Accelerating tests of general relativity with gravitational-wave signals using hybrid sampling

Noah E. Wolfe,[1,2,*] Colm Talbot,[3,4,†] and Jacob Golomb[5]

[1]*Department of Physics, North Carolina State University, Raleigh, North Carolina 27695, USA*
[2]*Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695, USA*
[3]*LIGO Laboratory, Massachusetts Institute of Technology, 185 Albany Street, Cambridge, Massachusetts 02139, USA*
[4]*Department of Physics and Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA*
[5]*LIGO Laboratory, California Institute of Technology, Pasadena, California 91125, USA*

(Received 26 August 2022; accepted 20 April 2023; published 30 May 2023)

The Advanced LIGO/Virgo interferometers have observed ∼100 gravitational-wave transients enabling new questions to be answered about relativity, astrophysics, and cosmology. However, many of our current procedures for computing these constraints will not scale well with the increased size of future transient catalogs. We introduce a novel hybrid sampling method in order to more efficiently perform parameterized tests of general relativity with gravitational-wave signals. Applying our method to the binary black hole merger GW150914 and simulated signals we find that our method is approximately an order of magnitude more efficient than the current method with conservative settings for our hybrid analysis. While we have focused on the specific problem of measuring potential deviations from relativity, our method is of much wider applicability to any problem that can be decomposed into a simple and more complex model(s).

DOI: 10.1103/PhysRevD.107.104056

I. INTRODUCTION

General relativity (GR) is currently our most successful theory of gravity. Previous observations of sources within our solar system, including the Gravity Probe B experiment and time-delay measurements with the Cassini space probe, have placed constraints on deviations from general relativity in the nondynamical, weak-field regime [1]. These tests have been replicated and expanded with radio observations of pulsars, which probe similarly slow-motion, but strong-field gravitational physics [2,3] through measurements of the orbital decay rate of the first discovered binary pulsar system [4] to modern constraints on dipolar gravitational-wave emission constructed with multiple such systems [5]. Simultaneously, probes of large-scale cosmological structure including weak gravitational lensing and the cosmic microwave background have provided complimentary weak-field tests of general relativity across cosmic epochs and length scales [6]. Over the past decade new observations have unlocked the strong-field regime for tests of general relativity, including measurements of two supermassive black hole shadows [7–9] and gravitational waves from stellar-mass compact object mergers observed by Advanced LIGO [10] and Advanced Virgo [11], using both single observations [12,13] and the burgeoning population of gravitational wave transients [14–17]. To date, none of these experiments have found significant disagreement with the predictions of general relativity.

However, alternative theories of gravity that could emerge in the strong-field regime may be relevant to constructing unified field theories or the understanding of unexplained phenomena like dark energy (e.g. in scalar-tensor theories, among others [18]). Modern developments in theoretical physics have generated testable predictions of modifications to gravitational-wave emission from compact binary coalescence under alternative formulations of gravity, both analytically [19–25] and numerically [26–33], further enabling tests of general relativity in the most extreme gravitational environments yet accessible to us.

The number of observed mergers will only continue to grow, and our gravitational-wave detectors will continue to become more sensitive, further enhancing our resolution on potential deviations from general relativity. However, this also necessitates that our statistical and computational techniques improve to support larger and more complex analyses. Since the first observation of gravitational waves from a compact binary merger, the TIGER (Test Infrastructure for General Relativity) formalism and related methods have been some of the flagship analyses

*noah.wolfe@ligo.org
†colm.talbot@ligo.org

performed by the LIGO-Virgo-KAGRA scientific collaborations [13–17,34–36]. These methods require performing many independent, but largely identical, analyses for each potential deviation from relativity as the parameters describing the GR signal must be inferred from scratch. The goal of this work is to improve this analysis procedure with a new method for parameter estimation: hybrid sampling. Here, our hybrid approach uses an analysis assuming that general relativity is correct to initialize the inference of deviations from general relativity. This method is more computationally efficient, allowing us to scale our analysis as the population of observed mergers grows, and further constrain deviations in gravitational-wave signals predicted by general relativity.

The remainder of the paper is structured as follows. In Sec. II, we provide relevant background and introduce our hybrid sampling method. After this, we provide a demonstration of our method on a simple toy model in Sec. III. We then describe our model for observed gravitational waves according to general relativity and the parametrized deviations we consider in Sec. IV. In Sec. V, we apply our hybrid sampling method to simulated and real gravitational-wave signals. Specifically, we demonstrate that our method returns equivalent results to the previous method at a fraction of the computational cost and introduce an extension to the previous method. Finally, we provide closing thoughts in Sec. VI.

II. METHODS

A. Bayesian inference for gravitational-wave transients

We begin with a brief review of Bayesian inference in the context of gravitational-wave astronomy. In Bayesian inference, we wish to infer a set of parameters θ of a model M given some data d; formally, we want to construct the posterior distribution $p(\theta|d,M)$. For example, in this work, we will have a set of parameters that include properties of binary black hole systems (e.g. mass and spin), with additional parameters to denote deviations from the predictions of general relativity that we wish to infer from observations of gravitational-wave transients. For additional details, see, e.g. [37].

Bayesian inference allows us to construct the posterior distribution via Bayes' theorem,

$$p(\theta|d,M) = \frac{\mathcal{L}(d|\theta,M)\pi(\theta|M)}{\mathcal{Z}(d|M)}, \tag{1}$$

where $\mathcal{L}(d|\theta,M)$ is the likelihood of observing the data given parameter values, and $\pi(\theta|M)$ is the prior distribution, which encodes our assumptions about the Universe before considering the data. The normalization factor $\mathcal{Z}(d|M)$ is known as the evidence and is the probability of observing the data given the parametric model we choose

$$\mathcal{Z}(d|M) \equiv \int d\theta \mathcal{L}(d|\theta,M)\pi(\theta|M). \tag{2}$$

We may suppress the model M in subsequent expressions, however, everything is conditioned on a model.

When analyzing gravitational-wave transients we assume that the noise in each of our interferometers is a stationary Gaussian process described by a power spectral density S in the frequency domain. Additionally, our analysis is triggered by matched filter search pipelines that tell us a coherent non-Gaussian transient that is most likely an astrophysical signal is present in the data. To model this, we use the Whittle likelihood approximation [38] for the residual noise after subtracting the response of each detector to our template h for the astrophysical signal

$$\mathcal{L}(d|\theta) = \prod_{i,j} \frac{1}{2\pi S_{ij}} \exp\left(-\frac{4}{T}\frac{|d_{ij} - h(\theta)_{ij}|^2}{S_{ij}}\right). \tag{3}$$

Here, the products run over the interferometers in the network, and frequencies for the data in each interferometer are (generally) assumed to be uncorrelated. We note that our parameters only describe the astrophysical template and the response of the detector; however, it is also possible to construct parametrized models for the power spectrum [39]. The quantity T is the duration of data being analyzed and is the inverse of the frequency resolution.

We observe that $p(\theta|d)$ provides a distribution on the entire (multidimensional) set of parameters θ. To extract information on specific parameters of interest θ_i, we must "marginalize," i.e. integrate, over the rest of the parameters:

$$p(\theta_i|d) = \int \left(\prod_{k\neq i} d\theta_k\right) p(\theta|d). \tag{4}$$

This integration may be difficult to compute through standard numerical methods, especially in a high-dimensional parameter space. One common method to approximate this integral is to utilize a Markov chain Monte Carlo (MCMC) method [40,41], wherein a "walker" explores the parameter space of θ under rules such that, given enough iterations, the combined steps along its path form a representative sample of the posterior distribution. Another, more recent method is nested sampling [42,43], which instead focuses on estimating the evidence $\mathcal{Z}$, from which the posterior distribution can then be calculated. In this work, we will utilize both of these approaches and, in turn, detail specific implementations of these methods in the following subsections.

B. Nested sampling

Nested sampling, as developed in [42,43], is an algorithm to estimate the evidence $\mathcal{Z}$ and posterior probability density by climbing up discrete contours on the likelihood

surface and has been widely adopted in astrophysics including gravitational-wave astronomy [44–47]. We direct interested readers to [48] for a recent review. The core insight of nested sampling is that the high dimensional integral to compute the evidence $\mathcal{Z}$ can be approximated as a one-dimensional integral over a quantity known as the "prior mass" X. The prior mass corresponding to a likelihood value λ is the fraction of the volume that has a likelihood greater than λ

$$X(\lambda) = \int_{\mathcal{L}(\theta)>\lambda} \pi(\theta) d\theta. \tag{5}$$

If the mapping from $\theta \to X$ can be found, then the evidence [Eq. (2)] can be rewritten as

$$\mathcal{Z} = \int_0^1 \mathcal{L}(X) dX. \tag{6}$$

The nested sampling algorithm constructs this mapping numerically by gradually climbing the likelihood surface and we approximate $\mathcal{Z}$ as a weighted sum of values $\mathcal{L}(X)$, e.g.,

$$\mathcal{Z} \approx \sum_{i=1}^{N} w_i \mathcal{L}_i \tag{7}$$

for some number of samples N, with the prior volume associated with each likelihood isocontour w_i. For a full derivation of the functional form of w_i, see, e.g., [43]. For this work, we use the implementation of nested sampling in DYNESTY [46].

Another widely used feature of nested sampling is that the elements in the sum in Eq. (7) are the posterior weights associated with nested sampling. We can therefore generate samples from the posterior distribution by weighting the nested samples according to a normalized version of that quantity

$$p_i = \frac{\mathcal{L}_i w_i}{\mathcal{Z}}. \tag{8}$$

We note that after a sufficient number of iterations, nested sampling no longer produces additional posterior samples. This is because the algorithm continually climbs the likelihood surface and eventually the reduction in prior volume overcomes the increase in the likelihood and the posterior weights begin to decrease. This means that the number of posterior samples generated by a nested sampling analysis is completely determined by the shape of the likelihood surface.

We note that the values of the $\mathcal{L}_i$ are irrelevant to the nested sampling algorithm, and that only their order matters.[1] Therefore, we are free to perform any monotonic operation on the likelihood and can then trivially recompute the evidence and generate samples from the posterior distribution. In this work, we will focus on a specific family of operations that change the effective inverse "temperature" β_T of the posterior distribution[2]

$$\mathcal{L} \to \mathcal{L}^{\beta_T}, \tag{9}$$

$$p_{i,\beta_T} = \frac{\mathcal{L}_i^{\beta_T} w_i}{\mathcal{Z}_{\beta_T}}. \tag{10}$$

This "athermal" property of nested sampling has been known since the first introduction of the algorithm but has not been widely utilized.

In Fig. 1, we show the posterior weight p_{i,β_T} as a function of iteration for various temperatures for a simple model described in Sec. III. We note that, for the $\beta_T = 1$ case, we recover the usual posterior weights. As we increase the temperature (decreasing β_T) the posterior weights peak earlier in the nested sampling chain.

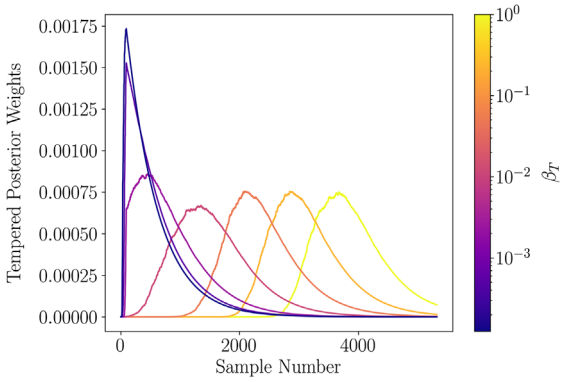

FIG. 1. Posterior weights p_i generated by DYNESTY, tempered according to the default method described in [49]. As $\beta_T \to 1$, we recover the original distribution of posterior weights. As $\beta_T \to 0$ we recover an exponential distribution, the result of choosing nested samples according to $X_i = \exp(-i/N)$.

C. Parallel-tempered Markov chain Monte Carlo methods

In contrast with nested sampling, MCMC methods directly explore the posterior and can be run as long as necessary, continually generating additional samples from the posterior distribution. Ensemble MCMC methods build upon existing MCMC methods by replacing a single walker, as used in traditional approaches [40,41], with an ensemble of walkers that explore the parameter space in parallel, e.g., [50]. A key feature of an ensemble approach is that we can reduce the number of iterations we need to

[1]Strictly speaking, this is only true if one has an infinite chain of nested samples.

[2]The temperature, in this case, is defined in analogy with statistical physics which historically shares strong links with Monte Carlo analyses.

evolve the MCMC to accurately resolve the posterior distribution, as ensembles of walkers have a far shorter autocorrelation length, measuring the correlation between sampling steps, than single walkers [50,51]. Additionally, at any step, the state of our ensemble is a representative estimate of the posterior distribution. For a recent review, we direct the readers to [52]. In further contrast with nested sampling techniques, we can choose the initial distribution of points in our ensemble. Common initialization schemes include drawing random samples from the prior distribution and initializing samples around a maximum likelihood estimate, however, the optimal initialization is a realization of the target distribution.

Further, ensemble MCMC methods can be parallel-tempered to explore the posterior distribution at arbitrary temperatures [53]. A parallel-tempered ensemble MCMC method then uses many walkers, in parallel, each exploring a tempered posterior surface

$$p_{\beta_T}(\theta|d) = \frac{\mathcal{L}^{\beta_T}(d|\theta)\pi(\theta)}{\mathcal{Z}_{\beta_T}(d)}. \tag{11}$$

We note that this is a continuous version of Eq. (10). For higher temperatures (smaller β_T), the ensemble can more easily explore the full prior volume, and by allowing walkers to jump between different temperature ensembles the convergence time of the $\beta_T = 1$ ensemble is greatly reduced. In this work, we use the PTEMCEE implementation of parllel-tempered ensemble MCMC method [49,51].

While MCMC methods may not always be as computationally efficient as nested sampling methods, once they have reached a stationary state, they have a relatively high computational efficacy; i.e. we can continually ask our MCMC sampler for more samples, increasing our resolution of the posterior as much as we desire. Therefore, if we can initialize our MCMC ensembles to closely approximate the target distributions, we can achieve very large efficiencies.

D. Hybrid sampling

We propose a hybrid sampling scheme to explore high-dimensional, degenerate parameter spaces that uses the exponential compression of the prior mass provided by nested sampling to seed a set of parallel-tempered MCMC ensembles. In particular, we seed each tempered ensemble by rejection sampling the nested samples, weighted by the tempered posterior weights defined by Eq. (10), which approximates the tempered posterior distribution defined by Eq. (11). In this work, we temper our MCMC ensembles over the default temperature ladder used by PTEMCEE, as detailed in Sec. 2.1 of [49]. In the case that the nested sampling is unbiased, this procedure initializes each ensemble from a realization of their target distribution. This is the optimal seeding for the MCMC ensembles. We can then run the tempered ensembles to generate an arbitrary number of samples from the target distribution (in contrast to nested sampling which can only generate a fixed number of samples, although dynamic nested sampling has been provides another solution to this problem [54]).

Our method can also be applied to more complex cases where the MCMC evolution explores an extended parameter space compared to the nested sampling analysis. We define two models M_1 and M_2 described by parameter sets $\theta_1 \subseteq \theta_2$ with likelihoods $\mathcal{L}_1$ and $\mathcal{L}_2$. We denote the extension parameters as $\bar{\theta}$; i.e. θ_2 is the union of θ_1 and $\bar{\theta}$. We first perform a nested sampling analysis of the data under model M_1, generating the posterior distribution $p(\theta_1|d, M_1)$. Since M_2 contains M_1, there must exist some value of $\bar{\theta}$ for which M_2 reduces to M_1, which we call $\bar{\theta}'$. Therefore, $p(\theta_1|d, M_1) = p(\theta_2|d, M_2, \bar{\theta}')$, and we can consider other realizations of $p(\theta_2|d, M_2)$ as an extension of the distribution for which $\bar{\theta} = \bar{\theta}'$. So, the posterior distribution for θ_1 that we achieve via nested sampling provides an efficient starting state for MCMC ensembles sampling in θ_2 under M_2. For the parameters of θ_2 included in θ_1, we seed each tempered ensemble as in the case where $M_1 = M_2$. In the remaining parameters $\bar{\theta}$, we initialize our chains from narrow distributions centered around $\bar{\theta}'$.

In the rest of this work, we will take M_1 to be a model of the gravitational-wave signal from a binary black hole merger in accordance with general relativity, and M_2 to be a model of the same phenomenon that allows for deviations from GR. Then, $\bar{\theta}$ are parameters of these deviations, and $\bar{\theta}'$ is zero in each deviation parameter.

III. HYBRID SAMPLING WITH A GENERALIZED GAUSSIAN DISTRIBUTION

As a demonstration of our hybrid sampling framework, we use a toy model where our reduced model in the first step is the standard Gaussian distribution characterized by mean μ and standard deviation σ and the complex model in the second step is a generalized Gaussian distribution characterized by mean μ, scale α, which can take alternative shapes parametrized by β. For comparison, the probability density function of the standard Gaussian is

$$P(x|\mu, \sigma) = \frac{1}{\sqrt{\pi\alpha^2}} e^{-(x-\mu)^2/\alpha^2}, \tag{12}$$

while that of the generalized Gaussian we employ is

$$P(x|\mu, \alpha, \gamma) = \frac{\beta}{2\alpha\Gamma(1/\gamma)} e^{-(|x-\mu|/\alpha)^\gamma}, \tag{13}$$

where Γ is the Gamma function. For consistency with the generalized model, we parametrize our Gaussian distribution with the parameter $\alpha = \sqrt{2}\sigma$. When the shape $\gamma = 1$,

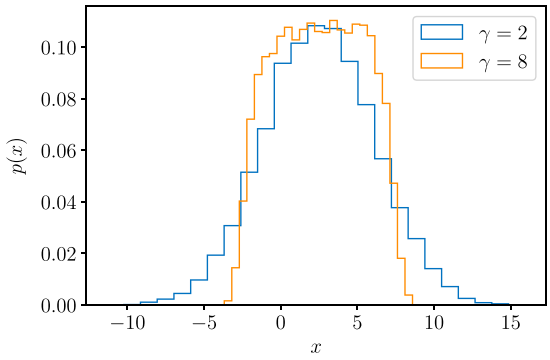

FIG. 2. Realizations of samples from the standard Gaussian distribution (blue) and generalized Gaussian distribution with $\gamma = 8$ (orange). The $\gamma = 8$ case is closer to a tophat function that the standard Gaussian.

we recover the Laplace distribution, while as $\gamma \to \infty$, we recover a tophat distribution on $(\mu - \alpha, \mu + \alpha)$. When $\gamma = 2$, we recover the standard Gaussian distribution.

As an example of these two distributions, we show the data used in the two examples considered in this section in Fig. 2. In blue we show samples from the standard normal distribution, while the orange shows samples from the generalized distribution with $\gamma = 8$. Thus, if the data follow a distribution with $\gamma \neq 2$, the value of α will be incorrectly estimated. In the remainder of this section, we verify that our hybrid sampling method can recover μ, α, and γ when we correctly assume that the data follows a standard Gaussian distribution during the first step of hybrid sampling and when the underlying data do not follow a Gaussian distribution. For our analyses, we use prior distributions as described in Table I.

A. Well-specified initial model

First, we verify that hybrid sampling can recover model parameters when the data follows a normal distribution with $\mu = 3$ and $\alpha = 5$. Our data d are $N = 10000$ random samples from this distribution. In the first step of hybrid sampling, we use DYNESTY to sample in $\{\mu, \alpha\}$, assuming that the data follows a standard Gaussian distribution, generating a posterior distribution we denote $p_1(\mu, \sigma|d)$ using 500 live points. Before the second step of hybrid sampling, we prepare initial points $\{\mu_0, \alpha_0\}$ for an ensemble of 200 walkers at seven temperatures as described in Sec. II D. We generate initial values of the shape parameter γ_0 by sampling from a standard Gaussian distribution with a standard deviation of 0.01 and centered on the value of γ assumed in the first hybrid step, $\gamma = 2$. We then evolve the ensemble using PTEMCEE for 128 iterations, discarding the first 100 iterations as burn-in.

TABLE I. Prior distributions for the parameters of the generalized Gaussian model. We denote a uniform distribution over $[a, b]$ as $\mathcal{U}(a, b)$. We note that the initial phase of the hybrid sampling fixes $\gamma = 2$.

Parameter	Distribution
μ	$\mathcal{U}(0, 5)$
α	$\mathcal{U}(0, 10\sqrt{2})$
γ	$\mathcal{U}(0, 10)$

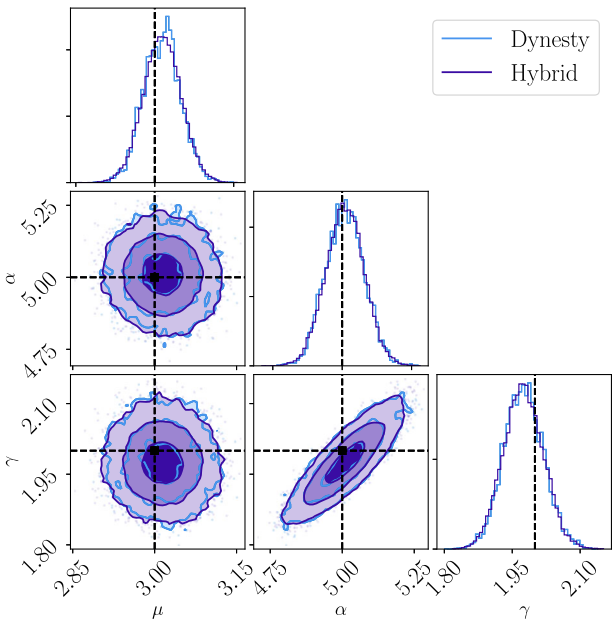

FIG. 3. Posteriors for μ, α, γ generated by our hybrid sampling method with the model well-specified during the first step, shown in purple, and DYNESTY sampling under a generalized Gaussian distribution for verification, shown in blue. True parameter values are shown in black. Contour level curves in this figure and all following two-dimensional distribution plots denote, from inside out, 39.3% (1-sigma level), 86.4% (2-sigma), and 98.8% (3-sigma) of the distribution volume.

For comparison, we also analyze the data under the generalized model with DYNESTY directly. In Fig. 3, we compare the posterior generated by DYNESTY alone (blue) to the one generated by hybrid sampling (purple). The dashed black lines show the true values of the three parameters. We see that both methods recover equivalent posterior distributions indicating that the hybrid sampling method is well converged.

B. Misspecified initial model

Next, we verify that hybrid sampling can recover model parameters when the model likelihood used in the first step has been inappropriately specified for the data. We repeat the first step of hybrid sampling as in the previous section. However, the input data we generate is not Gaussian. Instead, we generate $N = 10000$ random samples from a generalized Gaussian distribution with $\mu = 3$, $\alpha = 5$, and $\gamma = 8$, shaped approximately like a tophat function. We perform the same analyses as in Sec. III A, including the

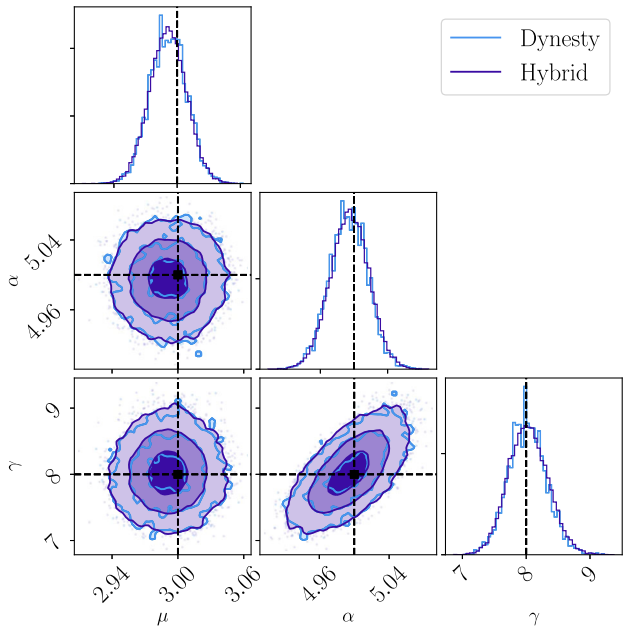

FIG. 4. Posteriors for μ, α, γ generated by our hybrid sampling method with the model misspecified during the first step, shown in purple, and DYNESTY sampling under a generalized Gaussian distribution for verification, shown in blue. True parameter values are shown in black.

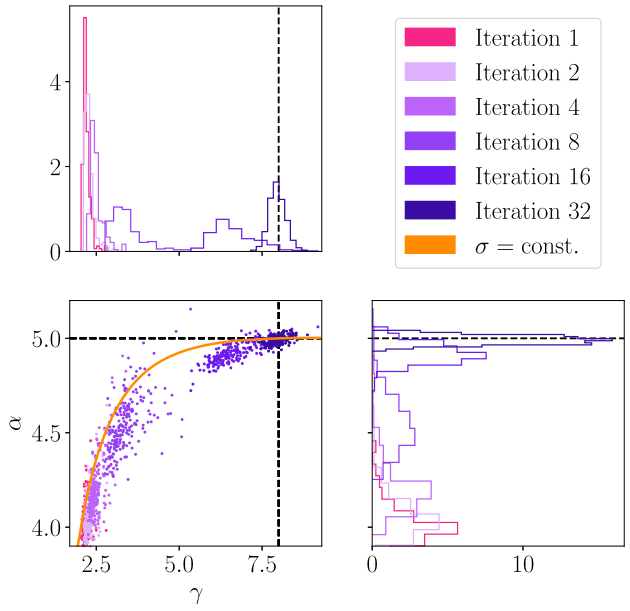

FIG. 5. Visualization of the evolution of the $\beta_T = 1$ ensemble for our generalized Gaussian model during the MCMC sampling stage. The different colored scatter plots and histograms correspond to different iterations of the ensemble. The orange curve shows a line of constant σ [Eq. (14)] which describes the degeneracy between the α and γ. We note that the ensemble approximately evolves along a curve of constant σ. We show the full one-dimensional evolution in Fig. 6.

discard of the first 100 iterations as burnin. In Fig. 4, we show the posterior distributions estimated using our two methods. Once again, we see that both methods recover equivalent results.

We are additionally interested in understanding the performance of the hybrid sampling stage. In Fig. 5 we provide two-dimensional snapshots from the evolution. The top and left-hand panels show the one-dimensional marginal distributions for γ and α at each iteration. The colors are consistent between the panels and darker shades correspond to later iterations of the evolution. We note that α and γ are strongly correlated, specifically the variance of the distribution

$$\sigma^2 = \alpha^2 \frac{\Gamma(\frac{3}{\gamma})}{\Gamma(\frac{1}{\gamma})} \tag{14}$$

is approximately conserved. The orange curve is constant σ intersecting the true values of α and γ. We note that the parameters α and γ are correlated and so the evolution follows the direction of the correlation. We observe that as the ensemble evolves, it follows this curve of constant σ. This is suggestive that ensemble sampling can readily explore problems when the extended parameter space is strongly correlated with the initial parameter space.

In Fig. 6, we show trace plots for μ, α, γ, and σ from the second step of hybrid sampling. At each iteration, we show the current state of the β_T ensemble. The color at each iteration matches the colors in Fig. 5. The dashed lines indicate the true value of each parameter. We see that in this case, the ensemble evolves from the initial state to the stationary distribution containing the true value within <100 iterations.

IV. GRAVITATIONAL WAVE SOURCE MODEL

A. Modeling gravitational waveforms from black hole mergers

To infer the properties of the source of a gravitational-wave signal, we require a model for the gravitational waveform. A quasicircular binary black hole (BBH) merger can be described by 15 source parameters, divided between eight "intrinsic" parameters (the masses and spins of the component black holes) and seven "extrinsic" parameters (the location and orientation of the source with respect to an observer). As these parameters describe the signal predicted by general relativity, we refer to these as "GR parameters," later denoted $\theta_{\rm GR}$. When modeling the signal from a binary black hole merger, the coalescence is typically broken down into three temporally distinct regimes: the *inspiral*, an *intermediate* phase, and the *merger ringdown* [16,34,55].

The inspiral begins when the black holes have formed a binary system; however, we typically only model the waveform after the emission has surpassed the lowest sensitivity frequency of our instruments (typically 20 Hz for current detectors). This regime is typically characterized

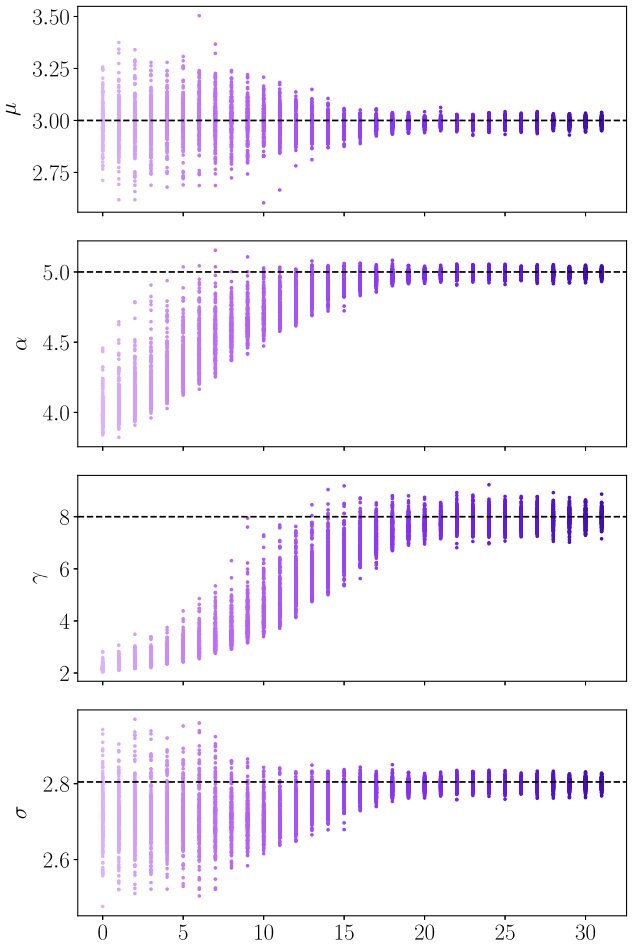

FIG. 6. Trace plots showing the evolution of samples taken in μ, α, γ, and σ during the second step of hybrid sampling for the first 32 steps of sampling, with black lines denoting the true values of these parameters. The samples plotted at each iteration of sampling are collated from each of the 200 walkers in the ensemble at temperature $\beta_T = 1$. The color scheme matches the state of the ensemble shown in Fig. 5. Throughout sampling, we observe that γ and α achieve the correct values while not biasing our correct estimate of μ from the first sampling step. Additionally, we see that the convergence of σ follows the convergence of the ensemble state in Fig. 5 around a line of constant σ.

using the post-Newtonian expansion [56] with higher-order corrections tuned to numerical relativity simulations. During the intermediate regime, the orbital frequency of the binary increases to a point where the post-Newtonian expansion breaks down and the binary "plunges" and the horizons merge. This regime can only be accurately described through numerical methods and is usually modeled using a fit to numerical relativity simulations (for example, public catalogs like [57]). Finally, after the merger, the remnant black hole undergoes a "ringdown" phase, in which gravitational-wave emission is modeled via quasinormal modes [58,59]. This regime is well described by analytical models and provides strong tests of the "no-hair" theorem [60] and the black hole area law [61].

In this work, we use IMRPHENOMPV2, a computationally efficient, phenomenological model of the gravitational waveform [55,62–64]. For a given set of BBH source parameters, IMRPHENOMPV2 returns a frequency-domain representation of the gravitational wave signal, taking the form

$$\tilde{h}(f) = A(f)e^{-i\Psi(f)}, \tag{15}$$

where $\tilde{h}(f)$ is the gravitational-wave strain as a function of frequency, $A(f)$ is the amplitude, and Ψ is the phase of the signal. Both A and Ψ depend on the intrinsic and extrinsic parameters of the BBH, although in general, the intrinsic parameters have a larger impact on the phase, while the extrinsic parameters primarily determine the amplitude. In this work, we focus on modifications to the phase Ψ as the first test of our hybrid sampling method for gravitational-wave signals, as current detectors are more sensitive to the phase of the signal [17].

During the inspiral regime, Ψ is approximated as a modified version of the post-Newtonian expansion:

$$\Psi_{\rm ins}(f) = 2\pi f t_c - \phi_c - \frac{\pi}{4} + \frac{3}{128\eta v^5}\sum_{i=0}^{7}(\varphi_i + \varphi_{iL}\ln v)v^i + \frac{1}{\eta}\left(\sigma_0 + \sigma_1 f + \frac{3}{4}\sigma_2 f^{4/3} + \frac{3}{5}\sigma_3 f^{5/3} + \frac{1}{2}\sigma_4 f^2\right). \tag{16}$$

Here $\eta = m_1 m_2/(m_1+m_2)^2$ is the symmetric mass ratio of the binary, $v = (\pi M f G c^{-3})^{1/3}$ is the dimensionless orbital frequency of the system, the phase coefficients φ_i are determined by the post-Newtonian expansion and the σ_j are tuned to numerical relativity waveforms. Terms φ_{iL} are those post-Newtonian coefficients leading $\ln v$ at order i. Both φ_i and σ_j depend on the intrinsic parameters of the source. The parameters t_c and ϕ_c are the coalescence time and the orbital phase at coalescence respectively.

In the intermediate phase, IMRPHENOMPV2 adopts the following form for Ψ,

$$\Psi_{\rm int}(f) = \frac{1}{\eta}\left(\beta_0 + \beta_1 f + \beta_2 \log(f) - \frac{\beta_3}{3} f^{-3}\right), \tag{17}$$

where β_0 and β_1 are chosen to require a smooth continuation of Ψ in the coalescence time and phase from the inspiral to intermediate phases. The parameters β_2 and β_3 depend on the intrinsic parameters of the source.

Finally, in the merger-ringdown phase, IMRPHENOMPV2 adopts another parametrized form for Ψ,

$$\Psi_{\rm MR} = \frac{1}{\eta}\left[\alpha_0 + \alpha_1 f - \alpha_2 f^{-1} + \frac{4}{3}\alpha_3 f^{3/4} + \alpha_4 \tan^{-1}\left(\frac{f - f_{\rm RD}}{f_{\rm damp}}\right)\right]. \quad (18)$$

As with the intermediate phase, α_0 and α_1 are chosen so that Ψ continues smoothly from the intermediate phase to the merger-ringdown phase, and α_{2-4} depend on the intrinsic parameters of the source. The frequencies $f_{\rm RD}$ and $f_{\rm damp}$ describe the complex ringdown frequency and are computed from the mass and spin of the remnant black hole [64]. We note that the above discussion applies to the aligned-spin IMRPHENOMD model; the IMRPHENOMPV2 phasing is obtained by "twisting-up" the IMRPHENOMD phasing to account for precession of the orbital plane as described in [64].

B. Generating beyond-GR waveforms

Following [12,15,16,34], we model deviations from general relativity as fractional corrections to the parameters described above, specifically, we define

$$p^{\rm BGR} = (1 + \delta p) p^{\rm GR} \quad (19)$$

for $p^{\rm GR} \in \{\varphi_{1-4}, \varphi_{5L,6L}, \varphi_{6,7}, \alpha_{2-4}, \beta_{2,3}\}$. Since under GR $\varphi_1 = 0$, we model $\delta\varphi_1$ as an absolute rather than fractional deviation. We note that the parameters describing global phase and time shifts are not modified as any such modification is indistinguishable from changes in phase or time. In principle, any combination of these parameters could have nonzero deviations, however, most previous analyses modify just a single parameter at a time. We use the implementation of IMRPHENOMPV2 provided by LALSUITE [65], which allows us to apply these fractional deviations when generating our waveforms.

Although these fractional deviations can take any real value, we do not expect them to be large as at worst we expect the general relativistic description of gravitational wave emission to be "mostly" correct. Thus, we limit the space of allowed deviations by limiting the allowed deviation between a beyond-GR waveform we generate with $p_{i,\rm BGR}$ and the associated waveform generated with $p_{i,\rm GR}$. We measure this deviation between waveforms via the overlap $\mathcal{O}$,

$$\mathcal{O} = \max_{\phi_c} \frac{\langle \tilde{h}_{\rm GR}(f), \tilde{h}_{\rm BGR}(f)\rangle}{\sqrt{\langle \tilde{h}_{\rm GR}(f), \tilde{h}_{\rm GR}(f)\rangle \langle \tilde{h}_{\rm BGR}(f), \tilde{h}_{\rm BGR}(f)\rangle}}. \quad (20)$$

Here $\tilde{h}_{\rm GR}$ and $\tilde{h}_{\rm BGR}$ are the GR frequency-domain waveform and associated beyond-GR waveform with the same intrinsic parameters. We maximize over the merger phase of the signal by taking the absolute value of the overlap (e.g., [66]). One can similarly maximize over the merger time. However, as detailed in Appendix D, we find that an overlap cut maximized over the merger phase and time introduces sufficient flexibility that the GR parameters can deviate significantly from the corresponding value with no beyond-GR deviation. Finally, $\langle \cdot, \cdot \rangle$ denotes a discrete inner product between the frequency-domain waveforms, weighted by the detector spectral power density, as

$$\langle \tilde{h}_1(f), \tilde{h}_2(f)\rangle = \frac{4}{T}\sum_i^N \frac{\tilde{h}_{1,i}\tilde{h}^*_{2,i}}{S_i}, \quad (21)$$

between two generic frequency-domain waveforms $\tilde{h}_1$ and $\tilde{h}_2$, where i enumerates N discrete sampling frequencies spaced by $1/T$. In practice, we use the $+$ polarization of the waveform for computing the overlaps. The quantity S is the harmonic sum of the power spectral densities for each of the interferometers in the network. In this work, for some cases of hybrid sampling, we enforce a cut on the priors of δp_i by enforcing that all beyond-GR waveforms we generate must have an overlap $\mathcal{O} > 0.9$ with their associated GR waveform. This manifests in practice as a cut on the prior bounds of the initial points provided to PTEMCEE, as well as an added acceptance condition for MCMC proposals, where any proposal with $\mathcal{O} < 0.9$ is rejected.

V. HYBRID SAMPLING IN GRAVITATIONAL WAVE SIGNALS

We now apply our method to real and simulated gravitational-wave signals. We follow the procedure described in Sec. II to jointly infer $\theta_{\rm GR}$ and each of the δp parameters. For each analyzed signal, we first analyze the data using DYNESTY under the GR model.[3] Unless otherwise specified, we then perform 28 subsequent analyses with PTEMCEE, two each allowing one of the δp parameters to vary either applying the condition that $\mathcal{O} > 0.9$ or no overlap cut. For all analyses, we numerically marginalize the likelihood over distance and the coalescence phase using standard methods [37]. Full details of the sampler configurations can be found in Appendix B.

The prior distribution we use for $\theta_{\rm GR}$ is given in Table II. We note that throughout we work with detector-frame mass quantities which differ from the source mass by a distance-dependent factor due to cosmological redshifting. For the PTEMCEE stage, we initialize the $\theta_{\rm GR}$ from the tempered posterior distribution obtained with DYNESTY. The prior and initialization distributions for the δp are shown in Table III.

[3]We note that, in practice, this analysis is typically performed by the LIGO/Virgo/Kagra collaboration and so would not be required in production scenarios if the nested samples are released for future analyses.

TABLE II. Prior distributions for θ_{GR} used in both steps of hybrid sampling to estimate the source properties of the gravitational-wave signals we consider. We denote a uniform distribution over $[a, b]$ as $\mathcal{U}(a, b)$, $\mathcal{P}(\alpha, a, b)$ is a power-law distribution with spectral index α over the same domain. The sine distribution for a quantity x is equivalent to a uniform distribution of $\cos(x)$. The notation $[a, b]$ denotes a parameter that is constrained to lie within that interval with the functional form defined in terms of other parameters. The prior for the coalescence time is centered on either the trigger time from the matched filter search pipelines for GW150194 or the known injection time for simulated signals. Parameter definitions follow [47].

Parameter	Distribution	Unit
m_1, m_2	$\mathcal{U}(1, 1000)$	$M_\odot$
$\mathcal{M}$	[21.418, 41.974]	$M_\odot$
q	[0.05, 1.0]	–
a_1, a_2	$\mathcal{U}(0, 0.99)$	–
$\theta_1, \theta_2, \theta_{\rm JN}, \kappa$	Sin	rad
$\phi_{12}, \phi_{jl}, \phi_c, \epsilon$	$\mathcal{U}(0, 2\pi)$	rad
ψ	$\mathcal{U}(0, \pi)$	rad
t_c	$\mathcal{U}(t_0 - 0.1, t_0 + 0.1)$	s
d_L	$\mathcal{P}(2, 10, 10^4)$	Mpc

TABLE III. Prior (center) and initialization (right) distributions for the post-Newtonian deviation parameters δp_i used in the PTEMCEE step of hybrid sampling for GW150914. The prior distributions were chosen to fully include the δp_i posteriors for GW150914 in [12]. The initialization distributions were chosen to be narrower than the expected posterior distributions. Here, $\mathcal{U}(a, b)$ denotes a uniform distribution in $[a, b]$ and $\mathcal{N}(\mu, \sigma)$ a normal distribution with mean μ and standard deviation σ.

Parameter	Prior	Initialization
$\delta\varphi_0$	$\mathcal{U}(-1, 1)$	$\mathcal{N}(0, 10^{-2})$
$\delta\varphi_1$	$\mathcal{U}(-2, 2)$	$\mathcal{N}(0, 10^{-1})$
$\delta\varphi_2, \delta\varphi_3, \delta\varphi_4, \delta\varphi_{5l}$	$\mathcal{U}(-5, 5)$	$\mathcal{N}(0, 1)$
$\delta\varphi_6$	$\mathcal{U}(-10, 10)$	$\mathcal{N}(0, 1)$
$\delta\varphi_{6l}, \delta\varphi_7$	$\mathcal{U}(-30, 30)$	$\mathcal{N}(0, 5)$
$\delta\alpha_2, \delta\alpha_3, \delta\alpha_4$	$\mathcal{U}(-5, 5)$	$\mathcal{N}(0, 1)$
$\delta\beta_2, \delta\beta_3$	$\mathcal{U}(-5, 5)$	$\mathcal{N}(0, 1)$

A. Analysis of a real signal—GW150914

First, we apply our hybrid sampling method on GW150914, the first observed gravitational-wave signal [67]. This signal was produced by the coalescence of a binary black hole system with a detector-frame chirp mass of $\mathcal{M} \sim 30M_\odot$ and a network signal-to-noise ratio of ∼25. These properties mean it is still one of the highest SNR signals to date and also lies at the mode of the observed binary black hole mass distribution [68] making it an excellent representative test case. Following [67], we analyze $8s$ of data ending $2s$ after the trigger time produced by matched-filter search pipelines for both of the Advanced LIGO interferometers. We use the power spectral densities and calibration envelopes used in the LIGO/Virgo collaboration analyses available at [69]. Marginalizing over uncertainty in the detector calibration adds 40 free parameters to the analysis, and we use the same prior distribution for those parameters as [70]. We downsample the data to 2048 Hz and analyze the data from 20–1024 Hz.

In Fig. 7, we show the posterior distributions for each of the δp obtained with (purple) and without (magenta) a cut in the GR vs non-GR overlap, respectively. We also overlay the results from the LIGO/Virgo collaboration analysis in blue obtained with LALINFERENCE [15,44]. The differences between the blue and magenta are likely due to sampler differences.

We note that for the inspiral deviation parameters the requirement that $\mathcal{O} > 0.9$ imposes a significant constraint compared to the constraining power of the data. This is because the inspiral deviation parameters are strongly degenerate with the chirp mass, and also show correlations with the mass ratio as can be seen in Fig. 8. However, for the $\delta\alpha$ and $\delta\beta$ parameters, the posteriors are unaffected by the requirement that $\mathcal{O} > 0.9$. This is because these parameters are not strongly correlated with the GR parameters and so an equivalent waveform cannot be obtained by changing, e.g., the black hole masses and $\delta\alpha_2$.

In Fig. 8, we show joint posterior distributions on the beyond-GR deviation parameters $\delta\varphi_2$, intrinsic parameters chirp mass $\mathcal{M}$ and mass ratio q, and the extrinsic sky parameters right ascension and declination from our estimation of $\delta\varphi_2$ when enforcing an overlap cut of $\mathcal{O} > 0.9$ (purple) as well as enforcing no overlap cut (magenta). We also compare these distributions to the posteriors in $\mathcal{M}$, q, right ascension, and declination generated during the first step of hybrid sampling, where we do not yet sample in deviations from general relativity. From the construction of the post-Newtonian inspiral phase coefficients, we expect deviations from φ_2 to be correlated with changes in the mass parameters, particularly $\mathcal{M}$, and we can observe this correlation in both results. Since the extrinsic parameters do not affect the phase evolution of the signal, we do not expect a correlation between $\delta\varphi_2$ and the extrinsic parameters. As expected, we do not see a correlation between $\delta\varphi_2$ and the extrinsic sky parameters. We also observe the effect of the overlap cut, which prevents our ensemble from exploring far away from the GR solution for the mass parameters.

In Fig. 9, we examine the evolution of the ensemble sampler for our analysis allowing $\delta\varphi_2$ to vary with no minimum allowed overlap. We show the distribution of $\mathcal{M}$ and $\delta\varphi_2$ at various iterations of the PTEMCEE analysis. As in Sec. III B, the hybrid analysis method is well able to capture the correlation between chirp mass and the new parameter added in the second stage of our hybrid analysis. We find that, by iteration 1024, the ensemble has converged to the correct solution. In Appendix C, we provide

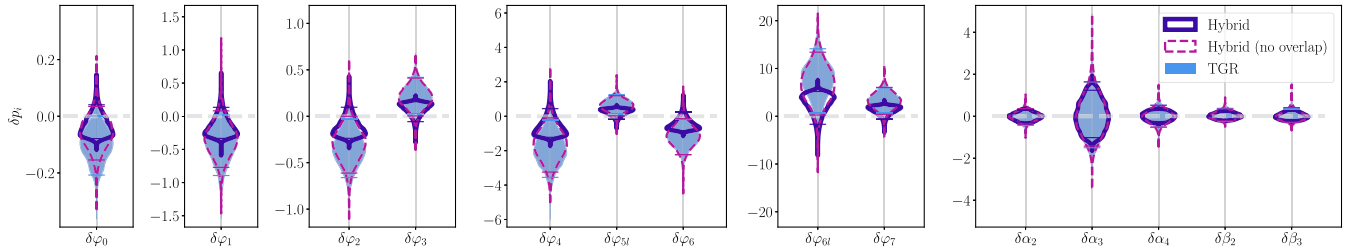

FIG. 7. Violinplot, showing the posterior distributions on each post-Newtonian deviation parameter δp_i, comparing the results of [15] with the results of our hybrid sampling method in the "overlap cut" (purple, $\mathcal{O} > 0.9$) and "no overlap cut" (magenta) cases. These results were obtained through 14 independent analyses, in each case, we vary only one deviation parameter at a time. Colored horizontal bars denote the 5th and 95th percentiles of the posteriors.

additional plots showing the evolution of the $\beta_T = 1$ ensemble for each of the analyses of GW150914. In general, we find ∼1000 iterations are sufficient to ensure convergence of the algorithm.

We now assess whether our hybrid sampling method is more computationally efficient than the previously employed direct sampling method. To do this, we compare the number of likelihood evaluations needed to produce well-converged results. The computational cost for hybrid sampling scales linearly with the number of extensions to the base model. A fixed number of likelihood evaluations are necessary for the first step of sampling with DYNESTY, followed by additional evaluations for each second step analysis performed with PTEMCEE. Using DYNESTY alone in a "standard" methodology also scales linearly without an

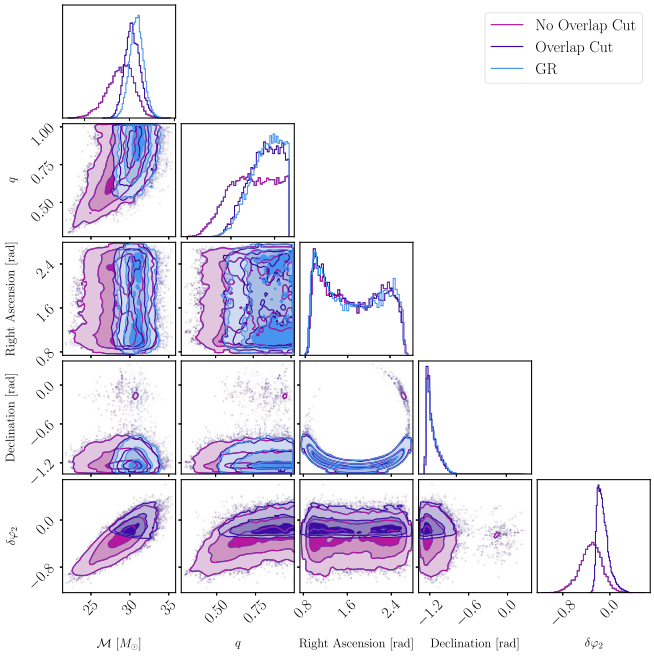

FIG. 8. Joint posterior distributions on chirp mass $\mathcal{M}$, mass ratio q, right ascension, declination, and deviation $\delta\varphi_2$ during our estimation of $\delta\varphi_2$ in GW150914 with hybrid sampling. In magenta, we plot the posteriors with no overlap cut enforced, whereas in purple we enforce an overlap cut of $\mathcal{O} > 0.9$. In blue, we show posteriors for $\mathcal{M}$ and q from the first step of hybrid sampling, with no deviations from general relativity. In both hybrid results, we see a correlation between $\delta\varphi_2$ and the intrinsic parameters, particularly $\mathcal{M}$, and no correlation between $\delta\varphi_2$ and the extrinsic sky parameters. We also observe the prior boundary imposed by the overlap cut that prevents our sampler from exploring values of the mass parameters more distant from the initial, general relativity-only result.

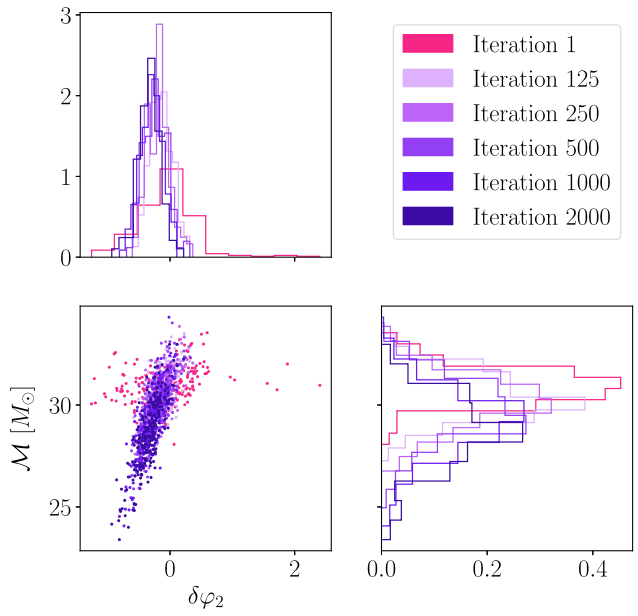

FIG. 9. Snapshots of the $\beta_T = 1$ ensemble of our parallel-tempered ensemble MCMC analysis at various iterations for our analysis of GW150914 allowing the inspiral deviation parameter $\delta\varphi_2$ to vary. We display chirp mass $\mathcal{M}$ against inspiral phase deviation coefficient $\delta\varphi_2$ with marginal distributions for $\mathcal{M}$ in the right column and those for $\delta\varphi_2$ in the top row. In pink, we show the state of the ensemble after the first MCMC step, near its initialization from the posterior generated with DYNESTY. As the ensemble evolves, represented with darkening shades of purple, the posterior expands to fill the extended posterior space. We note that the analysis correctly captures the expected correlation between the two parameters. In this analysis, we do not apply any condition on the overlap between the beyond-GR waveform and the corresponding GR waveform.

TABLE IV. The number of likelihood evaluations required to estimate δp_i in GW150914 using hybrid sampling. For reference, we include the number of likelihood evaluations required for the initial GR analysis. With the (optimistic) assumption that performing nested sampling to infer the δp_i requires the same number of likelihood evaluations as with the GR model, our method is ∼8× more efficient.

Parameter	$n_{\text{likelihood}}$
GR	23,200,000
$\delta\varphi_0$	2,955,000
$\delta\varphi_1$	2,952,500
$\delta\varphi_2$	2,952,500
$\delta\varphi_3$	2,952,500
$\delta\varphi_4$	2,911,250
$\delta\varphi_{5l}$	2,951,250
$\delta\varphi_6$	2,953,750
$\delta\varphi_{6l}$	2,951,250
$\delta\varphi_7$	2,490,000
$\delta\alpha_2$	2,951,250
$\delta\alpha_3$	2,952,500
$\delta\alpha_4$	2,951,250
$\delta\beta_2$	2,951,250
$\delta\beta_3$	2,952,500

initial fixed cost but, in general, each analysis with DYNESTY is more expensive than the same second-step hybrid analysis. Thus, if we only seek to estimate a small number of δp_i, using DYNESTY alone may be more efficient, but we expect hybrid sampling to be more efficient after some break-even number of deviation parameter estimations.

We summarize the computational cost of each of the analyses we performed for our analysis of GW150914 in Table IV. For the initial GR-only inference we required 23.6 million likelihood evaluations and each PTEMCEE analysis required <3 million likelihood evaluations. We do not have access to the computational cost for the LIGO/Virgo analysis, however, we can conservatively estimate that direct DYNESTY sampling for each non-GR parameter will be at least as expensive as the GR-only analysis. Additionally, in Appendix A, we directly sample in the GR parameters plus one non-GR parameter using PTEMCEE initialized at the maximum likelihood point, as well as with random samples from the prior, and find that these analyses had not converged after >6 million likelihood evaluations. We can therefore estimate that our hybrid sampling scheme is between ∼2 and ∼10× more efficient than the direct MCMC sampling method for this event.

B. Simulated non-GR signals

Analyses of real gravitational-wave transients have not revealed significant deviations from relativity, however, it is important to test whether our method will be sensitive to such effects if they are present. To accomplish this, we analyze a simulated signal with a nonzero value of $\delta\varphi_2$ with our hybrid method; the specific injection parameters are described in Table V. We add this signal to the Advanced LIGO Livingston and Hanford interferometers assuming their design sensitivities [71] resulting in an injection with a network signal-to-noise ratio SNR ≈ 370.

TABLE V. Parameters of the simulated signal injected into the Advanced LIGO gravitational wave detector network.

Parameter	Value	Unit
$\mathcal{M}$	30	$M_\odot$
q	0.8	⋯
a_1, a_2	0	⋯
$\theta_1, \theta_2, \phi_{12}, \phi_{jl}, \theta_{\text{JN}}, \phi_c, \psi$	0	rad
right ascension	1.35	rad
declination	−1.21	rad
$\delta\varphi_2$	0.2	⋯
other δp	0	⋯
t_c	0	s
d_L	100	Mpc

TABLE VI. Sampler arguments for hybrid sampling used in our analysis of injected signals as defined for the BILBY implementations of DYNESTY and PTEMCEE. For the DYNESTY-only analyses of injected signals, we also use DYNESTY sampler settings in this table. The variables in this table must be lower case and are shown here upper case to satisfy journal policy.

Sampling Argument	Value
DYNESTY	
NLIVE	500
SAMPLE	"RWALK"
WALKS	50
NACT	10
PTEMCEE	
NTEMPS	5
NWALKERS	250
BURN_IN_FIXED_DISCARD	2000

We follow the same hybrid sampling procedure as in our analysis of GW150914, and with the sampler settings found in Appendix B, Table VI. We also perform the beyond-GR analyses with DYNESTY without imposing an overlap cut to compare the results between the two methods.

In Fig. 10 we show the one- and two-dimensional marginal posterior probability distributions for three parameters for this simulated signal. From left to right (top to bottom) these are a non-GR deviation parameter $\delta\varphi_2$ and two intrinsic binary parameters, the chirp mass and mass ratio. We note again that the deviation parameter is correlated with the intrinsic parameters.

In Fig. 11, we consider snapshots of the $\beta_T = 1$ ensemble at various stages of the PTEMCEE analysis. For this analysis, the injected chirp mass is strongly excluded from the posterior distribution obtained after the first,

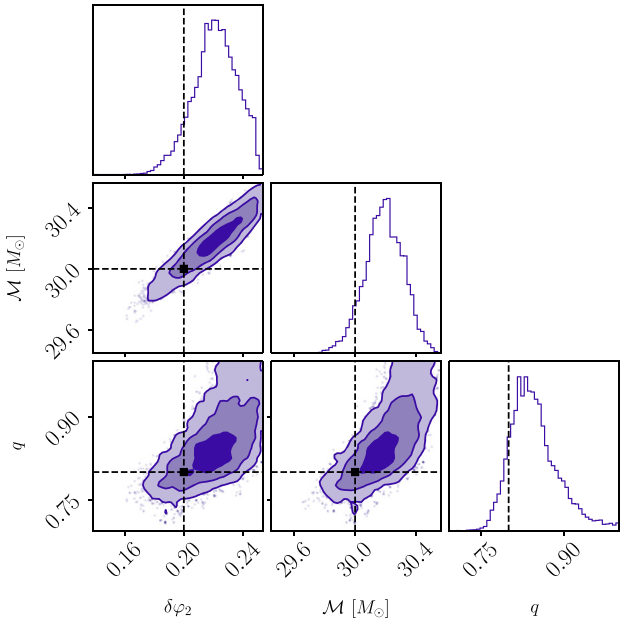

FIG. 10. Corner plot, displaying marginal and joint posterior distributions on the inspiral regime deviation parameter $\delta\varphi_2$, chirp mass $\mathcal{M}$, and mass ratio q from our injected signal generated by hybrid sampling with an overlap cut of $\mathcal{O} > 0.9$. As in Fig. 8, we note that $\delta\varphi_2$ is correlated with the intrinsic mass parameters.

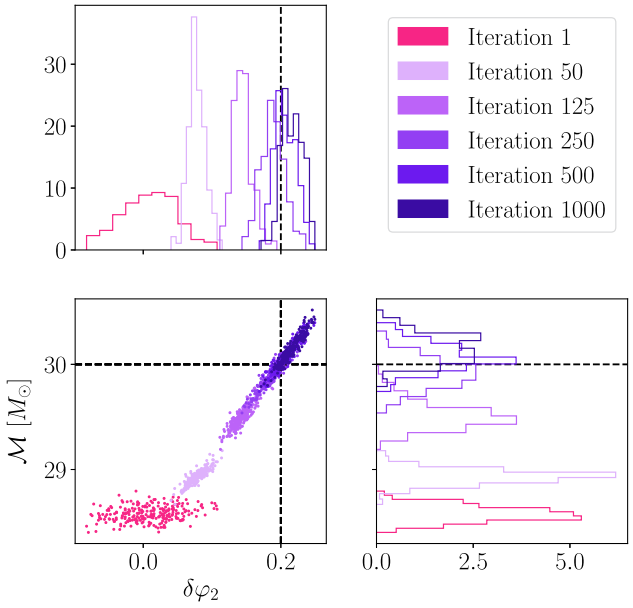

FIG. 11. We plot the evolution of the second step of our hybrid analysis of the injected signal, with an overlap cut of $\mathcal{O} > 0.9$ imposed on waveform generation. We display chirp mass $\mathcal{M}$ against inspiral phase deviation coefficient $\delta\varphi_2$ with marginal distributions for $\mathcal{M}$ in the right column and those for $\delta\varphi_2$ in the top row. In pink, we show the state of the ensemble after the first MCMC step, near its initialization from the posterior generated with DYNESTY. As the ensemble evolves, shown with darkening shades of purple, it evolves a tight correlation between $\mathcal{M}$ and $\delta\varphi_2$ at a roughly constant overlap. The ensemble converges to correct values for $\delta\varphi_2$ and $\mathcal{M}$, inconsistent with the initial estimate of $\mathcal{M}$ from the first step of hybrid sampling.

GR-only, analysis (pink) due to correlations between $\mathcal{M}$ and $\delta\varphi_2$. As in Fig. 5, the ensemble of walkers evolves to explore the extended parameter space and converge on the correct solution. As with our analysis of GW150914, ~1000 iterations are required until the ensemble converges.

In Fig. 12, we show the posteriors for δp_i for our simulated signal. Despite the deviation only being nonzero for $\delta\varphi_2$, the posterior distributions for all of the inspiral and intermediate deviation parameters are inconsistent with zero at high significance. For the merger-phase deviation parameters, the deviations from zero are less pronounced. This is consistent with previous work that has demonstrated that deviations at one post-Newtonian order can be identified with other deviation parameters [36,72] due to correlations between the parameters [73].

Additionally, in Fig. 13, we note that the posterior distributions for chirp mass we obtain while estimating δp_i are *only* consistent with the injected value when allowing

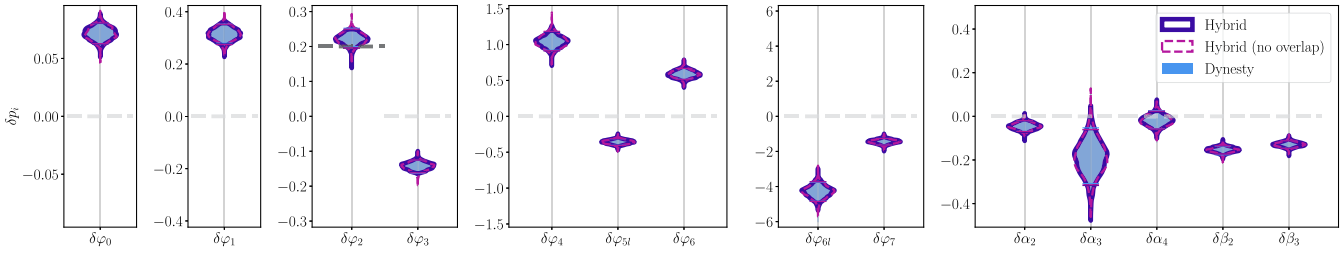

FIG. 12. Violinplot, showing the posterior distributions on each post-Newtonian deviation parameter δp_i for our injected signal, generated by hybrid sampling in the "no overlap cut" and "overlap cut" cases, with an additional solid-color posterior generated by using DYNESTY alone to check our results. Colored horizontal bars denote the 5th and 95th quantiles of the posteriors. In light gray, the injected value of $\delta p_i = 0$ is noted on each posterior with a dashed line, with a dark gray dashed line denoting the injected value of $\delta\varphi_2 = 0.2$. We observe that our hybrid sampling analysis agrees with DYNESTY-only analyses in all cases. Further, we observe that posteriors for $\delta\varphi_2$ generated by both methods are consistent with the injected value of 0.2, but those posteriors for other parameters are incorrectly inconsistent with 0.

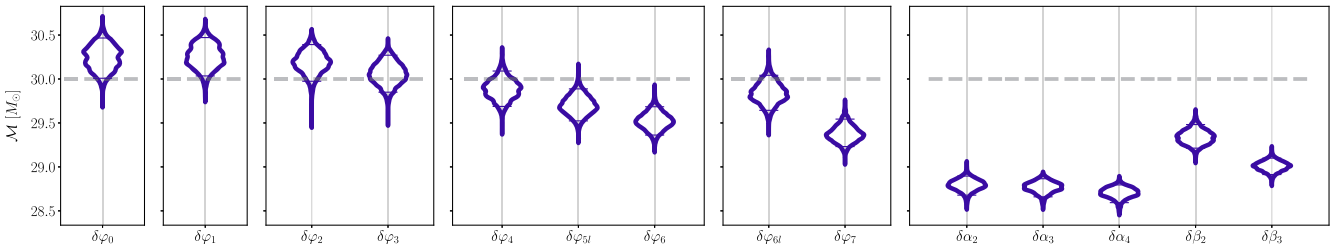

FIG. 13. Violinplot, showing the posterior distributions on chirp mass generated while estimating each post-Newtonian deviation parameter δp_i for our injected signal, generated by hybrid sampling in the "overlap cut" case. Colored horizontal bars denote the 5th and 95th quantiles of the posteriors. In dark gray, the injected value of $\mathcal{M} = 30M_\odot$ is noted on each posterior with a dashed line. We observe that our posterior distributions are *only* consistent with the injected value when varying inspiral phase coefficients φ_i, which follows from their definition as degenerate with the mass of the system. This indicates that, if we were to receive a signal whose phase evolution disagreed with that predicted by general relativity, then we would require a waveform model that admits deviations in δp_i degenerate with the mass of the system.

one of the inspiral phase deviation coefficients $\delta\varphi_i$ to vary. In general, corrections at similar post-Newtonian orders are more strongly correlated, and this is visible from our results. Thus, if we receive a signal whose phase evolution is inconsistent with general relativity we cannot trust our estimate of the chirp mass and require a model with additional degrees of freedom to capture the mass term accurately.

Comparing the number of likelihood evaluations for each analysis, we find that each DYNESTY analysis requires $\sim 10^7$ likelihood evaluations and each PTEMCEE analysis requires $\sim 3 \times 10^6$ likelihood evaluations. For each DYNESTY analysis we use only 500 live points in this case, compared to 2000 for our analysis of GW150914 and so we expect the DYNESTY analysis to require a factor of four fewer likelihood evaluations. Taking this into account, we see a comparable (or even larger) computational saving with our hybrid method as for our analysis of GW150914.

VI. CONCLUSIONS

In this work, we introduced a novel hybrid sampling method for exploring models that can be described as extensions of a simpler underlying model. By seeding a parallel-tempered ensemble MCMC with initial posterior estimates generated by performing nested sampling on a base model, hybrid sampling efficiently explores the extended parameter space of a more complex model. While previous methods have employed similar hybrid sampling methods, e.g., [74,75], we exploit the athermal property of the nested sampling algorithm to optimally seed the ensembles of walkers at each temperature.

First, we demonstrated our framework with a toy model, using hybrid sampling to estimate the parameters of a generalized Gaussian distribution. We saw that we are able to successfully recover the parameters of the true model, even when the base model is misspecified and the parameters of the extended model are correlated with those of the base model.

Following this, we applied our method to a widely performed test of general relativity with gravitational-wave transients: parametrized deviations from the waveform predicted by general relativity. Using our method, we accurately reproduced the tests of general relativity using GW150914 as performed by the LIGO/Virgo scientific collaborations and estimate that our method is approximately an order of magnitude more efficient than the current direct sampling method [12]. Finally, we analyzed a simulated signal with a measurable deviation from the prediction of relativity. We found that the efficiency of our hybrid sampling method is still far superior to direct sampling in this case.

Previous analyses have suffered from large computational costs as the parameters describing the waveform predicted by relativity are strongly correlated with the deviation parameters. In order to mitigate this, we introduced a "closeness" criterion between the non-GR waveform being considered and the corresponding GR signal. Specifically, this is implemented as a minimum overlap threshold between the two signals. This acts as an additional prior constraint that the signal must be similar to the GR prediction, given the previous success of relativity. This is particularly beneficial for lower signal-to-noise ratio systems where the data are less informative.

For the signals that we analyzed, we determined their consistency with GR by visual inspection of the marginal posteriors of each GR-deviation parameter returned by our hybrid analysis. Although beyond the scope of this work, one quantitative test of the consistency of our results with GR is Bayesian model selection, wherein one would compare the evidences assuming no deviations from GR and allowing a deviation from GR. The evidence from hybrid sampling is the evidence associated with the posterior generated by PTEMCEE; this code computes the evidence via thermodynamic integration [76,77] of the mean log-likelihood of each tempered chain [49]. In this

work, we only used five temperatures; however, an accurate calculation of the evidence would likely require more temperatures. With accurate estimation of the evidences, one could compute a Bayes factor between the GR and beyond-GR waveform models.

While we have focused on a narrow application of measuring single additional parameters describing deviations from relativity, the method presented here can be used for more exploratory analyses that allow multiple non-GR parameters simultaneously that otherwise have exploding computational costs due to the number of possible combinations of parameters to vary simultaneously. More generically, this method can be applied to any case where importance sampling to include a more physically realistic, but expensive model breaks down. For example, measuring eccentricity in compact binary mergers [78], estimating the impact of calibration uncertainty on inference [79], and analyzing pairs of potentially gravitationally lensed events [80].

ACKNOWLEDGMENTS

We thank Sylvia Biscoveanu, Max Isi, Nathan Johnson-McDaniel, Ralph Smith, Salvatore Vitale, and Alan Weinstein for helpful discussions and comments. C. T. is supported by an MKI Kavli Fellowship. J. G. is supported by Grants No. PHY-1764464 and No. 2207758. N. W. acknowledges support from the National Science Foundation (NSF) and the Park Scholarships program at NC State. We are grateful to the LIGO Caltech SURF program where this project began which is supported by the NSF REU program. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. The authors are grateful for computing resources provided by the California Institute of Technology and supported by National Science Foundation Grants No. PHY-0757058 and No. PHY-0823459. The analysis in this work made use of data available from the Gravitational Wave Open Science Center [81]. This analysis used the following software: NUMPY [82], SCIPY [83], MATPLOTLIB [84], CORNER.PY [85], PANDAS [86,87], LALSIMULATION [65], BILBY [45,47], DYNESTY [46], PTEMCEE [49]. We provide analysis scripts, notebooks, and some data.

APPENDIX A: COMPARISON OF INITIALIZATION METHODS

Hybrid sampling uses a posterior generated via nested sampling to initialize a set of tempered-ensembles of MCMC walkers, however, it is also possible to initialize MCMC walkers near the parameters that yield the maximum likelihood. Here, we repeat the analysis of GW150914 allowing $\delta\varphi_2$ to vary with no overlap cut, using PTEMCEE initialized with two common methods of MCMC initialization. In the "prior" method, we initialize our ensembles with random samples drawn from the prior distributions detailed in Tables II and III. In the "maximum likelihood" method, we initialize our ensembles near the maximum likelihood point, which we compute here as the peak of the likelihood function in our GR-only analysis of GW150914 using DYNESTY. For the initial values of $\delta\varphi_2$ in this run, we sample from a narrow truncated Gaussian distribution centered on zero. With both methods, we again employed 250 walkers at five temperatures, as in our hybrid analysis of GW150914.

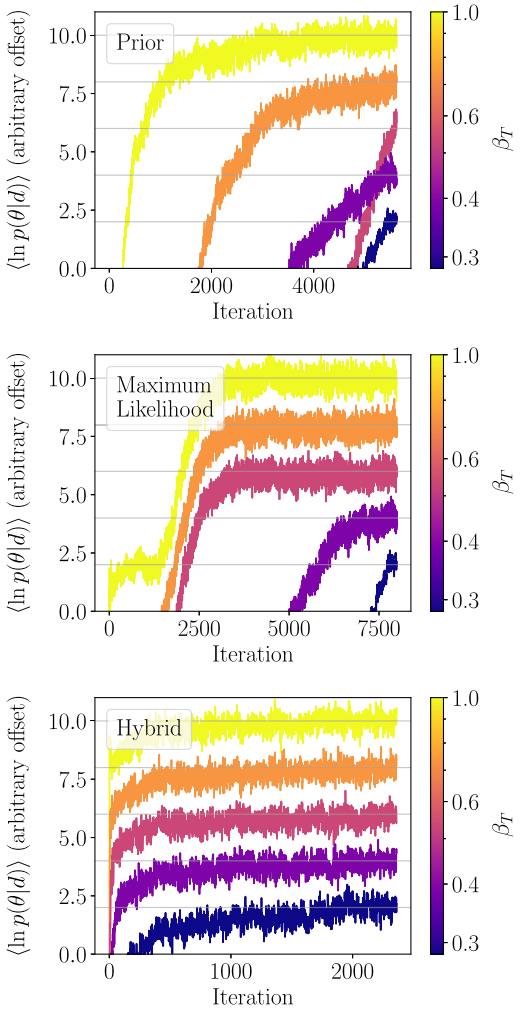

FIG. 14. The logarithm of the posterior probability averaged over all walkers at each temperature, $\langle \ln p(\theta|d) \rangle$, offset by an arbitrary value to simultaneously show $\langle \ln p(\theta|d) \rangle$ at each temperature. From left to right, we show $\langle \ln p(\theta|d) \rangle$ for the tempered-ensembles initialized with random samples from the prior, near the maximum likelihood point, and with our hybrid initialization method. We observe that only the hybrid initialized ensembles achieve convergence across all temperatures.

So that the lowest-temperature $\beta_T = 1$ ensemble may more efficiently explore the entire target distribution, PTEMCEE proposes swaps between ensembles of different temperatures throughout their evolution. While technically unlikely, this means in principle that swaps can occur

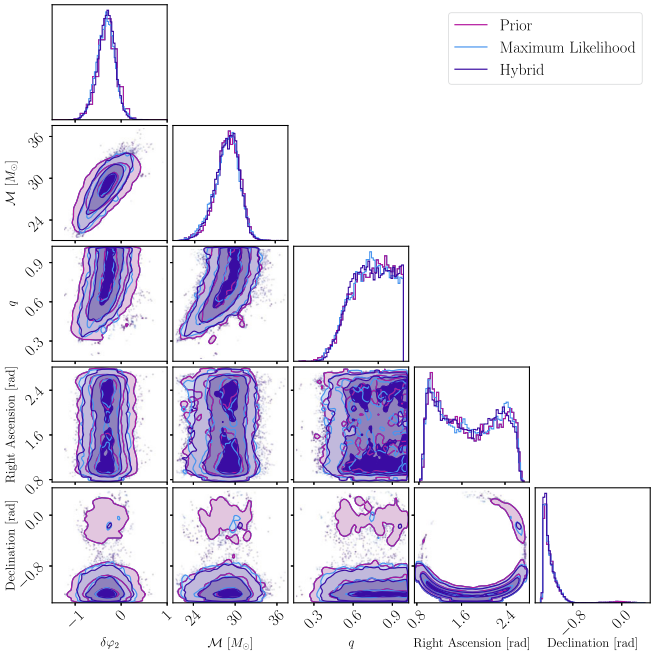

FIG. 15. The posterior distributions generated via PTEMCEE initialized with random samples from the prior (pink), near the maximum likelihood point (blue), and with our hybrid initialization method (purple). We observe that the results obtained with each initialization method are consistent with one another, however, the "prior" posterior distribution does not appear converged, consistent with the leftmost panel of Fig. 14.

between relatively hot and cold ensembles. Therefore, to formally consider a set of tempered ensembles converged, the ensemble at each temperature must be converged, or else swaps between ensembles of different temperatures do not satisfy detailed balance (see Ref. [49] and references therein for additional discussion). In Fig. 14, we show the mean logarithm of the posterior probability density, $\langle \ln p(\theta|d) \rangle$, at each MCMC iteration and each temperature. Generally, if $\langle \ln p(\theta|d) \rangle$ for a particular ensemble appears to be in a steady state, we expect that ensemble to have converged. With this perspective in mind, Fig. 14 indicates that the high-temperature (low β_T) ensembles initialized from the prior distribution or near the maximum likelihood point have failed to converge in over 4000 iterations. In half that time, the ensembles initialized with our hybrid methodology have converged at every temperature.

The convergence of these sets of tempered ensembles has direct consequences for the posterior distributions that they yield. In Fig. 15, we compare the posterior distributions generated with each method for initializing PTEMCEE. While the results obtained with each method are generally consistent, the ensembles initialized from the prior (pink) appear not converged with respect to the hybrid initialization (purple), particularly when looking at the marginal posteriors on declination. This particular result can be explained by the uniform priors adopted for the sky position parameters, as the MCMC ensembles initialized from the prior generally had a much larger distance to travel across the likelihood surface compared to the ensembles initialized with our hybrid method, or even near the maximum likelihood point. In total, we can conclude that our method yields results consistent with these standard methods of MCMC initialization while achieving convergence of the entire set of tempered MCMC ensembles in less than half the number of iterations.

APPENDIX B: SAMPLER SETTINGS

To enable reproducibility of our results, we provide the settings used during each stage of our hybrid sampling algorithm. These are listed in Table VI. All of the parameters are as defined in the BILBY implementation of the respective sampling code. Additionally, configuration files can be found at Ref. [88]. We note that for our analysis of GW150914, we used NLIVE $= 2000$ rather than 500 as for the simulated signals.

APPENDIX C: FURTHER RESULTS FOR GW150914 ANALYSIS

In this appendix, we provide trace plots showing the evolution of the $\beta_T = 1$ ensemble for the deviation parameters δp for our analysis of GW150914. In Fig. 16, we show the results of analyses without (left) and with (right) the requirement that $\mathcal{O} \geq 0.9$, respectively. In general, the sampler has converged to a steady-state after $\sim$1000 iterations and always after 2000 iterations. We note that in most cases implementing our overlap condition reduces the number of iterations required for the ensemble to converge to a steady state.

APPENDIX D: EFFECT OF TIME-MAXIMIZED OVERLAP CUT

In Sec. IV B, we introduced the overlap $\mathcal{O}$ to measure the deviation in a waveform induced by a beyond-GR deviation, maximized over the merger phase of the signal. Changing the merger time t_c introduces a frequency-dependent shift in the phase of the signal that is degenerate with a beyond-GR deviation; thus, some parametric tests of general relativity maximize t_c as well when calculating the overlap (see, for example, [24]). In Fig. 17, we present posterior distributions on the chirp mass, mass ratio, and the inspiral deviation parameter $\delta\varphi_2$ during our estimation of $\delta\varphi_2$ in GW150914, similar to the results presented in Fig. 8. Here, however, we have maximized $\mathcal{O}$ over both the merger phase and time. This reduces the cut on the prior for $\delta\varphi_2$ imposed by requiring $\mathcal{O} > 0.9$, as a larger range of deviations in the waveform induced by $\delta\varphi_2$ can be accounted for by varying the merger time. In turn, time maximization allows the mass parameters, in particular the chirp mass, to vary more widely as well, reducing the efficacy of an overlap cut.

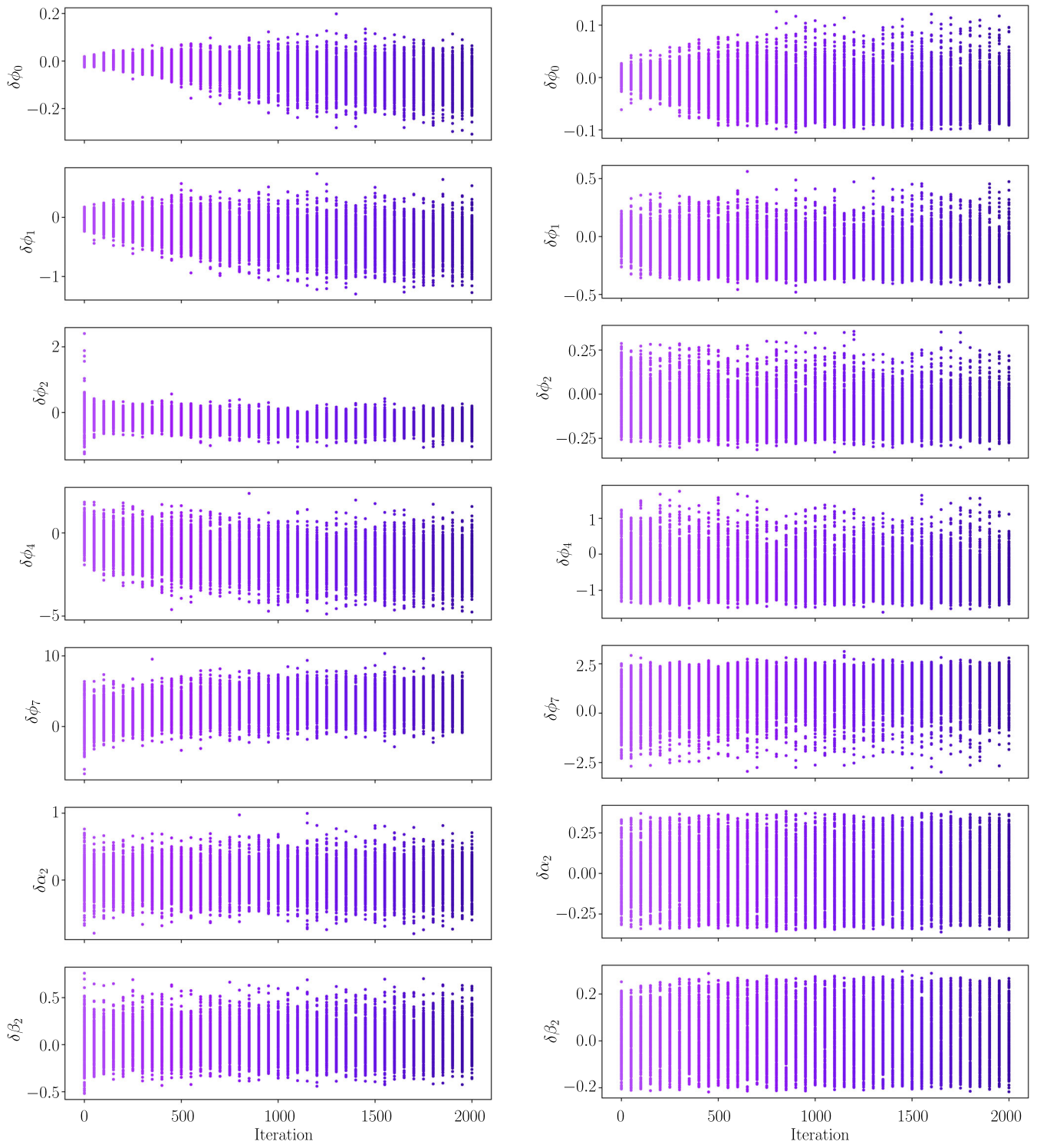

FIG. 16. Trace plots showing the evolution of samples taken in a subset of post-Newtonian deviation coefficients for GW150914 from both the inspiral ($\delta\varphi_0$, $\delta\varphi_1$, $\delta\varphi_2$, $\delta\varphi_4$, $\delta\varphi_7$) and postinspiral ($\delta\alpha_2$, $\delta\beta_2$) during the first 2000 steps of the second step of hybrid sampling. Traces in the left column are generated without any overlap cut, whereas traces in the right column are generated while $\mathcal{O} > 0.9$ is imposed. The samples plotted at each iteration of sampling are collated from each of the 250 walkers in the ensemble at temperature $\beta_T = 1$. The color scheme matches the state of the ensemble shown in Fig. 9. We observe that even in the most extreme case, when no overlap cut is applied, all ensembles converge within ∼1000 iterations, with many converging far sooner particularly when an overlap cut is applied.

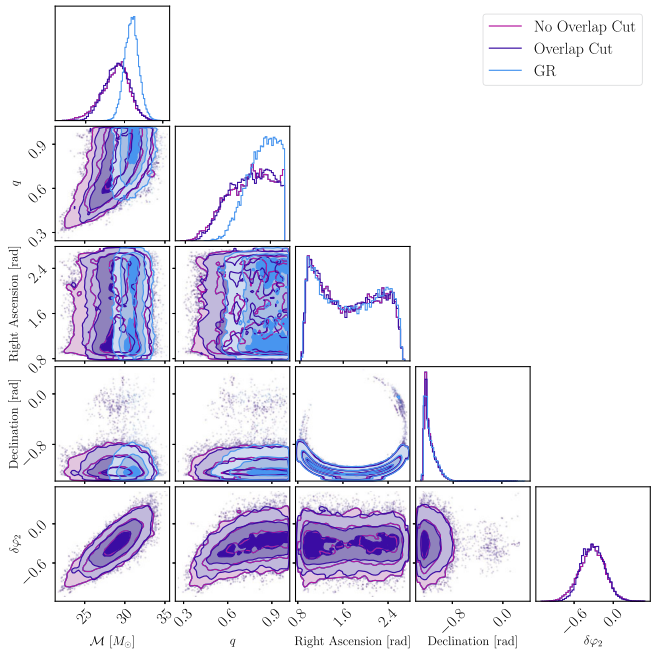

FIG. 17. Joint posterior distributions on chirp mass $\mathcal{M}$, mass ratio q, right ascension, declination, and deviation $\delta\varphi_2$ during our estimation of $\delta\varphi_2$ in GW150914 with hybrid sampling. In magenta, we plot the posteriors with no overlap cut enforced, whereas in purple we enforce an overlap cut of $\mathcal{O} > 0.9$ which has been maximized over both merger phase and time, in contrast to the results presented in Fig. 8. In blue, we show posteriors for $\mathcal{M}$ and q from the first step of hybrid sampling, with no deviations from general relativity. In both hybrid results, we see the expected correlation between $\delta\varphi_2$ and $\mathcal{M}$, and lack of correlation between $\delta\varphi_2$ and the extrinsic sky parameters. Although the hybrid result with an overlap cut is constrained relative to its counterpart without that cut, the difference between these posterior distributions is visually smaller than when $\mathcal{O}$ was only maximized over the merger phase.

[1] C. M. Will, The confrontation between general relativity and experiment, Living Rev. Relativity **17**, 4 (2014).

[2] I. H. Stairs, Testing general relativity with pulsar timing, Living Rev. Relativity **6**, 5 (2003).

[3] N. Wex, Testing relativistic gravity with radio pulsars, arXiv:1402.5594.

[4] T. Damour, 1974: The discovery of the first binary pulsar, Classical Quantum Gravity **32**, 124009 (2015).

[5] J. Zhao, P. C. C. Freire, M. Kramer, L. Shao, and N. Wex, Closing a spontaneous-scalarization window with binary pulsars, Classical Quantum Gravity **39**, 11LT01 (2022).

[6] P. G. Ferreira, Cosmological tests of gravity, Annu. Rev. Astron. Astrophys. **57**, 335 (2019).

[7] EHT Collaboration *et al.*, Gravitational Test Beyond the First Post-Newtonian Order with the Shadow of the M87 Black Hole, Phys. Rev. Lett. **125**, 141104 (2020).

[8] EHT Collaboration *et al.*, First sagittarius A* event horizon telescope results. VI. Testing the black hole metric, Astrophys. J. Lett. **930**, L17 (2022).

[9] EHT Collaboration *et al.*, First sagittarius A* event horizon telescope results. I. The shadow of the supermassive black hole in the center of the Milky Way, Astrophys. J. Lett. **930**, L12 (2022).

[10] LIGO Scientific Collaboration *et al.*, Advanced LIGO, Classical Quantum Gravity **32**, 074001 (2015).

[11] F. Acernese *et al.*, Advanced Virgo: A second-generation interferometric gravitational wave detector, Classical Quantum Gravity **32**, 024001 (2015).

[12] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Tests of General Relativity with GW150914, Phys. Rev. Lett. **116**, 221101 (2016).

[13] R. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Tests of General Relativity with GW170817, Phys. Rev. Lett. **123**, 011102 (2019).

[14] R. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Binary Black Hole Mergers in the First Advanced LIGO Observing Run, Phys. Rev. X **6**, 041015 (2016).

[15] R. Abbott *et al.* (The LIGO Scientific Collaboration and the Virgo Collaboration), Tests of general relativity with the

binary black hole signals from the LIGO-Virgo catalog GWTC-1, Phys. Rev. D **100**, 104036 (2019).
[16] R. a. Abbott, Tests of general relativity with binary black holes from the second LIGO-Virgo gravitational-wave transient catalog, Phys. Rev. D **103**, 122002 (2021).
[17] R. Abbott, H. Abe, F. Acernese, K. Ackley, N. Adhikari, R. Adhikari, V. Adkins, V. Adya, C. Affeldt, D. Agarwal *et al.*, Tests of general relativity with gwtc-3, arXiv:2112.06861 [Phys. Rev. D (to be published)].
[18] I. Quiros, Selected topics in scalar-tensor theories and beyond, Int. J. Mod. Phys. D **28**, 1930012–156 (2019).
[19] N. Yunes and F. Pretorius, Fundamental theoretical bias in gravitational wave astrophysics and the parametrized post-Einsteinian framework, Phys. Rev. D **80**, 122003 (2009).
[20] R. N. Lang, Compact binary systems in scalar-tensor gravity. II. Tensor gravitational waves to second post-Newtonian order, Phys. Rev. D **89**, 084014 (2014).
[21] R. N. Lang, Compact binary systems in scalar-tensor gravity. III. Scalar waves and energy flux, Phys. Rev. D **91**, 084027 (2015).
[22] N. Sennett, S. Marsat, and A. Buonanno, Gravitational waveforms in scalar-tensor gravity at 2PN relative order, Phys. Rev. D **94**, 084003 (2016).
[23] S. Tahura and K. Yagi, Parametrized post-Einsteinian gravitational waveforms in various modified theories of gravity, Phys. Rev. D **98**, 084042 (2018).
[24] G. S. Bonilla, P. Kumar, and S. A. Teukolsky, Modeling compact binary merger waveforms beyond general relativity, Phys. Rev. D **107**, 024015 (2023).
[25] B. Shiralilou, T. Hinderer, S. M. Nissanke, N. Ortiz, and H. Witek, Post-Newtonian gravitational and scalar waves in scalar-Gauss-Bonnet gravity, Classical Quantum Gravity **39**, 035002 (2022).
[26] E. Barausse, C. Palenzuela, M. Ponce, and L. Lehner, Neutron-star mergers in scalar-tensor theories of gravity, Phys. Rev. D **87**, 081506 (2013).
[27] M. Shibata, K. Taniguchi, H. Okawa, and A. Buonanno, Coalescence of binary neutron stars in a scalar-tensor theory of gravity, Phys. Rev. D **89**, 084005 (2014).
[28] M. Okounkova, L. C. Stein, M. A. Scheel, and D. A. Hemberger, Numerical binary black hole mergers in dynamical Chern-Simons gravity: Scalar field, Phys. Rev. D **96**, 044020 (2017).
[29] M. Okounkova, L. C. Stein, J. Moxon, M. A. Scheel, and S. A. Teukolsky, Numerical relativity simulation of GW150914 beyond general relativity, Phys. Rev. D **101**, 104016 (2020).
[30] M. Okounkova, Numerical relativity simulation of GW150914 in Einstein-dilaton-Gauss-Bonnet gravity, Phys. Rev. D **102**, 084046 (2020).
[31] R. Cayuso and L. Lehner, Nonlinear, noniterative treatment of EFT-motivated gravity, Phys. Rev. D **102**, 084008 (2020).
[32] W. E. East and J. L. Ripley, Evolution of Einstein-scalar-Gauss-Bonnet gravity using a modified harmonic formulation, Phys. Rev. D **103**, 044040 (2021).
[33] W. E. East and J. L. Ripley, Dynamics of Spontaneous Black Hole Scalarization and Mergers in Einstein-Scalar-Gauss-Bonnet Gravity, Phys. Rev. Lett. **127**, 101102 (2021).
[34] M. Agathos, W. Del Pozzo, T. G. F. Li, C. Van Den Broeck, J. Veitch, and S. Vitale, TIGER: A data analysis pipeline for testing the strong-field dynamics of general relativity with gravitational wave signals from coalescing compact binaries, Phys. Rev. D **89**, 082001 (2014).
[35] A. K. Mehta, A. Buonanno, R. Cotesta, A. Ghosh, N. Sennett, and J. Steinhoff, Tests of general relativity with gravitational-wave observations using a flexible–theory-independent method, Phys. Rev. D **107**, 044020 (2023).
[36] J. Meidam, K. W. Tsang, J. Goldstein, M. Agathos, A. Ghosh, C.-J. Haster, V. Raymond, A. Samajdar, P. Schmidt, R. Smith, K. Blackburn, W. Del Pozzo, S. E. Field, T. Li, M. Pürrer, C. Van Den Broeck, J. Veitch, and S. Vitale, Parametrized tests of the strong-field dynamics of general relativity using gravitational wave signals from coalescing binary black holes: Fast likelihood calculations and sensitivity of the method, Phys. Rev. D **97**, 044033 (2018).
[37] E. Thrane and C. Talbot, An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models, Pub. Astron. Soc. Aust. **36**, e010 (2019).
[38] J. D. Romano and N. J. Cornish, Detection methods for stochastic gravitational-wave backgrounds: A unified treatment, Living Rev. Relativity **20**, 2 (2017).
[39] T. B. Littenberg and N. J. Cornish, Bayesian inference for spectral estimation of gravitational wave detector noise, Phys. Rev. D **91**, 084034 (2015).
[40] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. **21**, 1087 (1953).
[41] W. K. Hastings, Monte Carlo sampling methods using markov chains and their applications, Biometrika **57**, 97 (1970).
[42] J. Skilling, Nested Sampling, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, American Institute of Physics Conference Series Vol. 735, edited by R. Fischer, R. Preuss, and U. V. Toussaint (American Institute of Physics, New York, 2004), pp. 395–405.
[43] J. Skilling, Nested sampling for general Bayesian computation, Bayesian Anal. **1**, 833 (2006).
[44] J. Veitch *et al.*, Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library, Phys. Rev. D **91**, 042003 (2015).
[45] G. Ashton *et al.*, BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy, Astrophys. J. Suppl. Ser. **241**, 27 (2019).
[46] J. S. Speagle, DYNESTY: A dynamic nested sampling package for estimating Bayesian posteriors and evidences, Mon. Not. R. Astron. Soc. **493**, 3132 (2020).
[47] I. M. Romero-Shaw *et al.*, Bayesian inference for compact binary coalescences with BILBY: validation and application to the first LIGO-Virgo gravitational-wave transient catalogue, Mon. Not. R. Astron. Soc. **499**, 3295 (2020).
[48] G. Ashton *et al.*, Nested sampling for physical scientists, Nature (London) **2**, 39 (2022).

[49] W. D. Vousden, W. M. Farr, and I. Mandel, Dynamic temperature selection for parallel tempering in Markov Chain Monte Carlo simulations, Mon. Not. R. Astron. Soc. **455**, 1919 (2015).
[50] J. Goodman and J. Weare, Ensemble samplers with affine invariance, Commun. Appl. Math. Comput. Sci. **5**, 65 (2010).
[51] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, emcee: The MCMC hammer, Publ. Astron. Soc. Pac. **125**, 306 (2013).
[52] D. W. Hogg and D. Foreman-Mackey, Data analysis recipes: Using Markov Chain Monte Carlo, Astrophys. J. Suppl. Ser. **236**, 11 (2018).
[53] R. H. Swendsen and J.-S. Wang, Replica Monte Carlo Simulation of Spin Glasses, Phys. Rev. Lett. **57**, 2607 (1986).
[54] E. Higson, W. Handley, M. Hobson, and A. Lasenby, Dynamic nested sampling: An improved algorithm for parameter estimation and evidence calculation, Stat. Comput. **29**, 891 (2019).
[55] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. J. Forteza, and A. Bohé, Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era, Phys. Rev. D **93**, 044007 (2016).
[56] L. Blanchet, Gravitational radiation from post-Newtonian sources and inspiralling compact binaries, Living Rev. Relativity **17**, 2 (2014).
[57] M. Boyle *et al.*, The SXS Collaboration catalog of binary black hole simulations, Classical Quantum Gravity **36**, 195006 (2019).
[58] E. Berti, V. Cardoso, and A. O. Starinets, TOPICAL REVIEW: Quasinormal modes of black holes and black branes, Classical Quantum Gravity **26**, 163001 (2009).
[59] M. Isi and W. M. Farr, Analyzing black-hole ringdowns, arXiv:2107.05609.
[60] M. Isi, M. Giesler, W. M. Farr, M. A. Scheel, and S. A. Teukolsky, Testing the No-Hair Theorem with GW150914, Phys. Rev. Lett. **123**, 111102 (2019).
[61] M. Isi, W. M. Farr, M. Giesler, M. A. Scheel, and S. A. Teukolsky, Testing the Black-Hole Area Law with GW150914, Phys. Rev. Lett. **127**, 011103 (2021).
[62] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. J. Forteza, and A. Bohé, Frequency-domain gravitational waves from nonprecessing black-hole binaries. I. New numerical waveforms and anatomy of the signal, Phys. Rev. D **93**, 044006 (2016).
[63] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Simple Model of Complete Precessing Black-Hole-Binary Gravitational Waveforms, Phys. Rev. Lett. **113**, 151101 (2014).
[64] A. Bohe, M. Hannam, S. Husa, F. Ohme, M. Puerrer, and P. Schmidt, PhenomPv2—Technical notes for the LAL implementation (2016), https://dcc.ligo.org/LIGO-T1500602/public.
[65] LIGO Scientific Collaboration, LIGO Algorithm Library—LALSuite, free software (GPL) (2018), 10.7935/GT1W-FZ16.
[66] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, FINDCHIRP: An algorithm for detection of gravitational waves from inspiraling compact binaries, Phys. Rev. D **85**, 122006 (2012).
[67] R. Abbott *et al.* (The LIGO Scientific Collaboration and the Virgo Collaboration), Properties of the Binary Black Hole Merger GW150914, Phys. Rev. Lett. **116**, 241102 (2016).
[68] The LIGO Scientific Collaboration, the Virgo Collaboration, and the KAGRA Collaboration *et al.*, Population of Merging Compact Binaries Inferred Using Gravitational Waves Through GWTC-3, Phys. Rev. X **13**, 011048 (2023).
[69] The LIGO/Virgo Collaboration, Parameter estimation sample release for gwtc-1, https://dcc.ligo.org/LIGO-P1800370/public.
[70] B. P. Abbott *et al.*, GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs, Phys. Rev. X **9**, 031040 (2019).
[71] B. P. Abbott *et al.*, Prospects for observing and localizing gravitational-wave transients with Advanced LIGO, Advanced Virgo and KAGRA, Living Rev. Relativity **21**, 3 (2018).
[72] C.-J. Haster, Pi from the sky—A null test of general relativity from a population of gravitational wave observations, arXiv:2005.05472.
[73] M. Saleem, S. Datta, K. G. Arun, and B. S. Sathyaprakash, Parametrized tests of post-Newtonian theory using principal component analysis, Phys. Rev. D **105**, 084062 (2022).
[74] M. C. Miller *et al.*, PSR J0030 + 0451 mass and radius from NICER data and implications for the properties of neutron star matter, Astrophys. J. Lett. **887**, L24 (2019).
[75] D. Psaltis, C. Talbot, E. Payne, and I. Mandel, Probing the black hole metric: Black hole shadows and binary black-hole inspirals, Phys. Rev. D **103**, 104036 (2021).
[76] P. M. Goggans and Y. Chi, Using thermodynamic integration to calculate the posterior probability in Bayesian model selection problems, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, American Institute of Physics Conference Series, Vol. 707, edited by G. J. Erickson and Y. Zhai (American Institute of Physics, New York, 2004), pp. 59–66.
[77] N. Lartillot and H. Philippe, Computing bayes factors using thermodynamic integration, Syst. Biol. **55**, 195 (2006).
[78] I. Romero-Shaw, P. D. Lasky, and E. Thrane, Signs of eccentricity in two gravitational-wave signals may indicate a subpopulation of dynamically assembled binary black holes, Astrophys. J. Lett. **921**, L31 (2021).
[79] E. Payne, C. Talbot, P. D. Lasky, E. Thrane, and J. S. Kissel, Gravitational-wave astronomy with a physical calibration model, Phys. Rev. D **102**, 122004 (2020).
[80] J. Janquart, O. A. Hannuksela, K. Haris, and C. Van Den Broeck, GOLUM: A fast and precise methodology to search for, and analyze, strongly lensed gravitational-wave events, arXiv:2203.06444.
[81] R. Abbott *et al.*, Open data from the first and second observing runs of Advanced LIGO and Advanced Virgo, SoftwareX **13**, 100658 (2021).
[82] C. R. Harris *et al.*, Array programming with NUMPY, Nature (London) **585**, 357 (2020).

[83] P. Virtanen *et al.* (SCIPY1.0 Contributors), SCIPY1.0: Fundamental algorithms for scientific computing in PYTHON, Nat. Methods **17**, 261 (2020).
[84] J. D. Hunter, MATPLOTLIB: A 2d graphics environment, Comput. Sci. Eng. **9**, 90 (2007).
[85] D. Foreman-Mackey, CORNER.PY: Scatterplot matrices in PYTHON, J. Open Source Softwaare **1**, 24 (2016).
[86] T. Pandas (Development Team), pandas-dev/pandas: PANDAS (2020), 10.5281/zenodo.3509134.
[87] Wes McKinney, Data Structures for Statistical Computing in PYTHON, in *Proceedings of the 9th PYTHON in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman (SciPy, Austin, Texas, 2010), pp. 56–61.
[88] https://github.com/noahewolfe/tgr-hybrid-sampling.