deformability, respectively. While the astrophysically restricted spin of neutron stars can make PE faster (which, however, is not always used in PE, e.g., Refs. [103,104]), the additional tidal parameter could counteract the speedup. Moreover, future analyses may incorporate eccentricity, adding two more dimensions. Detector calibration parameters [105] are also ignored in our simulation. Consequently, we are *underestimating* the total time cost for 3G. The scaling of sampling time with dimensionality is not entirely certain. Although it is shown that the number of likelihood evaluations scales exponentially with the dimensionality for NESSAI [50], the exact scaling depends on the new information (relevant to posterior and prior widths) brought by the new dimensions. Some samplers can achieve polynomial scaling [106], which may mitigate the increase in time cost due to the increased dimensionality.

(ii) Posterior shape: Adding/removing detectors may reform the shape of the posterior by reducing/increasing multimodalities, which affects the convergence of the sampler. Since the bilinear fitting is obtained using the ET-CE network, posteriors with ET only may have a more complex shape which takes the sampler longer to converge than the prediction. Similarly, the ET-2CE computational cost may be overestimated.

(iii) Sampling rate: We use a sampling rate of 2048 Hz in all experiments, but for 3G analysis this number could increase, especially for signals with higher modes and low mass systems. We are therefore *underestimating* the time cost.

(iv) Sampling algorithm and configuration: Our investigation a variant of the nested sampling algorithm, which is commonly used in current LVK analyses. Our sampler configuration is fixed in all experiments. However, for longer and louder signals, a larger nlive may be required, which slows down the PE. Since our experiments only extend to $T = 256$ s, we expect this could happen in signals that last for hours. In this sense, we are *underestimating* the total time cost. We also expect that results would vary with different choices of sampling method, for example MCMC does not require the code to sample the entire prior before converging on the posterior and so may run faster if initial parameter estimates are available [107].

(v) Overlapping signals: In realistic 3G detector scenarios, signals from different sources could overlap in the time domain [108–113], requiring a more sophisticated analysis such as joint PE. We do not consider this effect in the experiments, leading to an *underestimate* of time cost.

(vi) Long signal effects: For signals longer than ∼10 minutes, several additional effects need to be taken into consideration, such as the Earth's rotation [114] and variations in PSD [115]. These may bring more calculations during PE and thus we are *underestimating* of time cost.

(vii) *Ad hoc* ROQ: We restrict the prior of building ROQ to a small range, which reduces the required number of ROQ bases and lead to *underestimation* of time cost. However, on the other hand, ROQ could be improved with MB [69] and so compress the likelihood better, and this is not included in our experiments.

(viii) Duty cycle and evolving sensitivity: GW detectors may not operate continuously throughout the entire month, and their sensitivity may not reach the design level at the start of the observation period. As a result, we may *overestimate* the total time cost. However, the duty cycle can be easily included by applying a constant factor, and the sensitivity does not influence the validity of Eq. (11). Therefore, our methods and results can be readily extended to more realistic observational scenarios.

(ix) Hardware improvements: We employ AMD EPYC 7723 CPU model in this work, whose base clock is 2.85 GHz. Future hardware will be improved.

Although it appears we may have been underestimating the time cost, we expect that future developments in sampling techniques in coming years could further accelerate PE. For example, the DYNESTY sampler was tuned further in the analysis of GWTC-4 [116], which improved its performance. Given these caveats, we can say that with current technology, we provide an *optimistic* estimate of the future PE cost.

For each source in the MDC that is considered to be detected (network SNR > 8), we use Eq. (11) and the effective duration to estimate its PE time cost. The cumulative histogram of the PE time cost is shown in the top row of Fig. 4. Most events, including BNS and loud signals, can be analyzed within 100 CPU core hours using acceleration methods. However, when using the standard method, the time cost becomes prohibitively large, reaching up to 10^5 CPU core hours. The total time cost to analyze one month of observation is shown in the bottom row of Fig. 4. Without acceleration methods, the total CPU core

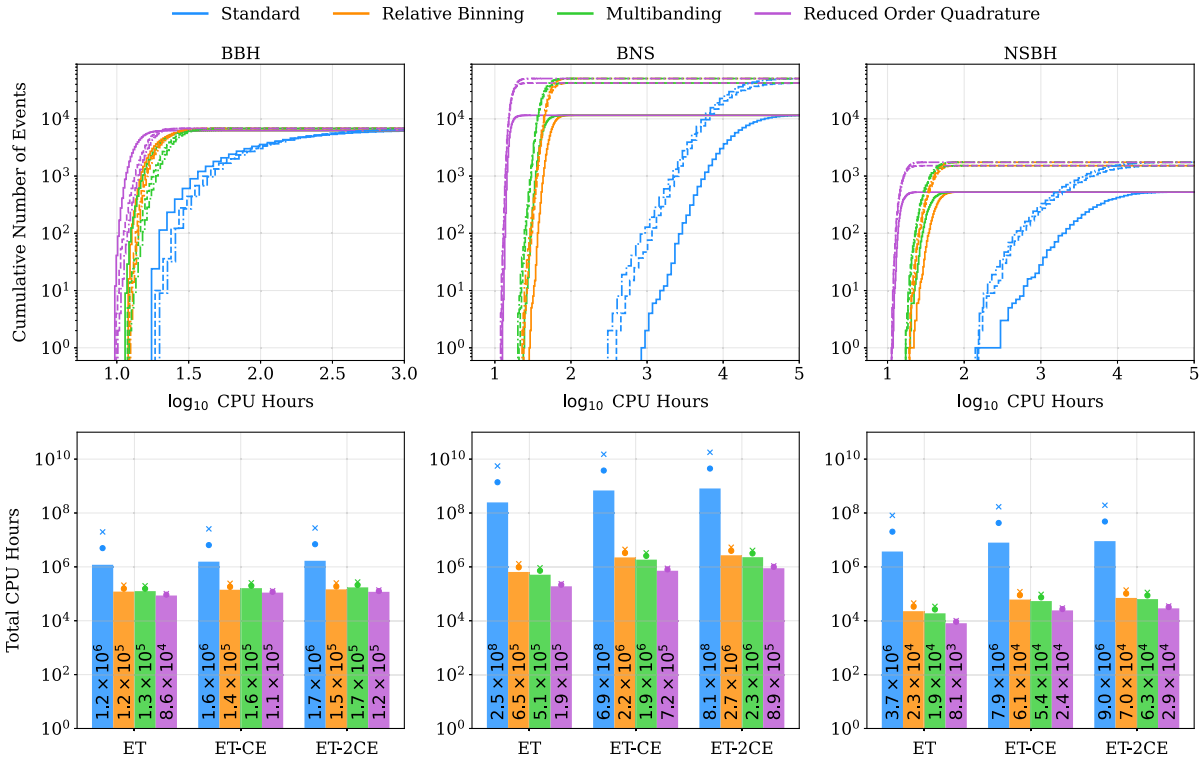


FIG. 4. Top: cumulative distributions of estimated PE cost for detected events (SNR > 8, in one-month observation) in CPU hours with different detector networks. The solid line represents ET-only network, dashed line represents the ET-CE network, and solid-dashed line represents the ET-2CE network. Different source types are in separate columns, with different colors representing various PE methods. Bottom: total CPU hours required to perform Bayesian PE for the one-month observation with different detector networks. The numbers indicate the corresponding CPU core hours for each bar. A dot and a cross above each bar represent the cost for $f_{\rm low} = 3$ Hz and 2 Hz, respectively.

hours required can reach 10^9, while acceleration methods can bring this number down to millions. Further reducing the $f_{\rm low}$ from 5 Hz to 3 Hz or 2 Hz, the cost for the standard method scales up drastically, while the acceleration methods do not. Therefore, we should expect the accelerated PE methods to play an important role in the future data analysis.

We do not fully explore different detector configurations in this study (e.g., L-shaped ET and CE-only networks). However, since the estimation of the PE time cost is purely a function of signal SNR and duration (i.e., the shape of the posterior is not considered), we do not expect the total PE time cost prediction to change significantly for different detector configurations. For example, for a single CE detector, the total time cost prediction should lie between the estimations of a single ET and ET-CE network, and a 2CE network should lie between the ET-CE and ET-2CE network. The lower frequency cutoff affects the total PE cost approximately by the factor given in Eq. (12), with examples of 3 Hz and 2 Hz shown in Fig. 4.

The LVK Computing Infrastructure [117,118] currently operates with fewer than 50 thousand CPU cores. Using the PE cost fitting for standard method, we could infer that at least 10^4–10^6 CPU hours is required to analyze all (over 200) detections over the past 10 years. It should be noted that this is the cost of performing PE only once; in LVK production, each event may be analyzed multiple times before the final results are obtained for publication. Ref. [117] shows that all data analysis activities during O3 cost 7.5×10^9 CPU hours under the HS06 benchmark [119], corresponding to $\sim 3 \times 10^8$ CPU core hours for the CPU model used in this work. Therefore, the millions of CPU hours required for the one-month observation in the future will be a heavy burden for current computing infrastructure. It would take several days to process on the entire computing cluster, consuming approximately 1 GWh of electricity and costing hundreds of thousands of USD in electricity charges, assuming 150W of CPU power and 0.15 dollar per kWh of electricity. Considering potential future upgrades to computing clusters (e.g., the Worldwide LHC Computing Grid, which operates with more than 500,000 CPU cores [120]), we conclude that while full Bayesian PE is technically feasible, it is neither budget-friendly nor environmentally sustainable. Therefore, more efficient PE methods, such as machine-learning-based methods that greatly outperforms traditional stochastic sampling methods discussed in this paper, will be crucial for GW astronomy in the 3G era.

V. CONCLUSIONS AND DISCUSSIONS

The computational challenge for the next-generation GW detectors has drawn significant attention in recent years, yet the precise computational cost has been largely an unexplored domain. In this paper, we investigated the computational cost associated with one of the most computationally intensive aspects of GW astronomy: Bayesian PE with stochastic sampling in the 3G era. We started from simulated signals that can be analyzed with a manageable cost and obtained a relationship between the sampling time and signal duration and SNR. With this relationship, we predicted the total PE cost for the event catalog from the ET MDC-1.

Our experiments include the standard PE method and three accelerated methods: relative binning, multibanding, and reduced order quadrature. The standard method has been shown to be impractically slow for the 3G catalog: using this approach, some events would require 10^5 CPU hours to analyze, and analyzing the entire one-month catalog could demand over billions CPU hours. In contrast, the acceleration methods significantly reduce the time cost. Most events can be analyzed within 100 CPU hours, and the total cost for the entire catalog observed in a month is reduced to millions of CPU hours.

Among the acceleration methods we tested, we found that ROQ provides the best speed while maintaining reasonable accuracy. However, the choice of ROQ prior may affect the speed [78]. While a narrower prior brings faster speed, more sets of ROQ bases would be needed to cover the entire parameter space, which would consume more computing resources to generate in advance. RB and MB have similar speeds, approximately twice as slow as ROQ. However, we observed accuracy issues with RB, potentially due to its dependency on fiducial parameters. Assuming not knowing the true parameters, how to assign accurate fiducial parameters needs to be addressed in future work (e.g., whether point estimates derived from detection pipelines are accurate enough.) We also emphasize that systematic errors in high SNR events need to be carefully controlled. For example, more ROQ bases may be required in high-SNR events for better accuracy. Further investigations into the accuracy of PE in the 3G era are necessary, and we plan to address this in future work.

As discussed in Sec. IV B, the numbers presented in this work represent an optimistic estimate of extending our analysis configurations to future data. However, the potentially underestimated computational cost for the future—millions of CPU hours per month—is sufficient to analyze all LVK detections over the past 10 years. It remains a substantial burden on computing infrastructure, electricity costs, and environmental impact. To address the computational burden in the 3G era, more efficient methods must be employed, such as advanced samplers, improved data compression, faster waveform evaluations, parameter marginalization, and machine learning techniques.

We hope that the methods and results in this work can serve as a reference and baseline benchmark for the research and development of 3G detectors. When novel data analysis methods are developed in the future, whether for more realistic data (e.g., including overlapped signals and noise variations) or a faster speed, their computational cost can be assessed and compared with the results presented here. For analyses at the catalog level, the computing efficiency must be carefully considered, as the cost for processing a single event will be multiplied across the entire 3G catalog.

ACKNOWLEDGMENTS

We would like to thank Tania Regimbau for helping with the use of ET MDC data and thank Michael Williams for helpful discussions about NESSAI. We also thank Filippo Santoliquido for helpful discussions and Steve Fairhurst for feedback. We are grateful for computational resources provided by Cardiff University, and funded by STFC Grant No. ST/I006285/1. This work was supported by STFC Grant No. ST/Y004256/1. Q. H. was also supported by CSC.

DATA AVAILABILITY

The data that support the findings of this article are openly available [99].

[1] J. Aasi *et al.* (LIGO Scientific Collaboration), Classical Quantum Gravity **32**, 074001 (2015).
[2] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **116**, 061102 (2016).
[3] F. Acernese *et al.* (VIRGO Collaboration), Classical Quantum Gravity **32**, 024001 (2015).
[4] T. Akutsu *et al.* (KAGRA Collaboration), Prog. Theor. Exp. Phys. **2021**, 05A101 (2021).
[5] A. G. Abac *et al.* (LIGO Scientific, VIRGO, and KAGRA Collaborations), arXiv:2508.18082.
[6] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. D **109**, 022001 (2024).
[7] R. Abbott *et al.* (KAGRA, VIRGO, and LIGO Scientific Collaborations), Phys. Rev. X **13**, 041039 (2023).
[8] S. L. Liebling and C. Palenzuela, Living Rev. Relativity **26**, 1 (2023).

[9] V. Cardoso and P. Pani, Living Rev. Relativity **22**, 4 (2019).
[10] B. P. Abbott *et al.* (LIGO Scientific, Virgo, 1M2H, Dark Energy Camera GW-E, DES, DLT40, Las Cumbres Observatory, VINROUGE, and MASTER Collaborations), Nature (London) **551**, 85 (2017).
[11] M. Soares-Santos *et al.* (DES, LIGO Scientific, and Virgo Collaborations), Astrophys. J. Lett. **876**, L7 (2019).
[12] B. P. Abbott *et al.* (LIGO Scientific, Virgo, and VIRGO Collaborations), Astrophys. J. **909**, 218 (2021).
[13] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Astrophys. J. Lett. **882**, L24 (2019).
[14] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Astrophys. J. Lett. **913**, L7 (2021).
[15] R. Abbott *et al.* (KAGRA, VIRGO, and LIGO Scientific Collaborations), Phys. Rev. X **13**, 011048 (2023).
[16] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **121**, 161101 (2018).
[17] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **118**, 121101 (2017); **119**, 029901 (E) (2017).
[18] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **118**, 121102 (2017).
[19] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **120**, 091101 (2018).
[20] L. Wade, J. D. E. Creighton, E. Ochsner, B. D. Lackey, B. F. Farr, T. B. Littenberg, and V. Raymond, Phys. Rev. D **89**, 103012 (2014).
[21] M. A. Shaikh, V. Varma, H. P. Pfeiffer, A. Ramos-Buades, and M. van de Meent, Phys. Rev. D **108**, 104007 (2023).
[22] A. Ramos-Buades, A. Buonanno, and J. Gair, Phys. Rev. D **108**, 124063 (2023).
[23] C. K. Mishra, K. G. Arun, B. R. Iyer, and B. S. Sathyaprakash, Phys. Rev. D **82**, 064010 (2010).
[24] R. Abbott *et al.* (LIGO Scientific, VIRGO, and KAGRA Collaborations), arXiv:2112.06861.
[25] J. Veitch *et al.*, Phys. Rev. D **91**, 042003 (2015).
[26] B. Zackay, L. Dai, and T. Venumadhav, arXiv:1806.08792.
[27] N. Leslie, L. Dai, and G. Pratten, Phys. Rev. D **104**, 123030 (2021).
[28] N. J. Cornish, arXiv:1007.4820.
[29] S. Vinciguerra, J. Veitch, and I. Mandel, Classical Quantum Gravity **34**, 115006 (2017).
[30] S. Morisaki, Phys. Rev. D **104**, 044062 (2021).
[31] P. Canizares, S. E. Field, J. Gair, V. Raymond, R. Smith, and M. Tiglio, Phys. Rev. Lett. **114**, 071104 (2015).
[32] R. Smith, S. E. Field, K. Blackburn, C.-J. Haster, M. Pürrer, V. Raymond, and P. Schmidt, Phys. Rev. D **94**, 044031 (2016).
[33] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. X **9**, 031040 (2019).
[34] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. X **11**, 021053 (2021).
[35] P. Couvares *et al.*, arXiv:2111.06987.
[36] S. Bagnasco *et al.*, EPJ Web Conf. **295**, 04015 (2024).
[37] M. Punturo *et al.*, Classical Quantum Gravity **27**, 194002 (2010).
[38] M. Branchesi *et al.*, J. Cosmol. Astropart. Phys. 07 (2023) 068.
[39] A. Abac *et al.*, arXiv:2503.12263.
[40] D. Reitze and *et al.*, Bull. Am. Astron. Soc. **51**, 035 (2019), https://baas.aas.org/pub/2020n7i035/release/1.
[41] M. Evans *et al.*, arXiv:2306.13745.
[42] S. Borhanian and B. S. Sathyaprakash, Phys. Rev. D **110**, 083040 (2024).
[43] R. Tania and S. Jishnu, arXiv:2506.12237.
[44] E. Thrane and C. Talbot, Pub. Astron. Soc. Aust. **36**, e010 (2019).
[45] N. Christensen and R. Meyer, Rev. Mod. Phys. **94**, 025001 (2022).
[46] L. S. Finn, Phys. Rev. D **46**, 5236 (1992).
[47] J. Skilling, AIP Conf. Proc. **735**, 395 (2004).
[48] J. S. Speagle, Mon. Not. R. Astron. Soc. **493**, 3132 (2020).
[49] M. J. Williams, J. Veitch, and C. Messenger, Mach. Learn. Sci. Tech. **4**, 035011 (2023).
[50] M. J. Williams, J. Veitch, and C. Messenger, Phys. Rev. D **103**, 103006 (2021).
[51] S. Koposov *et al.*, dynesty:v2.1.4 (2024).
[52] M. J. Williams, J. Veitch, and R. Tenorio, mj-will/nessai: v0.8.1 (2023).
[53] C. Cutler and E. E. Flanagan, Phys. Rev. D **49**, 2658 (1994).
[54] S. Fairhurst, New J. Phys. **11**, 123006 (2009); **13**, 069602 (E) (2011).
[55] G. Ashton *et al.*, Astrophys. J. Suppl. Ser. **241**, 27 (2019).
[56] K. Krishna, A. Vijaykumar, A. Ganguly, C. Talbot, S. Biscoveanu, R. N. George, N. Williams, and A. Zimmerman, arXiv:2312.06009.
[57] V. Villa-Ortega, T. Dent, and A. C. Barroso, Mon. Not. R. Astron. Soc. **515**, 5718 (2022).
[58] A. Buonanno, B. Iyer, E. Ochsner, Y. Pan, and B. S. Sathyaprakash, Phys. Rev. D **80**, 084043 (2009).
[59] N. J. Cornish, Phys. Rev. D **103**, 104057 (2021).
[60] J. Roulet, J. Mushkin, D. Wadekar, T. Venumadhav, B. Zackay, and M. Zaldarriaga, Phys. Rev. D **110**, 044010 (2024).
[61] L. Dai, T. Venumadhav, and B. Zackay, arXiv:1806.08793.
[62] M. Dax, S. R. Green, J. Gair, N. Gupte, M. Pürrer, V. Raymond, J. Wildberger, J. H. Macke, A. Buonanno, and B. Schölkopf, Nature (London) **639**, 49 (2025).
[63] K. W. K. Wong, M. Isi, and T. D. P. Edwards, Astrophys. J. **958**, 129 (2023).
[64] T. Dietrich, A. Samajdar, S. Khan, N. K. Johnson-McDaniel, R. Dudi, and W. Tichy, Phys. Rev. D **100**, 044003 (2019).
[65] T. Adams, D. Buskulic, V. Germain, G. M. Guidi, F. Marion, M. Montani, B. Mours, F. Piergiovanni, and G. Wang, Classical Quantum Gravity **33**, 175012 (2016).
[66] F. Aubin *et al.*, Classical Quantum Gravity **38**, 095004 (2021).
[67] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. Jiménez Forteza, and A. Bohé, Phys. Rev. D **93**, 044006 (2016).
[68] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez Forteza, and A. Bohé, Phys. Rev. D **93**, 044007 (2016).
[69] S. Morisaki, R. Smith, L. Tsukada, S. Sachdev, S. Stevenson, C. Talbot, and A. Zimmerman, Phys. Rev. D **108**, 123040 (2023).
[70] Q. Hu, J. Irwin, Q. Sun, C. Messenger, L. Suleiman, I. S. Heng, and J. Veitch, Astrophys. J. Lett. **987**, L17 (2025).

[71] Q. Hu and J. Veitch, Astrophys. J. Lett. **958**, L43 (2023).
[72] S. E. Field, C. R. Galley, J. S. Hesthaven, J. Kaye, and M. Tiglio, Phys. Rev. X **4**, 031006 (2014).
[73] M. Pürrer, Classical Quantum Gravity **31**, 195010 (2014).
[74] J. Blackman, S. E. Field, C. R. Galley, B. Szilágyi, M. A. Scheel, M. Tiglio, and D. A. Hemberger, Phys. Rev. Lett. **115**, 121102 (2015).
[75] L. M. Thomas, G. Pratten, and P. Schmidt, Phys. Rev. D **106**, 104029 (2022).
[76] R. Smith *et al.*, Phys. Rev. Lett. **127**, 081102 (2021).
[77] G. Morras, J. F. N. Siles, and J. Garcia-Bellido, Phys. Rev. D **108**, 123025 (2023).
[78] S. Morisaki and V. Raymond, Phys. Rev. D **102**, 104020 (2020).
[79] J. Lange, R. O'Shaughnessy, and M. Rizzo, arXiv:1805.10457.
[80] J. Wofford *et al.*, Phys. Rev. D **107**, 024040 (2023).
[81] L. P. Singer and L. R. Price, Phys. Rev. D **93**, 024013 (2016).
[82] Q. Hu, C. Zhou, J.-H. Peng, L. Wen, Q. Chu, and M. Kovalam, Phys. Rev. D **104**, 104008 (2021).
[83] T. Tsutsui, K. Cannon, and L. Tsukada, Phys. Rev. D **103**, 043011 (2021).
[84] M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno, and B. Schölkopf, Phys. Rev. Lett. **127**, 241103 (2021).
[85] M. Dax, S. R. Green, J. Gair, M. Pürrer, J. Wildberger, J. H. Macke, A. Buonanno, and B. Schölkopf, Phys. Rev. Lett. **130**, 171403 (2023).
[86] N. Gupte *et al.*, arXiv:2404.14286.
[87] A. Polanska, T. Wouters, P. T. H. Pang, K. K. W. Wong, and J. D. McEwen, in *Proceedings of the 38th Conference on Neural Information Processing Systems* (2024), arXiv:2410.21076.
[88] I. Kobyzev, S. J. Prince, and M. A. Brubaker, IEEE Trans. Pattern Anal. Mach. Intell. **43**, 3964 (2020).
[89] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, J. Mach. Learn. Res. **22**, 1 (2021).
[90] H. Gabbard, C. Messenger, I. S. Heng, F. Tonolini, and R. Murray-Smith, Nat. Phys. **18**, 112 (2022).
[91] A. Pagnoni, K. Liu, and S. Li, arXiv:1812.04405.
[92] J. Langendorff, A. Kolmus, J. Janquart, and C. Van Den Broeck, Phys. Rev. Lett. **130**, 171402 (2023).
[93] Y. Bouffanais and E. K. Porter, Phys. Rev. D **100**, 104023 (2019).
[94] D. Wysocki, R. O'Shaughnessy, J. Lange, and Y.-L. L. Fang, Phys. Rev. D **99**, 084026 (2019).
[95] S. Hild *et al.*, Classical Quantum Gravity **28**, 094013 (2011).
[96] C. Talbot *et al.*, BILBY-dev/BILBY:v2.2.2 (2024).
[97] The sampler NESSAI provides total sampling wall time (the wall time to finish the whole analysis) and wall time spent evaluating likelihood. Since the likelihood is evaluated in parallel while all other processes use a single CPU core, the total CPU time is calculated as (total sampling wall time-likelihood evaluation wall time)*1 + likelihood evaluation wall time* `ncpu`, which simplifies to the form in the main text.
[98] H. Qi and V. Raymond, Phys. Rev. D **104**, 063031 (2021).
[99] https://github.com/MarinerQ/3gpemethods.
[100] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011), https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.
[101] I. M. Romero-Shaw *et al.*, Mon. Not. R. Astron. Soc. **499**, 3295 (2020).
[102] ET MDC1 data are publicly available from http://et-origin.cism.ucl.ac.be.
[103] B. P. Abbott *et al.*, Phys. Rev. X **9**, 011001 (2019).
[104] B. P. Abbott *et al.*, Astrophys. J. **892**, L3 (2020).
[105] B. P. Abbott *et al.* (LIGO Scientific Collaboration), Phys. Rev. D **95**, 062003 (2017).
[106] W. J. Handley, M. P. Hobson, and A. N. Lasenby, Mon. Not. R. Astron. Soc. **453**, 4385 (2015).
[107] T. Wouters, P. T. H. Pang, T. Dietrich, and C. Van Den Broeck, Phys. Rev. D **110**, 083033 (2024).
[108] Y. Himemoto, A. Nishizawa, and A. Taruya, Phys. Rev. D **104**, 044010 (2021).
[109] E. Pizzati, S. Sachdev, A. Gupta, and B. Sathyaprakash, Phys. Rev. D **105**, 104016 (2022).
[110] P. Relton and V. Raymond, Phys. Rev. D **104**, 084039 (2021).
[111] P. Relton, A. Virtuoso, S. Bini, V. Raymond, I. Harry, M. Drago, C. Lazzaro, A. Miani, and S. Tiwari, Phys. Rev. D **106**, 104045 (2022).
[112] A. Samajdar, J. Janquart, C. V. D. Broeck, and T. Dietrich, Phys. Rev. D **104**, 044003 (2021).
[113] Q. Hu and J. Veitch, Astrophys. J. **945**, 103 (2023).
[114] R. Essick, S. Vitale, and M. Evans, Phys. Rev. D **96**, 084004 (2017).
[115] C. Palomba, P. Astone, and S. Frasca, Classical Quantum Gravity **22**, S1255 (2005).
[116] A. G. Abac *et al.* (LIGO Scientific, VIRGO, and KAGRA Collaborations), arXiv:2508.18081.
[117] S. Bagnasco, EPJ Web Conf. **295**, 04047 (2024).
[118] LIGO Scientific Collaboration, LIGO Data Grid (LDG), https://computing.docs.ligo.org/guide/computing-centres/ldg/ (2024).
[119] J. L. Henning, SIGARCH Comput. Archit. News **34**, 1 (2006).
[120] J. Albrecht *et al.* (HEP Software Foundation Collaboration), Comput. Software Big Sci. **3**, 7 (2019).