Posterior predictive checking for gravitational-wave detection with pulsar timing arrays. II. Posterior predictive distributions and pseudo-Bayes factors

Patrick M. Meyers,[1,*] Katerina Chatziioannou,[2,1,†] Michele Vallisneri,[3,1,‡] and Alvin J. K. Chua[4,5,§]

[1]*Department of Physics, California Institute of Technology, Pasadena, California 91125, USA*
[2]*LIGO Laboratory, California Institute of Technology, Pasadena, California 91125, USA*
[3]*Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA*
[4]*Department of Physics, National University of Singapore, Singapore 117551*
[5]*Department of Mathematics, National University of Singapore, Singapore 119076*

(Received 11 July 2023; accepted 4 October 2023; published 6 December 2023)

The detection of nanoHertz gravitational waves through pulsar timing arrays hinges on identifying a common stochastic process affecting all pulsars in a correlated way across the sky. In the presence of other deterministic and stochastic processes affecting the time-of-arrival of pulses, a detection claim must be accompanied by a detailed assessment of the various physical or phenomenological models used to describe the data. In this study, we propose *posterior predictive checks* as a model-checking tool that relies on the predictive performance of the models with regards to new data. We derive and study predictive checks based on different components of the models, namely the Fourier coefficients of the stochastic process, the correlation pattern, and the timing residuals. We assess the ability of our checks to identify model misspecification in simulated datasets. We find that they can accurately flag a stochastic process spectral shape that deviates from the common power-law model as well as a stochastic process that does not display the expected angular correlation pattern. Posterior predictive likelihoods derived under different assumptions about the correlation pattern can further be used to establish detection significance. In the era of nanoHertz gravitational wave detection from different pulsar-timing datasets, such tests represent an essential tool in assessing data consistency and supporting astrophysical inference.

DOI: 10.1103/PhysRevD.108.123008

I. INTRODUCTION

Building on millisecond-pulsar observations spanning decades, four international pulsar-timing-array (PTA) collaborations have recently reported varying levels of evidence for a low-frequency gravitational-wave (GW) background [1–4], which is broadly expected from the binaries of supermassive black holes at the centers of galaxies [5–9], but may also have been generated by "new physics" [6,10]. The PTAs are now collaborating to compare their estimates of the amplitude, shape, and significance of the background [11].

All PTAs use similar data models, which typically include a deterministic timing model characterizing the motion of each pulsar [12,13], stochastic noise that affects each pulsar individually (dispersion measure fluctuations [14,15] and intrinsic pulsar red noise [16–18]), a GW background common to all pulsars, as well as measurement noise. The intrinsic pulsar noise and the GW background are modeled phenomenologically as finite Gaussian processes with Fourier bases functions and power law priors [19–22], although more complex models have been proposed [14,23,24]. Given that GW searches rely crucially on these phenomenological models, it is important to develop methods to identify and assess model misspecification.

The most common model-checking approach consists of modifying parts of a model and then comparing the ratio of the marginal likelihoods (i.e., the Bayes factor) between the original and modified models. However, there are two problems with adopting the Bayes factor for this task. The first is a problem of principle: in addition to a Bayes factor, model comparison requires prior odds. However, it seems very hard to assign priors to hypotheses about the very existence of the GW background and its spectral shape, or to the unphysical null models used to establish detection significance. Furthermore, no set of models exhausts the space of relevant hypotheses, which should include alternatives that embody known and unknown systematics; indeed, a faithful model may be impossible to specify formally [25]. The second problem is one of interpretation:

*pmeyers@caltech.edu
†kchatziioannou@caltech.edu
‡Michele.Vallisneri@jpl.nasa.gov
§alvincjk@nus.edu.sg

even taking model comparison at face value, it remains unclear what confidence a Bayes factor actually conveys beyond arbitrary mappings [26,27] of Bayes factors to degree-of-belief descriptors ("strong," "decisive," etc.).

Such issues aside, the central idea of model checking through Bayesian model comparison has been thoroughly explored and employed in PTA analyses. In the parlance of hierarchical inference [28,29], the description of the pulsar noise and GW background by the Gaussian-process likelihood and decomposition onto sinusoids is the *model*, while the (complex) amplitude of each sinusoid is a *parameter* of the model. The assumption that the amplitudes follow a power law is the *hierarchical model* (or *hypermodel*), and the amplitude and spectral index of the power law are *hyperparameters*. In this context, the most straightforward check involves changing elements of the (hyper)model [30–33]. For example, Ref. [30] replaced the power law with a truncated power law and Ref. [33] explored the impact of the hyperparameter priors on the marginal likelihoods. However, since the model and the hypermodel for the stochastic processes are mainly phenomenological and unlikely to be perfectly representing reality, model comparisons between these extensions do not have a clear interpretation.

We propose a complementary approach to assessing model misspecification that hinges on the predictive power of our analysis with regards to new data. In a companion paper [34] we explore predictive tests in the context of null-hypothesis testing with the *optimal statistic* [35–38]; by contrast, this study focuses on Bayesian inference. The main idea behind predictive tests is to use inference based on current data to predict further data. Comparing the prediction with future or current data then allows us to probe different elements of the analysis. Compared to tests based on perturbing a model and comparing the marginalized likelihoods, predictive tests focus naturally on specific elements of the model or the hypermodel. For example, predictive checks of the GW spectrum allow us to directly assess whether specific frequency components have been over- or underestimated. In the context of GW analyses, such tests are a common step of estimating the populations of binary black hole and binary neutron star systems [39–42]. Similar posterior predictive tests have been used on individual pulsars [43]; our study here applies these tools to full PTA data analysis.

Following the discussion of [39] we identify three types of predictive tests, each targeting a different element of the analysis.

(i) The first and least explored test relies on the hyperparameters (e.g., the GW background amplitude and spectral slope). For example, the inferred hyperparameters from one PTA dataset (say, NANOGrav), can be used to predict data and inference products for another dataset (say, PPTA), which can then be compared with the actual data and products. We leave the detailed exploration of these to future work.

(ii) The second test is based on the model parameters and specifically the Gaussian-process coefficients (i.e., the Fourier-component amplitudes). We consider these coefficients under two probability distributions. *Predicted* coefficients are conditioned on the hypermodel and the posterior for the hyperparameters given the observed data: for instance, for a power-law background model, they would span the range of GW signals expected given the amplitude and spectral slope inferred from the data. *Inferred* coefficients are conditioned on both the posterior of the hyperparameters *and* the data: for a power-law background model, they would span the range of GW-induced residuals that are compatible with the data under the power-law assumption. By comparing predicted and inferred coefficients, we are considering whether the Fourier amplitudes actually follow a power law with an assumed correlation pattern.

(iii) The third test examines the pulsar-timing residual data directly through leave-one-out cross-validation on the population of pulsars. That is, we use $N_{\rm p}-1$ pulsars to calculate the (posterior predictive) likelihood of the data observed for the $N_{\rm p}^{\rm th}$ pulsar. We assess the likelihoods in the context of model criticism, (which pulsars are not predicted well by the model fit to the other pulsars?), and model comparison (which model, fit to $N_{\rm p}-1$ pulsars, does best at predicting the residuals of the left-out pulsar?). We further propose that a summary statistic built from the posterior predictive likelihoods can be used to establish detection significance, by comparing its observed value to a null distribution obtained from simulated datasets with no GW background.

We assess tests on the Gaussian process coefficients using simulated datasets that represent different levels of model misspecification. Simulations are based on the times-of-arrival (TOAs) and noise parameters of the NANOGrav 12.5-yr dataset [44] to create synthetic residuals and include a GW signal. We consider (i) a dataset that obeys our assumptions of a GW background with a power-law spectral shape and Hellings-Downs correlations; (ii) a dataset that breaks the power-law assumption, instead having a truncated power-law spectral shape; and (iii) a dataset that breaks the correlation-pattern assumption by adding monopolar correlations. Comparing inferred and predicted coefficients allows us to identify model misspecification for both (ii) and (iii).

Switching to predictive tests with the timing residuals, we introduce a "pseudo-Bayes factor" [45], defined as the ratio of the posterior predictive likelihoods of the observed data in a pulsar given all other pulsars under a model that includes Hellings-Downs correlations and a model that assumes no spatial correlations. We compute the

pseudo-Bayes factor for simulated datasets that contain a GW background and for "null" simulations with no signal. We show that, similar to the standard drop-out factor [46], the pseudo-Bayes factor is an indicator of Hellings-Downs correlations in most pulsars. However, even in the presence of a signal some pulsars show preference against Hellings-Downs. The latter seems to be an expected feature of PTA datasets. Finally, we compare the *total* pseudo-Bayes factor, i.e., the product over all pulsars, between the datasets with and without a GW and show that it can be used as a detection statistic.

The rest of this article is organized as follows. In Sec. II we summarize PTA analyses. In Sec. III we comment on posterior predictive checks using hyperparameters. In Secs. IV and IV we propose and test posterior predictive checks for model parameters and timing data respectively. In Sec. VI we conclude.

II. PULSAR TIMING ARRAY ANALYSIS

We begin with an overview of PTA analysis with an emphasis on the modeling choices we test in the subsequent sections. For a more detailed discussion on PTA physics and analyses, see Refs. [47–49].

A. PTA model and likelihood

The arrival times of radio pulses are influenced by both deterministic and stochastic processes. Deterministic effects include the apparent and proper motion of the pulsar, as well as its orbit in a binary. A first analysis step fits a *timing model* that describes the deterministic effects and subtracts it from the arrival times to obtain the *timing residuals* $\delta\mathbf{t}$ [12,13]. Recovery of the best-fit timing model is influenced by stochastic processes such as *spin noise* [16–18]—stochastic fluctuations of the pulsar rotation frequency intrinsic to each individual pulsar—and GWs, which induce a correlated stochastic signal common to all pulsars. For example, red noise affects (among others) the estimate of the pulsar rotation period and its derivative.

Assuming that the effect of stochastic processes on the timing solution is small, most PTA analyses are based on the timing residuals $\delta\mathbf{t}$, which we use here to denote timing residuals for all pulsars concatenated into a single vector. Stochastic processes are modeled in terms of their frequency content, expressed through a matrix $\mathbf{F}$ that contains sines and cosines of different frequencies and a vector of amplitudes $\mathbf{a}$ associated with each frequency [20].[1] Additionally, the presence of red noise in the original arrival times will have shifted the best-fit timing solution from its "true" value. We correct for this effect within a linear approximation, with a known design matrix $\mathbf{M}$ of partial derivatives mapping small changes in timing model parameters $\boldsymbol{\epsilon}$ onto changes in $\delta\mathbf{t}$. Defining

$$\mathbf{T} = [\,\mathbf{M} \quad \mathbf{F}\,], \tag{1}$$

$$\mathbf{b} = \begin{bmatrix} \boldsymbol{\epsilon} \\ \mathbf{a} \end{bmatrix}, \tag{2}$$

the full model residuals are

$$\mathbf{r} = \delta\mathbf{t} - \mathbf{T}\mathbf{b}, \tag{3}$$

and under the assumption of Gaussian measurement noise the likelihood is the Gaussian distribution

$$p(\delta\mathbf{t}|\mathbf{b}) = \frac{\exp\left(-\frac{1}{2}\mathbf{r}^T\mathbf{N}^{-1}\mathbf{r}\right)}{\sqrt{\det\left(2\pi\mathbf{N}\right)}}. \tag{4}$$

For "narrowband" timing campaigns, $\mathbf{N}$ is a block-diagonal noise matrix in which the dense blocks arise due to pulse profile "jitter" noise that is correlated across arrival times taken at different radio frequency channels during the same observation [50]. If TOAs across the measurement band are condensed into single TOAs, $\mathbf{N}$ is diagonal. In what follows, we assume $\mathbf{N}$ is characterized accurately and we do not consider relevant mismodeling.

At this stage, the model parameters are the sine and cosine spectral amplitudes $\mathbf{a}$ and the timing model corrections, $\boldsymbol{\epsilon}$, though we are primarily interested in the former. In order to separate the intrinsic pulsar noise and the common GW, we place a Gaussian hyperprior on $\mathbf{b}$ in terms of the hyperparameters $\boldsymbol{\Lambda}$

$$p(\mathbf{b}|\boldsymbol{\Lambda}) = \frac{\exp\left(-\frac{1}{2}\mathbf{b}^T\mathbf{B}^{-1}\mathbf{b}\right)}{\sqrt{\det(2\pi\mathbf{B})}}, \tag{5}$$

$$\text{with } \mathbf{B} = \begin{bmatrix} \infty & 0 \\ 0 & \boldsymbol{\varphi}(\boldsymbol{\Lambda}) \end{bmatrix}. \tag{6}$$

The top-corner entries of $\mathbf{B}$ express an improper prior of infinite variance on the timing-model corrections $\boldsymbol{\epsilon}$. The matrix $\boldsymbol{\varphi}(\boldsymbol{\Lambda})$ includes the correlation of different elements of $\mathbf{b}$ via power spectra $\boldsymbol{\eta}(\boldsymbol{\Lambda})$ and $\boldsymbol{\rho}(\boldsymbol{\Lambda})$ that encode the intrinsic pulsar noise and the GW signal, respectively. Furthermore, GWs induce correlations in the same frequency bin for different pulsars based on their angular separation as prescribed by the Hellings-Downs curve. Overall for each of the sine and cosine coefficient in $\mathbf{a}$,

$$\boldsymbol{\varphi}(\boldsymbol{\Lambda})_{(ai,bj)} = \Gamma_{ab}\rho_i^2(\boldsymbol{\Lambda})\delta_{ij} + \eta_{ai}^2(\boldsymbol{\Lambda})\delta_{ab}\delta_{ij}, \tag{7}$$

where a and b label pulsars, and i and j label frequencies. The GW power spectrum at a given frequency is captured by $\rho_i(\boldsymbol{\Lambda})$, the Hellings-Downs curve by Γ_{ab}, and the power spectrum of the intrinsic pulsar noise associated with each individual pulsar at each individual frequency by $\eta_{ai}(\boldsymbol{\Lambda})$.

[1] Time-domain approaches have also been considered [22].

A stronger assumption is that both $\eta_{ai}(\mathbf{\Lambda})$ and $\rho_i(\mathbf{\Lambda})$ follow a power law

$$\rho_i^2(\mathbf{\Lambda}) = \frac{A_{\rm gw}^2}{12\pi^2}\left(\frac{f_i}{f_{\rm y}}\right)^{-\gamma_{\rm gw}}\frac{f_{\rm y}^{-3}}{T}, \tag{8}$$

$$\eta_{ai}^2(\mathbf{\Lambda}) = \frac{A_{a,\rm int}^2}{12\pi^2}\left(\frac{f_i}{f_{\rm y}}\right)^{-\gamma_{a,\rm int}}\frac{f_{\rm y}^{-3}}{T}, \tag{9}$$

where $A_{\rm gw}$ is the amplitude of the GW background at $f_{\rm y}$, $f_i = i/T$ is the frequency of the ith bin, $f_{\rm y} = (1\ {\rm y})^{-1}$, and T is the dataset duration. Throughout, we use $i \in [1\text{–}10]$ for the GW background ($f = 2.5\text{–}24.6$ nHz) and $i \in [1-30]$ for the intrinsic red noise.[2]

Under the power-law assumption, the model *hyperparameters* $\mathbf{\Lambda}$ are the GW amplitude $A_{\rm gw}$ and the spectral index $\gamma_{\rm gw}$, and an intrinsic pulsar noise amplitude $A_{a,\rm int}$ and spectral index $\gamma_{a,\rm int}$ for each of the $N_{\rm p}$ pulsars. The posterior on these hyperparameters is obtained by marginalizing over the model parameters $\mathbf{b}$,

$$\begin{aligned} p(\mathbf{\Lambda}|\delta\mathbf{t}) &= \int d\mathbf{b}\, p(\delta\mathbf{t}|\mathbf{b})p(\mathbf{b}|\mathbf{\Lambda})p(\mathbf{\Lambda}) \\ &= \frac{p(\mathbf{\Lambda})}{\sqrt{\det(2\pi\mathbf{C})}}\exp\left(-\frac{1}{2}\delta\mathbf{t}^T\mathbf{C}^{-1}\delta\mathbf{t}\right), \end{aligned} \tag{10}$$

where the new covariance matrix is $\mathbf{C} \equiv (\mathbf{N} + \mathbf{TBT}^T)$, and $p(\mathbf{\Lambda})$ is the prior on the hyperparameters. Alternatively, the first two terms in the integrand of Eq. (10) can be written as a posterior, $p(\mathbf{b}|\delta\mathbf{t}, \mathbf{\Lambda})$, which is normal with mean and covariance given respectively by

$$\hat{\mathbf{b}} = \mathbf{\Sigma}\mathbf{T}^T\mathbf{N}^{-1}\delta\mathbf{t}, \tag{11}$$

$$\mathbf{\Sigma} = (\mathbf{T}^T\mathbf{N}^{-1}\mathbf{T} + \mathbf{B}^{-1})^{-1}. \tag{12}$$

Given the large dimensionality ($2N_{\rm p} + 2$ hyperparameters for a typical analysis), most GW analyses estimate the marginalized posterior on the hyperparameters $\mathbf{\Lambda}$ through stochastic sampling, resulting in N_s samples $\{\mathbf{\Lambda}^s\}_{s=1}^{N_s}$ drawn from their posterior,

$$\mathbf{\Lambda}^s \sim p(\mathbf{\Lambda}|\delta\mathbf{t}). \tag{13}$$

In Sec. III, we propose methods to assess how well the models and assumptions of this section fit the data based on having obtained $\mathbf{\Lambda}^s$.

[2]We use 10 frequencies for the GW background as opposed to the 5 frequencies used in [46] because we have injected a signal that is stronger than the common process observed in that analysis.

B. Simulated datasets

We experiment with our proposed methods by analyzing simulated datasets. We consider a total of four datasets, each spanning 12.9 years of data over 45 pulsars, and produce one realization for each of those datasets.

(i) HellingsDowns-PowerLaw: Constructed in accordance with the assumptions described in Sec. II A, this dataset contains a GW signal described by a power law with $\log_{10} A_{\rm gw} = -14$ and $\gamma_{\rm gw} = 13/3$, see Eq. (8). The Hellings-Downs correlations are detectable with an optimal-statistic signal-to-noise ratio (SNR) of 5.5.

(ii) HellingsDowns-Turnover: Constructed to test the power-law assumption, this dataset contains a GW signal described by the broken power law

$$\rho^2(f) = \frac{A_{\rm gw}^2}{12\pi^2}\left(\frac{f}{f_{\rm yr}}\right)^{-\gamma_{\rm gw}}\left[1+\left(\frac{f_{\rm b}}{f}\right)^{\kappa}\right]^{-1}\frac{f_{\rm y}^{-3}}{T}, \tag{14}$$

with $\gamma_{\rm gw} = 13/3$, $\log_{10} A_{\rm gw} = -13.5$, $f_{\rm b} = 7.9$ nHz, and $\kappa = 26/3$.[3] The optimal statistic SNR is 4.4.

(iii) HellingsDownsMonopole-PowerLaw: The third dataset focuses on spatial correlations and includes a power-law GW signal with $\log_{10} A_{\rm gw} = -14$ and $\gamma_{\rm gw} = 13/3$ as well as a stochastic process with $\log_{10} A_{\rm m} = -14.3$ and $\gamma_{\rm m} = 13/3$ that induces monopolar correlations across the pulsars ($\Gamma_{ab} = 1$). The optimal-statistic SNR is 6.[4]

(iv) NoGravitationalWave: Finally, we consider a dataset without any common process between the pulsars, setting $A_{\rm gw} = 0$.

Hyperparameters for the intrinsic pulsar noise are chosen from the posteriors of the NANOGrav 12.5-yr dataset [44,51]. We simulate data by first drawing from the posterior distribution on the intrinsic pulsar noises $A_{a,\rm int}^{\rm sim}, \gamma_{a,\rm int}^{\rm sim} \sim p(A_{a,\rm int}, \gamma_{a,\rm int}|\delta\mathbf{t}^{\rm NG12.5})$. The GW parameters are specified independently and listed above, thus completing the list of simulated hyperparameters $\mathbf{\Lambda}^{\rm sim}$. We then draw Gaussian process coefficients as $\mathbf{a}^{\rm sim} \sim p(\mathbf{a}|\mathbf{\Lambda}^{\rm sim})$ and set the timing parameters $\boldsymbol{\epsilon}^{\rm sim} = 0$. Finally, we draw simulated timing residuals from the Gaussian likelihood, $\delta\mathbf{t}^{\rm sim} \sim p(\delta\mathbf{t}|\boldsymbol{b}^{\rm sim})$.

Each dataset $\delta\mathbf{t}^{\rm sim}$ is analyzed with the standard model that assumes a GW signal with a power-law spectrum. The only quantity that the predictive tests rely on is $p(\mathbf{\Lambda}|\delta\mathbf{t}^{\rm sim})$, i.e., the posterior for the hyperparameters, which we estimate through stochastic sampling with ENTERPRISE [52]. For computational efficiency, we ignore

[3]This value is chosen for illustrative purposes, as it produces a noticeable turnover at low frequencies. It does not correspond to a specific astrophysical scenario.

[4]This SNR is calculated assuming only Hellings-Downs correlations.

Hellings-Downs correlations during sampling as the posterior for the hyperparameters is dominated by the auto-correlation terms [53–56].

III. PREDICTIVE CHECKS ON HYPERPARAMETERS

The most straightforward posterior predictive test performs comparisons directly at the level of the hyperparameters $\mathbf{\Lambda}$. In practise, this entails analyzing subsets of the data, for example by splitting the data of one PTA into two parts, or by analyzing data from one PTA only. The inferred GW amplitude and spectral slope are then used to predict the properties of the remaining data. However, given that current datasets are merely on the brink of making detections, splitting the data on one PTA will likely yield two uninformative datasets.

Such predictive tests are related to consistency tests that directly contrast results across different PTAs, for example the posterior comparisons between EPTA, PPTA, and NANOGrav [57]. That comparison used the Mahalanobis distance [58] for the σ deviations between two >1-dimensional distributions, and found at most a 2.6σ deviation between different PTAs. We do not consider such tests in this study any further, instead leaving them to future work.

IV. PREDICTIVE CHECKS ON MODEL PARAMETERS

The second posterior predictive test is based on the model parameters, and specifically the Gaussian process coefficients $\mathbf{a}$. The comparison of the predicted and the inferred coefficients allows us to evaluate the power-law assumption of Eqs. (8) and (9), as well as the assumption that the spatial correlations between pulsars follows the Hellings-Downs curve.

The *inferred* Gaussian-process coefficients are simply the inferred coefficients of the data. Stated differently, they are the Gaussian-process coefficients conditioned on the observed residuals, under the hypermodel prior. Given the full posterior for model and hypermodel parameters $p(\mathbf{\Lambda}, \mathbf{b}|\delta\mathbf{t})$, Eq. (10) marginalizes over the parameters $\mathbf{b}$ to obtain the posterior for the hyperparameters. Here we instead marginalize over the hyperparameters (and the timing-model parameters) to obtain the posterior for the Gaussian-process coefficients of the stochastic processes,

$$p_{\rm inf}(\mathbf{a}|\delta\mathbf{t}) = \int d\mathbf{\Lambda} d\boldsymbol{\epsilon}\, p(\mathbf{a}, \boldsymbol{\epsilon}|\mathbf{\Lambda}, \delta\mathbf{t}) p(\mathbf{\Lambda}|\delta\mathbf{t}). \quad (15)$$

The first term in the integral is the posterior on $\mathbf{b} = [\boldsymbol{\epsilon}, \mathbf{a}]$ conditioned on both the timing residuals (i.e., the data $\delta\mathbf{t}$) and the hyperparameters $\mathbf{\Lambda}$. In other words, $p_{\rm inf}(\mathbf{a}|\delta\mathbf{t})$ is the posterior of the Gaussian-process coefficients under the hyperprior assumption that the observed data are subject to a common stochastic process and (optionally) Hellings-Downs-induced correlations from the inferred GW background.[5]

The *predicted* coefficients instead are only conditioned on the hyperparameter posterior, and not on the data directly:

$$\begin{aligned} p_{\rm pre}(\mathbf{a}|\delta\mathbf{t}) &= \int d\mathbf{\Lambda} d\boldsymbol{\epsilon}\, p(\mathbf{a}, \boldsymbol{\epsilon}|\mathbf{\Lambda}) p(\mathbf{\Lambda}|\delta\mathbf{t}) \\ &= \int d\mathbf{\Lambda}\, p(\mathbf{a}|\mathbf{\Lambda}) p(\mathbf{\Lambda}|\delta\mathbf{t}). \end{aligned} \quad (16)$$

Compared to Eq. (15), the first term in the integral is *not* conditioned on $\delta\mathbf{t}$.

The various terms in the integrands of Eqs. (15) and (16) can be computed as follows. The hyperparameter posterior $p(\mathbf{\Lambda}|\delta\mathbf{t})$ is obtained by stochastic sampling via the analysis described in Sec. II. The Gaussian-process coefficients conditioned on the hyperparameters are, by definition, given by a simplification of Eq. (5)

$$p(\mathbf{a}|\mathbf{\Lambda}) = \frac{\exp\left(-\frac{1}{2}\mathbf{a}^T \boldsymbol{\varphi}^{-1}(\mathbf{\Lambda})\mathbf{a}\right)}{\sqrt{\det(2\pi\boldsymbol{\varphi}(\mathbf{\Lambda}))}}. \quad (17)$$

The Gaussian-process coefficients and timing parameters conditioned on the hyperparameters and the data are

$$p(\mathbf{a}, \boldsymbol{\epsilon}|\mathbf{\Lambda}, \delta\mathbf{t}) = \frac{p(\delta\mathbf{t}|\mathbf{a}, \boldsymbol{\epsilon}, \mathbf{\Lambda}) p(\mathbf{a}, \boldsymbol{\epsilon}|\mathbf{\Lambda})}{p(\delta\mathbf{t}|\mathbf{\Lambda})} = \mathcal{N}(\hat{\mathbf{b}}, \mathbf{\Sigma}), \quad (18)$$

where in the first equality we have used Bayes' theorem and $\mathcal{N}(\hat{\mathbf{b}}, \mathbf{\Sigma})$ indicates a normal distribution with mean and covariance given by Eqs. (11) and (12).

To construct the predicted coefficients we sample Eq. (16) by first drawing $\mathbf{\Lambda}^s \sim p(\mathbf{\Lambda}|\delta\mathbf{t})$, then using the sample $\mathbf{\Lambda}^s$ to construct $\boldsymbol{\varphi}^s(\mathbf{\Lambda})$ and draw from Eq. (17). The amplitude of these coefficients should, on average, be consistent with the assumed power-law model.[6] To construct the inferred coefficients we sample Eq. (15) by first drawing $\mathbf{\Lambda}^s \sim p(\mathbf{\Lambda}|\delta\mathbf{t})$, then using the sample $\mathbf{\Lambda}^s$ to construct $\boldsymbol{\varphi}^s(\mathbf{\Lambda})$, $\hat{\mathbf{b}}^s$, and $\mathbf{\Sigma}^s$ and draw from Eq. (18). The amplitude of the inferred coefficients has a power-law hyperprior, but is also conditioned on the data and can thus deviate from a pure power law.

Besides the assumption of a power-law common process, we can further use the inferred and predicted distributions

[5]In certain cases, stochastic sampling might yield the full posterior $p(\mathbf{b}, \mathbf{\Lambda}|\delta\mathbf{t})$, in which case $p(\mathbf{a}|\delta\mathbf{t})$ can be obtained by marginalizing over $\mathbf{\Lambda}$ and $\boldsymbol{\epsilon}$. This is typically not the case for PTA analyses that sample from the marginalized posterior of Eq. (10), we therefore have to reconstruct $p(\mathbf{a}|\delta\mathbf{t})$ using Eq. (15).

[6]This is true if γ for the power-law model is fixed. If the spectral index is sampled over then the power reconstructed from an individual draw for $\mathbf{a}$ will, on average, be consistent with a power law associated with the γ for that specific draw.

to test the nature of the spatial correlations. Both Eqs. (17) and (18) depend on $\boldsymbol{\varphi}(\boldsymbol{\Lambda})$, whose nondiagonal terms encode the inter-pulsar correlations. We can therefore evaluate the inferred and predicted distributions by assuming a correlation pattern, such as Hellings-Downs or monopolar correlations. On average, the predicted coefficients will have the assumed correlation pattern. The inferred coefficients will have a correlation pattern informed by the data, but subject to the hyperprior of a power-law common process with the assumed correlation pattern. A discrepancy between these predicted and inferred distributions would signal that the assumed pattern is not consistent with the data. In this work, we focus on visual discrepancies that can be seen from the figures, however, one could also consider constructing associated p-values [34].

A. Intrinsic noise model

We begin by applying the above methodology to pulsar intrinsic noise, which is modeled with Eq. (9). The relevant model parameters are the sine and cosine amplitudes associated with each frequency, $a_{i,a}^{(s)}$ and $a_{i,a}^{(c)}$ respectively for pulsar a and frequency bin i. Specifically, we use Eqs. (17) and (18) to draw from the inferred and predicted distribution of the intrinsic noise in pulsar a and frequency bin i and then obtain the total power as the square-sum of the sine and cosine components,

$$\eta_{ai}^2 = \frac{1}{2}\left\{\left[a_{i,a}^{(s)}\right]^2 + \left[a_{i,a}^{(c)}\right]^2\right\}. \tag{19}$$

Each of $a_{i,a}^{(s)}$ and $a_{i,a}^{(c)}$ is normally distributed according to the intrinsic-pulsar-noise power spectrum, Eq. (17), so the total power at each frequency follows a χ^2 distribution with 2 degrees of freedom for a given $\boldsymbol{\Lambda}^s$.

Results for a representative pulsar are shown in Fig. 1 using the HellingsDowns-PowerLaw simulated dataset. We show inferred (blue) and predicted (orange) spectra as a function of frequency. For reference, we also show the injected and maximum *a posteriori* spectrum. The inferred power is only significantly constrained away from zero at the fourth frequency bin, while the predicted power are wider. In most bins, the inferred and predicted distributions have comparable width (given the logarithmic scale on the y axis), suggesting that the data are not strongly informative. The inferred and predicted distributions overlap for all frequencies, as expected since the simulated dataset includes intrinsic noise that obeys the power-law assumption.

B. GW-background model

We now turn our attention to arguably the most important part of the analysis: the GW background. Detection of the GW background hinges on establishing that the data follow the Hellings-Downs correlation pattern, while the astrophysical interpretation of the signal relies on its spectral shape, specifically the amplitude and slope of the assumed power law [30,59–61]. Below we apply posterior predictive checks to assess both elements.

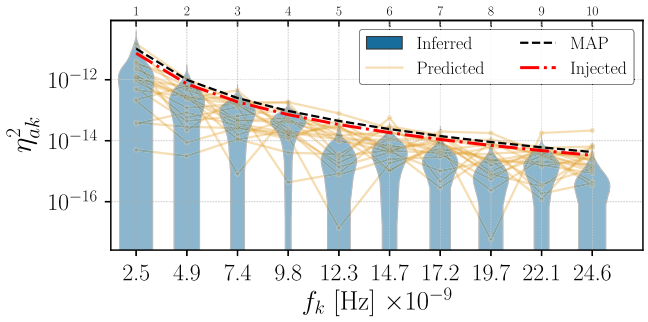

FIG. 1. Intrinsic pulsar noise power, η_{ak}^2 from Eq. (9), as a function of frequency f_k (bottom x-axis, k index on the top x-axis) for B1937 + 21 for the HellingsDowns-PowerLaw simulated dataset. Sample predicted power spectra are shown in orange. The blue violins show the posterior for the inferred power at each frequency, which is a combination of the data and the power-law prior. For reference, we plot the injected and maximum *a posteriori* power-law spectra in red dot-dashed and black dashed lines respectively.

1. GW power spectrum of individual pulsars

While the GW background has a single power spectrum across all pulsars as in Eq. (8), the exact realization in each pulsar is unique,[7] and this results in different Gaussian process coefficients. We therefore begin by considering the inferred and predicted GW power in individual pulsars. Figure 2 shows power spectra (left) and power distributions for frequency bins of interest (right) for an "informative" pulsar with detectable GW power in some bins. The top panels show results for the HellingsDowns-PowerLaw dataset, while the bottom panels correspond to HellingsDowns-Turnover. Both datasets are analyzed with the same GW model, hence the maximum *a posteriori* draw and the predicted spectra are power laws.

The posterior predictive test proceeds as follows. First, we analyze the data assuming a power-law model and (inevitably) infer power-law parameters that fit the data as well as possible. The predicted spectra are draws from this inferred power law. The maximum *a posteriori* draw is essentially the power-law model's best attempt to match the true spectrum. Second, the inferred spectra are the power in the data inferred under a GW spectrum prior that is the inferred power-law posterior. The final inferred spectra are thus a combination of the data and the prior. For informative pulsars, in a few of the frequency bins the data dominate over the power-law prior. For uninformative pulsars, on the other hand, the inferred spectra would be consistent with the power law imposed by the prior in all bins.

[7]This is in part, but not solely, due to the "pulsar term" that Hellings-Downs correlations do not capture.

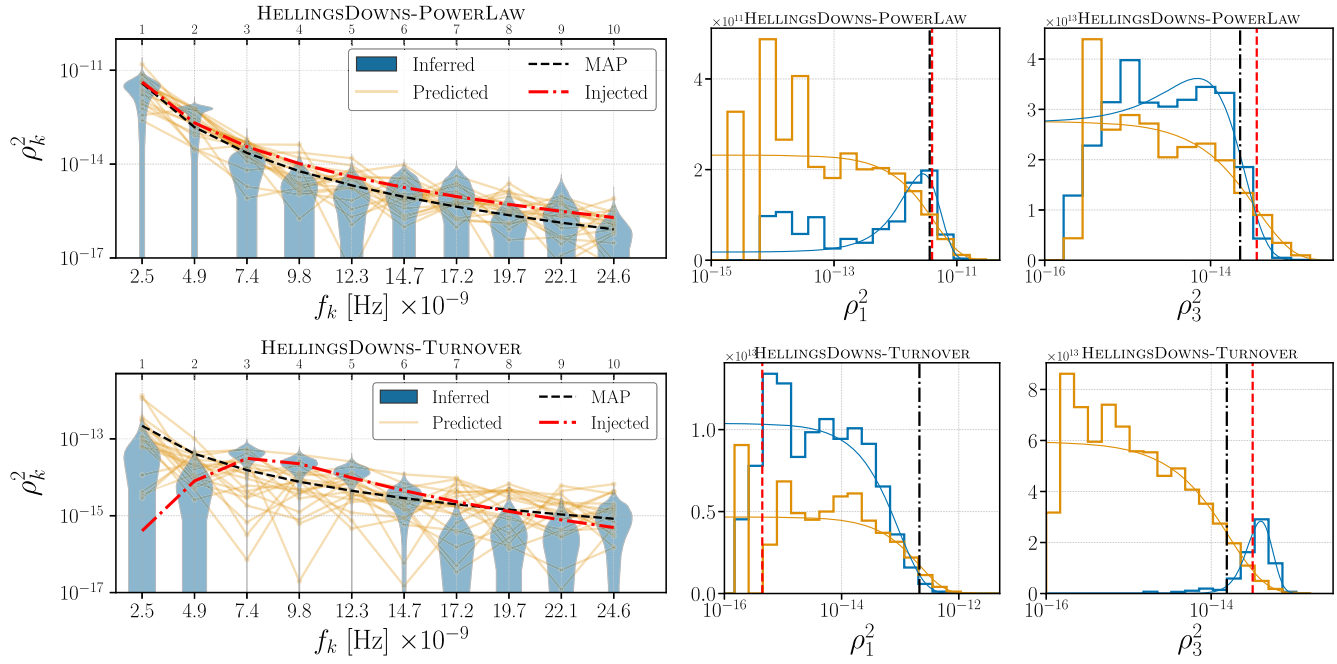

FIG. 2. GW power spectrum, ρ_k^2 from Eq. (8), as a function of the frequency f_k (left) and power distributions for select frequency bins (right) for an "informative" pulsar, J1909–3744, for the HellingsDowns-PowerLaw (top) and the HellingsDowns-Turnover (bottom) dataset. In the left panels, sample predicted power spectra are shown in orange and blue violins show the posterior for the inferred power at each frequency. For reference, we plot the injected and maximum *a posteriori* spectra in red dot-dashed and black dashed lines respectively. In the right panels, we show histograms of the inferred and predicted power for the 1st and 3rd bins, along with a fit to a χ^2 distribution with two degrees of freedom. In the top panels, we find agreement between the predicted and inferred spectra for the data-informed frequency bins, i.e., the ones constrained away from zero. In the bottom panel, data-informed bins contain systematically higher power than the prediction, as expected from the injected spectra.

Indeed, in Fig. 2 the 1st–2nd (top) and 3rd–6th (bottom) frequency bins have inferred spectra that are constrained away from zero. The inferred spectra in these bins are narrower than the predicted ones, suggesting informative data. In the top panel the inferred and predicted spectra fully overlap since the model matches the simulated spectrum. In the bottom panel, however, the inferred spectra are systematically higher than the predicted ones. Moreover, the 1st–2nd bins are consistent with zero, which is in tension with expectations from a power law. This behavior is due to the fact that the injection follows a power law with a turnover, which the GW power-law model cannot fully match, as manifest in the maximum *a posteriori* draw. The inferred spectra are therefore dominated by the data and reveal a tension with the predicted spectra.

Though not explicitly plotted, we have verified that for uninformative pulsars, i.e., pulsars with high intrinsic noise with no detectable GW power, the inferred and predicted distributions are nearly identical. This suggests that the total inference is dominated by the prior.

2. Total GW power spectrum

In order to obtain an estimate of the total GW power spectrum, we use the optimal statistic [35–37], which is based on the timing residuals from all pulsars. The optimal statistic gives a noise-weighted average of the cross-correlation between pulsar pairs, and therefore allows us to synthesize the inferred or predicted coefficients from different pulsars into a single estimate of the GW background amplitude. Since we are testing the GW model, we reconstruct the optimal statistic using only the GW contribution to the timing residuals and ignore the timing model and intrinsic pulsar noise parts.

We obtain draws for the Gaussian process coefficients $\mathbf{a}^s$ of the GW background through Eqs. (15) or (16) as applicable, and construct timing residuals $\delta\mathbf{t}^s = \mathbf{F}\mathbf{a}^s$. We then use the optimal statistic to compute inter-pulsar cross-correlations $\xi^s_{ab,k}$ and GW background amplitude $A^s_{\rm gw}$ for each frequency bin k.[8] For a pair of pulsars a and b, the former is

$$\xi_{ab,k} = \frac{\delta\mathbf{t}_a^T \mathbf{D}_a^{-1} \tilde{\mathbf{\Phi}}^{\rm gw}_{ab,k} \mathbf{D}_b^{-1} \delta\mathbf{t}_b}{\mathrm{tr}(\mathbf{D}_a^{-1} \tilde{\mathbf{\Phi}}^{\rm gw}_{ab,k} \mathbf{D}_b^{-1} \tilde{\mathbf{\Phi}}^{\rm gw}_{ba,i})}, \tag{20}$$

$$\sigma^2_{ab,k} = [\mathrm{tr}(\mathbf{D}_a^{-1} \tilde{\mathbf{\Phi}}^{\rm gw}_{ab,k} \mathbf{D}_b^{-1} \tilde{\mathbf{\Phi}}^{\rm gw}_{ba,i})]^{-1}, \tag{21}$$

[8]This "per-frequency" optimal statistic as compared to the most common summed-over-frequencies version is studied in [62].

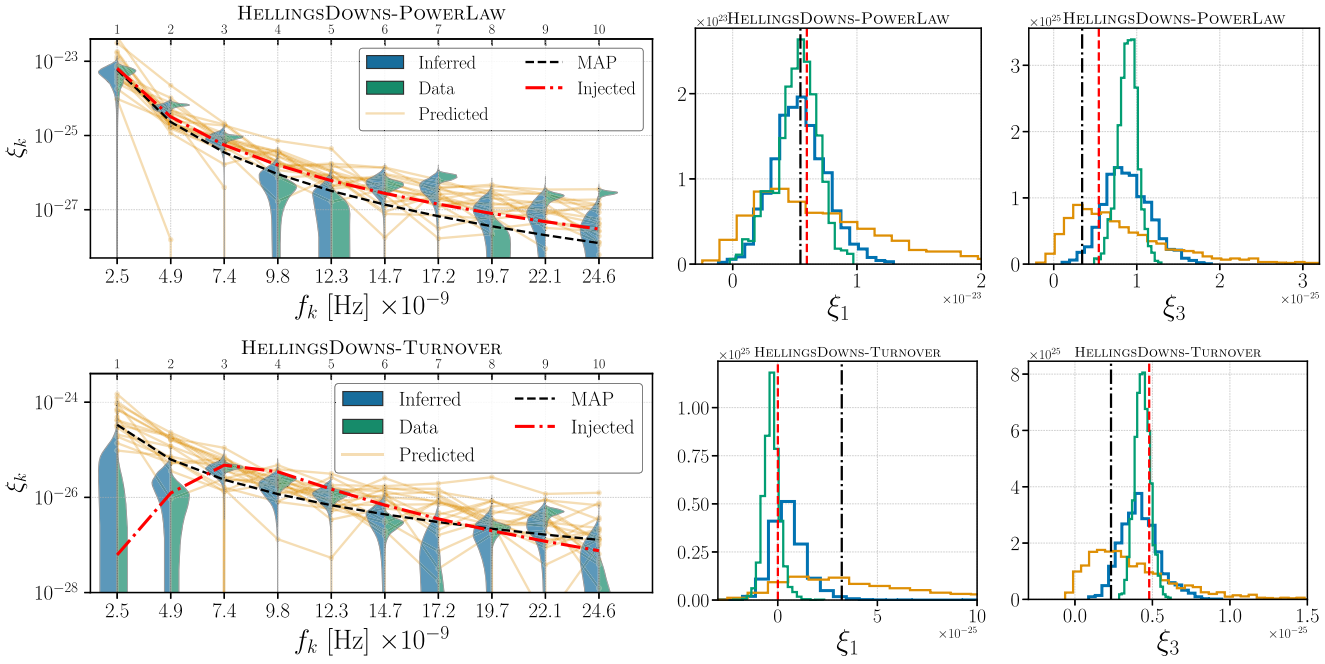

FIG. 3. Total GW power spectrum, ξ_k from Eq. (23), as a function of frequency f_k (left) and total power distributions for select bins (right) for the HellingsDowns-PowerLaw (top) and HellingsDowns-Turnover (bottom) datasets. In the left panels we show the inferred (blue left violins) and predicted (orange lines) distributions using the single-frequency optimal statistic. The injected and maximum *a posteriori* power-law spectra are shown in red dot-dashed and black dashed lines respectively. Right green violins show the power as inferred directly from the data without conditioning on a power-law spectrum. In the right panels, we show histograms of the inferred and predicted power for the 1st and 3rd bins, along with a fit to a χ^2 distribution with two degrees of freedom. Inferred and predicted spectra are consistent in the top panel. However, the inferred power in the 1st and 2nd frequency bins in the bottom panel is lower than what predicted under the power-law model.

where no summation is implied. In the above equations we have defined $\tilde{\mathbf{\Phi}}^{\rm gw}_{ab,k} = \mathbf{F}_{a,k}\tilde{\boldsymbol{\varphi}}^{\rm gw}_{ab,k}\mathbf{F}^T_{b,k}$ where

$$\tilde{\boldsymbol{\varphi}}^{\rm gw}_{ab,k} = \Gamma_{ab}\frac{1}{12\pi^2}\frac{f_{\rm y}^{-3}}{T}, \tag{22}$$

is a GW-only normalized version of Eq. (7). The subscripts in $\mathbf{F}_{a,k}$ denote that it is evaluated at the times for which pulsar a has data and for *only* frequency k. The matrix $\mathbf{D}_a = [\mathbf{C}(\mathbf{\Lambda})]_{(ai,aj)}$ is the autocorrelation block for pulsar a of the marginalized covariance matrix used in Eq. (10), and depends on the hyperparameters $\mathbf{\Lambda}$. It represents the total noise autocorrelation for pulsar a from both uncorrelated and correlated processes. The normalization in Eq. (22) is chosen such that $\xi_{ab,k}$ is an estimator for the GW background in each frequency bin.

Given $\xi_{ab,k}$ we construct a bin-by-bin estimator for the GW background obtained through a weighted average across all pulsar pairs,

$$\xi_k = \frac{\sum_{ab}\xi_{ab,k}\sigma^{-2}_{ab,k}}{\sum_{ab}\sigma^{-2}_{ab,k}}, \tag{23}$$

$$\sigma^2_k = \left[\sum_{ab}\sigma^{-2}_{ab,k}\right]^{-1}. \tag{24}$$

These equations assume independent frequency bins and pair correlations, which is not strictly true [62]. In the weak-GW limit, the frequency bins and paired correlations are approximately uncorrelated, but for strong signals such as those that we inject here the covariances between pair correlations become significant [62–67]. We nevertheless ignore them in this work for the sake of computational efficiency. Including them would broaden the green and blue violins for both the spectral and correlation reconstructions in Figs. 3 and 5 [62].

Figure 3 shows the total GW spectrum (left) and power distributions for select bins (right) for the HellingsDowns-PowerLaw (top) and HellingsDowns-Turnover (bottom) datasets. We present the same inferred, predicted, maximum *a posteriori*, and injected spectra as in Fig. 2. Additionally, we calculate Eqs. (23) and (24) directly using the original simulated data and obtain an estimate that is informed solely by the data without assumptions about the GW spectral shape. The various spectra represent the optimal statistic calculated on the predicted, inferred, and simulated data for the same set of posterior samples drawn from $p(\mathbf{\Lambda}|\delta\mathbf{t})$. For the inferred and predicted case, the hyperparameters are used to construct the GW coefficients $\mathbf{a}^s$ and $\mathbf{D}_a$, while for the data, the hyperparameters are only needed in the construction of $\mathbf{D}_a$. The predicted estimate

corresponds to power-law spectra whose amplitude and slope have been inferred by the data. The inferred estimate is a combination of data and prior: it corresponds to the GW spectrum as observed by all pulsars and under the assumption of a power law. Thus, the predicted estimate will always follow a power law, while the inferred estimate will shift the spectra as close to a power law as the data allow.

Starting with the top panel of Fig. 3 and the HellingsDowns-PowerLaw dataset, we find that the predicted and inferred data on average overlap with some scatter. In places where the data contain higher power than the injected power law, e.g., 6th and 7th frequency bins, the inferred estimate is wider and shifted down toward the power law. In some cases, such as the 9th and 10th bins, what looks like a GW detection from the data turns out to be insignificant when estimated in the context of the power-law model. Despite these, for the most informative 1st, 2nd, and 3rd bins, the observed data fully agree with the power-law model as expected.

Moving to the bottom panel of Fig. 3 and the HellingsDowns-Turnover dataset, the spectra comparison is drastically different. The most significant bins are now the 3rd, 4th, and 5th ones as expected from the injected spectrum shape. These bins agree with the predicted distribution, suggesting that they largely drive the inference of the power-law amplitude. However, the 1st and 2nd bins are consistent with no GW power and are systematically lower than the power-law model prediction. As expected, the inferred distribution is shifted upward compared to the data-only distribution, attempting to match the power-law model. However, the data place strong upper limits on the GW power in those bins and the tension between the predicted and inferred distributions is apparent.

Beyond the full distributions shown in Fig. 3, we compare the various spectra estimates on a draw-by-draw basis in Fig. 4. We show a scatter plot of ξ_1 for 300 posterior draws from the HellingsDowns-PowerLaw (top) and HellingsDowns-Turnover (bottom) datasets. The x-axis shows the value calculated on the measured data, while the y-axis shows the predicted and inferred ξ_1. In the top panel, inferred draws are narrower than predicted draws and stay close to the $x - y$ line, an outcome of the fact that the data are very informative in this bin. In the bottom panel the inferred draws are more weakly correlated with the data draws, and shifted upward due to the power-law prior. Additionally, the bulk of the predicted draws overlap with the inferred ones in the top panel, which we expect because the model used for the predicted draws matches the injected model. In the bottom panel the predicted draws have a larger tail toward higher values, as the power-law model overestimates the GW power in this frequency bin.

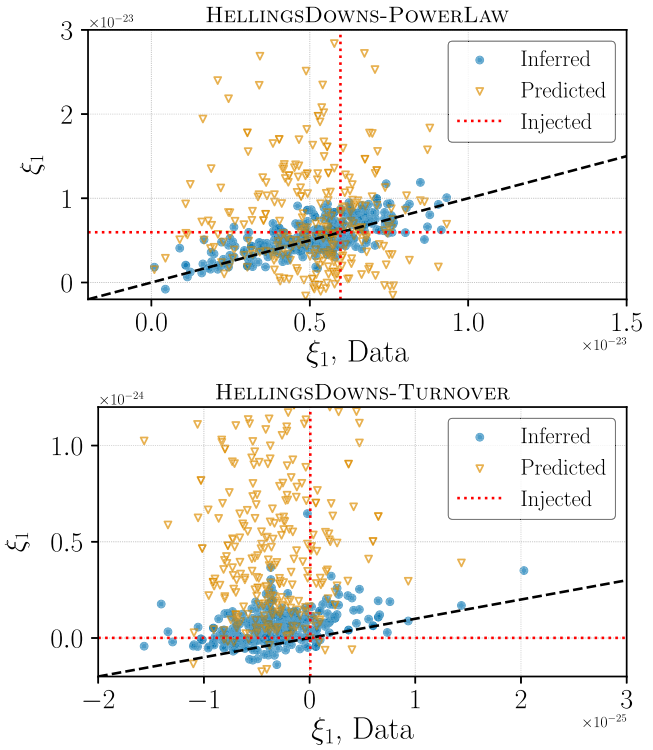

FIG. 4. Scatter plot comparison of the power in the first frequency bin for the data only vs. the predicted (orange) and inferred (blue) power for the HellingsDowns-PowerLaw (top) and HellingsDowns-Turnover (bottom) datasets. Each point is a draw from the distributions shown in Fig. 3. In the top panel the bulk of the predicted and inferred draws overlap, while the inferred draws follows the $x - y$ lines as expected from highly informative data. In the bottom panel, the predicted draws overestimate the GW power.

3. Spatial correlations

The predicted and inferred data can also be compared to assess consistency with the Hellings-Downs correlation pattern. We correlate data between pulsars using the full frequency band version of Eq. (20), i.e., we use the full $\boldsymbol{\varphi}^{\rm gw}_{ab}$ instead of $\boldsymbol{\varphi}^{\rm gw}_{ab,k}$, so we drop the subscript k and write ξ_{ab}. Additionally, since the Hellings-Downs model is already built in to the optimal statistic, we divide Eq. (20) by Γ_{ab} and Eq. (21) by Γ^2_{ab}. We denote these "normalized" correlations with $\tilde{\xi}_{ab} \equiv \xi_{ab}/\Gamma_{ab}$. Finally, we collect the $\tilde{\xi}_{ab}$'s into 8 bins (each containing approximately the same number of pulsar pairs) based on the pair angular separation θ_{ab} through an inverse noise weighted average.

Results are shown in Fig. 5 for the data, inferred, and predicted distributions. The top panel corresponds to the HellingsDowns-PowerLaw dataset, while the bottom panel to HellingsDownsMonopole-PowerLaw. In the top panel, the inferred and predicted distributions overlap, to within expected scatter. In the bottom panel, although the distributions overlap for any given angular bin, the predicted distributions are systematically shifted downward. This is because the inferred distributions contain a monopole, while the predicted ones are solely based on Hellings-Downs correlations.

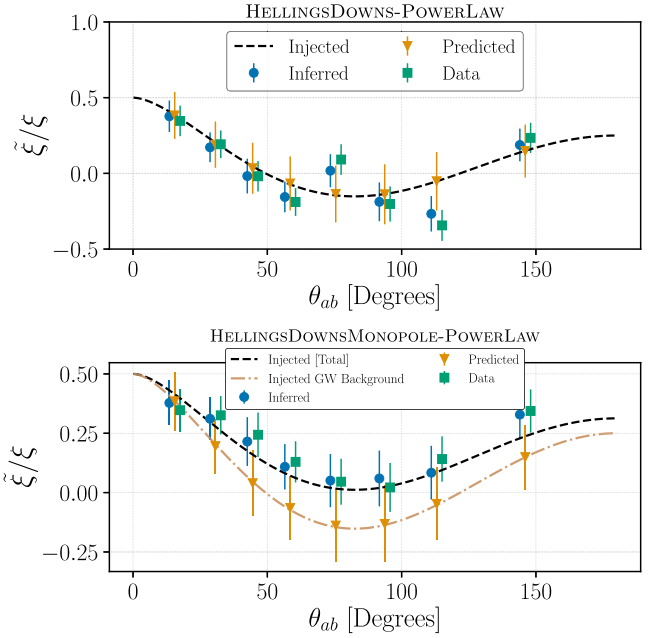

FIG. 5. Spatial correlations (median and 68% credible intervals) as a function of pulsar pair angular separation θ_{ab} for the HellingsDowns-PowerLaw (top) and HellingsDowns-Monopole-PowerLaw (bottom) datasets. We show the inferred (blue) and predicted (orange) correlations as a function of pulsar angular separation. We also show the correlations as inferred from solely the data (green). The black dashed line shows the injected correlation, while the Hellings-Downs correlations are shown in orange dot-dashed in the bottom panel. In the bottom panel, the predicted correlations are systematically lower than the inferred ones.

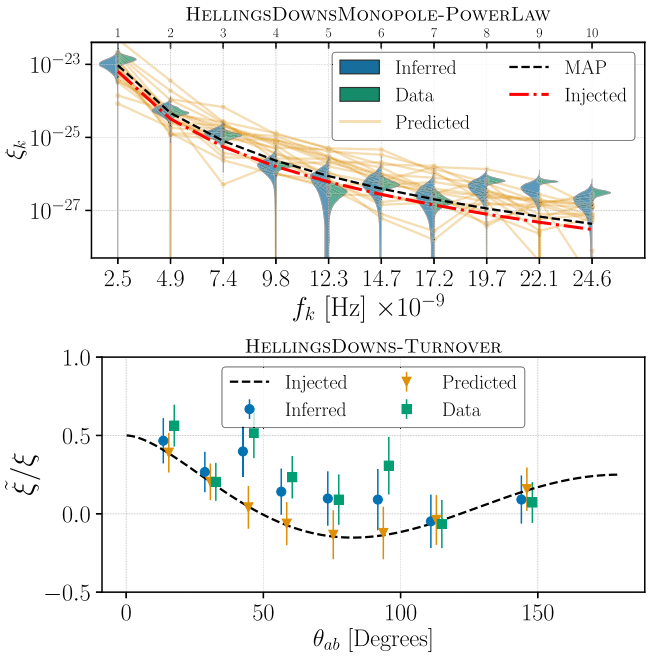

FIG. 6. Total GW spectrum for the HellingsDowns-Turnover (bottom) and spatial correlations for the HellingsDowns-Monopole-PowerLaw (top) datasets. Plotted quantities and colors are similar to Figs. 3 and 5 A correlation mismodeling does not manifest in the spectrum comparison (top). A spectrum mismodeling has a larger effect on the characterization of the spatial correlations (bottom).

4. Comparing spectrum and correlations mismodeling

The above tests demonstrate that spectral and spatial correlations mismodeling can be identified by their corresponding predictive tests. Though the spectrum and the correlation pattern of a stochastic process are separate elements of the GW model, it is not clear they are fully independent. This is because the pulsars are not uniformly distributed in the sky and the signal periods are comparable to the observation time. It is therefore possible that mismodeling in one element of the GW model appears in the test for another. To test for such mismodeling "leakage," we investigate whether using a Hellings-Downs model on the HellingsDownsMonopole-PowerLaw dataset can result in spectral mismodeling, and whether using a power-law model on the HellingsDowns-Turnover dataset can result in correlation mismodelling.

Figure 6 shows the posterior predictive comparison for the spectrum of HellingsDownsMonopole-PowerLaw (top) and the spatial correlations of HellingsDowns-Turnover (bottom). The top panel shows largely consistent inferred and predicted spectra distributions, suggesting that a mismodeling of the spatial correlations, i.e., assuming Hellings-Downs when the data also contain a monopole, does not strongly impact spectral characterization. This is likely due to the fact that spectral characterization is dominated by autocorrelations, at least for weak signals such as the ones considered here. The bottom panel shows that the predicted correlations are systematically lower than the inferred ones, which exhibit signs of a monopole, i.e. a constant upward shift. This suggests that a spectrum mismodeling can affect the inferred correlations pattern. Indeed, a misestimated GW power spectrum will affect the pulsar noise weighting in the optimal-statistic calculation, especially for informative pulsars with low intrinsic noise.

V. PREDICTIVE CHECKS ON TIMING RESIDUALS

The final posterior predictive tests are based directly on the timing residuals $\delta\mathbf{t}$. We first consider visual checks, where we use the model to predict our residuals. As in [14], we isolate contributions from different parts of our model, showing how they sum together to model the timing residuals. Next, we discuss *leave-one-out* tests where we use data from $N_{\rm p} - 1$ pulsars to predict the data of the $N_{\rm p}^{\rm th}$ pulsar.

A. Visual data checks

We use the Gaussian process coefficients from Sec. IV to reconstruct expected residuals for each pulsar. We draw

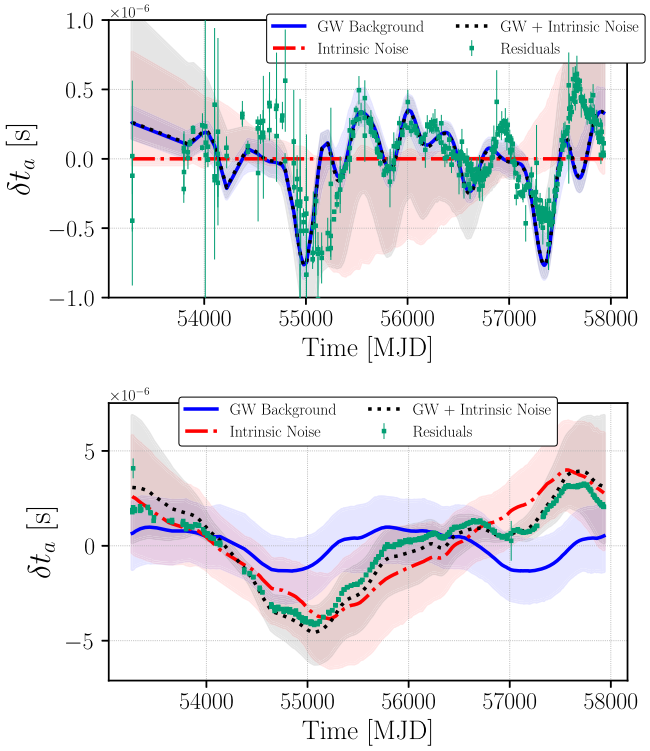

FIG. 7. GW background (blue, solid), intrinsic pulsar noise (red, dashed-dotted), and total noise (black, dotted) contribution to timing residuals for J1909 − 3744 (top) and B1937 + 21 (bottom), compared to the simulated residuals (green). The shaded regions indicate 90% credible intervals and the lines indicate the median. The residuals were simulated using the HELLINGSDOWNS-TURNOVER model. For J1909 − 3744, the residuals are dominated by the GW, while for B1937 + 21 the intrinsic noise dominates. In both cases, the total noise posterior tracks the residuals closely. We do not plot the timing model corrections for clarity, as they are small in this case.

$\mathbf{b}_a^s \sim p(\mathbf{b}_a|\mathbf{\Lambda}^s, \delta\mathbf{t})$, and use these to reconstruct predicted timing residuals in pulsar a, $\delta\mathbf{t}_a^s = \mathbf{T}_a\mathbf{b}_a^s$. This procedure allow us to separate contributions to the residuals from the GW background, the intrinsic pulsar noise, and from timing-model fluctuations.

Figure 7 plots the simulated timing residuals and the separate contributions from intrinsic pulsar noise, GW background, and the sum of the two for J1909 − 3744 (top, low intrinsic noise) and B1937 + 21 (bottom, high intrinsic noise) for the HELLINGSDOWNS-TURNOVER dataset. These reconstructions include frequencies $f_i > 3/T$, because the two lowest frequencies are degenerate with the frequency and spin down parameters in the timing model. In the J1909 − 3744 case (top) the median estimate of the intrinsic noise at each time is near zero, although there is a spread in potential values. Meanwhile, the GW background and total noise (GW plus intrinsic) track the residuals more closely. In the B1937 + 21 case (bottom) the residuals are dominated by intrinsic noise, while the GW background contribution is smaller.

We do not show the contribution from timing-model corrections as it is small in this case. However, their posterior is estimated and could be compared to the fiducial values used to create the original timing residuals. This could serve as a useful cross-check, especially for individual pulsars that are difficult to model.

B. Leave-one-out analysis: Hellings-Downs vs common noise model comparison

We construct predicted data distributions for each pulsar under different assumptions for the correlation pattern, and specifically assuming either Hellings-Downs correlations or an uncorrelated common process. Evaluating these distributions on the actual observed data, we introduce a *pseudo-Bayes factor* [45] for the presence of Hellings-Downs correlations. We compare the pseudo-Bayes factor to null distributions obtained from simulated data and show how they can be used to establish the presence of Hellings-Downs correlations, and equivalently the detection of a GW background.

In contrast to the parameter predictive tests of Sec. IV, here we perform per-pulsar tests conditioned on the data of the *other* pulsars. This distinction is driven by two main reasons. First, the tests of Sec. IV focus on GW model parameters, inference of which is informed by more than one pulsar. For example, the GW Gaussian process coefficients in one pulsar are informed by the other pulsars through Hellings-Downs correlations. There is therefore no clear sense in which GW parameters "belong" to one pulsar. Second, typically a small number of pulsars dominates the constraints. Therefore in-sample and out-of-sample data predictions can be quite distinct.

We begin by selecting a pulsar a to leave out. Quantities with a subscript of a correspond to this pulsar, while a subscript of $-a$ denotes the set of all the other pulsars in the array. We also explicitly break up all quantities into GW, pulsar a, and all other pulsars $(-a)$: $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}_a, \boldsymbol{\epsilon}_{-a}]$, $\mathbf{\Lambda} = [\mathbf{\Lambda}_{\rm gw}, \mathbf{\Lambda}_a, \mathbf{\Lambda}_{-a}]$, $\mathbf{a} = [\mathbf{a}_{{\rm gw},a}, \mathbf{a}_{{\rm gw},-a}, \mathbf{a}_a, \mathbf{a}_{-a}]$. This split is motivated by the fact that $\delta\mathbf{t}_{-a}$ offers no information about the intrinsic parameters of pulsar a, for example $p(\mathbf{\Lambda}|\delta\mathbf{t}_{-a}) = p(\mathbf{\Lambda}_{\rm gw}, \mathbf{\Lambda}_{-a}|\delta\mathbf{t}_{-a})p(\mathbf{\Lambda}_a)$.

The likelihood of residuals $\delta\mathbf{t}_a$ in pulsar a given the residuals $\delta\mathbf{t}_{-a}$ in all other pulsars is

$$p(\delta\mathbf{t}_a|\delta\mathbf{t}_{-a}) = \int {\rm d}\mathbf{\Lambda}{\rm d}\boldsymbol{\epsilon}{\rm d}\mathbf{a}\, p(\delta\mathbf{t}_a|\mathbf{\Lambda}, \boldsymbol{\epsilon}, \mathbf{a})p(\mathbf{\Lambda}, \boldsymbol{\epsilon}, \mathbf{a}|\delta\mathbf{t}_{-a}). \tag{25}$$

After a long derivation laid out in the Appendix we find

$$p_{\rm HD}(\delta\mathbf{t}_a|\delta\mathbf{t}_{-a}) \approx \frac{1}{N_s}\sum_s \int {\rm d}\mathbf{\Lambda}_a {\rm d}\mathbf{a}_{{\rm gw},a}\, p(\delta\mathbf{t}_a|\mathbf{\Lambda}_a, \mathbf{a}_{{\rm gw},a}) \times p(\mathbf{a}_{{\rm gw},a}|\mathbf{\Lambda}_{\rm gw}^s, \mathbf{\Lambda}_{-a}^s, \delta\mathbf{t}_{-a})p(\mathbf{\Lambda}_a), \tag{26}$$

$$p_{\rm CN}(\delta\mathbf{t}_a|\delta\mathbf{t}_{-a}) \approx \frac{1}{N_s}\sum_s \int d\mathbf{\Lambda}_a p(\delta\mathbf{t}_a|\mathbf{\Lambda}_a, \mathbf{\Lambda}^s_{\rm gw}) p(\mathbf{\Lambda}_a). \tag{27}$$

where the "HD" subscript signifies that we have assumed Hellings-Downs correlations and "CN" subscript signifies that we ignore the Hellings-Downs correlations and assume that the pulsars are only subject to an uncorrelated common process. Equations (26) and (27) are evaluated over N_s draws from the hyperparameter posterior

$$\mathbf{\Lambda}^s_{\rm gw}, \mathbf{\Lambda}^s_{-a} \sim p(\mathbf{\Lambda}_{\rm gw}, \mathbf{\Lambda}^s_{-a}|\delta\mathbf{t}_{-a}), \tag{28}$$

from the analysis of Sec. II B. The integral over $d\mathbf{a}_{{\rm gw},a}$ is performed analytically as it involves a product of Gaussian distributions, while the one over $d\mathbf{\Lambda}_a$ is performed numerically.

Comparing Eqs. (26) and (27) can provide an estimate of how much each pulsar supports the presence of Hellings-Downs correlations. We introduce the "pseudo-Bayes factor" (PBF) [45] between Hellings-Downs and common noise in pulsar a as

$$\mathrm{PBF}^{\rm HD}_{{\rm CN},a} \equiv \frac{p_{\rm HD}(\delta\mathbf{t}_a|\delta\mathbf{t}_{-a})}{p_{\rm CN}(\delta\mathbf{t}_a|\delta\mathbf{t}_{-a})}, \tag{29}$$

where the numerator and denominator are defined in Eqs. (26) and (27) respectively, and are *posterior predictive likelihoods* that are calculated on the observed $\delta\mathbf{t}_a$. The total pseudo-Bayes factor is then the product over all pulsars

$$\mathrm{PBF}^{\rm HD}_{\rm CN} = \prod_{a=1}^{N_{\rm p}} \mathrm{PBF}^{\rm HD}_{{\rm CN},a}. \tag{30}$$

The pseudo-Bayes factor shares some similarities with the traditional Bayes factor (i.e., the marginal likelihood ratio), but there are also important differences. First, both traditional and pseudo-Bayes factors are a ratio of likelihoods. Second, unlike traditional Bayes factors, the pseudo-Bayes factor is insensitive to the existence of parameter space regions of little likelihood support, which reduce Bayes factors by the so-called Occam factors. In that sense, the pseudo-Bayes factor does not suffer from interpretation issues related to the extent of parameter priors or the presence of improper priors [31,33,68]. Third, by definition $\mathrm{PBF}^{\rm HD}_{{\rm CN},a}$ is a measure of how well the model predicts new data. This means that it can be estimated on a per-pulsar basis, thereby assessing which pulsar is more consistent with each model, and identifying outliers. Specifically, $\mathrm{PBF}^{\rm HD}_{{\rm CN},a}$ tests whether certain pulsars are poorly understood compared to others, potentially signaling issues with their intrinsic noise modeling.

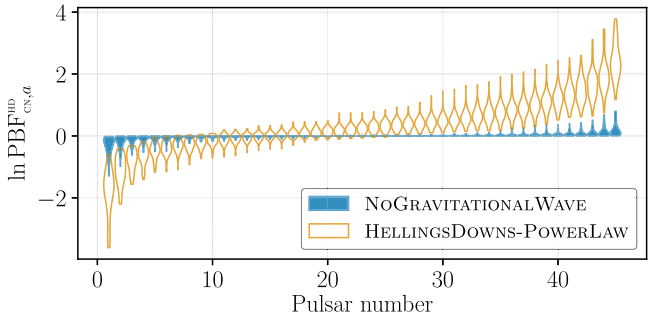

FIG. 8. Pseudo-Bayes factor comparing the Hellings-Downs and common noise models for each pulsar. We consider 59 realizations of simulated HellingsDowns-PowerLaw (orange) and 45 NoGravitationalWave (blue) datasets and show the distribution of obtained pseudo-Bayes factors in violins. Pulsars are ordered from lowest to higher value of the pseudo-Bayes factor. Most pulsars support the presence of Hellings-Downs correlations when a signal is present, though a minority displays the opposite behavior. All pulsars have uninformative pseudo-Bayes factors when no signal is injected.

The pseudo-Bayes factor, however, does suffer from *calibration* issues just as the traditional Bayes factor. That is, how are we to interpret its value in terms of statistical confidence? Rather than relying on arbitrary classifications schemes [26,27], a common procedure to interpret Bayes factors involves using a large set of simulations to estimate a false-alarm probability for the measured value [69–74].[9]

Figure 8 shows the (natural logarithm of the) pseudo-Bayes factors for individual pulsars ordered from lowest to highest. We produce 59 simulated datasets using HellingsDowns-PowerLaw, and 45 using NoGravitationalWave. The following result should be interpreted only as a demonstration of our method as the simulated GW background amplitude of $\log_{10} A_{\rm gw} = -14$ is higher than the one inferred from real data by a factor of ∼5 [46]. Such a high value was chosen so that we have a detectable signal in 12 years of simulated data and thus we can meaningfully test the proposed methods.

For each simulated dataset, we compute $\ln \mathrm{PBF}^{\rm HD}_{{\rm CN},a}$ for each pulsar a, we sort the pulsars from the smallest to the largest value, and we plot the distribution over data realizations.[10] In the HellingsDowns-PowerLaw case, we regularly find ∼20 pulsars with positive $\ln \mathrm{PBF}^{\rm HD}_{{\rm CN},a}$. This means that data from the other pulsars can predict the observed data in pulsar *a better* if Hellings-Downs correlations are present. The test is uninformative for ∼10 pulsars with $\ln \mathrm{PBF}^{\rm HD}_{{\rm CN},a} \sim 0$, while a similar number of

[9]Another recommendation is to compare $\mathrm{PBF}^{\rm HD}_{\rm CN}$ to the variance of $\mathrm{PBF}^{\rm HD}_{{\rm CN},a}$ over the pulsars [75]; we leave this to future work.

[10]This procedure means that the pulsar order is different for each simulated dataset. Therefore the x-axis of Fig. 8 is not a *specific* pulsar, but instead the nth pulsar as ranked by its pseudo-Bayes factor in each dataset.

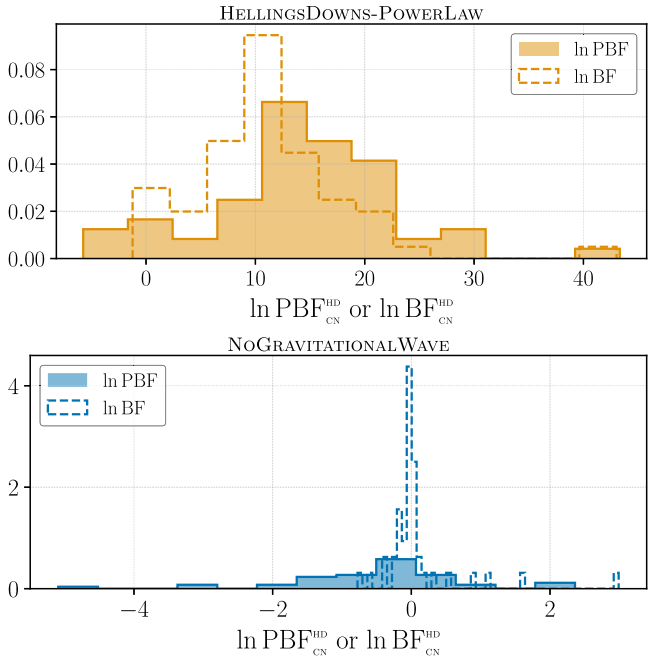

FIG. 9. Distribution of total pseudo-Bayes factors (solid histograms) and traditional Bayes factors (dashed histograms) for repeated simulations with the HellingsDowns-PowerLaw (top) and NoGravitationalWave (bottom) datasets comparing the Hellings-Downs and common noise hypotheses. Note the different x-axis scales on the two panels. Using the results of the bottom panel as a null distribution, 89% of the simulated datasets in the top panel have detectable Hellings-Downs correlations at $> 2\sigma$ significance.

pulsars has $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CN},a} < 0$. The latter means that these pulsars support no Hellings-Downs correlations even if these exist in the data. Such behavior is also encountered in the "drop-out factors" calculated by sampling an indicator variable that switches between the common process and no common signal hypotheses for each individual pulsar [46]. *Some* negative $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CN},a}$ are therefore to be expected even in simulated data and they are not immediately an indication of mismodeling.[11] In the NoGravitationalWave case, all pulsars have $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CN},a} \sim 0$, suggesting no preference either way. This is to be expected as no signal is present, so there should be no information about its correlation pattern.

Even though individual pulsars can have $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CN},a} < 0$, the total pseudo-Bayes factor is in favor of Hellings-Downs correlations for the majority of the simulated datasets with a signal. Figure 9 shows distributions of $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CN}}$ over 59 data realizations for HellingsDowns-PowerLaw (top) and 45 for NoGravitationalWave (bottom). In the top panel, we find $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CN}} > 0$ for 92% of the realizations, with most datasets resulting in a strong preference for Hellings-Downs correlations and $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CN}} \sim 10$–$20$. However, as discussed above, the absolute scale of the pseudo-Bayes factor has no definite statistical interpretation, and results should instead be calibrated to simulations. The bottom panel shows the null distribution of $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CN}}$. All datasets have $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CN}} < 2$ and 61% of them have $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CN}} < 0$. Given this null, Hellings-Downs correlations would have been detected in 89% of the HellingsDowns-PowerLaw simulations with a significance of $> 2\sigma$. With 59 background simulations the significance estimate is limited to $\sim 1/59 \sim 2\sigma$.

Figure 9 shows also the distributions of traditional Bayes factors between the Hellings-Downs and common noise hypotheses for the same simulations computed via likelihood reweighting [55]. On average the HellingsDowns-PowerLaw dataset results in larger pseudo-Bayes factors than traditional Bayes factors, while the trend is reversed for the NoGravitationalWave datasets. However, due to the high GW signal amplitude we still find that 90% of the simulated datasets in the top panel have detectable Hellings-Downs correlations at $> 2\sigma$ significance when using the traditional Bayes factor as a detection statistic. These results suggest that pseudo-and traditional Bayes factors can act as complementary model-checking tools. We leave the determination of their relative sensitivity as detection statistics to future work, since this demonstration is based on only 45 simulations and a loud injected GWB.

VI. DISCUSSION AND CONCLUSIONS

PTA analyses assume that a GW background results in arrival time residuals that are subject to a common power-law process among pulsars and Hellings-Downs spatial correlations between them. While the correlation pattern is robust under a tensorial GW background, systematic errors can induce further monopolar or dipolar correlations [14,77–81]. Moreover, the GW spectral shape is subject to astrophysical, statistical, and even cosmological uncertainties [60,61,72]. Here we propose to test these assumptions using *posterior predictive checks* that assess how well predicted data based on the inferred model parameters match the observed data. Predictive tests based on different quantities allow us to assess different aspects of the model or pulsars in the array separately and thus can offer insights about model extensions if a discrepancy is identified.

We propose and study two types of tests. The first type concerns the Gaussian-process coefficients of the GW and intrinsic-noise stochastic processes. Comparing predicted and inferred coefficients on simulated datasets, we can identify frequency bins where the power-law model under- or over-predicts the observed power. Moreover, by comparing the inferred and predicted spatial correlations we can assess the presence of non-Hellings-Downs correlations. The second type of test concerns the timing residuals themselves, and specifically the likelihood of the observed data in a select pulsar given all other pulsars. We compute

[11] In fact down-selecting pulsars based on arbitrary metrics can lead to biased estimates [76].

the *pseudo-Bayes factor* as the ratio of these likelihoods under the Hellings-Downs and the uncorrelated common process hypotheses. We show that among all the pulsars in the array it is expected for a handful to show preference *against* Hellings-Downs correlations. However, the total pseudo-Bayes factor over the entire array can be used as a detection statistic to establish the presence of Hellings-Downs correlations.

Our study adds to existing efforts that explore extensions of PTA analyses. A common extension to the power-law spectrum (and one of our simulated datasets) is the truncated power law that arises when astrophysical hardening mechanisms accelerate the inspiral of the black hole binaries that source the GW background [30]. A different kind of broken power law flattens the spectrum at high frequencies [46]. Such flattening is interpreted as being caused by modeling systematics related to the intrinsic pulsar noise, and it is used to limit the number of frequency bins analyzed [46]. Doing away with a parametric model, "free spectral" analyses instead allow for independent amplitudes at each frequency bin [46]. Beyond the details of the spectral shape, a GW background has a unique spectrum, even though the exact realization will differ between pulsars. A test of this assumption involves allowing for some scatter in the GW amplitude inferred from each pulsar, whose probable origin would be mismodeling [32]. Applying the test to PPTA data, Ref. [32] found no evidence for such a scatter.

Moving on to spatial correlations, proposed checks include reconstructing the correlations as interpolated functions, sums of Legendre polynomials [82], or perturbed Hellings-Downs patterns [83]. These tests proceed with the observed data alone and compare the reconstructed generic correlation pattern with the expected Hellings-Downs pattern. A related test replaces or augments the Hellings-Downs correlations with nontensorial correlations expected for certain theories of gravity beyond general relativity [84,85].

The tests proposed in this study offer complementary ways to assess PTA models. We expect such tests to become increasingly important as PTA datasets expand in sensitivity, and move toward detection of the GW background. Furthermore, our tests can be used to assess consistency between different PTA datasets. For example, we could use NANOGrav data to predict PPTA data and then compare to the actual observed PPTA data. Such tests would generalize the comparisons performed in [57] and help establish consistency between datasets, thus strengthening astrophysical conclusions.

ACKNOWLEDGMENTS

We thank Will Farr for discussions on posterior predictive checks in the LIGO context. Our analyses make use of ENTERPRISE [52,86], SCIPY [87], MATPLOTLIB [88], NUMPY [89], PANDAS [90], and SEABORN [91]. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), supported by NSF award No. ACI-1548562, and specifically the Bridges-2 system at the Pittsburgh Supercomputing Center, supported by NSF award No. ACI-1928147. P. M. M., M. V., and K. C. were supported by the NANOGrav Physics Frontiers Center, National Science Foundation (NSF), Grant No. 2020265. Part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004).

APPENDIX: DETAILED DERIVATION OF THE POSTERIOR PREDICTIVE LIKELIHOOD FOR SINGLE-PULSAR DATA

The starting point of the derivation is the likelihood of the residuals $\delta\mathbf{t}_a$ in pulsar a given the residuals $\delta\mathbf{t}_{-a}$ in all other pulsars, reproduced here from Eq. (25):

$$p(\delta\mathbf{t}_a|\delta\mathbf{t}_{-a}) = \int \mathrm{d}\mathbf{\Lambda}\mathrm{d}\boldsymbol{\epsilon}\mathrm{d}\mathbf{a}\, p(\delta\mathbf{t}_a|\mathbf{\Lambda},\boldsymbol{\epsilon},\mathbf{a})\, p(\mathbf{\Lambda},\boldsymbol{\epsilon},\mathbf{a}|\delta\mathbf{t}_{-a}). \tag{A1}$$

The first term in the integrand of Eq. (A1) reduces to

$$p(\delta\mathbf{t}_a|\mathbf{\Lambda},\boldsymbol{\epsilon},\mathbf{a}) = p(\delta\mathbf{t}_a|\boldsymbol{\epsilon}_a,\mathbf{a}_{\mathrm{gw},a},\mathbf{a}_a), \tag{A2}$$

as the data of pulsar a depend on the parameters of this pulsar only, as given by Eq. (4). The second term in the integrand of Eq. (A1) is

$$\begin{aligned} &p(\mathbf{\Lambda},\boldsymbol{\epsilon},\mathbf{a}|\delta\mathbf{t}_{-a}) \\ &\quad = p(\mathbf{\Lambda}_a,\boldsymbol{\epsilon}_a,\mathbf{a}_a)\, p(\mathbf{\Lambda}_{\mathrm{gw}},\mathbf{\Lambda}_{-a},\boldsymbol{\epsilon}_{-a},\mathbf{a}_{-a},\mathbf{a}_{\mathrm{gw},a},\mathbf{a}_{\mathrm{gw},-a}|\delta\mathbf{t}_{-a}), \end{aligned} \tag{A3}$$

where the first term includes all properties of pulsar a that do not depend on the data of the other pulsars. The only property of pulsar a that remains in the second term are the GW Gaussian process coefficients $\mathbf{a}_{\mathrm{gw},a}$, since those are informed by $\delta\mathbf{t}_{-a}$ through the Hellings-Downs correlations. Returning to the full predictive likelihood in Eq. (A1), the integrals over $\mathbf{a}_{\mathrm{gw},-a}, \boldsymbol{\epsilon}_{-a}, \mathbf{a}_{-a}$ are now trivial. Performing those and substituting Eqs. (A2) and (A3) in Eq. (A1) we get

$$
\begin{aligned}
p_{\rm HD}(\delta\mathbf{t}_a|\delta\mathbf{t}_{-a}) &= \int \mathrm{d}\mathbf{\Lambda}_{\rm gw}\mathrm{d}\mathbf{\Lambda}_a\mathrm{d}\mathbf{\Lambda}_{-a}\mathrm{d}\boldsymbol{\epsilon}_a\mathrm{d}\mathbf{a}_{{\rm gw},a}\mathrm{d}\mathbf{a}_a p(\delta\mathbf{t}_a|\boldsymbol{\epsilon}_a,\mathbf{a}_{{\rm gw},a},\mathbf{a}_a)p(\mathbf{\Lambda}_a,\boldsymbol{\epsilon}_a,\mathbf{a}_a)p(\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a},\mathbf{a}_{{\rm gw},a}|\delta\mathbf{t}_{-a}) \\
&= \int \mathrm{d}\mathbf{\Lambda}_{\rm gw}\mathrm{d}\mathbf{\Lambda}_a\mathrm{d}\mathbf{\Lambda}_{-a}\mathrm{d}\boldsymbol{\epsilon}_a\mathrm{d}\mathbf{a}_{{\rm gw},a}\mathrm{d}\mathbf{a}_a p(\delta\mathbf{t}_a|\boldsymbol{\epsilon}_a,\mathbf{a}_{{\rm gw},a},\mathbf{a}_a)p(\mathbf{\Lambda}_a)p(\boldsymbol{\epsilon}_a,\mathbf{a}_a|\mathbf{\Lambda}_a)p(\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a},\mathbf{a}_{{\rm gw},a}|\delta\mathbf{t}_{-a}) \\
&= \int \mathrm{d}\mathbf{\Lambda}_{\rm gw}\mathrm{d}\mathbf{\Lambda}_a\mathrm{d}\mathbf{\Lambda}_{-a}\mathrm{d}\mathbf{a}_{{\rm gw},a} p(\delta\mathbf{t}_a|\mathbf{a}_{{\rm gw},a},\mathbf{\Lambda}_a)p(\mathbf{\Lambda}_a)p(\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a},\mathbf{a}_{{\rm gw},a}|\delta\mathbf{t}_{-a}) \\
&= \int \mathrm{d}\mathbf{\Lambda}_{\rm gw}\mathrm{d}\mathbf{\Lambda}_a\mathrm{d}\mathbf{\Lambda}_{-a}\mathrm{d}\mathbf{a}_{{\rm gw},a} p(\delta\mathbf{t}_a|\mathbf{a}_{{\rm gw},a},\mathbf{\Lambda}_a)p(\mathbf{\Lambda}_a)p(\mathbf{a}_{{\rm gw},a}|\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a},\delta\mathbf{t}_{-a})p(\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a}|\delta\mathbf{t}_{-a}) \\
&= \int \mathrm{d}\mathbf{\Lambda}_{\rm gw}\mathrm{d}\mathbf{\Lambda}_{-a}\left[\int \mathrm{d}\mathbf{\Lambda}_a\mathrm{d}\mathbf{a}_{{\rm gw},a} p(\delta\mathbf{t}_a|\mathbf{a}_{{\rm gw},a}\mathbf{\Lambda}_a)p(\mathbf{\Lambda}_a)p(\mathbf{a}_{{\rm gw},a}|\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a},\delta\mathbf{t}_{-a})\right]p(\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a}|\delta\mathbf{t}_{-a}),
\end{aligned}
\tag{A4}
$$

where the "HD" subscript signifies that we have assumed Hellings-Downs correlations. In the second line we have used the definition of conditional probabilities $p(\mathbf{\Lambda}_a,\boldsymbol{\epsilon}_a,\mathbf{a}_a) = p(\mathbf{\Lambda}_a)p(\boldsymbol{\epsilon}_a,\mathbf{a}_a|\mathbf{\Lambda}_a)$ and in the third line we have marginalized over $\boldsymbol{\epsilon}_a,\mathbf{a}_a$ following Eq. (10). In the third line we have again used conditional probabilities $p(\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a},\mathbf{a}_{{\rm gw},a}|\delta\mathbf{t}_{-a}) = p(\mathbf{a}_{{\rm gw},a}|\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a},\delta\mathbf{t}_{-a})(\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a}|\delta\mathbf{t}_{-a})$ and in the last line we reorganize the integrals. The first term in the integral, $p(\delta\mathbf{t}_a|\mathbf{a}_{{\rm gw},a},\mathbf{\Lambda}_a)$, is given by Eq. (4) after (analytically) marginalizing over the intrinsic noise Gaussian process coefficients, $p(\mathbf{a}_{{\rm gw},a}|\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a},\delta\mathbf{t}_{-a})$ is a Gaussian with mean and covariance given by Eqs. (11) and (12), $p(\mathbf{\Lambda}_a)$ is the prior on $\mathbf{\Lambda}_a$, while $p(\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a}|\delta\mathbf{t}_{-a})$ is the posterior of the hyperparameters.

A simplified version of Eq. (A4) can be obtained if we ignore the Hellings-Downs correlations and assume that the pulsars are only subject to an uncorrelated common process, denoted as "CN" in equations below. Then

$$
\begin{aligned}
p_{\rm CN}(\delta\mathbf{t}_a|\delta\mathbf{t}_{-a}) &= \int \mathrm{d}\mathbf{\Lambda}_{\rm gw}\mathrm{d}\mathbf{\Lambda}_{-a}\left[\int \mathrm{d}\mathbf{\Lambda}_a\mathrm{d}\mathbf{a}_{{\rm gw},a} p(\delta\mathbf{t}_a|\mathbf{a}_{{\rm gw},a}\mathbf{\Lambda}_a)p(\mathbf{\Lambda}_a)p(\mathbf{a}_{{\rm gw},a}|\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a},\delta\mathbf{t}_{-a})\right]p(\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a}|\delta\mathbf{t}_{-a}) \\
&= \int \mathrm{d}\mathbf{\Lambda}_{\rm gw}\mathrm{d}\mathbf{\Lambda}_{-a}\left[\int \mathrm{d}\mathbf{\Lambda}_a\mathrm{d}\mathbf{a}_{{\rm gw},a} p(\delta\mathbf{t}_a|\mathbf{a}_{{\rm gw},a}\mathbf{\Lambda}_a)p(\mathbf{\Lambda}_a)p(\mathbf{a}_{{\rm gw},a}|\mathbf{\Lambda}_{\rm gw})\right]p(\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a}|\delta\mathbf{t}_{-a}) \\
&= \int \mathrm{d}\mathbf{\Lambda}_{\rm gw}\left[\int \mathrm{d}\mathbf{\Lambda}_a p(\delta\mathbf{t}_a|\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_a)p(\mathbf{\Lambda}_a)\right]p(\mathbf{\Lambda}_{\rm gw}|\delta\mathbf{t}_{-a}),
\end{aligned}
\tag{A5}
$$

where in the second line we have simplified $p(\mathbf{a}_{{\rm gw},a}|\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a},\delta\mathbf{t}_{-a}) = p(\mathbf{a}_{{\rm gw},a}|\mathbf{\Lambda}_{\rm gw})$ due to the lack of Hellings-Downs correlations, and in the third line we have marginalized over $\mathbf{a}_{{\rm gw},a},\mathbf{\Lambda}_{-a}$ following Eq. (10).

Equations (A4) and (A5) are estimated as follows. The integral over $\mathrm{d}\mathbf{\Lambda}_{\rm gw}$ (and $\mathbf{\Lambda}_{-a}$ if applicable) is performed through Monte-Carlo integration using N_s samples

$$
\mathbf{\Lambda}^s_{\rm gw},\mathbf{\Lambda}^s_{-a} \sim p(\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a}|\delta\mathbf{t}_{-a}), \tag{A6}
$$

from the analysis of Sec. II B:

$$
\begin{aligned}
p_{\rm HD}(\delta\mathbf{t}_a|\delta\mathbf{t}_{-a}) \approx \frac{1}{N_s}\sum_s \int \mathrm{d}\mathbf{\Lambda}_a\mathrm{d}\mathbf{a}_{{\rm gw},a} p(\delta\mathbf{t}_a|\mathbf{\Lambda}_a,\mathbf{a}^s_{{\rm gw},a}) \\
\times p(\mathbf{a}^s_{{\rm gw},a}|\mathbf{\Lambda}^s_{\rm gw},\mathbf{\Lambda}^s_{-a},\delta\mathbf{t}_{-a})p(\mathbf{\Lambda}_a),
\end{aligned}
\tag{A7}
$$

$$
p_{\rm CN}(\delta\mathbf{t}_a|\delta\mathbf{t}_{-a}) \approx \frac{1}{N_s}\sum_s \int \mathrm{d}\mathbf{\Lambda}_a p(\delta\mathbf{t}_a|\mathbf{\Lambda}_a,\mathbf{\Lambda}^s_{\rm gw})p(\mathbf{\Lambda}_a). \tag{A8}
$$

The integral over $\mathrm{d}\mathbf{\Lambda}_a$ is performed numerically. The integral over $\mathrm{d}\mathbf{a}_{{\rm gw},a}$ is performed analytically as both terms involving $\mathbf{a}_{{\rm gw},a}$ are Gaussians.

The above equations require estimating $N_{\rm p}$ posteriors $p(\mathbf{\Lambda}|\delta\mathbf{t}_{-a})$—one for each individual pulsar, a. This results in a heavy computational cost that may be unfeasible. Instead, if the hyperparameter posterior is not strongly affected by any individual pulsars, we can approximate Eq. (A6) with

$$
\begin{aligned}
p(\mathbf{\Lambda}|\delta\mathbf{t}_{-a}) &= p(\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a}|\delta\mathbf{t}_{-a})p(\mathbf{\Lambda}_a) \\
&\approx p(\mathbf{\Lambda}_{\rm gw},\mathbf{\Lambda}_{-a}|\delta\mathbf{t})p(\mathbf{\Lambda}_a).
\end{aligned}
\tag{A9}
$$

Crucially, while we use the data from pulsar a to constrain $\mathbf{\Lambda}_{\rm gw}$ by assuming that the effect is small, we do not use the same data to constrain $\mathbf{\Lambda}_a$, instead still integrating over the prior. We have checked that this approximation has a minor impact on our results while greatly reducing computational cost, so we adopted it to produce the results in Secs. IV and V.

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

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

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

[4] 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).

[5] 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).

[6] J. Antoniadis *et al.* (EPTA Collaboration), The second data release from the European Pulsar Timing Array: V. Implications for massive black holes, dark matter and the early Universe, arXiv:2306.16227.

[7] J. Antoniadis *et al.* (EPTA Collaboration), The second data release from the European Pulsar Timing Array IV. Search for continuous gravitational wave signals, arXiv:2306.16226.

[8] G. Agazie *et al.* (NANOGrav Collaboration), The NANOGrav 15-year data set: Search for anisotropy in the gravitational-wave background, arXiv:2306.16221.

[9] G. Agazie *et al.* (NANOGrav Collaboration), The NANOGrav 15 yr data set: Bayesian limits on gravitational waves from individual supermassive black hole binaries, Astrophys. J. Lett. **951**, L50 (2023).

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

[11] G. Agazie *et al.* (International Pulsar Timing Array Collaboration), Comparing recent PTA results on the nanohertz stochastic gravitational wave background, arXiv:2309.00693.

[12] R. T. Edwards, G. B. Hobbs, and R. N. Manchester, Tempo2, a new pulsar timing package. 2. The timing model and precision estimates, Mon. Not. R. Astron. Soc. **372**, 1549 (2006).

[13] J. Luo *et al.*, PINT: A Modern Software Package for Pulsar Timing, Astrophys. J. **911**, 45 (2021).

[14] B. Goncharov *et al.*, Identifying and mitigating noise sources in precision pulsar timing data sets, Mon. Not. R. Astron. Soc. **502**, 478 (2021).

[15] M. L. Jones *et al.*, The NANOGrav Nine-year data set: Measurement and analysis of variations in dispersion measures, Astrophys. J. **841**, 125 (2017).

[16] E. J. Groth, Timing of the Crab Pulsar III. The slowing down and the nature of the random process, Astrophys. J. Suppl. Ser. **29**, 453 (1975).

[17] J. M. Cordes, Pulsar timing. II. Analysis of random walk timing noise: Application to the Crab pulsar., Astrophys. J. **237**, 216 (1980).

[18] R. M. Shannon and J. M. Cordes, Assessing the role of spin noise in the precision timing of millisecond pulsars, Astrophys. J. **725**, 1607 (2010).

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

[20] 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).

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

[22] R. van Haasteren, Y. Levin, P. McDonald, and T. Lu, On measuring the gravitational-wave background using Pulsar Timing Arrays, Mon. Not. R. Astron. Soc. **395**, 1005 (2009).

[23] J. S. Hazboun *et al.* (NANOGrav Collaboration), Bayesian solar wind modeling with Pulsar Timing Arrays, Astrophys. J. **929**, 39 (2022).

[24] M. T. Lam *et al.*, A second chromatic timing event of interstellar origin toward PSR J1713 + 0747, Astrophys. J. **861**, 132 (2018).

[25] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory* (Wiley, Chichester, England, 2009).

[26] H. Jeffreys, *The Theory of Probability*, Oxford Classic Texts in the Physical Sciences (OUP, Oxford, 1998).

[27] R. E. Kass and A. E. Raftery, Bayes factors, J. Am. Stat. Assoc. **90**, 773 (1995).

[28] T. J. Loredo, Accounting for source uncertainties in analyses of astronomical survey data, AIP Conf. Proc. **735**, 195 (2004).

[29] I. Mandel, W. M. Farr, and J. R. Gair, Extracting distribution parameters from multiple uncertain observations with selection biases, Mon. Not. R. Astron. Soc. **486**, 1086 (2019).

[30] 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).

[31] J. S. Hazboun, J. Simon, X. Siemens, and J. D. Romano, Model dependence of bayesian gravitational-wave background statistics for Pulsar Timing Arrays, Astrophys. J. Lett. **905**, L6 (2020).

[32] B. Goncharov *et al.*, Consistency of the Parkes Pulsar Timing Array signal with a nanohertz gravitational- wave background, Astrophys. J. **932**, L22 (2022).

[33] A. Zic *et al.*, Evaluating the prevalence of spurious correlations in pulsar timing array data sets, Mon. Not. R. Astron. Soc. **516**, 410 (2022).

[34] M. Vallisneri, P. Meyers, K. Chatziioannou, and A. J. K. Chua, preceding paper, Posterior predictive checking for gravitational-wave detection with pulsar timing arrays: I. The optimal statistic, Phys. Rev. D **108**, 123007 (2023).

[35] M. Anholm, S. Ballmer, J. D. E. Creighton, L. R. Price, and X. Siemens, Optimal strategies for gravitational wave stochastic background searches in pulsar timing data, Phys. Rev. D **79**, 084030 (2009).

[36] S. J. Chamberlin, J. D. E. Creighton, X. Siemens, P. Demorest, J. Ellis, L. R. Price, and J. D. Romano, Time-domain implementation of the optimal cross-correlation statistic for stochastic gravitational-wave background searches in pulsar timing data, Phys. Rev. D **91**, 044048 (2015).

[37] S. J. Vigeland, K. Islo, S. R. Taylor, and J. A. Ellis, Noise-marginalized optimal statistic: A robust hybrid frequentist-Bayesian statistic for the stochastic gravitational-wave background in pulsar timing arrays, Phys. Rev. D **98**, 044003 (2018).
[38] J. S. Hazboun, P. M. Meyers, J. D. Romano, X. Siemens, and A. M. Archibald, Analytic distribution of the optimal cross-correlation statistic for stochastic gravitational-wave-background searches using pulsar timing arrays, arXiv:2305.01116.
[39] M. Fishbach, W. M. Farr, and D. E. Holz, The most massive binary black hole detections and the identification of population outliers, Astrophys. J. Lett. **891**, L31 (2020).
[40] T. A. Callister, S. J. Miller, K. Chatziioannou, and W. M. Farr, No evidence that the majority of black holes in binaries have zero spin, Astrophys. J. Lett. **937**, L13 (2022).
[41] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Binary black hole population properties inferred from the first and second observing runs of Advanced LIGO and Advanced Virgo, Astrophys. J. Lett. **882**, L24 (2019).
[42] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Population properties of compact objects from the second LIGO-Virgo gravitational-wave transient catalog, Astrophys. J. Lett. **913**, L7 (2021).
[43] H. Wang, S. R. Taylor, and M. Vallisneri, Bayesian cross validation for gravitational-wave searches in pulsar-timing array data, Mon. Not. R. Astron. Soc. **487**, 3644 (2019).
[44] NANOGrav Scientific Collaboration, Nanograv data, https://data.nanograv.org/.
[45] S. Geisser and W. F. Eddy, A predictive approach to model selection, J. Am. Stat. Assoc. **74**, 153 (1979).
[46] Z. Arzoumanian *et al.* (NANOGrav Collaboration), The NANOGrav 12.5 yr Data Set: Search for an isotropic stochastic gravitational-wave background, Astrophys. J. Lett. **905**, L34 (2020).
[47] S. R. Taylor, The nanohertz gravitational wave astronomer, arXiv:2105.13270.
[48] R. van Haasteren and Y. Levin, Understanding and analysing time-correlated stochastic signals in pulsar timing, Mon. Not. R. Astron. Soc. **428**, 1147 (2013).
[49] Z. Arzoumanian *et al.* (NANOGrav Collaboration), The NANOGrav nine-year data set: Limits on the isotropic stochastic gravitational wave background, Astrophys. J. **821**, 13 (2016).
[50] Z. Arzoumanian *et al.* (NANOGrav Collaboration), The NANOGrav nine-year data set: Observations, arrival time measurements, and analysis of 37 millisecond pulsars, Astrophys. J. **813**, 65 (2015).
[51] M. F. Alam *et al.* (NANOGrav Collaboration), The NANOGrav 12.5 yr data set: Observations and narrowband timing of 47 millisecond pulsars, Astrophys. J. Suppl. Ser. **252**, 4 (2021).
[52] J. A. Ellis, M. Vallisneri, S. R. Taylor, and P. T. Baker, Enterprise: Enhanced numerical toolbox enabling a robust pulsar inference suite, Zenodo (2020).
[53] N. S. Pol *et al.* (NANOGrav Collaboration), Astrophysics milestones for Pulsar Timing Array gravitational-wave detection, Astrophys. J. Lett. **911**, L34 (2021).
[54] 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).
[55] 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).
[56] W. G. Lamb, S. R. Taylor, and R. van Haasteren, The need for speed: Rapid refitting techniques for bayesian spectral characterization of the gravitational wave background using PTAs, arXiv:2303.15442.
[57] 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).
[58] P. C. Mahalanobis, On the generalized distance in statistics, Proc. Indian National Sci. Acad. **2**, 49 (1936).
[59] 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).
[60] 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 yr results, Mon. Not. R. Astron. Soc. **502**, L99 (2021).
[61] B. Bécsy, N. J. Cornish, and L. Z. Kelley, Exploring realistic nanohertz gravitational-wave backgrounds, Astrophys. J. **941**, 119 (2022).
[62] K. Gersbach, P. Meyers, S. Taylor, R. Joseph *et al.* (to be published).
[63] B. Allen, Variance of the Hellings-Downs correlation, Phys. Rev. D **107**, 043018 (2023).
[64] B. Allen and J. D. Romano, Hellings and Downs correlation of an arbitrary set of pulsars, Phys. Rev. D **108**, 043026 (2023).
[65] R. C. Bernardo and K.-W. Ng, Pulsar and cosmic variances of pulsar timing-array correlation measurements of the stochastic gravitational wave background, J. Cosmol. Astropart. Phys. 11 (2022) 046.
[66] R. C. Bernardo and K.-W. Ng, Hunting the stochastic gravitational wave background in pulsar timing array cross correlations through theoretical uncertainty, J. Cosmol. Astropart. Phys. 08 (2023) 028.
[67] A. D. Johnson *et al.* (NANOGrav Collaboration), The NANOGrav 15-year gravitational-wave background analysis pipeline, arXiv:2306.16223.
[68] M. Isi, W. M. Farr, and K. Chatziioannou, Comparing Bayes factors and hierarchical inference for testing general relativity with gravitational waves, Phys. Rev. D **106**, 024048 (2022).
[69] J. Veitch and A. Vecchio, Assigning confidence to inspiral gravitational wave candidates with Bayesian model selection, Classical Quantum Gravity **25**, 184010 (2008).
[70] M. Vallisneri, Testing general relativity with gravitational waves: A reality check, Phys. Rev. D **86**, 082001 (2012).
[71] N. J. Cornish and L. Sampson, Towards robust gravitational wave detection with Pulsar Timing Arrays, Phys. Rev. D **93**, 104047 (2016).
[72] S. R. Taylor, L. Lentati, S. Babak, P. Brem, J. R. Gair, A. Sesana, and A. Vecchio, All correlations must die:

Assessing the significance of a stochastic gravitational-wave background in pulsar-timing arrays, Phys. Rev. D **95**, 042002 (2017).

[73] T. B. Littenberg, J. B. Kanner, N. J. Cornish, and M. Millhouse, Enabling high confidence detections of gravitational-wave bursts, Phys. Rev. D **94**, 044050 (2016).

[74] M. Isi, R. Smith, S. Vitale, T. J. Massinger, J. Kanner, and A. Vajpeyi, Enhancing confidence in the detection of gravitational waves from compact binaries using signal coherence, Phys. Rev. D **98**, 042007 (2018).

[75] A. Vehtari, A. Gelman, and J. Gabry, Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, arXiv:1507.04544.

[76] 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).

[77] M. Tinto, Gravitational wave searches with pulsar timing arrays: Cancellation of clock and ephemeris noises, Phys. Rev. D **97**, 084047 (2018).

[78] C. Tiburzi, G. Hobbs, M. Kerr, W. Coles, S. Dai, R. Manchester, A. Possenti, R. Shannon, and X. You, A study of spatial correlations in pulsar timing array data, Mon. Not. R. Astron. Soc. **455**, 4339 (2016).

[79] M. Vallisneri, S. R. Taylor, J. Simon, W. M. Folkner, R. S. Park, C. Cutler, J. A. Ellis, T. J. W. Lazio, S. J. Vigeland, K. Aggarwal *et al.*, Modeling the uncertainties of solar system ephemerides for robust gravitational-wave searches with Pulsar-timing Arrays, Astrophys. J. **893**, 112 (2020).

[80] R. N. Caballero *et al.*, Studying the solar system with the International Pulsar Timing Array, Mon. Not. R. Astron. Soc. **481**, 5501 (2018).

[81] E. Roebber, Ephemeris errors and the gravitational wave signal: Harmonic mode coupling in pulsar timing array searches, Astrophys. J. **876**, 55 (2019).

[82] J. Gair, J. D. Romano, S. Taylor, and C. M. F. Mingarelli, Mapping gravitational-wave backgrounds using methods from CMB analysis: Application to pulsar timing arrays, Phys. Rev. D **90**, 082001 (2014).

[83] 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).

[84] Z.-C. Chen, C. Yuan, and Q.-G. Huang, Non-tensorial gravitational wave background in NANOGrav 12.5-year data set, Sci. China Phys. Mech. Astron. **64**, 120412 (2021).

[85] Z. Arzoumanian *et al.* (NANOGrav Collaboration), The NANOGrav 12.5-year data set: Search for non-Einsteinian polarization modes in the gravitational-wave background, Astrophys. J. Lett. **923**, L22 (2021).

[86] S. R. Taylor, P. T. Baker, J. S. Hazboun, J. Simon, and S. J. Vigeland, enterprise extensions (2021), v2.3.3.

[87] P. Virtanen *et al.*, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nat. Methods **17**, 261 (2020).

[88] J. D. Hunter, Matplotlib: A 2d graphics environment, Comput. Sci. Eng. **9**, 90 (2007).

[89] C. R. Harris *et al.*, Array programming with NumPy, Nature (London) **585**, 357 (2020).

[90] The pandas development team, pandas-dev/pandas: Pandas (2020).

[91] M. L. Waskom, seaborn: Statistical data visualization, J. Open Source Softwaare **6**, 3021 (2021).