Rapid refitting techniques for Bayesian spectral characterization of the gravitational wave background using pulsar timing arrays

William G. Lamb,[1,*] Stephen R. Taylor,[1,†] and Rutger van Haasteren[2,‡]

[1]*Department of Physics and Astronomy, Vanderbilt University, 2301 Vanderbilt Place, Nashville, Tennessee 37235, USA*

[2]*Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut), Callinstrasse 38, D-30167 Hannover, Germany*

(Received 5 April 2023; revised 2 October 2023; accepted 4 October 2023; published 13 November 2023)

Pulsar timing arrays (PTAs) have recently found evidence for a nanohertz-frequency stochastic gravitational-wave background (SGWB). Constraining its spectral characteristics will reveal its origins. In order to achieve this, we must understand how data and modeling conditions in each pulsar influence the precision and accuracy of SGWB spectral recovery. These goals typically require many Bayesian analyses on real data sets and large-scale simulations that are slow and computationally taxing. To combat this, we have developed several new rapid approaches that instead operate on intermediate SGWB analysis products. These techniques refit SGWB spectral models to previously-computed Bayesian posterior estimates of the timing power spectra. We test our new techniques on simulated PTA data sets and the NANOGrav 12.5-year data set, where in the latter our refit posterior achieves a Hellinger distance—bounded between 0 for identical distributions and 1 for zero overlap—from the current full production-level pipeline that is $\lesssim 0.1$. Our techniques are $\sim 10^2$–10^4 times faster than the production-level likelihood, and scale much more favorably (sub-linearly) as a PTA is expanded with new pulsars or observations. Our techniques also allow us to demonstrate conclusively that SGWB spectral characterization in PTA data sets is driven by the longest-timed pulsars and the best-measured power spectral densities, which is not necessarily the case for SGWB detection that is predicated on correlating many pulsars. Indeed, the common-process spectral properties found in the NANOGrav 12.5-year data set are given by analyzing only the ∼14 longest-timed pulsars out of the full 45 pulsar array, and we find that the "shallowing" of the common-process power-law model occurs when gravitational-wave frequencies higher than ∼50 nanohertz are included. The implementation of our techniques is openly available as a software suite to allow fast and flexible PTA SGWB spectral characterization and model selection.

DOI: 10.1103/PhysRevD.108.103019

I. INTRODUCTION

Pulsar timing array (PTA) experiments [1] across the world have now reported compelling evidence for a nanohertz-frequency stochastic gravitational wave background (SGWB) [2–5]. This new insight into the gravitational wave (GW) spectrum was achieved by measuring small deviations between the expected and observed radio-pulse times-of-arrival (TOAs) from a set of Galactic millisecond pulsars, wherein the distinctive imprint of an SGWB is inferred through a quasiquadrupolar correlation signature imparted between pulsars in the PTA. This Hellings and Downs correlation pattern [6] has now been inferred with varying levels of significance by most regional PTA Collaborations, with the promise of higher significance and enhanced scientific returns when these are synthesized into an updated International Pulsar Timing Array dataset [7].

While PTA detection statistics are centered around the cross-correlation of distinct pulsars, it is an interesting consequence of the PTA data model that spectral characterization of the SGWB is dominated by pulsar auto-correlations [8,9]. In fact, multiple PTA Collaborations saw the first hints of the SGWB through emerging common spectral behavior in many pulsars, which was modeled as a common uncorrelated red noise (CURN) signal [10–12]. Even now, with evidence of GW-induced cross-correlations, SGWB spectral characteristics derived from a CURN data model provide an excellent approximation to a full Hellings and Downs–correlated model (HD), yet at a fraction of the computational cost. Adequately modeled, the shape of the inferred SGWB spectrum encodes information about the emitting source, e.g., the dynamics and demographics of a black-hole binary population, or the

*william.g.lamb@vanderbilt.edu
†stephen.r.taylor@vanderbilt.edu
‡rutger.v.haasteren@aei.mpg.de

details of early-Universe processes. Assuming that the source is a population of circular supermassive black hole binaries (SMBHBs) evolving via GW radiation reaction, the SGWB's characteristic strain h_c as a function of frequency follows a power law, $h_c(f) = A(f/f_{yr})^{\alpha}$, where $\alpha = -2/3$ [13]. The amplitude A is the characteristic strain referenced to a frequency of $f_{yr} = 1/\text{year}$, and determined by the demographics of the binary population, e.g., the number density of emitting systems per redshift, primary mass, and mass ratio e.g., [14] and references therein.

While measuring this amplitude parameter A can provide interesting constraints on the SMBHB population, the inward migratory dynamics of supermassive black holes after a galaxy merger are likely much more complicated ([15] and references therein). Binary orbital eccentricity and interactions of binaries with gas and stars (particularly at wider orbital separations) will attenuate the expected SGWB characteristic strain at lower frequencies [14,16]. This causes a deviation from a power law [17] and as such, the SGWB may carry information about the dynamical interactions of the SMBHB population within the final parsec of orbital evolution [18]. Even finiteness of the emitting population under the most simplified conditions above may cause spectral deviations from $f^{-2/3}$ [19–21], rendering fixed-α studies of limited utility for astrophysical inference.

Beyond SMBHBs, searches are underway for relic signatures of processes in the early Universe [22], e.g., cosmic strings [23], primordial gravitational waves [24], and cosmological phase transitions [25]. While it is thought that these signals are likely to be an additional, weaker contribution to the SMBHB signal, current searches can not yet arbitrate on the dominant contributing source of the SGWB. Kaiser *et al.* [26] investigated the separability of a SGWB signal into its component sources; a circular-SMBHB-population signal, and a background from primordial gravitational waves. Using simulated datasets developed by Pol *et al.* [9], they found that they could begin to distinguish two injected power-law spectra with 45 pulsars and 17 years of data, while after 20 years they could begin to characterize the subdominant power-law GWB signal. However, their subdominant injected GWB spectrum had a cosmological energy density that was still rather strong, comparable to upper limits on primordial gravitational waves (e.g., [24,27]).

The goal for SGWB spectral characterization is to be a bridge between pulsar timing data and the physics of these sources. Our guiding principle is for spectral characterization to be scalable and modular; testing a new spectral model should not need the analysis to be started from scratch back at the level of timing residuals, nor should adding a new pulsar to the PTA require us to ignore that the analysis has already been successfully performed on the other pulsars. With this being said, the current production-level pipelines do indeed start from scratch whenever a new model is tested or a pulsar is added. The current PTA data model is formulated in the time domain. Uneven observational sampling of the pulsars, and concerns over the potential for spectral leakage from windowing, renders fast searches directly in the Fourier domain impractical. The computational bottleneck of this time-domain Gaussian likelihood is the required inversion of the data covariance matrix containing the SGWB signal contributions. Elegant accelerations can be achieved simply by modeling low-frequency processes (like the SGWB or per-pulsar red noise) with only a small number of Fourier basis functions [28,29]. However, even with these accelerations and optimized sparse linear algebra routines, Bayesian SGWB parameter estimation with the PTA likelihood via Markov Chain Monte Carlo (MCMC) sampling can require several days to weeks of computation. This is the status quo, and will worsen as more pulsars are added, and further observations of existing pulsars are incorporated into datasets.

There is a tremendous need for robust, efficient, and flexible analysis methods for PTAs that follow our previously mentioned guiding principles of scalability and modularity. For example, high-cadence timing campaigns from telescopes such as CHIME [30] generate large data volumes that will slow current pipelines. More pressing is that the synthesis of all current regional PTA datasets will result in a combined IPTA dataset with more than 100 pulsars, which will tax existing pipelines. Significant acceleration of parameter estimation was achieved by Taylor *et al.* [31], who modeled the SGWB as a CURN, which thereby allows the PTA likelihood to be factorized into parallelized per-pulsar analyses (see e.g.,[7,10,11,32]). This factorized likelihood (FL) method shows excellent agreement with the full production-level PTA likelihood. Unfortunately, the FL method assumes a power-law model with a fixed spectral index, which limits its usefulness for spectral model selection and source inference. A more general approach would maintain the likelihood computational speed-up, parallelization over pulsars, and the intended modularity of this FL technique while permitting analyses of SGWB models with arbitrary spectral parametrizations.

In this paper, we introduce the aforementioned generalization of the FL approach, allowing for rapid SGWB spectral characterization under arbitrary parametrized models, rather than just a fixed-index power law. This is made possible by condensing the pulsar timing data down to what we call Bayesian periodograms; probability density reconstructions of the pulsar timing-residual power spectral density at each frequency. Models are then refit to combinations of these Bayesian periodograms. In Sec. II, we discuss current analysis methods before introducing our new analysis techniques. We present the results of our comparison tests between the current and new methods on simulated and real data in Sec. III, before sharing our conclusions and goals for further developments in Sec. IV.

II. METHODS

Here we describe current PTA data-analysis techniques as they pertain to SGWB spectral characterization, and discuss expected future limitations as PTA datasets expand. We then introduce the factorized likelihood (FL) approach [31], and the concept of refitting spectral models to Bayesian periodograms of PTA timing residuals.

A. Current spectral characterization methods

PTA analyses model pulsar timing residuals $\overrightarrow{\delta t}$ as the sum of a deterministic pulsar timing model and stochastic red and white noise components,

$$\overrightarrow{\delta t} = \mathbf{M}\vec{\epsilon} + \mathbf{F}\vec{a} + \vec{n}. \tag{1}$$

The $(N_{\rm TOA} \times m)$-shaped design matrix $\mathbf{M}$ is a matrix of partial derivatives of the TOAs with respect to m timing-ephemeris parameters evaluated at an initial fitting solution, with a vector of linear offsets from the initial fit $\vec{\epsilon}$. Red-noise processes, such as the common gravitational wave signal and red noise intrinsic to each pulsar, are modeled as a Fourier sum over N_f sampling frequencies such that, for the ith timing residual observed at time t_i,

$$[\mathbf{F}\vec{a}]_i = \sum_{k=1}^{N_f} \left\{ a_{s,k} \sin\left(\frac{2\pi k t_i}{T}\right) + a_{c,k} \cos\left(\frac{2\pi k t_i}{T}\right) \right\}, \tag{2}$$

where T is the timing baseline (typically the total time span of the dataset being analyzed). As such, $\mathbf{F}$ is a $N_{\rm TOA} \times 2N_f$ matrix of sines and cosines evaluated at observation times, and $\vec{a} = (a_{s,1}, a_{c,1}, a_{s,2}, a_{c,2}, \ldots, a_{s,N_f}, a_{c,N_f})^{\rm T}$ is a vector of Fourier coefficients. We model intrinsic red noise (IRN) as independent between pulsars, and the SGWB as a common signal to all pulsars. For a single pulsar p, its total red noise is the sum, $(\mathbf{F}\vec{a})_p = (\mathbf{F}\vec{a})_p^{\rm IRN} + (\mathbf{F}\vec{a})^{\rm SGWB}$. Finally, $\vec{n}$ is uncorrelated white noise due to radiometer noise, instrumental effects, and pulsar phase jitter. We rearrange Eq. (1) to model residual noise as $\vec{r}$,

$$\vec{r} = \overrightarrow{\delta t} - \mathbf{M}\vec{\epsilon} - \mathbf{F}\vec{a} = \overrightarrow{\delta t} - \mathbf{T}\vec{b}, \tag{3}$$

where $\mathbf{T} = [\mathbf{M}\,\mathbf{F}]$ and $\vec{b} = [\vec{\epsilon}\,\vec{a}]^{\rm T}$. Other contributions to the timing residuals include correlated white noise between TOAs within the same timing epoch, and radio-frequency dependent red noise due to time-dependent variation in dispersion from the interstellar medium (see e.g., [33,34]).

We place a zero-mean Gaussian prior on $\vec{b}$ such that, for model hyperparameters $\vec{\eta}$,

$$p(\vec{b}|\vec{\eta}) = \frac{\exp\left(-\frac{1}{2}\vec{b}^{\rm T}\mathbf{B}^{-1}\vec{b}\right)}{\sqrt{\det\left(2\pi\mathbf{B}\right)}}, \tag{4}$$

where $\mathbf{B} \equiv \mathbf{B}(\vec{\eta}) = \langle \vec{b}\vec{b}^{\rm T} \rangle = {\rm diag}(\infty, \phi)$. The matrix ϕ is the Fourier-domain covariance on red-noise processes, while the ∞-block effectively converts the Gaussian prior into a improper uniform prior on the timing model. Given that we will eventually only deal with the inverse of $\mathbf{B}$, we need not worry about the practicalities of treating infinities.

The full hierarchical likelihood of the timing residuals given the model hyperparameters and b-coefficients is given by $p(\overrightarrow{\delta t}|\vec{b},\vec{\eta}) = p(\overrightarrow{\delta t}|\vec{b}) \times p(\vec{b}|\vec{\eta})$. However, we are only interested in the model hyperparameters $\vec{\eta}$ that describe the statistical properties of various stochastic processes; thus we marginalize over the Gaussian b-coefficients to recover a more concise likelihood,

$$p(\overrightarrow{\delta t}|\vec{\eta}) = \frac{\exp\left(-\frac{1}{2}\overrightarrow{\delta t}^{\rm T}\mathbf{C}^{-1}\overrightarrow{\delta t}\right)}{\sqrt{\det\left(2\pi\mathbf{C}\right)}}. \tag{5}$$

Here, $\mathbf{C} = \mathbf{N} + \mathbf{T}\mathbf{B}\mathbf{T}^{\rm T}$ is the full time-domain covariance matrix of the data model, with white-noise covariance $\mathbf{N}$, where

$$[\mathbf{C}]_{(pi),(qj)} = [\mathbf{N}]_{p,(ij)}\delta_{pq}\delta_{ij} + [\mathbf{C}^{\rm IRN}]_{p,(ij)}\delta_{pq} + \Gamma_{pq}[\mathbf{C}^{\rm SGWB}]_{(ij)}. \tag{6}$$

Equation (6) indexes over pulsars (p, q) and TOAs (i, j). $[\mathbf{N}]_{p,(ij)}$ and $[\mathbf{C}^{\rm IRN}]_{p,(ij)}$ are the white noise and intrinsic red noise covariance matrix components respectively for pulsar p and ith TOA, while $[\mathbf{C}^{\rm SGWB}]_{(ij)}$ is the covariance matrix components for the SGWB between the ith and jth TOAs. The expected GW-induced cross-correlation in timing residuals between pulsars is given by the overlap reduction function (ORF) coefficient Γ_{pq}, which, for an isotropic SGWB is the aforementioned Hellings and Downs (HD) curve [6].

All current spectral characterization techniques involve computing the PTA likelihood in Eq. (5) under different models or assumptions [10,35]. When cross-correlations between pulsars are modeled (hereafter referred to as interpulsar correlations), inverting $\mathbf{C}$ should scale as $\mathcal{O}(N_p^3 N_b^3)$. As more TOAs and more pulsars are added to the array, evaluation of this likelihood will slow down significantly because of this scaling. The autocorrelation blocks in the PTA data covariance matrix contain white noise, pulsar-intrinsic red noise, and the SGWB, while the interpulsar blocks only feature the SGWB. However, we now know that spectral characterization of an SGWB is dominated by PTA autocorrelation information [8,9]. Therefore, for the class of techniques below where the PTA likelihood is factorized over pulsars, we assume no interpulsar correlations (i.e., a CURN model) such that $\Gamma_{pq} = \delta_{pq}$.

Modeling only the diagonal blocks of the PTA data covariance matrix reduces the likelihood evaluation scaling

to $\mathcal{O}(N_p N_b^3)$. Physically speaking, this significant acceleration arises because the PTA likelihood is factorized as a product over pulsars,

$$p(\{\vec{\delta t}\}|\vec{\eta}) = \frac{\exp(-\frac{1}{2}\sum_{p=1}^{N_p} \vec{\delta t}_p^{\mathrm{T}} C_{pp}^{-1} \vec{\delta t}_p)}{\sqrt{\det(2\pi C)}} = \prod_{p=1}^{N_p} \frac{\exp(-\frac{1}{2}\vec{\delta t}_p^{\mathrm{T}} C_{pp}^{-1} \vec{\delta t}_p)}{\sqrt{2\pi C_{pp}}} = \prod_{p=1}^{N_p} p(\vec{\delta t}_p|\vec{\eta}), \quad (7)$$

where $\{\vec{\delta t}\}$ where is the set of timing residuals for all pulsars, $p(\vec{\delta t}_p|\vec{\eta})$ is the likelihood for a single pulsar p with a set of timing residuals $\vec{\delta t}_p$, and $\vec{\eta}$ are model hyperparameters describing variables like spectral-shape parameters, etc. However, this factorization is not exploited to full effect within the production-level `enterprise` analysis pipeline [36], which carries this out as a serialized calculation over pulsars. Parallelizing the computation over N_p processors would theoretically remove the likelihood computation's dependence on the number of pulsars, while being numerically equivalent to an analysis that uses the production-level PTA likelihood.

B. Factorized likelihood methods

The factorized likelihood (FL) approach makes possible a class of techniques where Eq. (7) is computed in parallel across pulsars, with reweighted posterior distributions from each pulsar combined in postprocessing to calculate the likelihood for the array. Evaluation of the likelihood becomes approximately scale invariant with N_p. Taylor *et al.* [31] modeled a power-law SGWB strain spectrum with a fixed spectral index of $\alpha = -2/3$ to recover posteriors on the SGWB strain amplitude for each pulsar. The posteriors on the strain amplitude were represented by histograms, re-weighted by the single-pulsar parameter priors, then multiplied across pulsars with a suitable prior for the final posterior calculation.

This fixed-index FL technique (along with variants) has already been adopted as a new tool in large analysis campaigns from NANOGrav [2,10], the Parkes Pulsar Timing Array [4,11], and IPTA [7], as well as other studies [32,37], to accelerate parameter estimation and cross-validation. However as discussed earlier in Sec. I, there are many reasons why the SGWB strain spectrum could deviate from this simple fixed-index power-law model. We therefore require a more flexible and generalized FL approach that would allow for inference of physically motivated SGWB spectral models.

A general factorized likelihood (GFL) approach is possible by fitting spectral models to the free spectrum, a minimally modeled Bayesian spectral characterization of pulsar timing data [38,39]. The free spectrum recovers the joint posterior of the red-noise power spectrum at all sampling frequencies, parametrized by the coefficient ρ, such that

$$\rho_k^2 := \frac{\langle \vec{a}_k^{\mathrm{T}} \vec{a}_k \rangle}{T} = \frac{S(f_k)}{T}, \quad (8)$$

where k is the Fourier frequency-bin index and S is the power spectral density of the timing residuals induced by red processes. Typically, a free-spectrum analysis is conducted with a uniform prior on $\log_{10}\rho$, with posteriors jointly recovered at all sampling frequencies. In the following we assume that there is independence between sampling frequencies, thus no covariance between them. Pulsar-timing analyses deal with unevenly sampled observations, so we will assess the strength of this assumption in our tests.

Refitting spectral models to Bayesian free spectra can be done at various levels; one can (i) perform a PTA Bayesian free-spectrum analysis, followed by refitting on the frequency-factorized PTA free spectrum, or (ii) perform free-spectral analysis on individual pulsars, which are then combined into a frequency- and pulsar-factorized likelihood against which spectral models are fit. The general scheme for (i) is as follows. Translating h_c into ρ-space gives

$$\rho_k^2 = \frac{h_c(f_k)^2}{12\pi^2 f_k^3 T} = \frac{A^2}{12\pi^2 T}\left(\frac{f_k}{f_{1\ \mathrm{yr}^{-1}}}\right)^{-\gamma}, \quad (9)$$

where $\gamma = 3 - 2\alpha = 13/3$ for the idealized SMBHB population. We form a likelihood by computing the probability that a given model is supported by the free spectrum at each frequency,

$$p(\{\vec{\delta t}\}|\vec{\eta}) = \int \mathrm{d}\vec{\rho}\, p(\{\vec{\delta t}\}|\vec{\rho}) p(\vec{\rho}|\vec{\eta}) \propto \int \mathrm{d}\vec{\rho}\, \frac{p(\vec{\rho}|\{\vec{\delta t}\})}{p(\vec{\rho})} \times p(\vec{\rho}|\vec{\eta}) \approx \prod_{k=1}^{N_f} \int \mathrm{d}\rho_k \frac{p(\rho_k|\{\vec{\delta t}\})}{p(\rho_k)} \times p(\rho_k|\vec{\eta}), \quad (10)$$

where $p(\rho_k)$ is the prior probability of ρ_k in the free-spectrum analysis, $p(\rho_k|\{\vec{\delta t}\})$ is the marginal posterior probability density of ρ_k that is sampled using MCMC techniques, and $p(\rho_k|\vec{\eta})$ is the probability of ρ_k under a parametrized spectral model, such as Eq. (9). In all cases considered here, the spectral model maps precisely to a value of ρ at each frequency, in which case the integral in Eq. (10) is trivial. However, the more general form shown allows for models that have intrinsic spread, e.g., where there is an expected form of the spectrum due to a population of SMBHBs, and population finiteness induces departures in a given realization [14,15]. We note that in (i),

the PTA free spectrum need not necessarily use only autocorrelation information in the PTA likelihood; $p(\rho_k|\{\overrightarrow{\delta t}\})$ is not yet factorized over pulsars, hence an interpulsar-correlated free-spectral analysis may be performed that accounts for HD correlations, and this would still allow a frequency-factorized refitting analysis to be subsequently performed.

Finally, (ii), the extension to factorize the likelihood over pulsars simply requires that the right-hand side of Eq. (10) is modified to have an additional product over pulsars. However, by doing so, we must explicitly make the usual FL assumption of conditioning spectral characterization on the PTA autocorrelation information under a CURN model,

$$p(\{\overrightarrow{\delta t}\}|\vec{\eta}) \underset{\sim}{\propto} \prod_{p=1}^{N_p} \prod_{k=1}^{N_f} \int \mathrm{d}\rho_k \frac{p(\rho_k|\overrightarrow{\delta t}_p)}{p(\rho_k)} \times p(\rho_k|\vec{\eta}). \quad (11)$$

To compute probabilities of a spectral model with a given set of hyperparameters $\vec{\eta}$ under the free-spectral likelihoods $p(\{\overrightarrow{\delta t}\}|\rho_k)$, we use optimized density estimation with MCMC samples. There are already a number of examples in the literature of fitting SGWB spectra to a free spectrum of a PTA (e.g., [40–42]). It is favored over analyzing the full likelihood because it is fast, since the data structures that we are fitting to are no longer the timing residuals themselves, but a compressed-data representation in terms of a red process at each GW frequency. Other timing-residual contributions from the timing model, uncorrelated and correlated white noise, and interstellar-medium effects, are marginalized over.

The simplest density estimation technique is to bin our free-spectrum MCMC samples as histograms, just like in the FL method. This recreates the probability densities of the free-spectra posteriors within bins of $\log_{10}\rho$. To faithfully reconstruct the original distribution, an appropriate choice of bin width must be made. If the width of the bins is too large, the histogram will be oversmoothed, perhaps removing important fluctuations in the actual distribution. In contrast, if the bin width is too narrow, the histogram will undersmooth the data, creating a density reconstruction that captures all of the fluctuations in the data that are a result of statistical sampling randomness and not due to the underlying distribution. There are several standard rules-of-thumb for finding the optimal bin width for a histogram given some data, such as by using Scott's normal reference rule [43] or the Freedman-Diaconis rule [44], which are tuned for an underlying normal distribution. Unfortunately, histograms do not result in a continuous distribution from which probability densities can be extracted, causing some loss of data in between bins, particularly if the bin width is wide.

An alternative method is to use kernel density estimators (KDEs) [45,46]. A KDE recreates a distribution by replacing each sample with a normalized, symmetric, strictly positive, real-valued function called a kernel (also known as a window function). If samples $(x_1, x_2, \ldots, x_N)$ are extracted from an unknown distribution f, the density estimate $\hat{f}$ of a KDE is given by

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right), \quad (12)$$

where K is the kernel function, and h is the bandwidth of the KDE. As with histograms, an appropriate bandwidth must be chosen to avoid creating an under or oversmoothed estimator. The kernel function itself is also a choice to be decided. In this paper, we use an Epanechnikov kernel [47], and select bandwidths using the Sheather-Jones plug-in selector [48]. Further details on these choices, and KDEs in general, are given in Appendix A.

Some free-spectrum posteriors may be poorly constrained and show non-negligible support for $\log_{10}\rho$ down to their lower prior boundary. The corresponding likelihood would effectively be constant if the boundary were lowered to $-\infty$. To ensure accurate KDE reconstruction at the boundary, we mirror the data about the boundary to create the KDE, and then cut off the KDE at the boundary. Any proposed spectral model in our refitting scheme that goes below the boundary is given the same probability as spectra at the boundary.

C. Refit pipelines

We refit parametrized spectral models against these optimized KDE representations of PTA and pulsar-free spectra using MCMC techniques. A typical algorithm is as follows: (1) an iteration of the MCMC proposes a set of parameters for the spectral model, from which we calculate our $\log_{10}\rho$ coefficients at all GW sampling frequencies; (2) we then find the probability of these model $\log_{10}\rho$ values under the free-spectrum likelihoods at each frequency—and, if applicable, for each pulsar—given our KDEs,[1] and take their product to compute the total likelihood. The employed Metropolis-Hastings algorithm will then reject or accept those parameters accordingly. We repeat this until the MCMC has sufficiently sampled the parameter space of the spectral model and converged to the target posterior.

In this paper, we explore two possible types of refits:

[1]KDE objects are memory intensive, and extracting the probability density function of a point from every KDE object at each MCMC iteration would slow down computation. However, KDEs are continuous, therefore before conducting the MCMC, we extract an array of probabilities along a grid of $\log_{10}\rho$ that is intentionally finer than the KDE bandwidth. This allows us to implement `numpy` vectorization techniques to accelerate the computation of the likelihood. When a set of $\log_{10}\rho$ is calculated, we look up its probability within the precalculated KDE density array across frequencies (and pulsars, where relevant).

(i) PTA free-spectrum refit: This involves refitting spectral models against the PTA free spectrum, which requires an initial analysis using the full PTA likelihood as a one-time cost. The PTA free-spectrum analysis describes each pulsar with a timing model, white noise, and power-law intrinsic red noise, with a free-spectrum common process across the entire array. Note that the model of interpulsar correlations for this common process can be CURN (uncorrelated), HD (SGWB-correlated), or other, since this refit technique involves only a factorization over frequencies. While we can refit SGWB spectral models to different numbers of frequencies with the PTA free spectrum, we cannot refit using different combinations of pulsars without recomputing the PTA free-spectrum.

(ii) Generalized factorized likelihood (GFL) Lite: In preparation for our goal of a complete generalization of the factorized likelihood technique, we introduce and study an intermediate analysis approach here called GFL Lite. Each pulsar is analyzed independently in parallel, with a model composed of a timing model, white noise, power-law intrinsic red noise, and a free-spectrum that acts as a proxy for the common process in each pulsar. Given the implicit factorization of the PTA likelihood over pulsars, GFL Lite assumes CURN as an interpulsar correlation model. We then refit a common spectral model to the free spectra of an ensemble of all (or a subset of) pulsars to recreate the PTA common process. This method allows us to fit a common signal to different combinations of pulsars and frequencies quickly. However, the per-pulsar intrinsic red noise model cannot be refit. This method is labeled as 'Lite' because the full GFL technique will also be capable of refitting per-pulsar intrinsic red noise models. Plans and prospects for full GFL are discussed in Sec. IV.

D. Pipeline profiling

In addition to these techniques being modular and flexible, we are also motivated by the prospects of significantly accelerating spectral characterization of the SGWB with PTA data, especially where many repeated studies and simulations are required. As discussed in Sec. II A, the full PTA likelihood with a CURN model should scale $\propto N_p$ because of the required inversion of a block-diagonal PTA data covariance matrix, and should scale as $\propto N_p^3$ for the full likelihood with an HD model because of the additional off-diagonal structure of the data covariance matrix.

Before carrying out a suite of simulations to compare the accuracy of our refit parameter estimation with the full PTA likelihood, we profiled our analyses on a simulated N_p-pulsar PTA dataset that contains an injected SGWB signal (as detailed later in Sec. III A). The timing profiles are shown in Fig. 1. For a simulated 45-pulsar dataset, the mean-likelihood evaluation time for the CURN full likelihood was 0.012 seconds, and 0.24 seconds for the HD full likelihood. The CURN full likelihood scales as expected with the number of pulsars. However the HD full likelihood scales $\propto N_p^2$, rather than the expected $\propto N_p^3$. As explained in Johnson *et al.* [49], this is due to the use of sparse matrix algebra. The exact scaling depends on details such as memory transfer, sparse matrix representation transforms, parallel computation across CPU cores, and matrix layout, all of which differ depending on the exact PTA analysis that is performed. But empirically, this typically results in a $\propto N_p^2$ dependence. The 45-pulsar PTA free-spectrum refit likelihood takes 53 microseconds while the GFL Lite likelihood takes 88 microseconds. These are 226 and 136 times faster than the CURN full likelihood respectively. The GFL Lite likelihood evaluation is sublinear as the number of pulsars increases. The PTA free spectrum is the fastest; however, a new free spectrum must be produced if we wish to change the number of pulsars in the array.

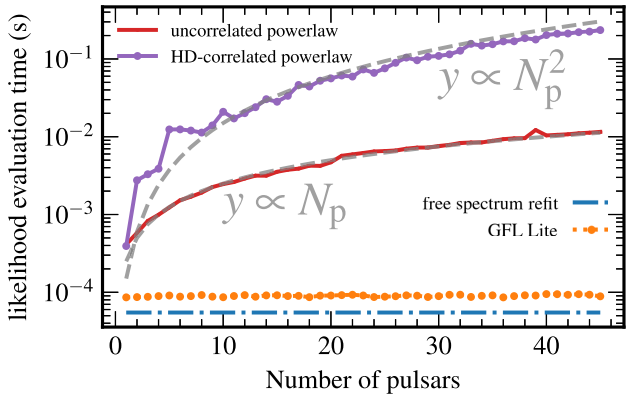

FIG. 1. The likelihood evaluation time as a function of the number of pulsars on a simulated dataset, ran on an AMD EPYC 7702 64-core processor. The CURN (red) and HD-correlated (purple) full PTA likelihoods scale $\propto N_p$ and $\propto N_p^2$ respectively. GFL Lite (orange) is scale independent with the number of pulsars. The PTA free-spectrum refit (dashed blue) is the most rapid method, being 10^2–10^4 times faster than the CURN and HD-correlated full PTA likelihoods.

III. RESULTS

We present the results of our analyses of 100 simulated PTA datasets that contain injected SGWB signals, comparing the performance of the full PTA likelihood to our refit techniques. In each analysis, we model intrinsic red noise as a 10-frequency power law in each pulsar in addition to a 10 frequency power-law common process, unless otherwise specified. These frequencies are linearly spaced from $1/T$ to $10/T$, where T is the total observing time of the array. We assess the ability of the PTA free-spectrum refit and GFL Lite techniques to recover SGWB parameter

posteriors that are comparable to the full likelihood, and investigate tolerance factors.

To quantify the difference between the posteriors recovered by our techniques compared to the full likelihood, we use the Hellinger distance [50], a measure of the similarity between two probability distributions. The Hellinger distance is bounded $0 \leq H \leq 1$, where $H = 0$ implies that distributions are identical, while $H = 1$ implies that they do not have any overlap and are completely different distributions. For our refit techniques to be robust and accurate, we seek Hellinger distances to be low with respect to results from using the full likelihood. See Appendix B for more details and guiding values for interpretation. We compare Hellinger distances between the 2D posteriors, as well as the 1D marginalized posteriors for each parameter in a power-law spectral model for the SGWB signal, γ and $\log_{10} A$, as defined in Eq. (9).

A. Simulations

Our simulated dataset creation follows Pol *et al.* [9]. The pulsar datasets are based on the observational timestamps and TOA uncertainties from the 45 pulsars of the NANOGrav 12.5 year dataset [10]. We extended the timespan of the dataset by drawing new TOAs and uncertainties from the distributions of the final year of each pulsar's observations to form a 15 year dataset. However, we kept the number of pulsars fixed, rather than adding new ones over time. We injected intrinsic red noise in each pulsar at linearly-spaced frequencies of $1/T$ to $10/T$, where $T = 15$ years. The injected spectral characteristics of a pulsar's intrinsic red noise were based on measured values taken from a CURN search in the NANOGrav 12.5 year dataset.

Finally, 100 SGWB signal realizations were injected into 100 copies of our simulated PTA dataset. We randomly drew SGWB spectra from a bank of 234,000 that had been fit to SMBHB population realizations [51] (see also [52]). Figure 2 shows the total distribution of SGWB spectral characteristics in blue, and the spectral characteristics injected into our simulations in red. Typical PTA analyses use priors of $\gamma \in [0, 7]$ and $\log_{10} A \in [-18, -12]$. Following this convention, we also ensured that the randomly-drawn SGWB spectral characteristics satisfied these prior constraints. Unless otherwise stated, all models search for a CURN process to ensure the most fair comparison between the full PTA likelihood and the refitting techniques.

B. Parameter estimation fidelity

We choose one of our simulations as a case study of our refitting techniques. The chosen simulation has spectral characteristics comparable to the CURN detected in the NANOGrav 12.5 year dataset, and has one of the smallest Hellinger distances between the uncorrelated full-likelihood power-law analysis and the PTA free-spectrum refit. As a first exploration, given that each GFL-Lite per-

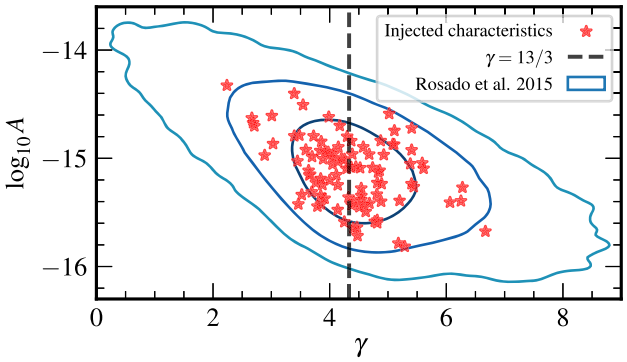

FIG. 2. The blue regions are the 68%, 95%, and 99% credible regions of the distribution of SGWB spectral characteristics from the 234,000 SMBHB population realizations of Rosado *et al.* [51]. We randomly selected 100 SGWB realizations from this distribution (red markers). The dashed black line designates $\gamma = 13/3$, the realization-averaged expected SGWB spectral index from a population of SMBHBs.

pulsar free spectrum has already been marginalized over intrinsic red noise parameters, the combined product of those likelihood distributions across pulsars should be consistent with the PTA free-spectrum. This is shown in the left panel of Fig. 3, where there is broad consistency between the techniques.

The comparison of power-law–model posterior distributions for our case-study simulation is shown in the right panel of Fig. 3, where credible regions correspond to 68% and 95% levels for the spatially-uncorrelated full-likelihood, the PTA free-spectrum refit, and the GFL Lite analysis. Both refit methods perform well, recovering posteriors consistent with the full production-level PTA likelihood, with both achieving 2D Hellinger distances of 0.10. The 1D-marginalized posteriors on $\log_{10} A$ and γ have distances with respect to the full PTA likelihood of 0.06 and 0.07 for the PTA free-spectrum, and 0.09 and 0.05 for GFL Lite. In this case, the PTA free-spectrum refit and GFL Lite performances are on par. We see a similar consistency when comparing the Hellinger distances of all 100 dataset realizations. The distributions of Hellinger distances for the 2D and 1D marginalized posteriors are shown in Fig. 4, from which we quote the median, 16th percentile, and 84th percentile values. The 2D Hellinger distances between the PTA free-spectrum refit and the full likelihood are $0.26^{0.40}_{0.17}$, while GFL Lite has 2D Hellinger distances of $0.27^{0.40}_{0.20}$. We conclude that the PTA free-spectrum refit and GFL Lite analysis are consistent with each other.

To better understand the origin of discrepancies between our refit methods and the full likelihood, we investigate the magnitude of interfrequency correlations in the Bayesian free-spectrum posteriors, using Pearson's correlation coefficient [53]. If interfrequency correlations are weak, the correlation matrix of the posterior samples should be mostly diagonal in structure. Pearson's correlation

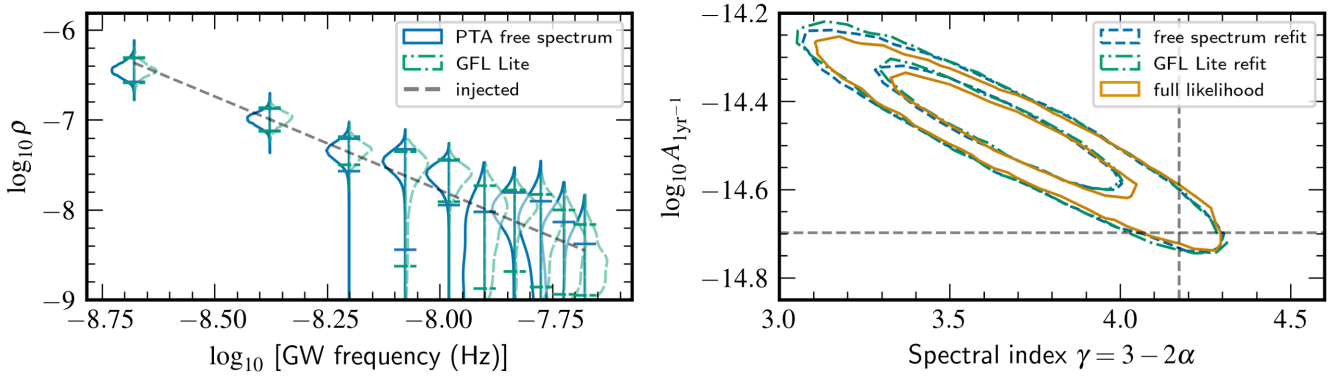

FIG. 3. *Left panel:* A comparison of the free-spectrum from a full PTA likelihood analysis (blue) with a product of the per-pulsar free-spectra from the GFL Lite pipeline (green) on a simulated dataset. The two violins are nearly identical and follow the injected SGWB power-law (gray line). *Right panel:* Posteriors of a 10 frequency power-law analysis with the full likelihood (orange), PTA free-spectrum refit (blue), and GFL Lite methods (green), for the simulated dataset shown in the left panel. Credible regions enclose 68% and 95% of the posterior. The injected SGWB spectral characteristics are shown as the dashed gray lines, with $\log_{10} A = -14.7$ and $\gamma = 4.17$. The PTA free-spectrum refit and GFL Lite posteriors match well to the full likelihood.

coefficient quantifies how 'diagonal' a correlation matrix is, with a coefficient of 1 indicating a perfectly diagonal matrix (i.e., no interfrequency correlations), and lower values indicating off-diagonal structure (i.e., interfrequency correlations). In the limit that there are no interfrequency correlations, Pearson's correlation coefficient becomes unity, and our approximation becomes an identity. The median, 16th, and 84th percentile values of this coefficient across all 100 realizations of the PTA free-spectra is $0.92^{0.98}_{0.86}$, suggesting weak correlations between GW frequencies. We also compute the coefficient for all 45 per-pulsar free-spectra from the GFL Lite pipeline across all 100 simulation realizations, giving $0.99^{1.0}_{0.91}$; per-pulsar free-spectra appear to be uncorrelated across frequencies. Hence our assumption throughout of independence between frequencies is justified, and suggests that information being lost from our refit pipelines is through the compounding of small inaccuracies in our density estimators.

Finally, we test the efficacy of Bayesian recovery between our proposed methods and the full likelihood with p-p plots, as shown in Fig. 5. If we were to draw our injected spectral characteristics from the same priors as employed in our Bayesian analysis, then we would expect to recover our injections within the p%-credible region for p% of our simulations. However in our analyses—even with the full likelihood—we see bias, causing deviation from the diagonal p-p plot, since we drew our injected characteristics from the SMBHB populations of Rosado *et al.* [51], and other analysis approximations. Instead, we compare the relative efficacy of our refit methods to the full likelihood analysis by taking the difference in p-p recovery between the full likelihood and our refit methods. A perfect comparison would give zero difference for all p. The PTA free-spectrum refit has the smallest differences from the full likelihood, showing deviations around zero mostly within a 1σ confidence interval, where $\sigma = \sqrt{p(1-p)/100}$ is the binomial standard error for a sample of 100 realizations [54].

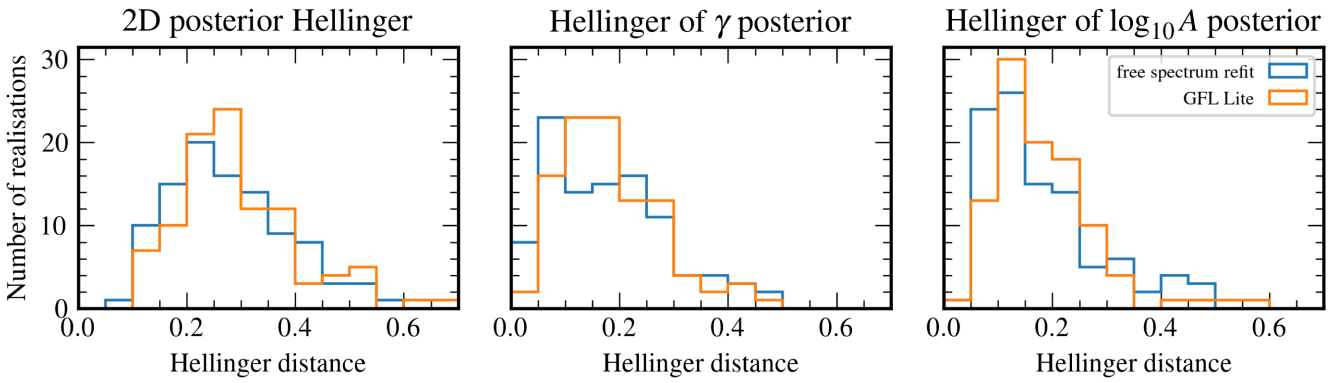

FIG. 4. Hellinger distances between the posteriors of the full likelihood and for each refitting technique for all 100 SGWB dataset realizations. Both methods have a similar distribution of Hellinger distances, thereby demonstrating similar performance when compared to the full PTA likelihood analysis.

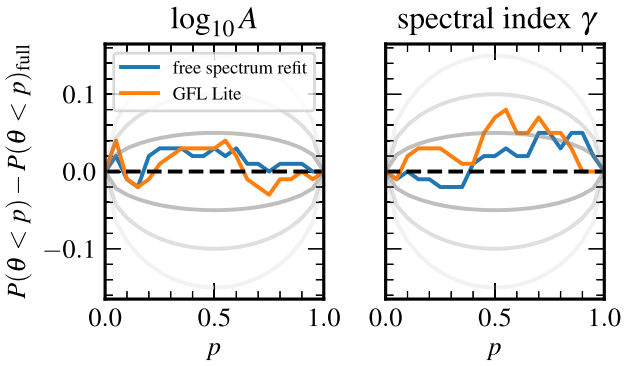

FIG. 5. The difference in p-p plots between the full likelihood and the PTA free-spectrum refit or GFL Lite. Equivalent recovery would show zero for all $p\%$ credible regions. The PTA free-spectrum refit is centered close to zero and mostly within the 1σ confidence region, where gray curves show 1σ, 2σ, 3σ regions. GFL Lite is also close to zero, and mostly within the 2σ confidence region for both parameters.

GFL Lite shows more deviation from the full likelihood than the PTA free-spectrum refit, but these remain typically within a 1σ confidence interval on $\log_{10} A$, and within 2σ for the spectral index.

For more context, we compare Fig. 5 to a gallery of toy univariate distribution comparisons in Fig. 13. For the PTA free-spectrum refit, the p-p plot for $\log_{10} A$ is similar to the center top panel of Fig. 13, which suggests this method may underestimate $\log_{10} A$ relative to the full PTA likelihood. Meanwhile, the spectral index recovery appears similar to the left-middle panel, suggesting that the width of the recovered γ posterior is narrower than the full PTA likelihood. By contrast, the p-p plot for $\log_{10} A$ and γ for GFL Lite appear similar to the right middle panel and top center panel of Fig. 13 respectively, suggesting a typically wider recovered $\log_{10} A$ posterior, and a slightly underestimated γ recovery. These are again likely due to compounding of small inaccuracies in our density estimators over many frequencies (and pulsars). However, overall these refitting methods achieve excellent parameter posterior recovery when compared to the full PTA likelihood.

C. Model selection

We now explore the efficacy of our refitting techniques for spectral model selection. The SGWB spectrum is typically modeled as a power-law, but other astrophysical and cosmological phenomena, and potentially even noise contamination, may influence its inferred shape. We would like to test whether these models better fit PTA data than a simple power-law, and make astrophysical and cosmological interpretations from their spectral characteristics.

Model selection with the current production-level PTA analysis pipeline is challenging given the relatively slow computation time of the PTA likelihood compared to the size of the parameter space that must be searched over. We must compare the Bayesian evidence of our data given our hypothesis models, $p(\vec{\delta t}|\mathcal{H})$, to derive a Bayes factor $\mathcal{B}_{12} = p(\vec{\delta t}|\mathcal{H}_1)/p(\vec{\delta t}|\mathcal{H}_2)$, and interpret those values to reject or accept $\mathcal{H}_1$ over $\mathcal{H}_2$. The interpretation is problem-specific, but some rules-of-thumb are given in Kass and Raftery [55]. In PTA analysis, model selection is typically conducted via calculating the Savage-Dickey density ratio [56] for low-contrast nested models, or with product-space sampling for mildly disjoint nested models [see, e.g., [57–59]].

One model selection technique that is currently impractical for production-level PTA analyses on large arrays ($\gtrsim$40 pulsars) is nested sampling, for which one analyzes each model separately to compute the Bayesian evidence [60,61]). Nested sampling is computationally expensive and cannot be realistically used with the full PTA likelihood given the combination of parameter dimensionality and slow evaluation time for larger arrays. In the PTA literature, nested sampling has been used before, but only for a small collection of pulsars [62]. Our new techniques now make spectral model selection via nested sampling feasible for larger PTAs.

Table I compares Bayes factors between various spectral models and the injected power-law behavior from the same case-study simulation as Fig. 3, using the PTA free-spectrum refitting technique. A broken power law has power-law behavior at low frequencies that then transitions into (in this case) a flat spectrum at higher frequencies in order to account for a white-noise floor in real data. This is used often in production-level analyses as a data-driven way of identifying the optimal number of frequencies with which to model a common red-noise process such that the inference is not biased by white noise [10]. A turnover model is similar in spirit to the broken power-law—in that it is effectively two power-laws connected by a bend—but motivated as a way to model low-frequency SGWB spectral attenuation from a binary population's interactions with their respective galaxy environments [17,63]. A t-process model has an underlying power-law behavior, but with per-frequency deviations that are constrained by an inverse-gamma prior. This is used to account for spectral fuzziness owing to noise conflation with the CURN, or potentially even binary-population finiteness influencing the spectral

TABLE I. Bayes factors for different 10-frequency CURN spectral models compared to a power-law when refitted to a PTA free-spectrum via the ULTRANEST nested sampler [65]. As expected, a power-law model is favored over every other tested model.

Disfavored model	Favored model	$\mathcal{B}$
Broken power-law	Power-law	21.1 ± 6.0
Turnover	Power-law	1.71 ± 0.44
t-process	Power-law	50.7 ± 11.8

shape [64]. Unsurprisingly, the power-law is the most favored model, since it is the injected spectrum. However, the power law is only slightly favored over the turnover model with $\mathcal{B} = 1.71$. This is because we allowed the range of turnover frequencies to be in any of the 10 modeled GW frequency bins. The model favored the lowest frequency bin, which made it behave mostly like a power law. The broken power-law's bend frequency was also allowed to vary across all frequencies, however it is much less favored than the power-law because its spectral index at frequencies greater than the bend frequency is fixed at zero, which the data do not support. Similarly, the injected power-law signal is so strong that any noise-induced deviations from it are small, thereby disfavoring the t-process model.

Using our spectral refitting techniques, it is now possible to systematically explore the evidence for various realistic SGWB spectra in PTA data. We however emphasize that this is currently only for spectral model selection; necessary developments for performing model selection between interpulsar correlated models (e.g., monopole, dipole, Hellings–Downs), or to assess evidence for the presence of a CURN process over only intrinsic per-pulsar noise, are discussed in Sec. IV.

D. Evolution of Bayesian spectral constraints with number of GW frequencies and pulsars

Given that spectral characterization is now trivial with our refitting techniques, we use our simulations to study how the Bayesian inference of spectral characteristics evolves with the number of modeled GW frequencies and pulsars.

1. Dependence on number of GW frequencies

In Fig. 6 we recover the Bayesian posterior for a CURN power-law process on our case-study simulation as a function of the number of modeled GW frequencies. Typically, we fit a common-process model to the N_f lowest GW frequencies; this is shown by the blue regions. However, we may also fit a power-law to our highest N_f frequencies, given by the gray contours. This gives a comparison between the information content of the highest versus lowest frequencies. The SGWB spectrum from

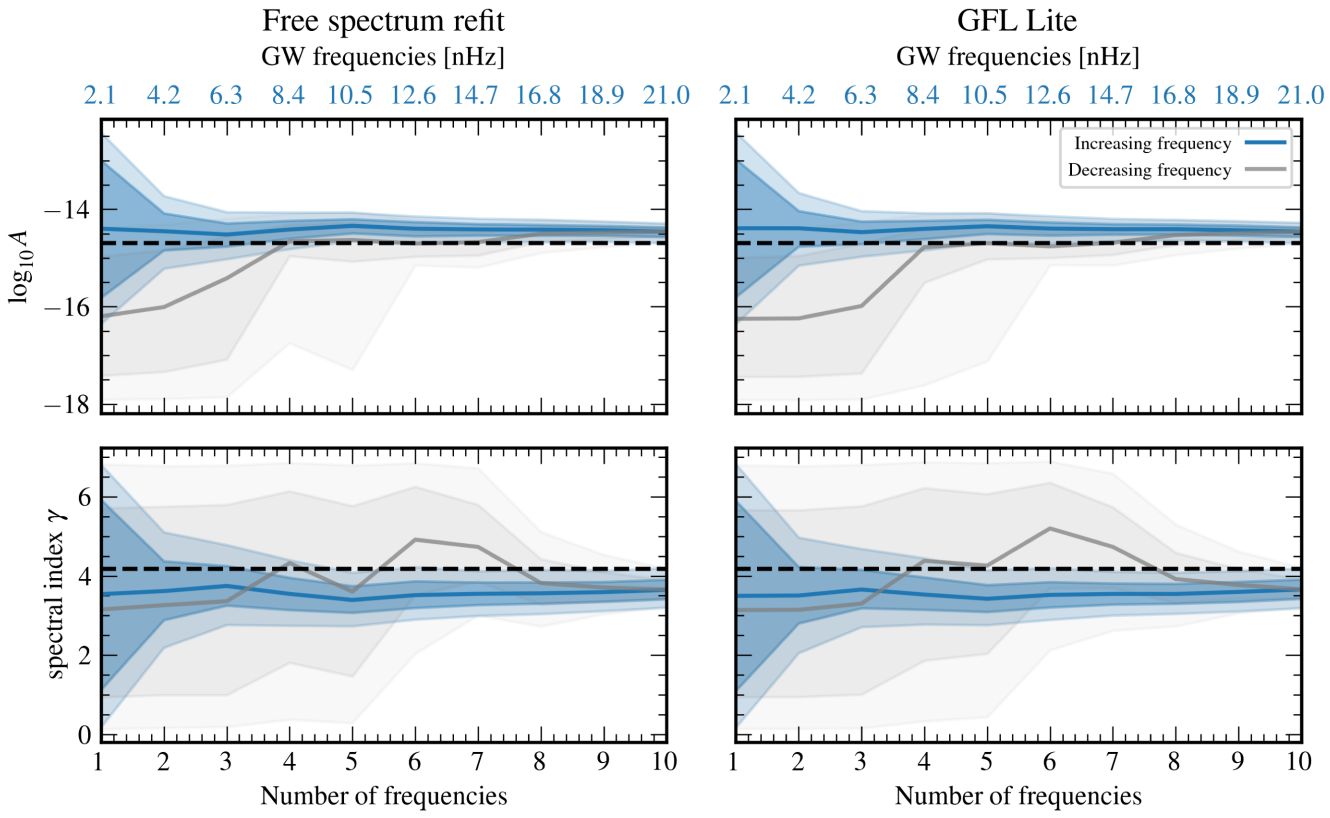

FIG. 6. The median, 1-σ, and 2-σ posterior credible constraints on $\log_{10} A$, γ for a power-law process as a function of the number of modeled frequencies, N_f. The blue regions signify the constraints from fitting to the lowest frequency upwards (where these frequencies are explicitly shown in blue on the top x-axis), while the gray signifies fitting from the tenth frequency downwards. As the number of frequencies increase, the posteriors become more constrained towards the injected parameter values (dashed black lines). For the PTA free-spectrum refit, we see the expected behavior of the blue contours constraining the parameters more quickly than the gray. Qualitatively, GFL Lite (right column) performs as well as the PTA free-spectrum refit (left column).

astrophysical or cosmological sources is expected to be red, with more power at lower frequencies. Hence, PTAs should be more sensitive to the SGWB at frequencies of $\sim 1/T$ than at higher frequencies, where intrinsic per-pulsar red noise and white noise can dominate [66,67]. Therefore, we expect the blue contours to converge toward the lines of injected values faster than the gray contours; we see this for both the PTA free-spectrum refit and GFL Lite techniques, where the posterior spread in recovered parameters decreases significantly after only two frequencies. The gray contours (representing fitting from higher frequencies downwards) remain wide for a larger number of modeled frequencies, where both techniques require eight frequencies to converge on the spectral index γ, while the recovered amplitude converges on the injection after only four frequencies. As expected, PTAs derive most information on SGWB spectral characteristics from the lowest analyzed GW frequencies, by virtue of the fact that red noise processes have more power there.

2. Dependence on number of pulsars

We may also analyze the SGWB parameter posterior recovery as a function of the number of pulsars N_p in our PTA (Fig. 7), this time using the GFL Lite technique. Given the large number of combinations with which N_p pulsars can be chosen from the array of 45, we only look at two sets of analyses, where we either add pulsars by decreasing or increasing timespan. Similar to Sec. III D 1, pulsars with longer observational timespans should be more informative of lower GW frequencies, where the signal is expected to be strongest. Therefore, we expect, and indeed observe, that the blue contours converge on the injected parameter values faster than the gray contours, requiring only the ∼8 longest-timespan pulsars before the median and posterior credible regions of the recovered spectral characteristics become approximately constant. By contrast, the ∼35 shortest-timed pulsars are required to recover the same precision as those eight longest-timed pulsars.

3. Characterization through the effective number of pulsars

From these analyses, it is clear that not all pulsars and frequencies contribute equally toward spectral characterization. Frequencies with more noise than others will be down-weighted in spectral model fitting, as will pulsars whose overall noise level exceeds that of others. Using the GFL Lite free spectrum of each pulsar, we can calculate the effective number of pulsars $N_{\rm eff}$ in an N_p-pulsar PTA searching for an N_f frequency power-law SGWB spectrum. We adapt and modify Eq. (8) in Cornish and Sampson [68] to the case of spectral characterization, also accounting for the uncertainty on the free-spectrum measurements,

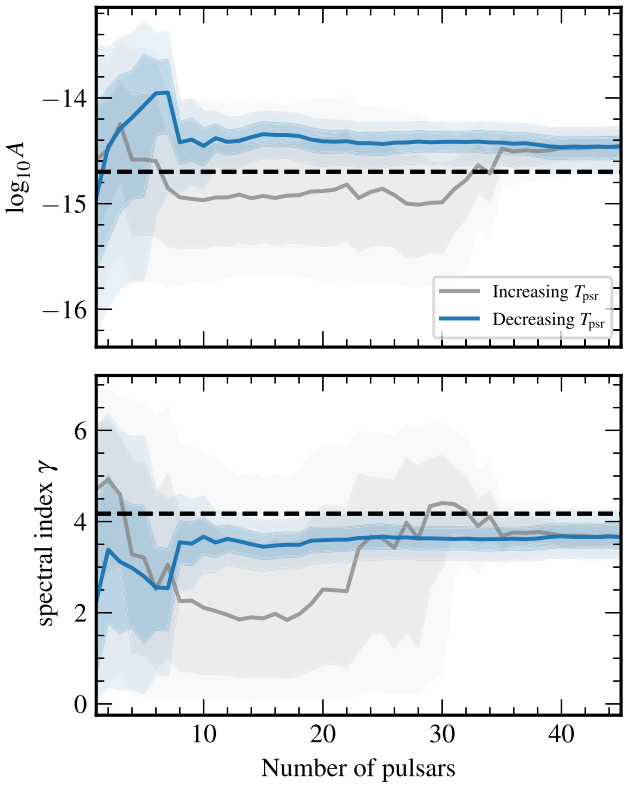

FIG. 7. The median, 1σ, and 2σ posterior credible constraints on $\log_{10} A, \gamma$ for a power-law process as a function of the number of modeled pulsars, N_p. The blue regions signify posterior constraints of a 10-frequency power-law CURN fitted to the N_p-pulsars with the longest observing time spans, while the gray regions are the corresponding constraints from the N_p shortest-timed pulsars. The black dashed line denotes the injected SGWB spectral characteristics.

$$N_{\rm eff} = \frac{\sum_{p=1}^{N_p} \sum_{k=1}^{N_f} 1/\sigma(\log_{10}\rho_{p,k})^2}{\max_{1\le p\le N_p} \sum_{k=1}^{N_f} 1/\sigma(\log_{10}\rho_{p,k})^2}, \qquad (13)$$

where σ is measurement uncertainty. The free-spectrum posteriors $\log_{10}\rho_{p,k}$ come from the pth pulsar and kth frequency of the GFL Lite free-spectrum pipeline. We estimate the measurement uncertainty of the posterior of the pth pulsar and kth frequency with $\sigma_{\rm G}$, a rank-based estimate of the standard deviation to account for distribution non-Gaussianity, $\sigma_{\rm G} \approx 0.7413 \times {\rm IQR}$, where IQR is the interquartile range, and the prefactor originates from computing the IQR of a Gaussian [69]. However, some posteriors are prior dominated and uninformative, and estimating the standard deviation will return, at worst, that of the prior. We determine which pulsar and frequency posteriors are uninformative by computing the Savage-Dickey density ratio [56], which, in this case, is used to estimate the Bayes factor between a model with and without a CURN process in a given pulsar, at a given frequency. $\mathcal{B} > 1$ suggests that a CURN process is

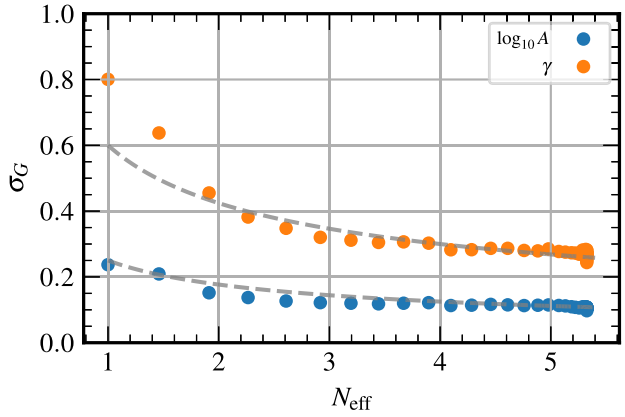

FIG. 8. The relationship between the effective number of pulsars in a PTA, $N_{\rm eff}$, and the uncertainty on the spectral parameters (see text for a definition of $\sigma_{\rm G}$), derived using GFL Lite. We increase N_p and keep the number of GW frequencies as 10. The recovered parameter uncertainties scale approximately as the expected $1/\sqrt{N_{\rm eff}}$ for both $\log_{10} A$ and γ.

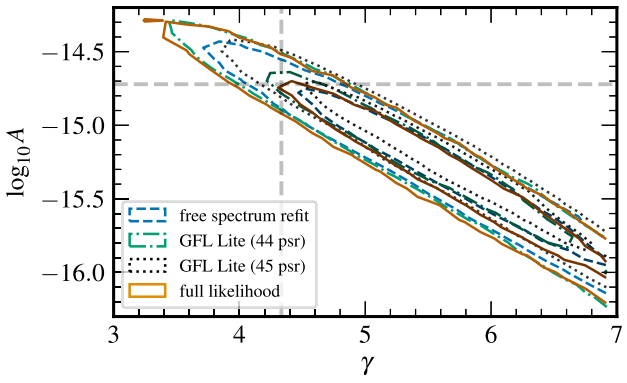

FIG. 9. Fitting a 5-frequency power law to the NANOGrav 12.5-year dataset via the PTA free-spectrum refit technique (blue) and GFL Lite analysis (green and dotted black). We compare our analyses to the published posterior (orange). We find excellent agreement between the published result and the PTA free-spectrum refit, which attains a Hellinger distance of $H = 0.13$. The 45-pulsar GFL Lite analysis does not recover the full-likelihood posterior as well ($H = 0.31$). However, removing one mismodeled pulsar results in better performance (green, $H = 0.12$)—see text for details.

supported, while if $\mathcal{B} < 1$, we determine that it is uninformative and set $\sigma = \infty$. The normalization of Eq. (13) ensures that $N_{\rm eff} \geq 1$ for all N_p and N_f. Therefore, $N_{\rm eff}$ is the effective number of pulsars relative to the most constrained (i.e., least noisy, and therefore most informative) pulsar for spectral characterization in the array. For a PTA with heterogeneous pulsar spectral uncertainties, $N_{\rm eff} < N_p$, while a PTA with homogeneous uncertainties would have $N_{\rm eff} = N_p$.

Figure 8 shows the relationship between the power-law parameter uncertainties as a function of $N_{\rm eff}$. We fitted a 10-frequency, N_p-pulsar power law with the GFL Lite pipeline, adding pulsars in order of the greatest to smallest value of $\sum_k^{N_f} 1/\sigma_{\rm G}(\log_{10}\rho_{p,k})^2$, i.e., in order of most-constrained to least-constrained pulsar spectrum. We computed the marginalized posterior uncertainty on both power-law parameters using the rank-based standard-deviation estimate $\sigma_{\rm G}$, defined es earlier. We see here that increasing the number of real pulsars increases the effective number of pulsars in the PTA, and decreases $\sigma_{\rm G}$ for both parameters. These studies allow us to posit a general relationship for spectral constraints in Bayesian PTA analyses, where $\sigma_{\rm G} \propto 1/\sqrt{N_{\rm eff}}$, as one may expect for a standard-deviation-type quantity computed from a data sample.

E. Recreating the results of the NANOGrav 12.5-year dataset

We now apply our refitting techniques to the NANOGrav 12.5-year dataset to assess performance against published results. Analysis of the NANOGrav 12.5-year dataset did not find significant evidence for Hellings and Downs interpulsar correlations, however, there was strong evidence for a CURN process. The posterior probability density for an analysis with a 5-frequency power-law CURN process (including 30-frequency power-law per-pulsar intrinsic red noise) is shown in Fig. 9, along with a PTA free-spectrum refit and GFL Lite analysis on this dataset. The PTA free-spectrum refit is consistent with the published full likelihood with a Hellinger distance of $H = 0.13$.

For GFL Lite, we modeled a 5-frequency free- pectrum and 30-frequency power law (to model the intrinsic red noise), and refit to the 5 free-spectrum posteriors. We found that, because there is excess unmodeled noise in the real dataset,[2] modeling a greater number of frequencies with the free-spectrum in each individual pulsar resulted in noise corruption, causing the free-spectrum to be conflated with intrinsic red noise in some pulsars. This is not a problem in the PTA free-spectrum refit, where the strength of the CURN from all of the pulsars inhibits the potential conflation with intrinsic red noise in pulsars that have misspecified noise models. Keeping the GFL Lite free spectrum to just 5 frequencies, and allowing the intrinsic red noise to be informed by 30 frequencies, attempts to limit this confusion. Unfortunately for pulsar B1855 + 09, the power law is a poor model for its intrinsic red noise, resulting in the free-spectrum posterior recovering the strong-intrinsic red noise of this pulsar rather than the CURN. When a 5-frequency GFL Lite refit is conducted, this pulsar is influential, causing the GFL Lite refit posterior to appear slightly offset from that of the full PTA likelihood in Fig. 9, with a Hellinger distance of $H = 0.31$. Removing this pulsar results in a more

[2]The potential for model misspecification in pulsar timing datasets when only simple noise models are used has now been recognized. Ameliorating this requires custom noise modeling. This has been challenging to incorporate in large-array studies, but is recognized as the correct path forward.

TABLE II. Bayes factors $\mathcal{B}$ for different 5-frequency common-process spectral models compared to a power law, when refitted to a PTA free spectrum for the NANOGrav 12.5-year dataset. The power law has varied spectral index γ unless stated.

Disfavored model	Favored model	$\mathcal{B}$
$\gamma = 13/3$ power-law	Power-law	1.17 ± 0.40
Broken power-law	Power-law	1.82 ± 0.34
Turnover	Power-law	2.23 ± 0.57
t-process	Power-law	1.83 ± 0.57

consistent refit, with a Hellinger distance of just $H = 0.12$. For the remainder, unless otherwise specified, we conduct the GFL Lite refit with just 44 pulsars. Improving the modeling of B1855 + 09 is beyond the scope of this paper and we discuss how we can improve analysis of pulsars like it in Sec. IV.

Table II shows the results of model selection for various 5-frequency spectral models with the PTA free-spectrum refit via nested sampling, a technique that estimates the Bayesian evidence of a model, and which we introduced in Sec. III C. We see that a varied-γ power-law is barely favored over a $\gamma = 13/3$ power-law, and $\gamma = 13/3$ is not ruled out by these data. There is a little more evidence to favor a power law over broken power-law, turnover, and t-process spectra, however none of these are substantial.

We also characterize the spectral recovery as a function of the number of modeled GW frequencies and pulsars. The PTA free-spectrum refit to increasing numbers of low GW frequencies (blue) in the left panel of Fig. 10 shows a similar "shallowing" of the spectrum as seen in Arzoumanian *et al.* [10], where γ trends toward ∼2–3, potentially due to coupling with unmodeled excess higher-frequency noise. Meanwhile, increasing the number of frequencies from $f = 30/T$ downwards tends to have a broad, unconstrained posterior for all frequencies consistent with low γ i.e., a flatter power spectrum typified by white noise. Fitting up to the first five frequencies, GFL Lite is consistent with the PTA free-spectrum refit. In the

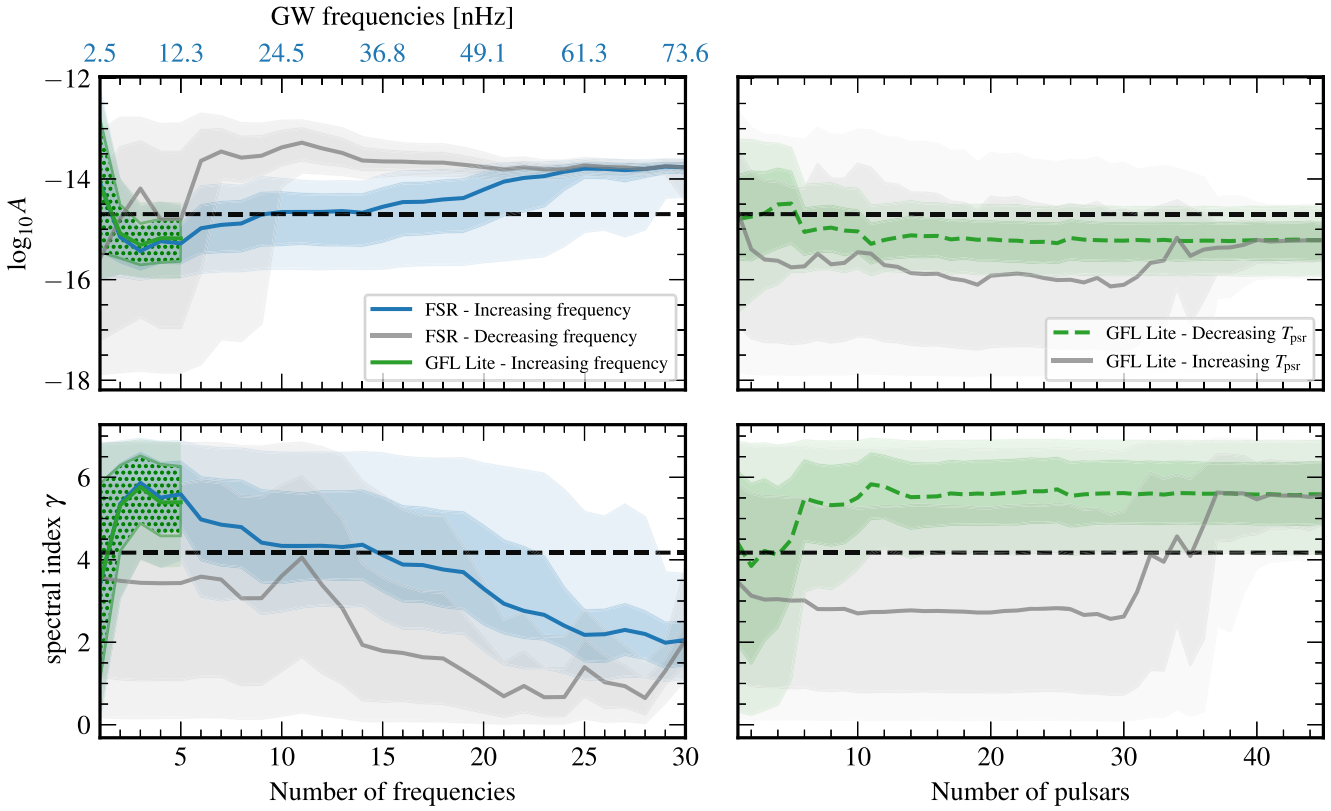

FIG. 10. *Left:* The median, 1σ, and 2σ posterior credible constraints on $\log_{10} A$, γ for an N_f frequency power law as a function of number of GW frequencies in the NANOGrav 12.5-year dataset. In blue, we show increasing number of frequencies from the lowest bin and increase upwards to $f = 30/T$ for the free-spectrum refit, and in green, we show the same analysis for GFL Lite upwards to $5/T$ which is consistent with the blue contour. In gray, we show addition of frequencies from $f = 30/T$ downwards for the free-spectrum refit. We observe a similar shallowing of the spectrum as Arzoumanian *et al.* [10] when a larger number of frequencies are modeled because of the contribution of white noise. *Right:* A 5-frequency power-law is fit to an increasing number of pulsars in the NANOGrav 12.5-year dataset, where green regions show constraints from adding pulsars in longest- to shortest-timed order. The blue posteriors are well-constrained after ∼14 pulsars, while the gray posterior require ∼36 pulsars out of 45 to be constrained. Hence, the longest 14 pulsars are the most important for spectral characterization in this dataset.

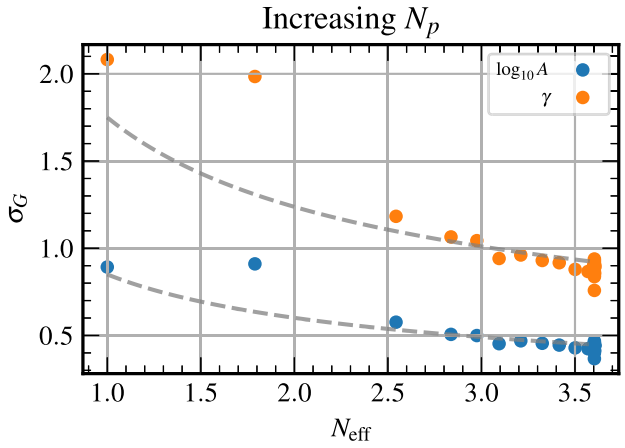

FIG. 11. Similar to Fig. 8, using the NANOGrav 12.5-year dataset, minus B1855 + 09. We fit a 5-frequency power law to an increasing number of pulsars in the array. About 12 pulsars ($N_{\rm eff} \sim 3.6$) inform the spectral characteristics of the CURN, with a scaling of $1/\sqrt{N_{\rm eff}}$. We use the single-pulsar free spectra from GFL Lite for calculating $N_{\rm eff}$.

right panel, we use GFL Lite to fit a 5-frequency power law to an increasing number of pulsars (including B1855 + 09), added by decreasing pulsar timespan (green) and increasing pulsar timespan (gray). As with Fig. 7, the green posterior is constrained after relatively few pulsars are added (∼14),[3] while the gray posterior is unconstrained, particularly for γ, until the final ten pulsars are added.

Finally, we also analyze the measurement uncertainty of $\{\log_{10} A, \gamma\}$ as a function of the effective number of pulsars, $N_{\rm eff}$, using the GFL Lite technique. Figure 11 shows $\sigma_{\rm G}$ of parameters for a 5-frequency power law model as a function of $N_{\rm eff}$ as we increase the number of pulsars in the array in order of greatest to the smallest value of $\sum_k^{N_f} 1/\sigma_{\rm G}(\log_{10}\rho_{p,k})^2$. We use the same methods as Sec. III D 3. Again, we observe an approximate $1/\sqrt{N_{\rm eff}}$ scaling relationship. We need at most four pulsars that are equivalent to the best modeled pulsar in order to effectively recover the spectral characteristics of the CURN from this dataset. We also notice that the real dataset has fewer numbers of effective pulsars than our earlier simulated datasets, due to the data model of the simulations being entirely known and prescribed.

IV. CONCLUSIONS AND FUTURE PROSPECTS

We have developed a set of rapid and robust spectral refitting techniques that operate on posterior samples from pulsar and PTA Bayesian periodogram analyses, where the power spectral density is jointly modeled by free parameters at each GW frequency (sometimes referred to in the PTA literature as free-spectrum analyses). This is a generalization of our previously developed factorized likelihood (FL) technique [31], where GW background amplitude posteriors for fixed power-law spectral index models are combined in postprocessing, under the assumption that spectral characterization is mostly driven by autocorrelation information in the PTA covariance matrix. The main limitation of FL was its conditioning on a GW-background spectral model with a fixed power-law spectral index. Our new formulations loosen that assumption, allowing for refitting and inference of arbitrary spectral models.

In order of generality, we assessed the performance of a model that refits on a Bayesian PTA free-spectrum (PTA free-spectrum refit) and one that refits on the combination of per-pulsar free spectra, which act as proxies for the CURN signal in each pulsar, with intrinsic per-pulsar red noise modeled separately (GFL Lite). These techniques are several orders of magnitude faster in evaluating their likelihood functions when compared to the production-level PTA pipeline, and also scale much more favorably when adding new pulsars. These gains in speed and scalability will be important in safeguarding PTA analyses from future bottlenecks, as significantly more data and pulsars are added to arrays through IPTA combinations and high-cadence observations in MeerTime [70], CHIME [71], and (farther in the future) the SKA [72].

We assessed the fidelity of parameter estimation using a set of 100 realistic PTA datasets based on the NANOGrav 12.5-year dataset that is extended into the future, and into which realizations of a GW background are injected with power-law spectral characteristics based on supermassive black hole binary population models. Through Hellinger-distance comparisons—which assess the distance between probability distributions—we found that the PTA free-spectrum refit and GFL Lite analyses are equivalent in performance, and consistent with the full production-level PTA likelihood analysis. While equivalent, we recommend using these refit methods in the following cases. For the PTA free-spectrum refit, it should be used when one is analyzing the evolution of spectral characterization with the number of GW frequencies, because combined influence of the PTA will make it less likely to confuse a CURN process with high-frequency white and/or mismodeled noise. If available, this technique can also be used to refit on PTA free-spectral posteriors that have an assumed interpulsar correlation signature. For example, in the case study presented in Fig. 3, the Hellinger distance (closer to zero is better) of the PTA free-spectrum refit was 0.06 when refitting on the HD-correlated free spectrum, compared to 0.10 from the CURN free-spectrum. Additionally, we showed that the PTA free-spectrum refit technique allows us to trivially perform spectral-model selection. The GFL

[3]The 14 pulsars are J1744-1134, J1455-3330, J1012 + 5307, B1937 + 21, J2145-0750, J1909-3744, J1918-0642, J1643-1224, J2317 + 1439, B1855 + 09, J1713 + 0747, J0030 + 0451, J1640 + 2224, and J0613-0200. It is likely that removing B1855 + 09 would result in better constraining power.

Lite technique should be used in studying how different subsets of frequencies and pulsars—e.g., long-baseline pulsars versus short-baseline pulsars—affect the spectral characterization of a CURN process (which is used as an approximate spectral model of the GW background). Care must be taken to ensure that, at the single pulsar level, the CURN process and intrinsic pulsar noise are not being conflated, as this will reduce the accuracy of the refitted posterior when compared to the full PTA likelihood. GFL Lite is more sensitive to noise misspecification than the PTA free-spectrum refit, since the latter has the benefit of many other pulsars to mitigate the impact of noise and CURN conflation.

We plan for a further generalization of the GFL Lite method, called GFL, which will have the advantage of allowing trivial changes to the spectral models and priors of the intrinsic red noise in each pulsar. This method will enable quick GW-background analyses in the presence of advanced per-pulsar red-noise models that are customized to each pulsar, which is currently not tractable with the production-level PTA pipeline for large arrays. As shown in Sec. III E, model misspecification of intrinsic red noise results in an inaccurate refitted posterior; advanced noise modeling, in concert with GFL, will improve SGWB spectral characterization for more pulsars and numbers of GW frequencies [73]. We also suspect that the main loss of information and fidelity at the moment is through the sampling and representation of the distribution tails of the per-pulsar free-spectral posteriors. A potential solution to this is to use Gibbs techniques and to draw directly from the analytic conditional posteriors of our free-spectral parameters, which has been shown to have better tail sampling [74,75]. This work will appear in a future publication.

Improvements to the representation of the posterior densities could be achieved through alternative KDE kernel functions that have more gradual drop-offs in support. If information is being lost in our density estimation, then performance gains may be made through multivariate KDEs across frequencies, or other higher-dimensional density estimation techniques based on neural network architectures, such as normalizing flows [76]. Another avenue is based on likelihood reweighting techniques, where an approximate distribution that is easier to sample is used to generate many random draws, then a subsequent reweighting stage updates these samples based on their support under the correct (potentially computationally-expensive) distribution [see, e.g., [77], for a recent PTA application]. Given the speed with which GFL refit analyses can be conducted, we could subsequently reweight these samples to match the full PTA likelihood. While this procedure will add extra computation time, it would still be quite a bit faster than a full pipeline analysis.

We also envision that future development of GFL-style refitting techniques will include interpulsar correlations, which would be the zenith of stochastic GW-background modeling through compressed sufficient statistics. While our current techniques are based on power-spectrum modeling, we would need to recover the Fourier coefficients of the timing residuals in order to retain phase information among the pulsars. We would then need to accurately represent the likelihood distribution of these Fourier coefficients, using density estimation techniques, to act as sufficient statistics for interpulsar correlation studies. There is ongoing development along these lines to replace the current production-level PTA pipeline and ensure that future Bayesian PTA analyses with significantly larger datasets will continue to be tractable.

The new techniques presented in this paper will have several immediate benefits for astrophysical- and cosmological-model testing with PTA data. The demographics and dynamics of supermassive black-hole binary populations is encoded in the amplitude and shape of the GW characteristic strain spectrum in the PTA band. Our techniques offer a path to use intermediate data products (i.e., Bayesian free-spectrum posteriors) for rapid spectral parameter estimation and model selection. Likewise, several potential sources of early-Universe GW-background signals give rise to strain spectra that deviate from the expected form of the supermassive black hole binary population signal, e.g., a phase transition may produce a more peaked spectrum than the power-law expected from binaries.

We plan to use our fast and flexible techniques to study milestones for PTA spectral estimation, such as what can be inferred in the near future about SMBHB populations, and the conditions under which cosmological background signals could be inferred beneath a dominant astrophysical signal. Answering these questions, and developing the spectral-estimation techniques with which they are addressed, are key to illuminating the path for PTA science in the next decade.

A. Software

The introduced refit methods are featured in a new analysis suite called CEFFYL for quick model selection and parameter estimation of spectra given PTA data. This is achieved by creating condensed data products representing the Bayesian spectra of a PTA's timing residuals. The data are represented by highly optimized KDEs from which we can extract probabilities to form Bayesian likelihoods to estimate our PTA likelihoods and to rapidly recover posteriors to our models. The suite employs code from `enterprise` [36], which was also used to create our free-spectra and the full-likelihood posteriors to which our analyses were compared. The PTA free-spectrum refit method is featured in the wrapper code, PTArcade [78]. We conducted parameter estimation via MCMC with `PTMCMC` [79], which utilizes parallel tempering and empirical proposal distributions for more efficient sampling

of the parameter space, while the nested sampler ULTRANEST [65] is used for model selection. To calculate the relevant KDE bandwidths, we translated the Sheather-Jones algorithm from an R implementation [80] into Python; this code is now contained within the CEFFYL suite. The KDEs are created using the FFTKDE method in KDEpy [81], and we use `ChainConsumer` [82] to create our corner plots to compare posteriors. The suite of PTA simulations were created with LIBSTEMPO [83].

ACKNOWLEDGMENTS

We thank our colleagues in NANOGrav and the International Pulsar Timing Array for fruitful discussions and feedback during the development of this technique. We particularly thank Alberto Sesana for providing the distributions of power-law fit parameters to characteristic strain spectra from SMBHB population realizations, based on Rosado *et al.* [51], Kyle Gersbach for conversations on using the Pearson correlation coefficient, David Wright for help to make the CEFFYL software suite easily installable, Xavier Siemens and Michele Vallisneri for stimulating discussions, and Joe Romano for creating a gallery of p-p plots that inspired our Fig. 13. S. R. T. acknowledges support from NSF AST-2007993, the NANOGrav NSF Physics Frontier Center No. 2020265, and an NSF CAREER No. 2146016. W. G. L. is supported by the NANOGrav NSF Physics Frontier Center No. 2020265, and acknowledges travel support from the Division of Gravitational Physics (DGRAV) to present this work at the APS April 2022 meeting, and from Vanderbilt University's Graduate Student Council. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, Nashville, Tennessee. This work was performed in part at Aspen Center for Physics, which is supported by National Science Foundation Grant No. PHY-2210452.

APPENDIX A: KERNEL DENSITY ESTIMATORS

Selecting the optimal kernel K and bandwidth h typically focuses on minimizing the asymptotic mean integrable squared error (AMISE) between the underlying distribution f and the reconstructed estimator $\hat{f}$ given N samples,

$$\mathrm{AMISE}(h) = \frac{R(K)}{Nh} + \frac{1}{4}\sigma_K h^4 R(f''). \tag{A1}$$

Here, $R(g) = \int g(x)^2 \mathrm{d}x$ for any function g, and $\sigma_K^2 = \int x^2 K(x)\mathrm{d}x > 0$ is the second moment of the kernel, at a given point x. The second derivative f'' is with respect to x. Note that if the kernel is normal with standard deviation σ, $\sigma_K^2 = \sigma^2$.

The optimal bandwidth h^* is found by minimizing Eq. (A1) with respect to h such that

$$h^* = \left(\frac{R(K)}{\sigma_K^4 R(f'')}\right)^{\frac{1}{5}} N^{-\frac{1}{5}}. \tag{A2}$$

If the kernel is normal with standard deviation $\sigma_K = 1$, and the underlying distribution f is known to be normal with standard deviation σ, bandwidth selection is trivial; $h^* = 1.06\hat{\sigma}N^{-1/5}$, where $\hat{\sigma}$ is the standard deviation of the samples.

However, f is not always known and a method is required to reduce the AMISE without prior knowledge of f. One such method is the Sheather-Jones plug-in selector [48]. It computes the optimal bandwidth h^* by estimating $R(f'')$ and iteratively solving Eq. (A2) with the Newton-Raphson method. This is a fast and effective bandwidth selector which we use in our KDE reconstructions.

After the optimal bandwidth is selected, substituting Eq. (A2) into Eq. (A1) finds the following relation between the AMISE and the kernel:

$$\mathrm{AMISE}(h^*) \propto [\sigma_K R(K)]^{\frac{4}{5}}. \tag{A3}$$

The optimal kernel is the kernel which minimizes this relation. This is the Epanechnikov kernel [47] which has the form

$$K(x) = \frac{3}{4}(1 - x^2), \qquad x \in [-1, 1]. \tag{A4}$$

We expect to collect samples at the lower boundary of the free-spectrum prior. To ensure accurate KDE representation of the samples at the boundary, we mirror the data at the boundary point and fit the KDE to the mirrored data. This reduces the bias induced at the boundary. We then compute probability densities along a grid of $\log_{10}\rho$ within the prior boundaries that is finer than the bandwidth size.

Figure 12 shows a toy model of using KDEs to recreate a distribution. We randomly drew 100, 1000, and 10000 points from a Rayleigh distribution, $f(x) = x\exp(-x^2/2)$, and recreate the distribution from those random samples using a KDE with the aforementioned optimizations. The reconstruction improves as the number of random draws in the training sample increases. Constructing a KDE with 10,000 random samples from the distribution more accurately estimates the original distribution than using less number of samples. Therefore, the more data points we draw from the original distribution, the smaller the absolute difference between the distribution and its reconstruction.

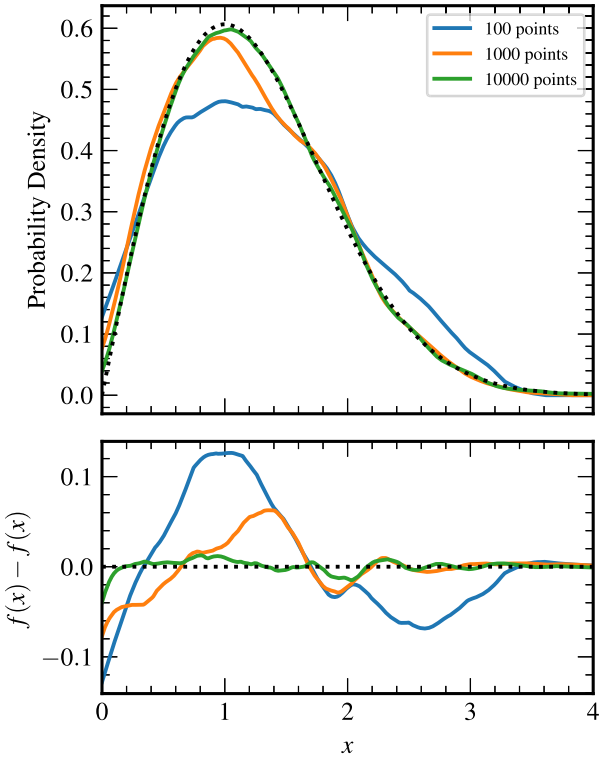

FIG. 12. A demonstration on the importance of good sampling to construct an accurate KDE. We randomly drew 100, 1000, and 10000 points from a Rayleigh distribution $f(x)$, and created the KDE $\hat{f}(x)$ with the Epanechnikov kernel [47] and a bandwidth selected by the Sheather-Jones method [48]. The bottom panel shows the absolute difference between the actual distribution and its reconstruction.

APPENDIX B: THE HELLINGER DISTANCE

Given probability density functions $f(\vec{x})$ and $g(\vec{x})$ in N-dimensional parameter space, the Hellinger distance H is defined as

$$H^2(f,g) = \frac{1}{2}\int \left(\sqrt{f(\vec{x})} - \sqrt{g(\vec{x})}\right)^2 \mathrm{d}^N x \tag{B1}$$

$$= 1 - \int \sqrt{f(\vec{x})g(\vec{x})}\,\mathrm{d}^N x. \tag{B2}$$

We choose the Hellinger distance as a metric for refitting accuracy over other distance measures—such as Jensen-Shannon—as it is bounded $0 \leq H \leq 1$, and valid for multivariate distributions. A value of $H = 0$ implies that distributions are identical, while $H = 1$ implies that they do not have any overlap and are completely different distributions.

TABLE III. The Hellinger distance, H, between two univariate normal distributions with equal standard deviations, yet with means offset by a certain number of standard deviations, n.

n	0.25	0.50	0.75	1.0	1.5	2.0	3.0	4.0
H	0.09	0.18	0.26	0.34	0.50	0.63	0.82	0.93

Our goal in building rapid and accurate refitting techniques is to ensure the Hellinger distance with respect to the posterior derived from the full PTA likelihood is sufficiently small. The interpretation of what sufficiently small means is problem specific, but some guiding intuition can be gleaned from simple analytic examples. One can show that the Hellinger distance between two univariate normal distributions, $f \sim \mathcal{N}(\mu_1, \sigma_1)$ and $g \sim \mathcal{N}(\mu_2, \sigma_2)$, is

$$H = \left\{1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}}\exp\left[-\frac{1}{4}\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}\right]\right\}^{1/2}, \tag{B3}$$

where μ and σ are the mean and standard deviation of the respective distributions. For distributions of equal standard deviation, but with their means offset from one another by a certain number, n, of these standard deviations, the Hellinger distance is

$$H = \left[1 - \exp\left(-\frac{n^2}{8}\right)\right]^{1/2}. \tag{B4}$$

A 1-σ offset between these normal distributions may not typically be regarded as a significant disparity, and corresponds to a Hellinger distance of 0.34. Values for other n are given in Table III.

In Fig. 13 we show some examples of univariate normal distributions with different means and standard deviations. Assuming we generate $n = 100$ realizations from these distributions, we show what the associated p-p plots would be, and the Hellinger distance between the distributions.

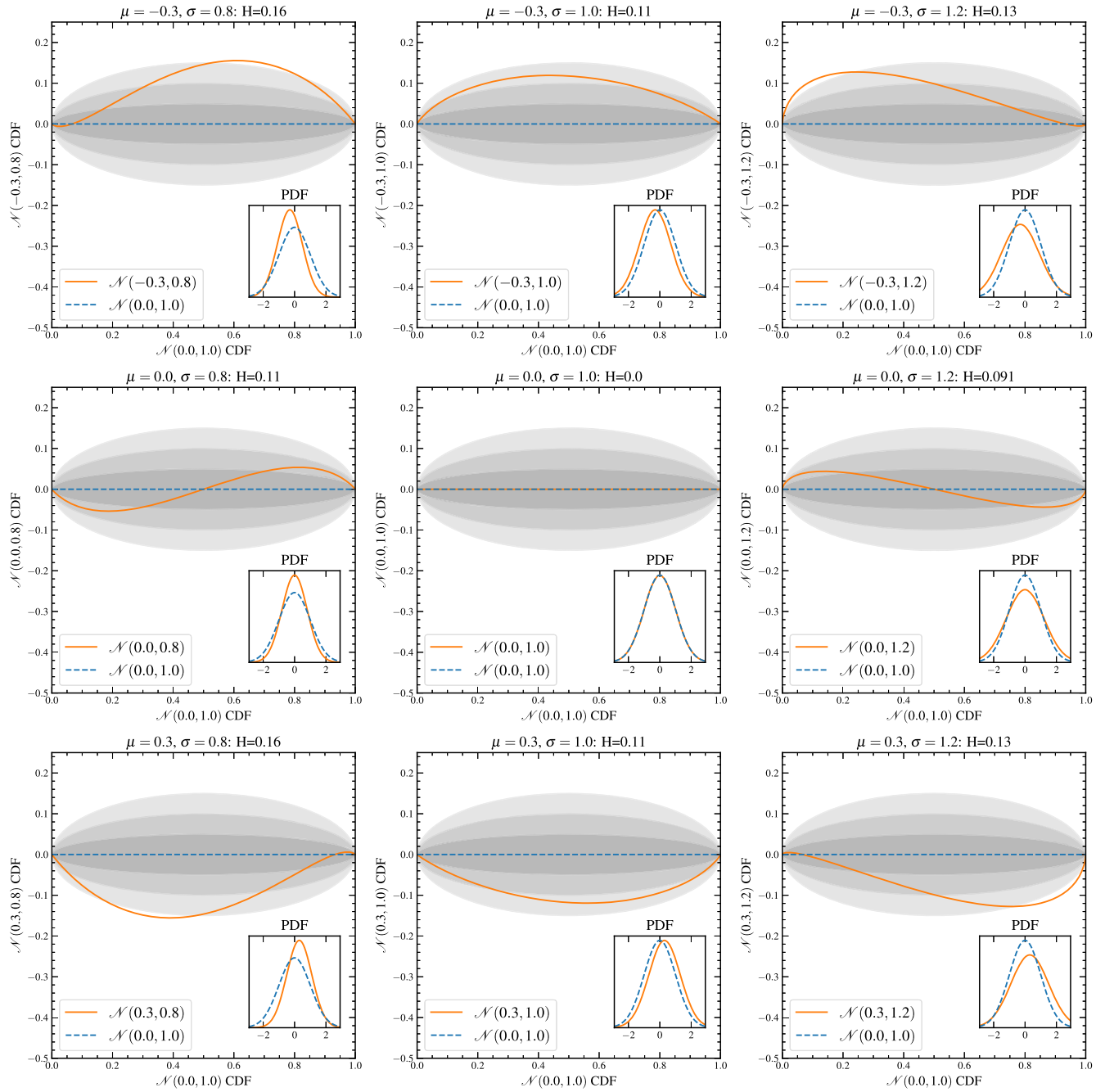

FIG. 13. Examples of what p-p plots look like for distributions that are approximately, but not entirely, equal to the true posterior (here the standard normal distribution $\mathcal{N}(0, 1)$—the blue curves in the insets), if we assume that we create $n = 100$ realizations of data. The orange curves in the insets show a modified normal distribution $\mathcal{N}(\mu, \sigma)$: our approximated posterior. The large figures show the corresponding p-p plots. At the top we have indicated the associated Hellinger distance between the two posteriors.

[1] R. S. Foster and D. C. Backer, Constructing a Pulsar Timing Array, Astrophys. J. **361**, 300 (1990).

[2] G. Agazie *et al.* (NANOGrav Collaboration), The NANOGrav 15 yr data set: Evidence for a gravitational-wave background, Astrophys. J. Lett. **951**, L8 (2023).

[3] J. Antoniadis *et al.*, The second data release from the European Pulsar Timing Array III. Search for gravitational wave signals, Astron. Astrophys. **678**, A50 (2023).

[4] D. J. Reardon *et al.*, Search for an isotropic gravitational-wave background with the Parkes Pulsar Timing Array, Astrophys. J. Lett. **951**, L6 (2023).

[5] H. Xu *et al.*, Searching for the nano-hertz stochastic gravitational wave background with the Chinese Pulsar Timing Array data release I, Res. Astron. Astrophys. **23**, 075024 (2023).

[6] R. Hellings and G. Downs, Upper limits on the isotropic gravitational radiation background from pulsar timing analysis, Astrophys. J. **265**, L39 (1983).

[7] J. Antoniadis *et al.*, The International Pulsar Timing Array second data release: Search for an isotropic gravitational wave background, Mon. Not. R. Astron. Soc. **510**, 4873 (2022).

[8] J. D. Romano, J. S. Hazboun, X. Siemens, and A. M. Archibald, Common-spectrum process versus cross-correlation for gravitational-wave searches using Pulsar Timing Arrays, Phys. Rev. D **103**, 063027 (2021).

[9] N. S. Pol, S. R. Taylor, L. Z. Kelley, S. J. Vigeland, J. Simon, S. Chen, Z. Arzoumanian, P. T. Baker, B. Bécsy, A. Brazier *et al.*, Astrophysics milestones for pulsar timing array gravitational-wave detection, Astrophys. J. Lett. **911**, L34 (2021).

[10] Z. Arzoumanian *et al.*, The NANOGrav 12.5-year data set: Search for an isotropic stochastic gravitational-wave background, Astrophys. J. Lett. **905**, L34 (2020).

[11] B. Goncharov *et al.*, On the evidence for a common-spectrum process in the search for the nanohertz gravitational-wave background with the Parkes Pulsar Timing Array, Astrophys. J. Lett. **917**, L19 (2021).

[12] S. Chen *et al.*, Common-red-signal analysis with 24-yr high-precision timing of the European Pulsar Timing Array: Inferences in the stochastic gravitational-wave background search, Mon. Not. R. Astron. Soc. **508**, 4970 (2021).

[13] E. S. Phinney, A practical theorem on gravitational wave backgrounds, arXiv:astro-ph/0108028.

[14] S. Burke-Spolaor, S. R. Taylor, M. Charisi, T. Dolch, J. S. Hazboun, A. Miguel Holgado, L. Z. Kelley, T. J. W. Lazio, D. R. Madison, N. McMann, C. M. F. Mingarelli, A. Rasskazov, X. Siemens, J. J. Simon, and T. L. Smith, The astrophysics of nanohertz gravitational waves, Astron. Astrophys. Rev. **27**, 5 (2019).

[15] G. Agazie *et al.* (Nanograv Collaboration), The NANOGrav 15 yr data set: Constraints on supermassive black hole binaries from the gravitational-wave background, Astrophys. J. Lett. **952**, L37 (2023).

[16] A. Sesana, Insights into the astrophysics of supermassive black hole binaries from pulsar timing observations, Classical Quantum Gravity **30**, 224014 (2013).

[17] L. Sampson, N. J. Cornish, and S. T. McWilliams, Constraining the solution to the last parsec problem with pulsar timing, Phys. Rev. D **91**, 084055 (2015).

[18] S. R. Taylor, J. Simon, and L. Sampson, Constraints on the dynamical environments of supermassive black-hole binaries using Pulsar-Timing Arrays, Phys. Rev. Lett. **118**, 181102 (2017).

[19] A. Sesana, A. Vecchio, and C. N. Colacino, The stochastic gravitational-wave background from massive black hole binary systems: Implications for observations with Pulsar Timing Arrays, Mon. Not. R. Astron. Soc. **390**, 192 (2008).

[20] E. Roebber, G. Holder, D. E. Holz, and M. Warren, Cosmic variance in the nanohertz gravitational wave background, Astrophys. J. **819**, 163 (2016).

[21] L. Z. Kelley, L. Blecha, L. Hernquist, A. Sesana, and S. R. Taylor, The gravitational wave background from massive black hole binaries in Illustris: Spectral features and time to detection with Pulsar Timing Arrays, Mon. Not. R. Astron. Soc. **471**, 4508 (2017).

[22] A. Afzal *et al.* (Nanograv Collaboration), The NANOGrav 15 yr data set: Search for signals from new physics, Astrophys. J. Lett. **951**, L11 (2023).

[23] T. W. B. Kibble, Topology of cosmic domains and strings, **9**, 1387 J. Phys. A (1976).

[24] P. D. Lasky *et al.*, Gravitational-wave cosmology across 29 decades in frequency, Phys. Rev. X **6**, 011035 (2016).

[25] P. Schwaller, Gravitational waves from a dark phase transition, Phys. Rev. Lett. **115**, 181101 (2015).

[26] A. R. Kaiser, N. S. Pol, M. A. McLaughlin, S. Chen, J. S. Hazboun, L. Z. Kelley, J. Simon, S. R. Taylor, S. J. Vigeland, and C. A. Witt, Disentangling multiple stochastic gravitational wave background sources in PTA datasets, Astrophys. J. **938**, 115 (2022).

[27] Z. Arzoumanian, P. Baker, A. Brazier, S. Burke-Spolaor, S. Chamberlin, S. Chatterjee, B. Christy, J. M. Cordes, N. J. Cornish, F. Crawford *et al.*, The NANOGrav 11 year data set: Pulsar-timing constraints on the stochastic gravitational-wave background, Astrophys. J. **859**, 47 (2018).

[28] R. van Haasteren and M. Vallisneri, New advances in the Gaussian-process approach to pulsar-timing data analysis, Phys. Rev. D **90**, 104012 (2014).

[29] R. van Haasteren and M. Vallisneri, Low-rank approximations for large stationary covariance matrices, as used in the Bayesian and generalized-least-squares analysis of pulsar-timing data, Mon. Not. R. Astron. Soc. **446**, 1170 (2015).

[30] M. Amiri *et al.* (CHIME/Pulsar Collaboration), The CHIME pulsar project: System overview, Astrophys. J. Suppl. Ser. **255**, 5 (2021).

[31] S. R. Taylor, J. Simon, L. Schult, N. Pol, and W. G. Lamb, A parallelized Bayesian approach to accelerated gravitational-wave background characterization, Phys. Rev. D **105**, 084049 (2022).

[32] A. D. Johnson, S. J. Vigeland, X. Siemens, and S. R. Taylor, Gravitational-wave statistics for Pulsar Timing Arrays: Examining bias from using a finite number of pulsars, Astrophys. J. **932**, 105 (2022).

[33] J. Cordes and R. Shannon, A measurement model for precision pulsar timing, arXiv:1010.3785.

[34] M. Keith, W. Coles, R. Shannon, G. Hobbs, R. Manchester, M. Bailes, N. Bhat, S. Burke-Spolaor, D. Champion, A. Chaudhary *et al.*, Measurement and correction of variations in interstellar dispersion in high-precision pulsar timing, Mon. Not. R. Astron. Soc. **429**, 2161 (2013).

[35] W. Coles, G. Hobbs, D. Champion, R. Manchester, and J. Verbiest, Pulsar timing analysis in the presence of correlated noise, Mon. Not. R. Astron. Soc. **418**, 561 (2011).
[36] J. A. Ellis, M. Vallisneri, S. R. Taylor, and P. T. Baker, Enterprise: Enhanced numerical toolbox enabling a robust pulsar inference suite, Zenodo (2020).
[37] J. Sun, P. T. Baker, A. D. Johnson, D. R. Madison, and X. Siemens, Implementation of an efficient Bayesian search for gravitational-wave bursts with memory in Pulsar Timing Array data, Astrophys. J. **951**, 121 (2023).
[38] L. Lentati, P. Alexander, M. P. Hobson, S. Taylor, J. Gair, S. T. Balan, and R. van Haasteren, Hyper-efficient model-independent bayesian method for the analysis of pulsar timing data, Phys. Rev. D **87**, 104021 (2013).
[39] S. R. Taylor, J. R. Gair, and L. Lentati, Weighing the evidence for a gravitational-wave background in the first International Pulsar Timing Array data challenge, Phys. Rev. D **87**, 044035 (2013).
[40] N. J. Cornish, L. O'Beirne, S. R. Taylor, and N. Yunes, Constraining alternative theories of gravity using Pulsar Timing Arrays, Phys. Rev. Lett. **120**, 181101 (2018).
[41] W. Ratzinger and P. Schwaller, Whispers from the dark side: Confronting light new physics with NANOGrav data, SciPost Phys. **10**, 047 (2021).
[42] D. Wang, Novel physics with International Pulsar Timing Array: Axionlike particles, domain walls and cosmic strings, arXiv:2203.10959.
[43] D. W. Scott, On optimal and data-based histograms, Biometrika **66**, 605 (1979).
[44] D. Freedman and P. Diaconis, On the histogram as a density estimator: L 2 theory, Z. Wahrsch. Verw. Geb. **57**, 453 (1981).
[45] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Stat. **33**, 1065 (1962).
[46] M. Rosenblatt, Remarks on some nonparametric estimates of a density function, Ann. Math. Stat. **27**, 832 (1956).
[47] V. A. Epanechnikov, Non-parametric estimation of a multivariate probability density, Theory Probab. Appl. **14**, 153 (1969).
[48] S. J. Sheather and M. C. Jones, A reliable data-based bandwidth selection method for Kernel density estimation, J. R. Stat. Soc. Ser. B **53**, 683 (1991).
[49] A. D. Johnson, P. M. Meyers, P. T. Baker, N. J. Cornish, J. S. Hazboun, T. B. Littenberg, J. D. Romano, S. R. Taylor, M. Vallisneri, S. J. Vigeland *et al.*, The NANOGrav 15-year gravitational-wave background analysis pipeline, arXiv:2306.16223.
[50] E. D. Hellinger, Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen J. Reine Angew. Math. **136**, 210 (1909).
[51] P. A. Rosado, A. Sesana, and J. Gair, Expected properties of the first gravitational wave signal detected with Pulsar Timing Arrays, Mon. Not. R. Astron. Soc. **451**, 2417 (2015).
[52] H. Middleton, A. Sesana, S. Chen, A. Vecchio, W. Del Pozzo, and P. A. Rosado, Massive black hole binary systems and the NANOGrav 12.5 year results, Mon. Not. R. Astron. Soc. **502**, L99 (2021).
[53] K. Pearson, VII. Note on regression and inheritance in the case of two parents, Proc. R. Soc. London **58**, 240 (1895).
[54] G. Ashton and C. Talbot, BILBY-MCMC: An MCMC sampler for gravitational-wave inference, Mon. Not. R. Astron. Soc. **507**, 2037 (2021).
[55] R. E. Kass and A. E. Raftery, Bayes factors, J. Am. Stat. Assoc. **90**, 773 (1995).
[56] J. M. Dickey, The weighted likelihood ratio, linear hypotheses on normal location parameters, Ann. Math. Stat. **42**, 1 204 (1971).
[57] B. P. Carlin and S. Chib, Bayesian model choice via Markov Chain Monte Carlo methods, J. R. Stat. Soc. Ser. B **57**, 473 (1995).
[58] S. J. Godsill, On the relationship between Markov Chain Monte Carlo methods for model uncertainty, J. Comput. Graph. Stat. **10**, 230 (2001).
[59] K. Aggarwal, Z. Arzoumanian, P. Baker, A. Brazier, M. Brinson, P. Brook, S. Burke-Spolaor, S. Chatterjee, J. Cordes, N. Cornish *et al.*, The NANOGrav 11 yr data set: Limits on gravitational waves from individual supermassive black hole binaries, Astrophys. J. **880**, 116 (2019).
[60] J. Skilling, Nested sampling, in *AIP Conference Proceedings* (American Institute of Physics, Garching, Germany, 2004), Vol. 735 pp. 395–405.
[61] J. Buchner, Nested sampling methods, Stat. Surv. **17**, 169 (2023).
[62] S. Chen, R. Caballero, Y. Guo, A. Chalumeau, K. Liu, G. Shaifullah, K. Lee, S. Babak, G. Desvignes, A. Parthasarathy *et al.*, Common-red-signal analysis with 24-yr high-precision timing of the European Pulsar Timing Array: Inferences in the stochastic gravitational-wave background search, Mon. Not. R. Astron. Soc. **508**, 4970 (2021).
[63] Z. Arzoumanian *et al.* (The NANOGrav Collaboration), The NANOGrav nine-year data set: Limits on the isotropic stochastic gravitational wave background, Astrophys. J. **821**, 13 (2016).
[64] Z. Arzoumanian, P. T. Baker, A. Brazier, P. R. Brook, S. Burke-Spolaor, B. Bécsy, M. Charisi, S. Chatterjee, J. M. Cordes, N. J. Cornish *et al.*, Multimessenger gravitational-wave searches with Pulsar Timing Arrays: Application to 3c 66b using the NANOGrav 11-year data set, Astrophys. J. **900**, 102 (2020).
[65] J. Buchner, ULtraNest—a robust, general purpose Bayesian inference engine, J. Open Source Softwaare **6**, 3001 (2021).
[66] C. J. Moore, S. R. Taylor, and J. R. Gair, Estimating the sensitivity of Pulsar Timing Arrays, Classical Quantum Gravity **32**, 055004 (2015).
[67] J. S. Hazboun, J. D. Romano, and T. L. Smith, Realistic sensitivity curves for Pulsar Timing Arrays, Phys. Rev. D **100**, 104028 (2019).
[68] N. J. Cornish and L. M. Sampson, Towards robust gravitational wave detection with Pulsar Timing Arrays, Phys. Rev. D **93**, 104047 (2016).
[69] Ź. Ivezić, A. J. Connolly, J. T. VanderPlas, and A. Gray, *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data* (Princeton University Press, Princeton, NJ, 2020).
[70] M. Bailes, E. Barr, N. Bhat, J. Brink, S. Buchner, M. Burgay, F. Camilo, D. Champion, J. Hessels, G. Janssen *et al.*, Meertime-the MeerKAT key science program on Pulsar Timing, Proc. Sci. MeerKAT2016 (**2018**) 011.

[71] M. Amiri, K. Bandura, P. Berger, M. Bhardwaj, M. Boyce, P. Boyle, C. Brar, M. Burhanpurkar, P. Chawla, J. Chowdhury *et al.*, The chime fast radio burst project: System overview, Astrophys. J. **863**, 48 (2018).

[72] C. Carilli and S. Rawlings, Science with the square kilometer array: Motivation, key science projects, standards and assumptions, New Astron. Rev. **48**, 979 (2004).

[73] G. Agazie *et al.* (The NANOGrav Collaboration), The NANOGrav 15 yr data set: Detector characterization and noise budget, Astrophys. J. Lett. **951**, L10 (2023).

[74] S. Geman and D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-6**, 721 (1984).

[75] N. Laal, W. G. Lamb, J. D. Romano, X. Siemens, S. R. Taylor, and R. van Haasteren, Exploring the capabilities of Gibbs sampling in Pulsar Timing Arrays, Phys. Rev. D **108**, 063008 (2023).

[76] D. Rezende and S. Mohamed, Variational inference with normalizing flows, in *Proceedings of the 32nd International Conference on Machine Learning* (PMLR, Lille, France, 2015), pp. 1530–1538, http://proceedings.mlr.press/v37/rezende15.html.

[77] S. Hourihane, P. Meyers, A. Johnson, K. Chatziioannou, and M. Vallisneri, Accurate characterization of the stochastic gravitational-wave background with Pulsar Timing Arrays by likelihood reweighting, Phys. Rev. D **107**, 084045 (2023).

[78] A. Mitridate, D. Wright, R. von Eckardstein, T. Schröder, J. Nay, K. Olum, K. Schmitz, and T. Trickle, PTArcade, arXiv:2306.16377.

[79] J. Ellis and R. van Haasteren, jellis18/ptmcmcsampler: Official release (2017), 10.5281/zenodo.1037579.

[80] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2018).

[81] T. Odland, tommyod/kdepy: Kernel density estimation in python (2018).

[82] S. R. Hinton, ChainConsumer, J. Open Source Softwaare **1**, 45 (2016).

[83] M. Vallisneri, LIBSTEMPO: Python wrapper for Tempo2, Astrophysics Source Code Library, record ascl:2002.017 (2020), ascl:2002.017.